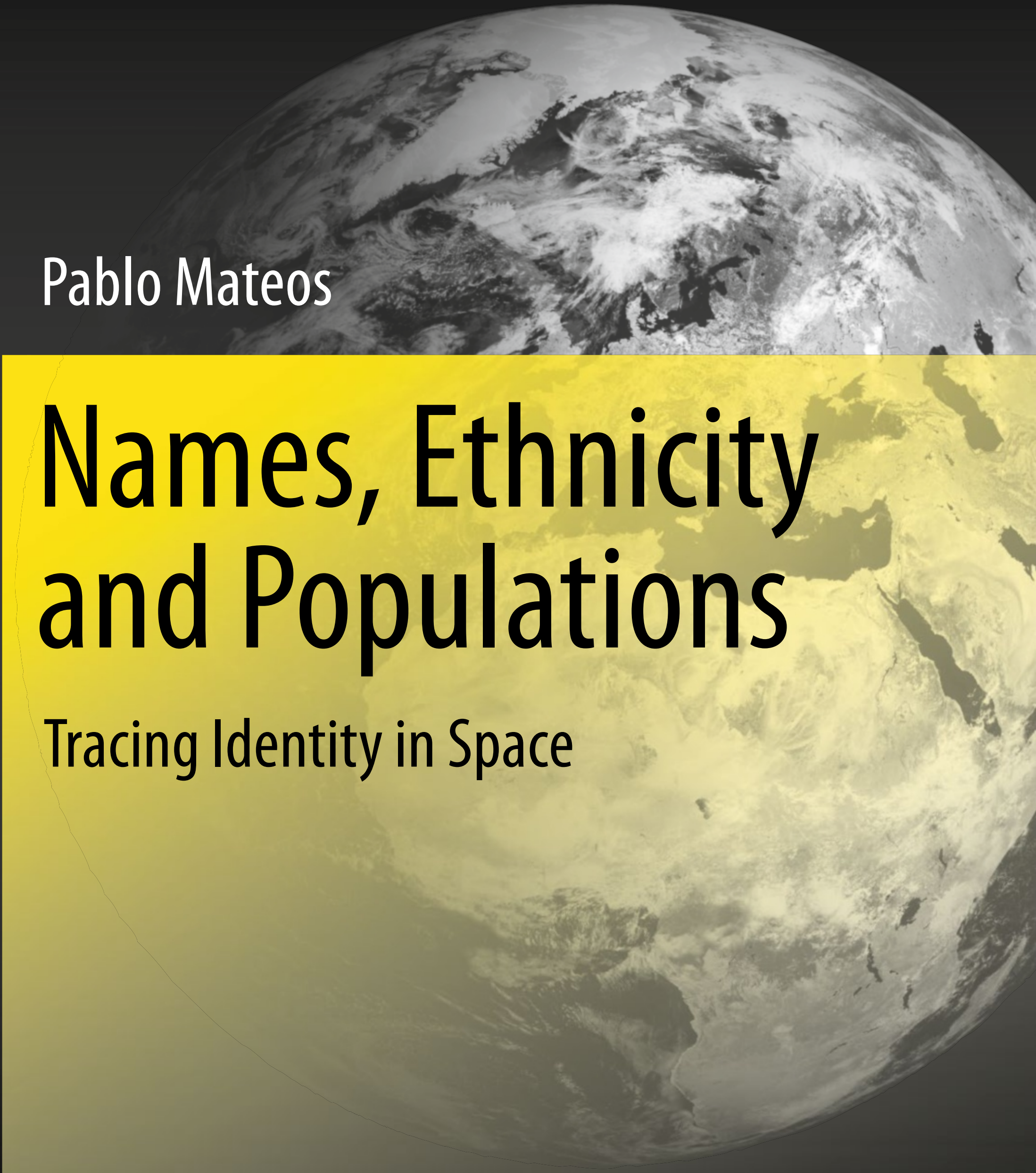
Advances in Spatial Science



Pablo Mateos

Names, Ethnicity and Populations

Tracing Identity in Space

Springer

Advances in Spatial Science

For further volumes:
http://www.springer.com/series/3302

Pablo Mateos

Names, Ethnicity and Populations

Tracing Identity in Space

Springer

Pablo Mateos
Center for Research and Advanced Studies
in Social Anthropology
CIESAS
Guadalajara
Mexico

Department of Geography
University College London
London
United Kingdom

ISSN 1430-9602
ISBN 978-3-642-45412-7 ISBN 978-3-642-45413-4 (eBook)
DOI 10.1007/978-3-642-45413-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014934671

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The research which conforms this book started in autumn 2004 in London, UK and its writing ended in early 2013 in Guadalajara, Mexico. In 2004 I used to live and work in the London Borough of Camden where this book's gestation took place. Walking a few hundred meters around Camden one seemed to be travelling around the world, given its diversity of people, smells, foods, clothes, or music, switching continents as you turn a street corner and listen to tens of languages in every bus ride. However, ethnic diversity in Camden also reflects stark wealth inequalities, home to some of the richest postcodes in the country lying next door to the poorest neighbourhoods in national rankings.

At the time I was analysing health inequalities by ethnic group in Camden, as part of a research project between geographers at University College London and epidemiologists at Camden National Health Service (NHS). It was then that I quickly became dissatisfied with the UK Census ethnic group classification, commonly used to produce all sorts of official statistics by population researchers in the UK. It was, and still is, a broad-brush classification of humanity into eight major groups of "ethnic origin". Its simplicity clearly fails to represent the wide range of ethnic groups present in Inner London. These are by all means no small population groups. For example, 40 % of pupils in London schools speak a language at home other than English, covering a total of 322 different languages (Von Ahn et al. 2010). The wide-spread use of the census ethnic group classification reified deeply-rooted stereotypes and expectations of ethnic disadvantage in British society. For example, a wealth of evidence in population studies points at Bangladeshis as the poorest, most segregated group, presenting the worst health outcomes in London. Because Bangladeshis comprises an ethnic group on their own in the census form, they get all the good and bad attention in academia and public policy. Meanwhile in Camden an equally sized group, the Somalis, complained of not getting the same level of resources because they were 'statistically hidden' under the all-encompassing 'Black African' group. As in-depth analyses of the 2001 Census unfolded at the time, the intrinsic characteristics and needs of tens if not hundreds of ethnic, national, linguistic, religious, and geographic groups were being ignored in London.

It was in this context that I turned to name-origin analysis, as I searched for unconventional ways to group finely grained sub-population 'labels' into alternative configurations of ethnicity that better reflected the diversity of London's population. This was not a gratuitous pastime, stark social inequalities by ethnic group that were clearly observable at the general practices and hospitals of the National Health Service, could not be addressed using population research methods because of a lack of statistics on those very same ethnic groups. However, I noticed that administrative datasets such as population and health registers contained millions of records of people's names and household addresses, partly reflecting that very rich cultural diversity that I was trying to grasp, and to map. "If only I could find a bunch of linguistic experts that could read all those hundreds of thousands of names and code them by ethnic origin…", I kept thinking at the time. I was surprised to find out that others before me have had that very same idea in various countries since at least the early twentieth century. I soon began conducting a literature review that formed part of what was then to become my PhD thesis. "There were 'tried and tested' ways to do this automatically and reliably", I found to my relief! Little did I know at the time that 9 years later I was going to be wrapping up a whole book on the subject.

Various projects and applications spun from that initial interest in mapping ethnic inequalities in health in Inner London. Within a year I managed to enthuse several other researchers at University College London's (UCL) Department of Geography in unveiling this intriguing and fascinating world of name analysis. The core team was comprised of Paul Longley, James Cheshire, Alex Singleton, Muhammad Adnan, Maurizio Gibin and myself. I must thank them all for their support, advice, guidance and co-authorship throughout these years, and more concretely for the materials they kindly gave me permission to publish in this book. Especial thanks go to Paul Longley, whose sense of humour and leadership was central to create the team spirit that made this research possible as well as the most enjoyable experience at UCL during almost a decade. Other researchers joined this team for specific projects and publications, weaving a world-wide network of researchers that over these 9 years has managed to compile and analyse name frequency statistics for the whole population of over 30 countries in four continents, representing at least a third of the World's population. I must first thank Richard Webber, at Experian, whose mentoring over my PhD research years was invaluable and to whom I owe a great deal of inspiration to bring name analysis to the next level. Special thanks must go to Michael Batty and colleagues at the Centre for Advanced Spatial Analysis (CASA), my intellectual home in London, at whose Wednesday seminars (which I never missed for at least 6 years) many of the ideas and collaborations in this book came to fruition. Thanks to Andrew Crooks, Junior Sinesio Alves (†), Oliver O'Brien, Jakob Petersen, Alan Wilson, Yi Gong, Richard Milton, Peter Wood, Adam Dennett, Juliana Cipa, Paloma Rojas, and many other researchers at CASA for their support and encouragement. All of these geographers, computer scientists, architects, planners, epidemiologists, and urban sociologists provided hints and knowledge that are somewhere present in this book.

Outside my *alma mater* at UCL, I was very fortunate to come across a number of researchers from various disciplines around the world that helped me mature the central idea of this book; linking forenames and surnames frequencies into groups of common origin using a network approach. I must specially thank Ken Tucker (computer scientist), at Carleton University, Ottawa Canada, a pioneer in the forename-surname linking approach, and David O'Sullivan (geographer), at the University of Auckland, New Zealand, for his mathematical ability to crack the clustering of huge networks and his key co-authorship that forms part of the materials in Chap. 7 of the book. Thanks must also go to Mario Cortina-Borja (statistician), at the Institute of Child Health, for very valuable insights into name frequency distributions, Franz Manni (geneticist), at the Musée de l'Homme in Paris, for his unorthodox thinking in applying names to population genetics, and to Ludi Simpson (demographer) at the University of Manchester, who read and commented many of my papers on this topic. I am indebted to Andres Moreno and Karla Sandoval (geneticists), at Stanford University, for furnishing my scarce knowledge on human population genetics, and opening up a whole new world that largely facilitated the transdisciplinary research accomplished in the book. My work in this book certainly stands upon these giants' shoulders who preceded me or walked alongside me in moving the research frontier in ethnicity classifications a little further.

Furthermore, my two PhD examiners, Paul Boyle, University of St. Andrews/ESRC Chief executive, and John Stillwell, University of Leeds, provided very useful feedback that allowed me to transform a dull PhD thesis into a much more amenable and coherent book monograph. Many others helped me to source valuable names data across the world or hints on how to get it in different languages, too many to mention here. I am indebted to every one of them.

This book also benefited from extended research visits to a number of institutions. I am indebted to Douglas Massey at Princeton University for hosting me over several months at the Office of Population Research, whose immense library allowed me to locate part of the evidence collected in this book. I must also thank the University of Auckland and the Royal Society for funding an academic stay at Auckland, New Zealand, forging the aforementioned crucial co-authorship work with David O'Sullivan. Repeated academic visits to the Geography Department at the Autonomous University of Madrid, Spain, facilitated by Antonio Moreno, as well as to the Institute of Geography at the National Autonomous University of Mexico, organised by Adrian Aguilar, provided another avenue of interesting research and teaching interactions with urban and population geographers, as well as the time and space to complete parts of this book.

A number of research grants funded the work that underlies the book. Various grants from the UK Economic and Social Research Council (ESRC) (Knowledge Transfer Partnership KTP-037) (RES-348-25-0015) (PTA-026-27-1521) (RES-172-25-0019), the Engineering and Physical Sciences Research Council (EPSRC), the Spanish Ministry of Science and Innovation Base Research Projects (CSO2011-26177), the Royal Society, and the UCL Graduate School, provided the funds for academic mobility and to conduct primary research that partially led to this book.

Finally, but most importantly, I would like to thank my family. My siblings and specially my parents, Manoli and Jose Maria, who laid the foundations that made me an inquisitive person, encouraging me to question all the time the world that surrounds me. Only my wife Brenda, has borne the burden side of writing this book. Without her continuous support, patience, sacrifices and enthusiasm to accomplish and finish this book, not a single chapter would have come to fruition. She also knows that our small children, Blas and Julian, both with multiple passports from birth, will probably laugh at their father's rather simplistic views on ethnicity in this book if they ever get to read it over the next decades. I only hope that in their adulthood, the importance currently placed on ethnicity as an essentialist dimension of a person's identity fades away, and people stop being judged by their physical appearance, accent, religion, kinship, place of birth or colour of their passport. May the reader take this book as a contribution to improving how we monitor progress towards this end.

Guadalajara, Mexico Pablo Mateos
September 2013

Reference

Von Ahn M, Lupton R, Greenwood C, Wiggins D (2010) Languages, ethnicity, and education in London. UPTAP. Working Papers 10–12. Institute of Education, University of London

Author Biography

Dr. Pablo Mateos is Associate Professor at the Centre for Research and Advanced Studies in Social Anthropology (CIESAS), Mexico. He was Lecturer in Human Geography at the Department of Geography, University College London (UCL) in the United Kingdom (2008–2012), where he currently holds an Honorary Lectureship. He obtained a PhD in Social Geography at the University of London (2007). At UCL he is a member of the Migration Research Unit (MRU), an associate of the Centre for Advanced Spatial Analysis (CASA) and Research Fellow at the Centre for Research and Analysis of Migration (CReAM). He is a member of the Mexican National System of Researchers (SNI Level II) and a member/fellow of the Population Association of America (PAA), American Association of Geographers (AAG), Royal Geographical Society, Royal Statistical Society, and British Society of Population Studies. He is a member of the UK *Economic and Social Research (ESRC) Peer Review College*, and a member of the editorial board of the journal *Human Biology*. His research interests lie at the intersections of Social, Urban and Population Geography and his work focuses on investigating ethnicity, identity, migration, citizenship and urban segregation primarily in the UK, Spain, US, Mexico and Latin America. He has published over 40 journal articles and book chapters, amongst others; *PLoS ONE, Journal of Ethnic and Migration Studies*, *Journal of Urban Affairs*, *Geoforum, Human Biology,* and *Population Space and Place*.

Contents

Part II Methods: Name-Based Ethnicity Classifications

Part III Applications: Mapping Names and Ethnicity

List of Figures

List of Tables

Chapter 1
Introduction

Over the last two decades, ethnicity has become one of the most studied, yet controversial, dimensions of populations in bio-medical and social sciences. Intrinsic to the very many definitions of ethnicity proposed in the literature, is the idea of a socially constructed collective identity that is always transient, contingent upon a particular socio-political context in space and time. These unstable boundaries make ethnicity one of the most contested categories of explanation in the social sciences, although it has passed largely uncontested in the bio-medical sciences (except for the work of authors cited in Chap. 2). Despite these challenges, the usefulness of the very many definitions and measurements of ethnicity has been amply demonstrated in studies that tackle discrimination and ethnic inequalities in all sorts of realms of social life. Consequently, there are important shortcomings in the means available to researchers to define, measure and classify rapidly changing population groups along the concept of ethnicity and its cognate multidimensional traits, such as; race, culture, language, phenotype, geography, migration, national identity and even religion.

Despite these shifting grounds, scholars interested in using quantitative measurements of ethnicity at a population level have not been very imaginative in adapting to the transient and fluid nature of ethnicity definitions. Conversely, they have passively accepted the official categorisations of ethnicity, coined by the censuses of population and crystallised into very rigid classifications that are perpetuated during several decades. These ethnicity classifications typically reflect broader societal perceptions on ethno-cultural diversity but are shaped under political agendas and priorities to "manage populations". Furthermore, albeit starting as fairly innocuous labels, over time ethnic groups' definitions on official forms get reified in a positive loop through which ethnic minority members end up identifying with these externally imposed labels. This is because previously unnoticed groups, see how official classification brings political power, resources and overall recognition in all spheres of public life. The wide adoption of these, slowly changing, official ethnicity classifications, has meant that scholars in this area have failed to propose unconventional classifications of human difference. In other words, there is a clear need for alternative ethnicity classifications built along

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_1, © Springer-Verlag Berlin Heidelberg 2014

various dimensions of diversity, and assembled at different levels and amalgamations of ethnic groups' hierarchies.

Because of these research gaps, a myriad of ethnic groups and identity collectives which fall outside the official categories find themselves completely "out of the radar" lacking public and political recognition. The author has discussed most of these aspects in a range of isolated publications that propose the use of personal names to produce alternative categorisations of human difference (Mateos 2007a, b, 2011, 2014; Mateos et al. 2007, 2009, 2011). This body of work reaches a climax in this book, which presents a theoretically and empirically supported line of argument fully developing a new methodological proposition.

This book investigates the use of people's names as alternative method to study ethnicity at a population level. The name-based ethnicity classification method can delineate human populations' contours and journeys across the world, and this book attempts to unveil the key most intriguing ties between ethnicity, geography, and populations in people's naming practices. The book follows the fascinating journey of personal names across the world, proposing the use of use of maps and naming networks to analyse the intriguing geographical and ethnicity patterns that we find today in name frequency datasets.

The core argument of the book goes as follows: Most of us acquired our surnames and first names from our immediate ancestors, either passed down to us over generations or chosen by our parents in ways that are by no means random. Linguistic, religious, regional, cultural and legal factors all shape the ways in which our names are chosen and transmitted over time and across space. Intriguingly, naming conventions usually adhere to unwritten social norms and customs that with time end up producing distinctive ethnic and geographic patterns in name frequency distributions over space. A sort of "name sediment" accretes over time that can be very distinctive of particular places, only altered by migration flows and inter-group marriage between different human groups. Furthermore, these mostly exceptional events can be disentangled in contemporary name distributions and sometimes traced back to their areas of origin. This book compiles evidence assembled from fields as diverse as linguistics, genetics, epidemiology, economics, geography, demography, sociology, anthropology, psychology, history, genealogy, physics, and computer science. This evidence is woven together into an innovative account of how personal name frequency distributions over space and time follow a set of regularities across societies that have hitherto not been studied from a joint, social science perspective on human difference over space.

The hereditary character and group identity function of surnames renders them useful to classify populations in demography (Mateos 2007b), health (Lauderdale and Kestenbaum 2000) and genetics research (Jobling 2001; King and Jobling 2009; Lasker 1985; Piazza et al. 1987; Scapoli et al. 2007), since they document ancestral proximity within and between populations and provide indicators of population structure (Lasker 1985), migration events (Piazza et al. 1987), intermarriage (Bugelski 1961), endogamy and genetic inheritance (Cavalli-Sforza et al. 1982; Jobling 2001). More generally, research has identified the potential usefulness of surnames to classify health and population registers according to ethno-cultural origin of sub-populations (Mateos 2007b), and even social on-line communities such as MySpace and Facebook (Chang et al. 2010) or Wikipedia (Ambekar et al. 2009). In surprising isolation from surname research, the cultural distinctiveness in *fore-naming* practices has attracted wide and interdisciplinary attention in sociology (Lieberson 2000; Lieberson and Bell 1992), geography (Zelinsky 1970), psychology (Seeman 1980), economics (Fryer and Levitt 2004) and linguistics (Bloothooft and Groot 2008; Hanks 1990) over recent decades. Such interest derives from the fact that parental selection of forenames is far from random since it arises out of the culture that a person is born into (Hanks 1990), alongside gender, class, ethnicity, religious affiliation, language and (post migration) identification with the host society (Lieberson 2000). The outcome is that distinctive naming practices in cultural and ethnic groups are persistent, often even long after immigration to different social contexts (Fryer and Levitt 2004; Tucker 2003).

Although widely exposed, such regularities in sur- and fore-naming practices have been largely exploited in isolation from each other. This book undertakes, for the first time, extensive international analysis of the *combined* effects of forenames and surnames as indicators of cultural or ethnic ties in studies of population structure using a network analysis approach. This has not hitherto received systematic focus at the international level. This book demonstrates the value of the linkages between forenames and surnames, through common bearers, to classify names into ethno-cultural groups. An innovative perspective is taken to represent these naming practices over space and time, as the outcome of the workings of a large "naming network", linking surnames to forenames for whole populations. This view leads us to conceive the ties between population groups and places as a complex web of ethno-cultural interactions, with parallels in linguistic and even slight genetic differences between population groups.

The book is structured in ten chapters, eight core chapters plus this introductory chapter and the conclusion. The eight core chapters are organised into three major parts that provide a coherent structure to the book's main thesis and line of argument.

Part I—Theory: Identity and names; surveys the theory and empirical evidence on ethnicity and naming. Chapter 2—"Ethnicity, language and populations", reviews definitions and constructs of ethnicity discussing the issues with its measurement in social and bio-medical sciences across a range of countries. Chapters 3–5, analyse the dynamics of personal naming as a social custom, each chapter picking apart a set of different aspects of naming. Chapter 3—"How we got our names: identity in personal names", reviews a brief history of naming introducing its key features as a basic human function, reviewing some of the most common Western naming systems developed since the Middle Ages, and presenting evidence on how they have been used to classify population origins in the early twentieth century. Chapter 4—"Surnames and genetics", is dedicated to surnames and their study in population genetics to unveil ancestral linkages and substructures within "populations". Chapter 5—"Forenames and social stratification", reviews key characteristics of forenaming practices in various countries and cultures, explaining how they have been used to analyse social stratification, especially with respect to ethnicity and discrimination.

Part II—Methods: Name-based ethnicity classifications; conforms the core methodological contribution of the book. Chapter 6—"Classifying ethnicity through peoples' names", summarises an extensive literature review of name-based ethnicity classifications and their application in population studies, to delineate the research frontier against which the book key developments are set. Chapter 7—"Naming networks and clustering", presents the major methodological contribution of the book, a new name-based ethnicity classification based on clustering naming networks, termed the *Onomap* classification.

Part III—Applications: Mapping names and ethnicity; presents a gallery of spatial applications of name analysis, in particular the mapping of populations sorted by such name-based ethnicity classifications. Chapter 8—"The Geography and ethnicity of peoples' names", proposes a typology of the methodological approaches to the spatial analysis of people's names, both historically and contemporarily, with an emphasis on applications on ethnicity. Chapter 9—"How segregated are peoples' names in London?" shows an in-depth example of one the potential geographic applications of name-based ethnicity classifications; the analyses of urban residential segregation. This chapter closes the book demonstrating how this innovative view of urban population diversity can go far beyond the entrenched stereotypes reified by studies that just rely on conventional, and rather coarse, ethnicity data sources.

The conclusion, in Chap. 10, wraps up the evidence presented throughout the book. Taken as a whole, this book presents evidence on the cultural, linguistic, religious and migration processes that gave rise to distinctive naming patterns, and analyses personal names to identify much of what is distinctive about the ethnic and geographic origins of contemporary population groups. The book argues that this innovative approach, is an important contribution towards building more nuanced understandings about the history and immediate future of our contemporary multicultural societies, at a time in the developed world, in which the predominant political discourse and public debates tend to challenge the very concept of population diversity.

References

Ambekar A, Ward C, Mohammed J, Male S, Skiena S (2009) Name-ethnicity classification from open sources. Presented at proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, Paris, France, June 28–July 1. Available at http://delivery.acm.org/10.1145/1560000/1557032/p49-ambekar.pdf?key1=1557032&key2=1502083521&coll=GUIDE&dl=GUIDE&CFID=53350992&CFTOKEN=96858509. Accessed 18 Dec 2010

Bloothooft G, Groot L (2008) Name clustering on the basis of parental preferences. Names 56:111–163

Bugelski BR (1961) Assimilation through intermarriage. Soc Forces 40(2):148

Cavalli-Sforza LL, Feldman MW, Chen KH, Dornbusch SM (1982) Theory and observation in cultural transmission. Science 218(4567):19–27

Chang J, Rosenn I, Backstrom L, Marlow C (2010) ePluribus: ethnicity on social networks. Association for the Advancement of Artificial Intelligence (AAAI) Washington, 23–26 May. Available at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1534/1828. Accessed 3 Feb 2011

Fryer RG, Levitt SD (2004) The causes and consequences of distinctively black names. Quart J Econ 119(3):767–805

Hanks P (1990) A dictionary of first names. Oxford University Press, Oxford

Jobling MA (2001) In the name of the father: surnames and genetics. Trends Genet 17(6):353–357

King TE, Jobling MA (2009) What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet 25(8):351–360

Lasker GW (1985) Surnames and genetic structure. Cambridge University Press, Cambridge

Lauderdale D, Kestenbaum B (2000) Asian American ethnic identification by surname. Popul Res Policy Rev 19(3):283–300

Lieberson S (2000) A matter of taste: how names, fashions, and culture change. Yale University Press, New Haven, CT

Lieberson S, Bell EO (1992) Children's first names: an empirical study of social taste. Am J Sociol 98(3):511–554

Mateos P (2007a) An ontology of ethnicity based upon personal names. Implications for neighbourhood profiling. Unpublished PhD Thesis, Department of Geography, University College London. Available at http://eprints.ucl.ac.uk/16145/

Mateos P (2007b) A review of name-based ethnicity classification methods and their potential in population studies. Popul Space Place 13(4):243–263

Mateos P (2011) Uncertain segregation: the challenge of defining and measuring ethnicity in segregation studies. Built Environ 37(2):226–238

Mateos P (2014) The international comparability of ethnicity classifications and its consequences for segregation Studies. In: Lloyd C, Shuttleworth I, Wong D (eds) Social-spatial segregation: concepts, processes and outcomes. Policy Press, Bristol, UK

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.bartlett.ucl.ac.uk/casa/publications/working-paper-116. Accessed 5 Mar 2007

Mateos P, Singleton A, Longley P (2009) Uncertainty in the analysis of ethnicity classifications: issues of extent and aggregation of ethnic groups. J Ethnic Migrat Stud 35(9):1437–1460

Mateos P, Longley PA, O'Sullivan D (2011) Ethnicity and population structure in personal naming networks. PLoS One 6(9):e22943

Piazza A, Rendine S, Zei G, Moroni A, Cavalli-Sforza LL (1987) Migration rates of human populations from surname distribution. Nature 329:714–716

Scapoli C, Mamolini E, Carrieri A, Rodriguez-Larralde A, Barrai I (2007) Surnames in Western Europe: a comparison of the subcontinental populations through isonymy. Theor Popul Biol 71:37–48

Seeman MV (1980) Name and identity. Can J Psychiatry 25(2):129–137

Tucker DK (2003) Surnames, forenames and correlations. In: Hanks P (ed) Dictionary of American family names. Oxford University Press, New York, pp xxiii–xxvii

Zelinsky W (1970) Cultural variation in personal name patterns in the eastern United States. Ann Assoc Am Geogr 60(4):743–769

Part I
Theory: Identity and Names

Chapter 2
Ethnicity, Language and Populations

Abstract In an increasingly globalised world, ethnicity, or the question of defining one's personal identity with reference to a "group of common descent" has become very prominent in political as well as scientific debates. Ethnicity is socially constructed and defined subjectively by a combination of aspects related to a group's ancestry, cultural customs, language, religion, national identity, kinship networks and even physical appearance. This slippery nature has made ethnicity the most difficult variable to conceptualise and measure in social as well as medical and biological research. This chapter reviews the main definitions of ethnicity, and the related concept of "race", as well as some of the approaches to measure it. It then proposes a multidimensional approach that conceives ethnicity as an outcome, disentangling some of the processes that end up constituting a group's identity. Finally, it justifies the need for alternative measures of ethnicity, one of them being the use of personal names' origins and forename-surname networks in attempting to disentangle such processes.

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 the 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 its 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 nineteenth century

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_2, © Springer-Verlag Berlin Heidelberg 2014

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' national censuses (Howard and Hopkins 2005; Kertzer and Arel 2002; Mateos 2014; 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 1999). Following an impassionate debate around the essentialism of ethnicity labels (Modood 2005), there seems to be a consensus, at least in population studies, 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 first chapter sets out the general context of the different intersections between the ontology of ethnicity and its measurement in social science. Section 2.1 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. Furthermore, Sect. 2.2 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.

2.1 Constructs of Race and Ethnicity

The study of ethnicity and race in multicultural societies and cities is probably one the most problematic phenomenon that social scientists face today. 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). Ethnicity relates to a person's inner sense of collective identity, and 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 have become increasingly multiculturally diverse and globally connected (Castles and Miller 2003). If "national identity requires a collective work of amnesia" Renan (1990 [1882]: 11), it could be argued that in today's context of globalisation and erosion of nineteenth century nation-state identities, ethnic identity requires a collective work of "remembrance and nostalgia".

The definition of ethnicity and race are controversial because identification is subjective, multi-faceted and changing in nature and because there is not a clear consensus on what constitutes an "ethnic or racial group" (Coleman and Salt 1996; Office for National Statistics 2003). Moreover, ethnicity and race 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 resources and solutions for locally perceived problems, financial aid, political representation, and benefits associated with positive discrimination initiatives (Kertzer and Arel 2002).

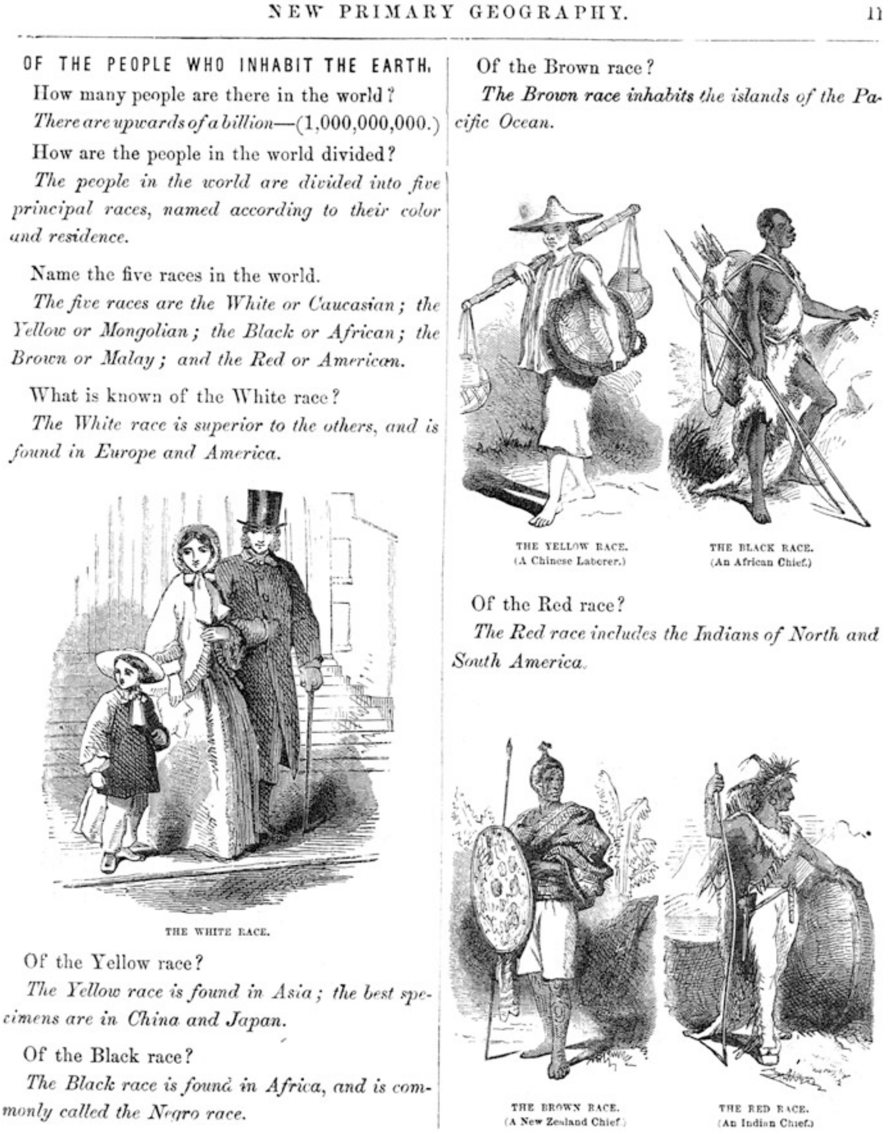
However, the main purpose for which race and ethnicity started to be officially classified and measured in national statistics in a number of developed countries in the last decades bore little correspondence with this identity politics struggle. It directly emanated from the need to monitor progress 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). Such perception of difference evolved into the term "visual minorities", commonly used in Canada and other countries, which goes beyond differences in phenotype and encompasses any other visual element of cultural difference such as religious symbols in clothing or hairstyle. Today, most countries' collective identity classifications go beyond visual or biologically rooted concepts, such as "race", and use the broader cultural term of ethnicity. Such issues of definition of race and ethnicity and their criticisms are discussed in the next paragraphs within this section.

2.1.1 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). These views are well depicted by the illustration shown in Fig. 2.1, taken from a 1968 US primary school Atlas.

This view of human races as discrete biological constructs, was dominant for most of the nineteenth 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). Consequently 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).



NEW PRIMARY GEOGRAPHY. 11

OF THE PEOPLE WHO INHABIT THE EARTH.

How many people are there in the world?

There are upwards of a billion—(1,000,000,000.)

How are the people in the world divided?

The people in the world are divided into five principal races, named according to their color and residence.

Name the five races in the world.

The five races are the White or Caucasian; the Yellow or Mongolian; the Black or African; the Brown or Malay; and the Red or American.

What is known of the White race?

The White race is superior to the others, and is found in Europe and America.

THE WHITE RACE.

Of the Yellow race?

The Yellow race is found in Asia; the best specimens are in China and Japan.

Of the Black race?

The Black race is found in Africa, and is commonly called the Negro race.

Of the Brown race?

The Brown race inhabits the islands of the Pacific Ocean.

THE YELLOW RACE.
(A Chinese Laborer.)

THE BLACK RACE.
(An African Chief.)

Of the Red race?

The Red race includes the Indians of North and South America.

THE BROWN RACE.
(A New Zealand Chief.)

THE RED RACE.
(An Indian Chief.)

Fig. 2.1 Nineteenth century illustrations of 'the races of the world'. *Source*: (Mitchell 1868:11)

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 special issue of *Science* magazine, published in 2005 to commemorate its 125th anniversary, two of the "125 big questions that face scientific inquiry over the next quarter-century" (Science 2005: 5) 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 scholarship as well as government in favour of that of "ethnicity". This trend has been observed in the last three decades of the twentieth 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).

2.1.2 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 1921. He defined ethnic groups as "[t]hose 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 1980 [1921]). 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). This brings to the fore the multi-dimensionality of the notion of ethnicity. Bulmer proposes one of the most widely accepted definitions of "an ethnic group [as] 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). 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 1996: 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 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, globalisation and the growth of nationalisms, and a growing number of different human groups living amongst each other in large numbers (Castles and Miller 2003).

2.1.3 Criticisms

Nonetheless, the characteristics that together define ethnicity are not fixed or easily measured, so ethnicity is considered 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, even overlapping 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 cultural/ethnicity factors, even more difficult and controversial.

Other authors such as David Harvey (2005) 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 in this book. Ethnicity is the concept most widely used to identity population groups that share an ancestral and cultural origin, and thus a much closer term to the main theme of this book, hence its prominence in this book's title.

2.2 Measurements of Ethnicity

Following from the complex definition of ethnicity presented in the previous section this section will review the issues around the difficult task of measuring ethnicity in order to classify people into ethnic groups.

2.2.1 Measurement 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, the measurement of ethnicity is today useful for a wide range of purposes in many countries, especially to reduce ethnic inequalities and to understand our recent past. 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 10 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 a halo of official statistical recognition that transcends the census enumeration exercise itself and determines all sorts of possibilities for an ethnic group 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 in 2000/2001, 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). At the time of writing, a similar evaluation exercise has not been conducted for the 2010/2011 round of censuses. In a thorough comparison of 20 countries 2010–2011 census questionnaires (Mateos 2014) we have identified 27 different questions on various identity aspects that could be summarised under six major themes: (1) Residency and migration, (2) Citizenship, (3) Country of birth, (4) Ethnicity/Race/Ancestry, (5) Language, and (6) Religion. However, international comparisons are highly limited because of the different ontologies of ancestral origin and identity that underlie each of the classifications. A detailed study of such national classifications would entail much more space than available in this chapter, and hence it will mainly focus on the United Kingdom experience, arguably the European country with a longer tradition in measuring ethnicity.

2.2.2 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 some Censuses, such us in the US, 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, in which 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.

2.2.3 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 eight core "ethnic groups". The detailed descriptions given by epidemiologists to these ethnic groups are worth exploring; hence they are included here in Table 2.1. Moreover, the authors denounce the frequent failure of researchers to explicitly define the ethnic categorizations and their context of use, to

Table 2.1 Terms used to refer to ethnicity and race in articles published in two American public health journals: 1996–1999

Asian ($n = 37$)	Black ($n = 16$)	Hispanic ($n = 46$)	White ($n = 32$)	Other ($n = 38$)	Unknown or missing ($n = 9$)	Mixed race or ethnicity ($n = 7$)	Additional terms ($n = 34$)
Asian	African	Caribbean Hispanic	Anglo American	All others	Missing	Biethnic	American
Asian/Oriental	African American	Central/South American	British Whites	Missing or other	Not indicated	Biracial	Ashkenazi Jews
Asian/Pacific Islander	Black Americans	Cuban	Caucasian	Multiple/other/unknown	Not ascertained	Black/White	Canadian
Asian American	Black and other	Cuban American	Caucasian/other	Neither Black nor White	Not classified	Mixed ethnicity	East Indian
Asian or Filipino	Black Hispanic	Dominican	European	Non-Black (all others)	Unknown/refused	Multiethnic	Egyptian
Asian or other races	Black, non-Hispanic	Foreign born Latina/Latino	Non-Hispanic White	Non-American Indian		Multiracial	Eskimo/Aleut/Alaskan
Asian or Pacific Islander	Black, other	Hispanic	Scandinavian	Non-Anglo cultural groups		Partly native American	Foreign born
Asian other	Black/other races	Hispanic/Puerto Rican	Southern European	Non-Latino			Hawaiian
Asian-Indians	Blacks and other minorities	Hispanic = White, Spanish surnamed	White/Anglo	Nonminority (other)			Jewish
Cambodian	United States/Blacks	Hispanic all races	White/other	Non-US born			Middle Easterns
Chinese		Hispanic American	White American	Non-White/Hispanics			Native American
Chinese (Americans)		Hispanic Black	White and Asian	Non-White ethnic minority			Pakistanis
Filipino		Hispanic ethnicity	White and other	Non-White or Hispanic			United States born
Japanese		Hispanic origin	White ethnicity	Other/mixed			

Japanese American		Hispanic surname	White Hispanic	Other minorities
Korean		Hispanic White	White non-Hispanic	Other races/ ethnicities
Korean (Americans)		Latin American	White Spanish surnamed	Other non-Hispanic
Laotian		Latino/Hispanic		Other non-White
Non-Hispanic Asian		Latino not Black		
Orientals		Mexican		
Pacific Islander		Mexican American		
Southeast Asian		Mexican immigrants		
Thai		Non-White Hispanic		
Vietnamese		Other Hispanic		
		Puerto Rican		
		South American		
		Spanish		
		US-born Hispanic		
		White-Hispanic of Mexican ancestry		

The number in the column headings (n=) refers to the total number of terms found per ethnic group, of which only the most common examples are reproduced in the table

Source: Adapted from Comstock et al. (2004: 614) 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, 2009), Raj Bhopal (2004, 2007), and Bhopal and Donaldson (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.

2.2.4 Alternative Measurements of Ethnic Difference

As a consequence of the lack of ethnicity data availability, other proxies, such as 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 2011 Census, less than half of the ethnic minority population was recorded as born outside the UK, and of those born abroad almost half of them hold UK passports. 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 was only recorded by the UK Census in 2011. This proxy is also problematic since it can change over time, some people retain more than one nationality, and usually second generation migrants acquire the host country’s nationality.

A third alternative method employed as a proxy for ethnicity is the analysis of personal names origins. Personal names are in principle good indicators of ethnicity, at least in relation to the immediately prior 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 book, and as such will be further reviewed in detail in Chaps. 3–6, and built upon through an innovative methodology in subsequent chapters. Therefore, repetition is avoided here.

The different dimensions that define ethnicity can be 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. Amongst common identity variables now available are: 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. In the UK, in addition to the traditional country of birth question, an ethnicity question was introduced in 1991, a religion question in 2001, and questions on language spoken at home, passports held, national identity, and year of arrival were introduced in the 2011 Census (Mateos 2014). The recent collection of these new variables in the UK Census will not only provide a richer insight into 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 an individual’s identity to be further revealed.

2.3 Conclusion: Ethnicity, Populations, Languages and Names

The evidence of ethnic and racial inequalities in most multicultural societies has grown strongly in the last decades (Finney and Simpson 2009; Mason 2003; Nazroo 2003). 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 2005; Finney and Simpson 2009; 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). 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 (Mateos 2011; Mateos et al. 2009).

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 book 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 book 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.

If this hypothesis can be proved correct and a suitable methodology can be developed for the purpose of studying ethnic group distribution 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.

References

Afshari R, 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. J Epidemiol Community Health 59(12):1014–1018

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 Sept 2006

Arday SL, Arday DR, Monroe S, Zhang J (2000) HCFA's racial and ethnic data: current accuracy and recent improvements. Health Care Financ Rev 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? Ethn 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 (2005) The operationalization of race and ethnicity concepts in medical classification systems: issues of validity and utility. Health Informatics J 11(4):259–274

Aspinall PJ (2007) Approaches to developing an improved cross-national understanding of concepts and terms relating to ethnicity and race. Int Sociol 22(1):41–70

Aspinall PJ (2009) The future of ethnicity classifications. J Ethnic Migrat Stud 35(9):1417–1435

Barrier NG (1981) The census in British India. Manohav, New Delhi

Bhopal R (1997) Is research into ethnicity and health racist, unsound, or important science? BMJ 314(7096):1751–1756

Bhopal R (2004) Glossary of terms relating to ethnicity and race: for reflection and debate. J Epidemiol Community Health 58(6):441–445

Bhopal R (2007) Ethnicity, race, and health in multicultural societies. Oxford University Press, Oxford

Bhopal R, Donaldson L (1998) White, European, Western, Caucasian, or what? Inappropriate labeling in research on race, ethnicity, and health. Am J Public Health 88(9):1303–1307

Brubaker R (2004) Ethnicity without groups. Harvard University Press, London

Bulmer M (1996) The ethnic group question in the 1991 census of population. In: Coleman D, Salt J (eds) Ethnicity in the 1991 census. Demographic characterisitics of the ethnic minority populations, vol 1. Office for National Statistics, HMSO, London, pp xi–xxix

Castells M (1997) The power of identity. Information age: economy, society and culture, vol 2. Blackwell, Oxford

Castles S, Miller MJ (2003) The age of migration, 3rd edn. Palgrave Macmillan, Basingstoke

Cavalli-Sforza LL (1997) Genes, peoples, and languages. Proc Natl Acad Sci USA 94(15): 7719–7724

Chapman RR, Berggren JR (2005) Radical contextualization: contributions to an anthropology of racial/ethnic health disparities. Health 9(2):145–167

Choi KH, 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 Feb 2005

Christopher AJ (2002) "To define the indefinable": population classification and the census in South Africa. Area 34(4):401–408

Coleman D, Salt J (eds) (1996) Ethnicity in the 1991 census. Demographic characteristics of the ethnic minority populations, vol 1. Office for National Statistics, HMSO, London

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, 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 Jan 2006

Couzin J (2005) To what extent are genetic variation and personal health linked? Science 309:81

Dorling D (2005) Why Trevor is wrong about race ghettos. The Observer, Sunday, 25 September. Available at http://www.guardian.co.uk/race/story/0,11374,1577790,00.html

Eriksen TH (2002) Ethnicity and nationalism, 2nd edn. Pluto Press, London

Finney N, Simpson L (2009) 'Sleepwalking into segregation'? Challenging myths about race and migration. Policy Press, Bristol

Frazier JW, Margai FM, Tettey-Fio E (2003) Race and place: equity issues in urban America. Westview Press, Boulder, CO

Nature Genetics (2001) Editorial. Genes, drugs and race. Nat Genet 29:265–269

Gerrish K (2000) Researching ethnic diversity in the British NHS: methodological and practical concerns. J Adv Nurs 31:918–925

Gill P, Bhopal R, Wild S, Kai J (2005) Limitations and potential of country of birth as proxy for ethnic group. Br Med J 330(7484):196

Gould S (1984) The mismeasure of man. Pelican, London

Graves JLJ (2002) The emperor's new clothes. Biological theories of race at the millennium. Rutgers University Press, New Brunswick, NJ

Harding S, Dews H, Simpson S (1999) The potential to identify South Asians using a computerised algorithm to classify names. Popul Trends 97:46–50

Harvey D (2005) A brief history of neoliberalism. Oxford University Press, New York

Hitler A (1925) Mein Kampf. Secker and Warburg, Munich

HM Government (1976) Race Relations Act 1976. HMSO, London

Howard D, Hopkins PE (2005) Editorial: race, religion and the census. Popul Space Place 11(2): 69–74

Kahn J (2005) Misreading race and genomics after BiDil. Nat Genet 37(7):655–656

Karlsen S, Nazroo JY, Stephenson R (2002) Ethnicity, environment and health: putting ethnic inequalities in health in their place. Soc Sci Med 55(9):1647–1661

Kertzer DI, Arel D (2002) Census and identity. The politics of race, ethnicity, and language in national censuses. Cambridge University Press, Cambridge

Marmot M, Adelstein A, Bulusu L (1984) Immigrant mortality in England and Wales 1970–78: causes of death by country of birth. OPCS. Her Majesty's Stationery Office, London

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

Mateos P (2011) Uncertain segregation: the challenge of defining and measuring ethnicity in segregation studies. Built Environ 37(2):226–238

Mateos P (2014) The international comparability of ethnicity classifications and its consequences for segregation Studies. In: Lloyd C, Shuttleworth I, Wong D (eds) Social-spatial segregation: concepts, processes and outcomes. Policy Press, Bristol, UK

Mateos P, Singleton A, Longley P (2009) Uncertainty in the analysis of ethnicity classifications: issues of extent and aggregation of ethnic groups. J Ethnic Migrat Stud 35(9):1437–1460

McAuley J, De Souza L, Sharma V, Robinson I, Main CJ et al (1996) Self defined ethnicity is unhelpful. Br Med J 313(7054):425b–426b

M'charek A (2005) The human genome diversity project. Cambridge University Press, Cambridge

Mitchell SA (1868) The new primary geography. E.H. Butler & Co., Philadelphia

Mitchell R, Shaw M, Dorling D (2000) Inequalities in life and death: what if Britain were more equal? Policy Press, Bristol

Modood T (2005) Multicultural politics: racism, ethnicity and muslims in Britain. Edimburg University Press, Edimburg

Morning A (2008) Ethnic classification in global perspective: a cross-national survey of the 2000 census round. Popul Res Policy Rev 27(2):239–272

Nazroo J (2003) Patterns of and explanations for ethnic inequalities in health. In: Mason D (ed) Explaining ethnic differences: changing patterns of disadvantage in Britain. Policy Press, Bristol

Nobles M (2000) Shades of citizenship: race and the census in modern politics. Stanford University Press, Stanford

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 (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 Feb 2006

Olson S (2002) Mapping human history: genes, race, and our common origins. First Mariner Books, New York

Oppenheimer GM (2001) Paradigm lost: race, ethnicity, and the search for a new population taxonomy. Am J Public Health 91(7):1049–1055

Peach C (1996) Ethnicity in the 1991 census. The ethnic minorities of Great Britain, vol 2. Office for National Statistics, HMSO, London

Peach C (1999) Social geography. Progr Hum Geogr 23(2):282–288

Peach C (2000) Discovering white ethnicity and parachuting plurality. Progr Hum Geogr 24(4): 620–626

Pearce N, Foliaki S, Sporle A, Cunningham C (2004) Genetics, race, ethnicity, and health. Br Med J 328:1070–1072

Pfeffer N (1998) Theories in health care and research: theories of race, ethnicity and culture. Br Med J 317(7169):1381–1384

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 4 Mar 2006

Rankin J, Bhopal R (1999) Current census categories are not a good match for identity. Br Med J 318(7199):1696

Renan E (1990 [1882]) What is a nation. In: Bhabha H (ed) Nation and narration. Routledge, London, pp 7–19

Science (2005) So much more to know.... Science 309(5731):78–102

Senior PA, Bhopal R (1994) Ethnicity as a variable in epidemiological research. Br Med J 309(6950):327–330

Skerry P (2000) Counting on the census? Race, group identity, and the evasion of politics. Brookings Institution Press, Washington

University of Cambridge (2004) English Dictionary. Cambridge University Press, Cambridge

Weber M (1980 [1921]) Wirtschaft und Gesellschaft (Eng. Tr. Economy and Society). Mohr, Tübingen

Whitehead M (1992) The health divide. In: Townsend P, Whitehead M, Davidson N (eds) Inequalities in health: the black report and the health divide. Penguin, London

Wild S, McKeigue P (1997) Cross sectional analysis of mortality by country of birth in England and Wales, 1970–92. Br Med J 314(7082):705–710

Chapter 3
How We Got Our Names: Identity in Personal Names

> *"For the first time in my life, I felt comfort, the firmness of identity that a name might provide, how it could carry an entire history in other people's memories [. . .] My name belonged and so I belonged"*. Barack Obama (2008: 305)

Abstract Language is an inherent human function and naming is just one of its multiple and inevitable consequences. Through naming we have defined ourselves through millennia, in ways that have involuntarily bounded human groups up through time and space. A brief history of naming is reviewed in this chapter, drawing from linguistic, historic and anthropological literatures, and illustrated with examples drawn from different countries and time periods both for surnames and forenames. The chapter introduces the idea that naming practices are not at all random, but indeed reflect the social norms and cultural customs of the group, and thus follow distinct geographical and cultural patterns. Although such naming patterns have been studied widely for particular groups of names, languages, religions or world regions, few previous attempts have been made to understand their socio-cultural effect on population structure at large, regardless of their individual historical or linguistic idiosyncrasies. The chapter concludes with an early controversial example of how people's name origins have been historically used to subdivide contemporary populations into ethnic groups: the use of historical surnames origins to determine the US immigration policy in the first half of the twentieth century.

Language above anything else is what truly makes us human. Furthermore, we could even describe ourselves as *homo nominus*, since naming anything new that we come across lies at the core of our existence—and survival. Indeed, the practice of naming ourselves has probably been with us since the early development of language. In all human groups every time a baby is born a name is bestowed upon him or her through a rite of passage that grants both personal identity as well as a tie of association with a kin and ethno-cultural group (Alford 1988). Such naming

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_3, © Springer-Verlag Berlin Heidelberg 2014

practices are not at all random, but indeed reflect the social norms and cultural customs of the group, and thus follow distinct geographical and cultural patterns (Hanks 1990). More importantly, these patterns are long preserved over generations, even in a contemporary globalised world (Tucker 2003). Although such naming patterns have been studied widely for particular groups of names, languages, religions or world regions, few previous attempts have been made to understand their socio-cultural effect on population structure at large, regardless of their individual historical or linguistic idiosyncrasies. This raises an important question: Can personal names provide definitive clues to classify populations according to their ethno-cultural origins?

Drawing upon this intrinsic human quality, analysis of the origin of people's names has been successfully used in order to ascribe population ethnic origins in the fields of public health, demography and genetics (Mateos 2007). Such analysis presents a number of advantages over conventional data sources on ethnicity, such as self-reported ethnicity statistics from censuses of population: it can facilitate more detailed and meaningful classification of people's ethno-cultural origins; it provides opportunities for temporal updating; it better accommodates changing perceptions of identity and self; it can be made available at the level of the individual, household or any other convenient geographic aggregation; and, most important of all, it offers the prospect of classification when self-reported ethnicity is not available, and at a fraction of the cost of other alternative research methods.

This chapter presents and introduction to the history of naming, particularly hereditary naming systems, and focusses on highlighting the identity function of names. This historical, linguistic and anthropological enquiry is inserted at this point in the book in order to introduce our object of study; names, and justify their potential secondary uses to signal a group's identity. The first Sect. 3.1, presents a very brief history of naming practices to this purpose, followed by Sect. 3.2 which classifies surnames according to a recurrent typology of origins in Western Europe. Section 3.3 focusses on a wider investigation of naming as a universal human function attached to language use, and its implications for individual and group identity, with subsections dealing specifically with both surnames and forenames. Finally, an early example of how people's name origins have been historically used to subdivide contemporary populations into ethnic groups is presented in Sect. 3.4. The case presented is the use of historical surnames origins to determine the US immigration policy in the first half of the twentieth century, and will serve as a historical precedent to more elaborate approaches to classification dealt with in the following chapters.

3.1 A Very Brief History of Naming Practices

From various linguistic and anthropological studies, it is clear that the practice of assigning a life-long forename or some sort of permanent nickname to identify individuals is as old as language itself. Proper names clearly appear on the very first records of literacy that exist, those found in the Middle East in Sumerian cuneiform language dating to circa 3000 BC, as a way to record legal and administrative transactions (Ostler 2005). However, these names were of the type of modern given names or forenames, i.e. not transmitted through generations. By contrast in China, the practice of using permanent family names is probably 5,000 years old, with some of these ancient Chinese surnames having survived to the present day (Hanks 2003: xiv). Japanese and Korean surnames are much more modern than Chinese surnames, but they still pre-date European surnames, originating in the fifth century AD and the first century BC respectively (Hanks 2003: xiv).

The best documented Western naming system is that used in the Roman Empire. Roman names were comprised of three parts; a forename (hereinafter abbreviated as "F") (*praenomen*), a family or clan name equivalent to a surname (S) (*nomen*), and a nickname (N) (*cognomen*) which could sometimes be inherited. Romans seemed to select their forenames and clan/surnames from a very small pool of names, probably just a few dozen (Hanks 2003). Hence the importance of having freedom to coin and select innovative nicknames (*cognomen*) that were used to distinguish individuals and sometimes lineages. Following the described pattern F-S-N, the name of the famous emperor *Gaius Julius Caesar* (100–44 BC), meant that the last element (N) was a nickname meaning "fine head of hair", while rather counter-intuitively *Julius* was his clan name (S). Christianity came to change the Roman naming system by introducing a biblical forenaming practice.

In the millennium between the Christianisation of the Roman Empire (c. 300 AD) to the High Middle Ages (c. 1000–1300 AD), a patronymic or genealogical system was the most commonly practiced custom across the whole European continent. Such system was based on parentage, through which a person would be known by his own forename, or given name at baptism (usually a Christian saint or biblical name), in addition to a genitive associated to one of his parents forenames or sometimes supplemented by a nickname or occupation. The main characteristic of this system is that such patronymic "second name" was not hereditary beyond one generation. In Iceland the patronymic system still survives today, i.e. a person's surname is derived from the father or mother's forename (e.g. *Sigurdardottir*, the daughter of *Sigurdar*). In the rest of Scandinavia the patronymic system survived until the nineteenth century, when the existing patronymic surnames became fixed and hereditary. Hence the high frequency of names ending in "-son" in Sweden or "-sen" in Denmark, meaning "son of" amongst many other common suffixes.

A reduction in the pool of forenames approved by the Christian church for baptism, was observed during the High Middle Ages in various European countries, displacing older vernacular forenames (Faure et al. 2001). This in turn put pressure on a patronymic naming system that increasingly failed to comply with the identification function of names, as defined by Alford (1988). This was specially the case at local level, since a large majority of people carried the same saint names as forenames. Because of this problem of identification, a system of fully hereditary surnames was gradually introduced in Europe, evolving from the early medieval patronymic tradition. In fact, Hanks (2003: xvi) notes that "there is not a country in

Europe that does not have surnames derived from forms of John, Matthew, Mark, Luke, Peter, Paul, and other saints, apostles, and missionaries of the Christian Church". With time, many regions developed the tradition of adding a family name in addition to the long-standing custom of a forename (Beech et al. 2002). This custom usually started with the nobiliary classes, because of the need to unequivocally identify individuals in property titles and inheritance, and was slowly but gradually adopted by more common individuals. Such linkage to transmission of nobiliary titles and property, probably explains why surnames are patrilineally transmitted in all European countries as opposed to a matrilineal tradition. After all, motherhood has been always much easier to demonstrate than fatherhood, and hence the need of a system to identify the parent of a child (Alford 1988).

As a result, current European hereditary surnaming conventions date from different periods over the last ten to eight centuries. Most Anglo Saxon family names, for example, date from the twelfth and thirteenth centuries (Hanks 1992), Spanish surnames from the thirteenth to fifteenth centuries (Mateos and Tucker 2008), while in The Netherlands between 1796–1811 (Manni et al. 2005), and in Turkey and Iran as late as the 1920s–1930s (Razum et al. 2001). Regional differences also existed within some countries. For example in Wales the patronymic system was still in use up to the nineteenth century (Hanks 2003), hence the peculiar surname frequency distribution found today in Wales (Longley et al. 2011).

Apart from some exceptions, such as the aforementioned of Scandinavia, Netherlands, Wales and Turkey, most European modern naming systems were widely established by the fourteenth to fifteenth century, and are thus at least five centuries old (Hanks 2003). As a consequence, we can safely conclude that in most European countries a great majority of surnames have been "mechanically" passed down the family (paternal) line for about 20 generations. In Ireland, clan names such as *Ó Néill* have been used as an hereditary surname since at least the tenth century (Hanks 2003). This means that the contemporary surname O'Neill, and its alternative spellings, is probably over a thousand years old, having been transmitted over 40 generations or more. Such patrilineal linkage between the first bearer or bearers of a name and all his descendants over centuries is what makes surnames such a unique resource to study vertical processes of biological and cultural inheritance over space and generations.

Outside Europe a variety of naming systems exist, typically following a linguistic or religious tradition. In some areas with a history of European colonisation, naming systems and the names themselves were established in the European metropolis tradition. Therefore, the surnames of British plantation owners in the Caribbean were imposed to former slaves, Spanish double paternal and maternal surnames bestowed upon indigenous and mixed populations across Latin America and the Philippines (Tibón 2001), and Portuguese surnames remain the norm in places as far apart as East Timor, Goa in India, Mozambique and Brazil, just to cite a few examples of ex-colonies that inherited both a custom and a stock of names. However, other non-European naming systems survived in various continents and countries, primarily in Asia, Africa, and parts or Oceania. As mentioned before, some Chinese surnames are 50 centuries old, and most countries influenced by Chinese language adopted the tradition of surnames around 2,000–1,000 years ago. However, in the majority of countries outside Europe and areas of historical European colonisation, the tradition of using surnames in addition to a forename has been introduced with the creation of modern state administrative systems over the last two centuries. As compulsory birth registration practices have slowly spread to the most remote rural areas of such countries, the custom of adopting an hereditary surname has been gradually accepted by a majority of the population. Even so, in countries such as Mexico, an OECD member, 1 % of the adult population entitled to vote still does not have a legal surname, according to its electoral register (Instituto Federal Electoral 2006).

Although some name examples and data from various countries will be presented throughout this book, reviewing the history and patterns of a representative range of naming systems worldwide is clearly beyond the scope of this investigation. The reader is referred to specialist reference works such as the Dictionary of American Family Names (Hanks 2003), and specialist literature in each language.

3.2 How We Got Our Surnames: A Typology

In most European naming systems, the great majority of surnames originated from seven broad types of proper names; patronyms, locative/toponyms, occupational, nicknames, diminutive, ornamental, and other types of surnames. It is useful to review what these types are and how they originated in medieval times.

(a) Patronyms, and their female equivalent Matronyms, as mentioned earlier, are surnames that originate from a parental forename. The early patronymic tradition of adding a prefix or suffix to a parent's forename to indicate the "son or daughter of" or a genitive form, later became completely fixed and hereditary regardless of a father's forename. For example, the English *Johnson;* originally meant the son of *John*, the Danish *Eriksen;* the son of *Erik*, the Spanish *Rodríguez;* the son of *Rodrigo*, the Irish/Scottish *McDonald;* the son of *Donald,* in Russian; *Sergeyevich* son of Sergey. A few examples of patronyms inherited by daughters are the Scandinavian suffixes for "daughter of" (Icelandic -*dóttir*, Swedish and Norwegian -*dotter*, Danish and Norwegian–*datter*), the Gaelic *Mac* as in *MacDonald*, daughter of *Donald*, and the Polish *Malinowska* daughter of *Malinow*. Matronyms, which were taken from a female forename, such as *Beaton, Marguerite, Margetson, Tillotson*, are much rarer, typically heiresses in their own right or long-term widows (Hanks 2003).

(b) Locative surnames, are surnames that refer to a place (toponyms) or feature of the landscape (topographical names). A large proportion of surnames in various European countries are toponyms, and were originally bestowed upon outsiders to denote the place where they came from. The scale of a toponym could range from a local name such as a neighbourhood, village, town (e.g. *Saint-Germain*, *Lancaster*, *Paris*), to a whole county (e.g. *Cheshire, Toledano*), region (e.g. *Breton, Norman*) or even country (e.g. *French*). As for topographical surnames, they typically refer to a feature on the landscape, either natural such as *Forest, Rose, Green, Hill, Dale, Vale, Field, Orchard* or human-made general places, such as *Gates, Hall, Church, Bridge, Hilton* (town on the hill), *Park, Wall or Port*. Again, many such examples exist in all European languages.

(c) Occupational surnames are derived from professional occupations. The most famous metonym is *Smith* (and its equivalent in other languages such as *Schmidt*, *Smit)* someone who worked in metal forgery. Since every village used to have a smith, this surname had many original "founders" and hence its exponential diffusion. Many medieval professions and crafts are well reflected in contemporary European surnames, such as *Archer, Baker, Cooper, Harper, Miller, Thatcher, Skinner, Taylor*, or *Turner* to name a few (only the English versions are provided here). Some occupational surnames present a distinct regional distribution, such as the profession of someone who softened freshly woven cloth by beating and tramping on it in water, which in England was called *Tucker* in the Southwest, *Walker* in the West and North, and *Fuller* in the Southeast and East Anglia. Such unique regional distributions of a surname's origin have been exploited to establish geographical patterns in surname frequencies and migration, as will be discussed in detail in Chap. 8. An example of such patterns is shown in Fig. 3.1, mapping the frequency distribution of occupational surnames ending in "*-man*" in Great Britain

(d) Nicknames are surnames derived from a moniker that was originally assigned as a humorous, sometimes cruel, local identifier of a person. Nicknames were sometimes passed down to children, in addition to the official name, until they became fixed into hereditary surnames. They commonly referred to a certain respectable quality of the person, such as in *Fairchild, Good, Smart* and *Trueman*, and equivalent examples in other languages. There are other more humorous nickname surnames, such as *Wild*, *Beard*, *Frost*, *Little*, or *Smallman* and much more cruel and offensive ones, such as *Dolittle*, *Hasty*, *Idle*, *Slow*, *Smellie*, *Pigg*, or *Bottom*. These last examples of embarrassing surnames have indeed fallen in frequency in Britain over the last century as people have managed to move away from undesirable surnames (Daily Mail 2006).

(e) Diminutive surnames were derived from abbreviating or adding a suffix (very rarely a prefix) to a forename to make it a surname (also called its "pet form"). In medieval English the most common diminutive suffixes were *-cock* (young man)*, -et, -lett, or -kin*, producing *Bartlett, Dykin,* or *Hitchcock. Huggins* for example is derived from *Hugh-kins*, while *Jenkins* from *Jan-kins*. *Littlejohn* is an example of a diminutive from a forename using a prefix. Diminutive surnames are specially common in Italian and Czech surnames (Hanks 2003).

(f) Ornamental surnames are "made up arbitrarily from vocabulary words with more a less pleasant associations" (Hanks 2003: 15). They are common in

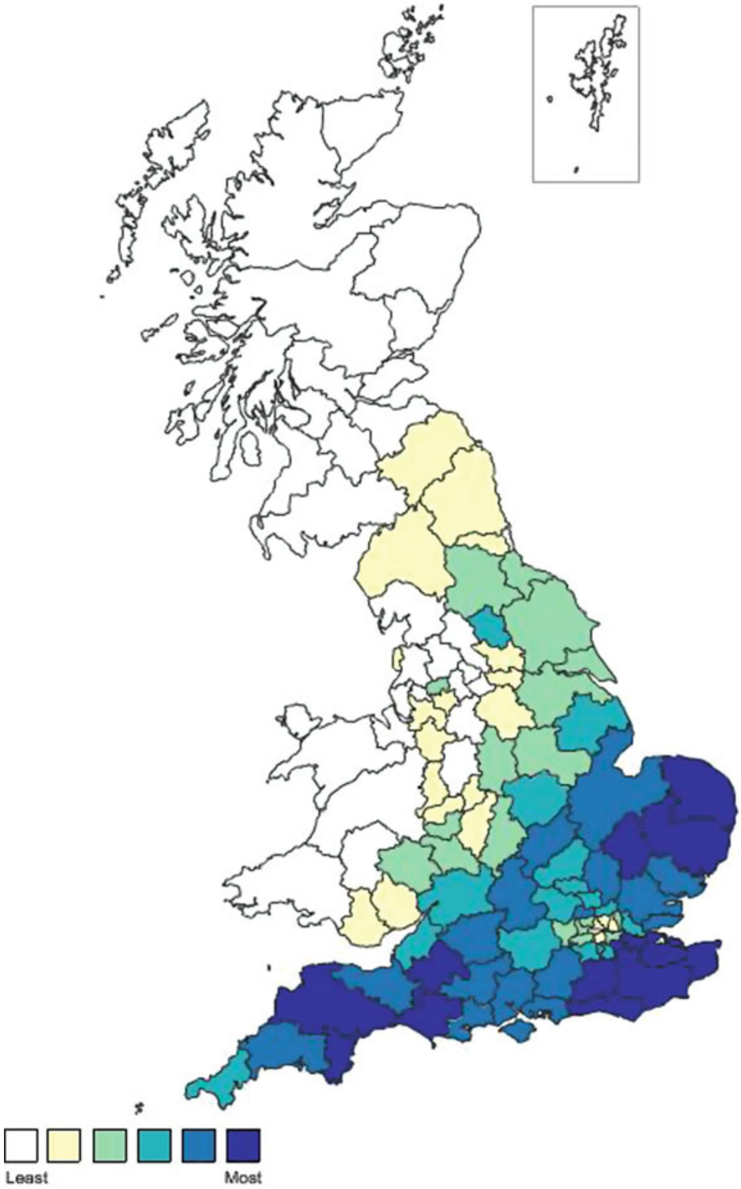


Fig. 3.1 Frequency distribution of occupational surnames ending in "-man" in Great Britain (1998). The colours depict most (*darker*) to least (*lighter*) concentrations of surname frequencies calculated from the 1998 Electoral Register. *Source*: GB Surname Profiler http://gbnames.publicprofiler.org/

Turkish, Jewish and Swedish surnames. For example, modern Turkish surnames were introduced by a 1934 law stating that all surnames should be free from foreign (i.e. Arabic) or religious (i.e. Muslim) connotations and must have a meaning in the Turkish language (Razum et al. 2001). Therefore, many new surnames were taken from the existing vocabulary and hence ornamental names are very frequent.

(g) Other types of surnames do not fit neatly into the above categories. For example surnames that indicated lack of kin, such as those assigned to children that grew in foundling institutions, illegitimate, or abandoned children. Some surnames derived from saint names belong to this type, as well as the Dutch *Weese*, Polish *Serota*, and French *Jetté* which literally mean "thrown out". Other surnames given to children reared at the expense of the community, such as Italian *Innocenti*, *Comunale,* the English *Parrish*, and others such as Italian and Spanish *D'Amore*, *Amor* ("of love"), *Di Dio, De Dios* ("of god"), and *Esposito* ("exposed"). Another type of surnames half-way between occupational and nicknames are status names. These make reference to social status, such as a particular role in medieval society, servant of a particular nobiliary title or more humorous assignments by local folk such as a role in a pageant or other festivities. Examples of these status names are; *Bachelor*, *Knight*, *King*, *Prince*, *Duke*, *Earl*, *Bishop*, *Kaiser*, *Graf*, *Herzog*, *Alderman*, *Beadle*, *Sherriff*, or *Reeve*.

Beyond its onomastic and linguistic interest, this classification of surnames into seven broad groups is useful for the purpose of this book, since it allows to group surnames into types that clearly present distinctive regional patterns in the geographic distribution of its contemporary frequencies. Such patterns in turn are helpful to reveal population distribution and dynamics over the last three to five centuries, as shown by the example in Fig. 3.1 and fully discussed in Chap. 8.

3.3 Identity in Naming Systems and Practices

There is no single human culture or society that does not bestow personal names on its members (Alford 1988; Hanks 1990). As such, personal names are considered to be "cultural universals" or "human universals", defined as a list of common traits present in all cultures (Brown 1991; Murdock 1945). Thus, personal naming is an inherently human activity, forming part of the broader linguistic function of assigning names to places, objects and many other physical phenomena as well as human constructs and abstractions. As Zabeeh (1968: 56) puts it "not only can there be no histories, geographies, biographies, novels, myths, etc., but more basically there can be no family relations, tribal institutions, or political organisations, even at the most primitive level without the existence of some linguistic expressions by means of which significant persons, places, times and objects are uniquely identified and referred to."

From a linguistic standpoint, a distinction must be made in how we assign names. When we coin a new noun to designate an object, a place or an entity we can either create a whole class of entities (e.g. town, hamlet, village, city) or we can assign a unique name to a particular instance of those entities (e.g. New York, Oxford, Guadalajara, Pisa). The former is usually termed a common noun while the latter a proper noun. A proper noun therefore serves the purpose of reference and identification of a unique entity. The minor distinction between a proper noun and a "proper name" is that the former is comprised of one word (the noun, such as *France*), while the latter can also be comprised of several words (e.g. *The United States*). The broader term proper name will be used in this book, and particularly one specific type of proper names that refer to individual persons, *personal names*.

Proper names are considered a non-descriptional "reference fixer" (Kripke 1972). That is, they have lost their original meaning as a word or group of words and its sole purpose is to fix a reference to the particular entity they identify. Therefore, the name of the city of *Grand Rapids,* in the US State of Michigan, is a proper name that does not refer anymore to a natural water feature that might have existed in the landscape when the city was founded. That is, as a proper name is non-descriptional and it only refers to a unique place. By the same token personal names are also non-descriptional. For example, Steve Bishop does not anymore describe the religious profession of a person whose forename is Steve, but its sole purpose is to identify a unique person (even when there might be several people with that combination of names) and not to provide meaning or refer to a class of entities (Bishops).

Following this line of argument, personal names are deemed not to "belong" to any language in that their "exchange value" is not on its meaning, but on its identification function, and hence they are not translatable to other languages (Recanati 1993). That is, when referring to the names of people from other countries or languages, journalists and writers do not usually translate their names to their equivalent in the language they are writing in (although translating forenames was a common practice until a few decades ago). Otherwise, the identity of the person (as in Kripke's (1972) reference-fixer) would be compromised. This permanence characteristic of personal names in turn improves our ability to detect the ethno-linguistic origin of names when they migrate to other areas. In fact, family names are generally preserved long after migrants and their descendants have been integrated into host societies (Tucker 2003), and a preference for certain forenames remains over several generations (Lieberson 2000). Historically, there has been complete translations of migrants names upon registration on the country of immigration (e.g. German *Shumacher* to English *Shoemaker*), but these cases are generally the exception rather than the rule. Translation is of course different from transliteration, or the transformation of a name in another linguistic context, to adapt sounds, pronunciation or an alphabet that does not exist in another language. Transliteration and other abbreviations or alterations in names spellings do transform names, but they generally remain distinct from the receiving society pre-existing names, i.e. they are not automatically fused with other names.

Despite the aforementioned consensus that personal names do not belong to a particular language, their overall permanence over space and time makes them preserve a link with the particular language in which they were first coined. Such linkages can be phonologically, morphologically, and syntactically established (Abbott 2005). Therefore, personal names are always rooted in a language, and they are deemed central to individual and group identities (Beech et al. 2002). In order to understand the nature of the linkage between names and language or group identity, it is necessary to investigate the common patterns found in naming systems.

In an extensive analysis of worldwide personal naming systems in 60 societies, Alford (1988) concludes that all serve two central antithetical but complementary functions: differentiation and categorization of individuals: "[t]he need to distinguish individuals for clarity in communication exists alongside an equally powerful need to categorize people, to fit people into a social matrix by highlighting their similarities, rather than their differences" (Alford 1988: 69). In a similar vein, Elias (1991) argues that the double construction of a person's name, through the forename plus surname formula, is part of a balance between what he terms "I-identity" and the "We-identity" in human societies (individual versus collective identification). Hence, a name is both "a symbol of uniqueness" of the individual and a "visiting card which indicates who one is in the eyes of others" (Elias 1991:184). Related to these two primary functions of naming, Alford recognises that in the wide range of societies he studied, personal names perform a variety of social and psychological functions. These are: to distinguish people, to emphasise family membership and continuity, to signal parenthood or social belonging, to express conceptions of personal identity, to reflect ethno-psychological conceptions of the self, sometimes to link an individual to a place or caste, to reflect a cultural dualism in societies in transition, and to distinguish between the sexes (Alford 1988). These seven universal naming functions denote the tension mentioned above between; an innate human desire to differentiate individuals but, at the same time, to classify people into identity groups. Furthermore, Nwigwe (2001) observes that for philosophers such as Locke and Hegel "naming must follow some historical and cultural regulations, such that names given in a culture, stay within the culture's frame of meaningfully accepted modes of designation" (Nwigwe 2001: 63). This view is corroborated by Alford, who concludes that "naming systems both reflect and help to create the conceptions of personal identity that are perpetuated within any society" (Alford 1988: 167).

The evidence presented so far clearly points towards the existence of a set of certain cultural regulations that create and preserve distinctive means of "accepted modes of designation" within a society or cultural group, through the aforementioned processes of differentiation and categorisation. These two primary functions of personal naming, alongside the other mentioned sociological and psychological functions, are exploited in the investigation presented in this book since, as we should see, the seemingly inevitable outcome is to create and preserve distinct cultural naming practices in every ethnic group even long after migration.

Until this point we have intentionally avoided the distinction between the different components of personal names as commonly found in most societies.

Hereinafter we distinguish between two types of personal names: surnames (also known in English 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 in English as first names, given names, or Christian names), which refer to the proper name given to a person, usually at birth. Unless qualified, our use of the term “names” will henceforth refer to both forenames and surnames.

3.3.1 Surnames and Intergenerational Identity

Although not all contemporary societies use surnames, they are by far the most common practice to emphasise family unity and continuity, and have been described as “conspicuous manifestations of kin-group-embedded conceptions of identity” (Alford 1988: 55), or as “historical—though recent—signs of identity in social groups” (Manrubia and Zanette 2002: 461). Their hereditary character and group identity function have made them the subject of study in demographic, historical, health and genetic research. There are two broad types of studies of surnames for this purpose of investigating population groups’ identity; (a) historical investigations of the degree of ancestral proximity within and between populations over the last five to ten centuries, used as indicators of population structure, migration events, intermarriage, endogamy and vertical transmission of culture (Bugelski 1961; Cavalli-Sforza and Feldman 1981; Jobling 2001; Lasker 1985); and (b) the study of contemporary or recent migration episodes, using surnames to classify ethnicity in contemporary health and population registers in the field of demography and public health amongst others. The latter type of studies are summarised in Chap. 6, which is devoted to investigate the value of names as general indicators of group identity in classifications of sub-populations into ethnocultural origins, while the former is dealt with in the next Chap. 4, on surnames and genetics.

3.3.2 Forenames and Parental Identity Choice

While the links between surnames and ethnicity have been amply demonstrated in the aforementioned demographic, health and genetics literature, and need no further justification here, those between forenames and ethnicity have been largely ignored by these research fields (Mateos 2007). However, the potential of distinctive forenaming practices in identifying cultural groups has attracted the attention of researchers in sociology (Lieberson 2000; Lieberson and Bell 1992), geography (Zelinsky 1970), psychology (Seeman 1980), economics (Fryer and Levitt 2004) and linguistics (Bloothooft and Groot 2008; Hanks 1990) over recent decades. These studies point of departure is the simple observation that parents do not select forenames for their children at random (Bloothooft and Groot 2008; Lieberson 2000). These and other authors see forenames as encoding personal and group identity since their selection arises out of the culture that a person is born into (Hanks 1990). As such, they may be seen as a “stamp of the namers’ traditions and their hopes for the child” (Seeman 1980: 129). Hanks (1990) observes that the two key factors that overtly or subliminally operate in the choice of a forename are religious identity and native language, since common forenames in the socio-cultural milieu in which people live are always preferred to alien or invented names. This trait also relates to the aforementioned tension between the naming functions of differentiation and categorization of individuals (Alford 1988). As Woods (1984) states: “the given name in any culture is a unique possession often connoting ethnicity, religious tradition, age, and a degree of adherence to a dominant culture” (1984: xiii) and advocates for their more profound study.

From these perspectives, forenames are widely seen as encoding personal and group identity since parental selection arises out of the culture that a person is born into (Bloothooft and Groot 2008; Hanks 1990). As such, forenames signal gender, social class, race and ethnicity, religious identity, and for migrant families, native language and the degree of assimilation and identification with the mainstream society (Hanks et al. 2003; Lieberson 2000). In this respect, forename preferences within ethnic groups in the US have been found to reflect both internal mechanisms (religion, language, convention) as well as external influences (acculturation, social attitudes), whereas there is ample evidence that group cultural traits persist even after their causal conditions have attenuated or even disappeared (Lieberson 2000).

Overall, such prominent socio-cultural features of forenames prompted Zelinsky (1970) to conclude that the “choice of [forenames] comes closer to fulfilling the criteria for an ideal cultural measure than any other known item” (1970: 746). Are we in front of a long-searched for proxy for socio-cultural preference? As a matter of fact, this is the premise upon which Lieberson’s (2000) extensive empirical study on forenames and social taste is based. As he puts it “fashions provide an extraordinary opportunity to study issues of social change more generally [. . .] [, and first] names provide an exceptional opportunity to study internal mechanisms of taste without the need to disentangle the powerful commercial forces that strive to mold fashion” (Lieberson 2000: xiii). Lieberson’s research into the principles driving social change through individuals’ decisions on social taste is summarised in detail in Chap. 5. That chapter also presents an unprecedented multidisciplinary investigation into the identity consequences of forenames.

3.4 From President Washington to Obama: Surnames, Identity and US Immigration Policy

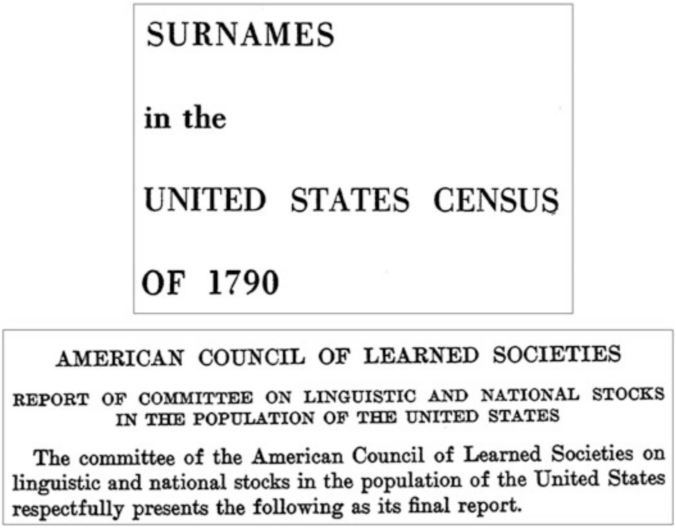
The use of people’s name origins to subdivide contemporary populations into ethnic groups has a long history, probably as old as the presence of large collectives of migrants started to become visible by the uniqueness of their names. Chap. 6 includes a review of such approaches in contemporary academic research over the last four decades. However, a well-documented—yet not widely publicized- historic precedent is the use of historical surnames origins to determine the US immigration policy for four decades of the twentieth century. This case will be introduced here to lay the ground for a more careful use of names and origins in subsequent chapters.

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. After the perceived success of this study and growing concerns for the growth of non-Anglo-Saxon migrants in the 1920s, the US Federal Government attempted to control the size and character of the streams of immigration. The 1924 Immigration Act 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 at the time for contemporary's concept of ethnicity (McDonald and McDonald 1980). Therefore the government needed to establish a baseline determining what were the nationalities of the original stock of white population in the eighteenth century that formed the country at the time of the US independence (US Senate 1928). The only method available at the time, was to examine the origins of surnames in the individual responses to the first ever US Census of Population, that of 1790 carried out under Washington's presidency (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) (see Fig. 3.2). 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.

The numerical 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). Furthermore, underlying these



SURNAMES

in the

UNITED STATES CENSUS

OF 1790

AMERICAN COUNCIL OF LEARNED SOCIETIES

REPORT OF COMMITTEE ON LINGUISTIC AND NATIONAL STOCKS IN THE POPULATION OF THE UNITED STATES

The committee of the American Council of Learned Societies on linguistic and national stocks in the population of the United States respectfully presents the following as its final report.

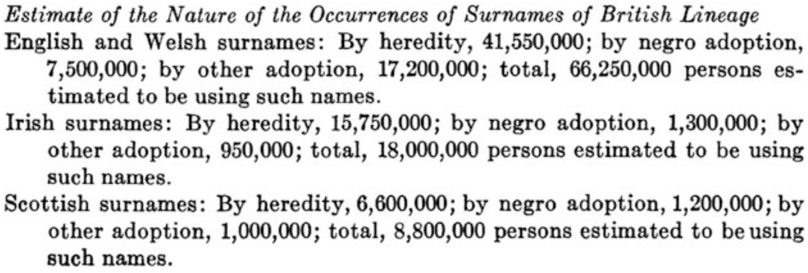
Fig. 3.2 Report of the committee on linguistic and national stocks in the 1790 population of the U.S.: Extracts from cover and first page (American Council of Learned Societies 1932)

Table 3.1 Estimates of national origin breakdown of the US white population in 1790, using surnames

	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

Source: created by the author based on data from Rossiter's (1909) reported in Akenson (1984: 103) and the American Council of Learned Societies (ACLS) (1932: 124)

studies was a preoccupation with the changing composition of the US population in the 1920s, as an anti-immigration sentiment clearly set in across the political spectrum. An extract from one of Baker's publications (1928) shown in Fig. 3.3, denoting the racial preoccupation with demarcating the number of African-Americans who bore British surnames as well as other persons that acquired such surnames "by adoption" as opposed to "heredity". After all, this was a time at which Polish and Italian migrants were undesirable and not conceived as white.



Estimate of the Nature of the Occurrences of Surnames of British Lineage

English and Welsh surnames: By heredity, 41,550,000; by negro adoption, 7,500,000; by other adoption, 17,200,000; total, 66,250,000 persons estimated to be using such names.

Irish surnames: By heredity, 15,750,000; by negro adoption, 1,300,000; by other adoption, 950,000; total, 18,000,000 persons estimated to be using such names.

Scottish surnames: By heredity, 6,600,000; by negro adoption, 1,200,000; by other adoption, 1,000,000; total, 8,800,000 persons estimated to be using such names.

Fig. 3.3 Barker's (1928) estimation of populations with British surnames in the U.S.

Based on such figures on the "original national stock" an immigration quota system was established in 1924 (The Immigration Act, including the National Origins Act, and Asian Exclusion Act) that lasted until 1965. By the early sixties, in the midst of the Civil Rights movement, large collectives of Greeks, Poles, Portuguese and Italian migrants and their descendants were denouncing that the immigration quota system discriminated them against Western Europeans. President Kennedy gave a speech in June 1963 to the American Committee on Italian Migration, describing the quota system "nearly intolerable". Two congress members, Peter Rodino and Dan Rostenkowski, respectively of Italian and Polish descent, were chairs of powerful committees in Congress and wanted to reverse the 1920s laws against people like their grandparents (Massey 1995). As a result of this process of de-racialisation of US policy, the Immigration and Naturalization Act (1965) was signed by president Lyndon Johnson at the foot of the Statute of Liberty, a symbolic place in US immigration history (see photograph in Fig. 3.4). The new law introduced an equal immigration numerical quota for nationals from each country in the world, regardless of national or racial origins. Johnson's speech emphasised the links between U.S. nation building, identity and fairness, replacing a quota system designed using western European surnames:

> *"This [old] system violates the basic principle of American democracy, the principle that values and rewards each man on the basis of his merit as a man. It has been un-American in the highest sense, because it has been untrue to the faith that brought thousands to these shores even before we were a country"* (Government Printing Office 1966: 1037)

This well documented example of the use of surname origins to classify populations according to migrant or ethnic groups by the US government spans almost two centuries; from the 1790 Census to the 1965 Immigration Act and the Civil Rights movement, from Washington to Lyndon Johnson. But after four decades after the introduction of a wide range of anti-discrimination policies, another US president has come to illustrate the heavy identity baggage carried by personal names. The strong identity implications of a person's name, especially when a name has been a source of puzzlement and even discrimination in a different cultural context, is extremely well encapsulated by the following passage from U.S. president Barack Obama's autobiography "Dreams from my father", an extract of which opens this chapter:

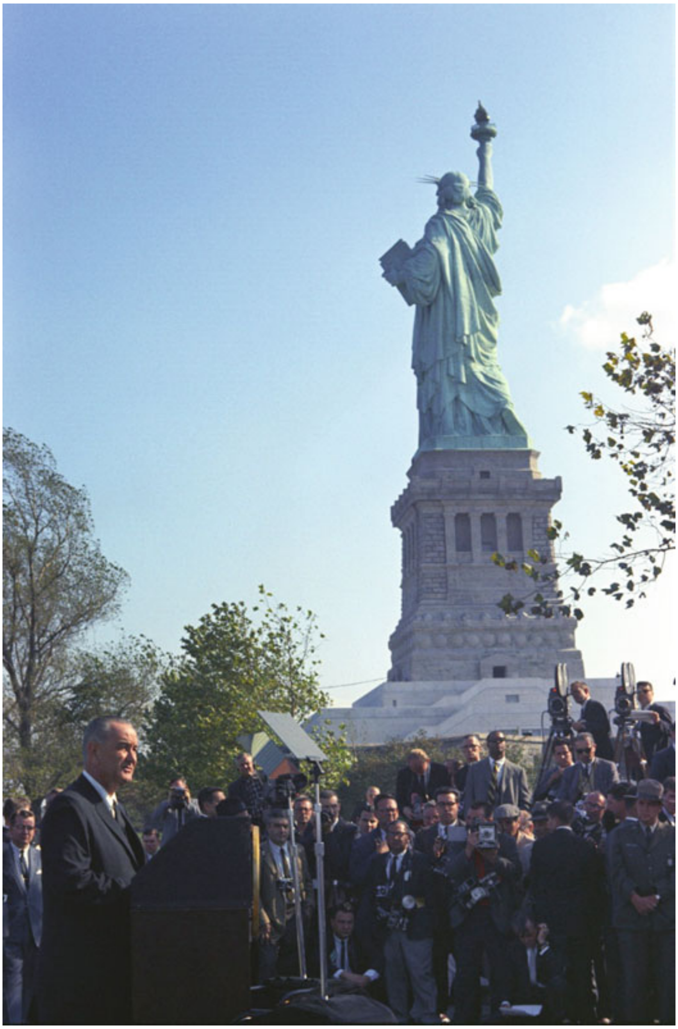


Fig. 3.4 U.S. President Lyndon Johnson signs the Immigration and Nationality Act of 1965 at the foot of the Statute of Liberty. *Source*: Photo taken by Yoichi R. Okamoto on 3 October 1965 (photo in the public domain in the United States—taken by an employee of the United States Government as part of that person's official duties)**.** http://www.lbjlibrary.org/collections/photo-archive/photolab-detail.html?id=1259

> *"You would't be related to Dr. Obama, by any chance?"* [asked a British Airways staff member at Nairobi airport in Kenya] *"Well, yes – he was my father."* [answered Barack Obama who had just arrived to Africa for the first time in his life] *"I found myself trying to prolong the conversation, encouraged [. . .] by the fact that she'd recognised my name. That had never happened before, I realized; not in Hawaii, not in Indonesia, not in L.A. or New York or Chicago. For the first time in my life, I felt comfort, the firmness of identity that a name might provide, how it could carry an entire history in other people's memories, so*

> *that might nod and say knowingly, ‘ Oh you are so and so’s son’. No one here in Kenya would ask how to spell my name, or mangle it with an unfamiliar tongue. My name belonged and so I belonged, drawn into a web of relationships, alliances, and grudges that I did not yet understand”.* (Obama 2008: 305)

3.5 Conclusion

Personal naming is an inherently human feature, most surely as old as the existence of language itself. Hence, personal names are considered to be “cultural universals”, a list of common traits present in all cultures. This chapter has reviewed a brief history of naming practices across the world, albeit generally focussing on Western European naming systems and examples. It has tried to answer the question “how we got our names?” referring to both forenames, chosen by our parents, and surnames, generally imposed through a family line and a legal system. In this review it has become clear that names in all societies come to serve two central antithetical, but complementary functions: differentiation and categorization of individuals. As a result, the latter function of names, the categorisation of individuals into a social matrix, serves well the purpose of classifying individual people into groups of communal origin, according to how both of their forenames and surnames fit within such a social matrix, framed within a global perspective. As such, this Chap. 3 has established a common point of departure of various threads that will be further explored and separately analysed in subsequent chapters. Hence, the introductory review of naming practices, together with the case study of the last section of this Chap. 3, form the basis of a more profound analysis of the identity implications of personal to classify populations in ulterior chapters. The next two chapters expand these implications for both surnames and forenames, in Chaps. 4 and 5 respectively, while Chaps. 6 and 7 bring together this parallel evidence into a joint analysis of both types of names, an obvious approach but scarcely addressed by the literature.

References

Abbott B (2005) Proper names and language. In: Carlson GN, Pelletier FJ (eds) Reference and quantification: the Partee effect. CSLI Publications, Stanford, CA, pp 63–81

Akenson DH (1984) Why the accepted estimates of ethnicity of the American people, 1790, are unacceptable. William Mary Q 41(1):102–119

Alford R (1988) Naming and identity: a cross-cultural study of personal naming practices. Hraf Press, New Haven, CT

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, DC

Barker HF (1926) Our leading surnames. Am Speech 1(9):470

Barker HF (1928) How we got our surnames. Am Speech 4(1):48

Beech GT, Bourin M, Chareille P (2002) Personal names studies of Medieval Europe: social identity and familial structures. Medieval Institute Publications, Western Michigan University, Kalamazoo, MI

Bloothooft G, Groot L (2008) Name clustering on the basis of parental preferences. Names 56: 111–163

Brown DE (1991) Human universals. McGraw-Hill, New York, NY

Bugelski BR (1961) Assimilation through intermarriage. Soc Forces 40(2):148

Cavalli-Sforza LL, Feldman M (1981) Cultural transmission and evolution: a quantitative approach. Princeton University Press, Princeton, NJ

Daily Mail (2006) Smellies and Bottoms named in surname list of shame. Daily Mail. Available at http://www.dailymail.co.uk/news/article-402920/Smellies-Bottoms-named-surname-list-shame.html

Elias N (1991) The society of individuals. Blackwell, Oxford

Faure R, Ribes MA, García A (2001) Diccionario de apellidos españoles. Espasa Calpe, Madrid

Fryer RG, Levitt SD (2004) The causes and consequences of distinctively black names. Q J Econ 119(3):767–805

Government Printing Office (1966) Public papers of the Presidents of the United States, Lyndon B. Johnson 1965. Government Printing Office, Washington, DC. Available at http://www.lbjlib.utexas.edu/Johnson/archives.hom/speeches.hom/651003.asp

Hanks P (1990) A dictionary of first names. Oxford University Press, Oxford

Hanks P (1992) The present-day distribution of surnames in the British Isles. Nomina 16:79–98

Hanks P (2003) Dictionary of American family names. Oxford University Press, New York, NY

Hanks P, Hardcastle K, Hodges F (2003) Oxford dictionary of first names. Oxford University Press, Oxford

Instituto Federal Electoral (2006) Estadísticas Lista Nominal y Padrón Electoral a 1 de Junio 2006. http://www-site.ife.org.mx/portal/site/ife/menuitem.f45fd5b18d4a2e55169cb731100000f7/. Accessed 15 July 2006

Jobling MA (2001) In the name of the father: surnames and genetics. Trends Genet 17(6):353–357

Kripke S (1972) Naming and necessity. Blackwell, Malden, MA

Lasker GW (1985) Surnames and genetic structure. Cambridge University Press, Cambridge

Lieberson S (2000) A matter of taste: how names, fashions, and culture change. Yale University Press, New Haven, CT

Lieberson S, Bell EO (1992) Children's first names: an empirical study of social taste. Am J Sociol 98(3):511–554

Longley PA, Cheshire JA, Mateos P (2011) Creating a regional geography of Britain through the spatial analysis of surnames. Geoforum 42(4):506–516

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

Manrubia SC, Zanette DH (2002) At the boundary between biological and cultural evolution: the origin of surname distributions. J Theor Biol 216(4):461

Massey DS (1995) The new immigration and ethnicity in the United States. Popul Dev Rev 21(3): 635–652

Mateos P (2007) A review of name-based ethnicity classification methods and their potential in population studies. Popul Space Place 13(4):243–263

Mateos P, Tucker DK (2008) Forenames and surnames in Spain in 2004. Names 56(3):165–184

McDonald F, McDonald ES (1980) The ethnic origins of the American people, 1790. William Mary Q 37(2):179–199

Murdock GP (1945) The common denominator of cultures. In: Linton R (ed) The science of man in the world crisis. Columbia University Press, New York, NY, p 123

Nwigwe BE (2001) Naming and being: a philosophical investigation on names and objects, with special reference to Igbo anthroponyms. Lit, Munster

Obama B (2008) Dreams from my father: a story of race and inheritance. Canongate, Edinburgh

Ostler N (2005) Empires of the word: a language history of the world. 1st American. HarperCollins Publishers, New York, NY

Petersen W (2001) Surnames in US population records. Popul Dev Rev 27(2):315

Purvis TL (1984) The European ancestry of the United States population, 1790. William Mary Q 41(1):85

Razum O, Zeeb H, Akgun S (2001) How useful is a name-based algorithm in health research among Turkish migrants in Germany? Trop Med Int Health 6(8):654–661

Recanati F (1993) Direct reference: from language to thought. Blackwell, Oxford

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

Seeman MV (1980) Name and identity. Can J Psychiatr 25(2):129–137

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

Tibón G (2001) Diccionario etimológico comparado de los apellidos españoles, hispano-americanos y filipinos, Fondo de Cultura Económica, México DF

Tucker DK (2003) Surnames, forenames and correlations. In: Hanks P (eds) Dictionary of American family names, Oxford University Press, New York, pp xxiii–xxvii

Woods RD (1984) Hispanic first names: a comprehensive dictionary of 250 years of Mexican-American usage. Greenwood Press, Westport, CN

Zabeeh F (1968) What is in a name? An inquiry into the semantics and pragmatics of proper names. Martinus Nijhoff, The Hague

Zelinsky W (1970) Cultural variation in personal name patterns in the Eastern United States. Ann Assoc Am Geogr 60(4):743–769

Chapter 4
Surnames and Genetics

Abstract There are fascinating parallels between surnames and genetics, and the extent and scale of them has only been recently revealed thanks to the availability in digital form of full population registers and new mapping methodologies. Surnames are typically patrilinearly inherited, and so are some of our genes. If families don't move from an area nor mix with newcomers over generations, surname frequencies will reflect clusters of population isolation that have a clear correspondence with lack of genetic diversity, and sometimes the development of unique dialects and language features. Starting with Darwin, this chapter weaves together the evidence of how languages, names, genes and human origins all seem to tell a similar story about our ancestral origin and the way populations have mixed, isolated themselves, or migrated over the last few centuries.

It might seem fairly obvious to the reader why surnames might be useful indicators of ethnicity. As introduced in Chap. 2, an ethnic group has been defined as "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" (Bulmer 1996: 35). A surname links the bearer with her or his ancestors for many generations, at least those on the paternal lineage. Therefore, a surname "automatically" signals an individual's ancestry and identity. However, when individuals are aggregated together into population groups according to the "cultural proximity" of their surnames, they can be very useful as proxies for such shared past and group identity, hence helping to delineate collectivities that could potentially approximate to ethnic groups.

Yet a necessary shift is required to build such approaches in name analysis, moving from the anecdotal study of individual surnames to groups of surnames, from personal and family histories to surname spatio-temporal frequencies, and ultimately to ethnic group dynamics. This shift has not been taken on board by most linguists or historians, but it was actually introduced and fully developed by human geneticists. In fact, this academic field has undoubtedly been the leader in such collective analysis of surnames as proxies to understand population structure. As it

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_4, © Springer-Verlag Berlin Heidelberg 2014

will be reviewed in this chapter, since the early work by George Darwin (son of Charles Darwin) in 1875, surname frequencies over space have been used by geneticists to untangle population socio-cultural and ancestral strata. The argument was simple, in the absence of detailed individual DNA data up to the late 1980s, surname frequencies comprised useful proxies to understand genetic structure, since both genes and surnames are passed down from parents to offspring following a set of known rules of inheritance. As they investigated further the validity of this analogy, population geneticists developed a set of materials, tools and methods of analysis throughout the Western world, that have passed the test of time and spread to other disciplines, such as statistics, public health, history, and geography. The aim of this chapter is to synthesise this outstanding work on surnames in population genetics, making it accessible to a wide audience, and building a basic understanding that will be very useful in establishing a multitude of transdisciplinar links throughout the rest of the book.

The chapter starts in Sect. 4.1 with a discussion of the concept of "populations" in human genetics. Perhaps a rather intuitive term but with very fuzzy boundaries, even by the "hard" scientific standards of human genetics, and which presents some remote resemblance (from distance) to the concept of ethnicity in social science. Section 4.2 introduces the history of "the human family", how we all evolved out of a common ecological niche in Africa, and diverged into bio-cultural lineages, best represented by linguistic diversity, and later intermixed forming a complex genetic spectrum across the world. The parallel between language and human evolution is then put in the context of the appearance and evolution of hereditary surnames since the middle ages. Section 4.3 aims to justify why surname structure might be good proxies for genetic structure in population studies. It does so by proposing a set of seven overarching principles or characteristics that summarise the value of surnames in genetics, with important parallels in the biological study of non-human populations. Section 4.4 introduces a particular technique that has predominated the analysis of surname frequencies in population genetics; termed "isonymy". A range of applications using isonymy are then presented and discussed in the last Sect. 4.5, with important implications for the spatial analysis of population structure using surnames, which will be finally established in Chap. 8. The chapter closes with a brief conclusion that brings the elements of the chapter together; surnames, genetics, language, geography and population structure, pointing towards how this proposition also applies to forenames in the following chapter.

4.1 Defining “Populations” in Genetics Studies of Human Difference

Underpinning human genetic studies is a contentious unit of observation; what geneticists repeatedly term “populations”, following a commonly accepted conception in non-human biology. All of the conclusions arrived at in population genetics or molecular anthropology are based on the assumption that such a basic unit of analysis has some validity representing a group of “similar” individuals. A population is thus conceived as “a group of individuals that may be defined according to some shared characteristic which may be social or physical. Sometimes used in a theoretical sense to mean a group of individuals in which there is random mating” (Jobling et al. 2004: 507). This last sense is generally referred to in biology as a “Mendelian population” (Cavalli-Sforza et al. 1994), characterised by a state known as “panmixia”, that is; general random mating within a population. Panmixia is easy to demonstrate when comparing different species, since the very definition of a species rests upon the idea of a community capable of producing fertile offspring. However, the mentioned concept of population is obviously much more subtle than a species, and it refers to sub-structures or strata present within a species, produced by non-random mating between individuals because for example, of physical barriers to interaction that separate strata.

Population geneticists have demonstrated that theoretical models of gene frequency distributions in populations fit well with the actual sample data analysed, except in four possible cases. According to Cavalli-Sforza et al. (1994), this situation can only happen when (a) the model is incorrectly formulated; (b) laboratory procedures are erroneous; (c) natural selection has an effect; or (d) when the population is not homogeneous, that is, it segmented into socio-economic or geographic strata that do not mate randomly with each other. Hence, in order to eliminate this last effect, the definition of population explicitly incorporates the requirement of panmixia as its main governing principle.

In another definition, populations are conceived “as collectivities of people living together and sharing a number of biological and social characteristics. Such populations differ from one another in many respects” (Boyce and Reynolds 1995: vi). These authors then describe that such differences can be: *genetic*, as a result of natural selection, migration or random genetic drift; *morphological*, as a result of genetic factors but also influenced by nutrition and other environment features; *cultural*, which can shape human fertility and mixing; and finally, *epidemiological*, through varying disease associations which lead to differential patterns of morbidity and mortality (Boyce and Reynolds 1995). It is therefore a presumption of important differences in these types of outcomes between human groups that actually justifies the whole classification effort in creating such population divisions in the first place.

Jobling et al. (2004: 276) propose a set of multidimensional criteria to define human populations: *geographical proximity*, as contact between individuals is required; a *common language*, as they must be able to communicate with each other; and shared *ethnicity, culture or religion*, as intermarriage is more likely if history and values are shared. This triad, geography, language and ethnicity (taken in a broad sense) hence form the key dimensions of human difference and identity that prevent random mating to operate at a universal level. Jobling et al. (2004) acknowledge that according to these criteria any individual may claim membership of, or classification into, more than one population. This can be especially true at different points in time and space since, as reviewed in Chap. 2, ethno-linguistic groups' boundaries are blurry and identity is a multidimensional and continuous concept. However, for the purpose of meaningful analysis in population genetics, some sort of thresholds need to be established to such continuum in order to distinguish communities of individuals which are much more likely to intermix within a population than between populations. Because of these issues, population geneticists remind readers that "a rigorous analysis should start with a definition of the 'population' to be sampled" (Cavalli-Sforza et al. 1994: 20). Despite this recommendation, many population geneticists are forced to use secondary data from studies that use different definitions of populations or set different criteria to allocate individuals to them, and which in many cases are not consistent across studies (Jobling et al. 2004).

Defining a population is thus a key problem when analysing the genetic linkages or dissimilarities between human groups, and any results on whether genetic similarities within the group and differences between groups are significant must always be interpreted in light of how precisely those populations were defined from the outset. This is an ontological problem that will persist through this book, but the core argument made in this and the next chapters attempts to make a positive contribution in helping to manage its effects.

4.2 Human and Language Evolution

> *"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).

The quote above is from Charles Darwin's *On the Origin of the Species* (1859) in which he drew a parallel between the evolution of languages and of humans, suggesting that the genealogical arrangement of—what he then called—the "races of man", necessarily had to follow a taxonomy of language families. This was an interesting proposition, probably reflecting upon previous research on language evolution and language family tree representations in the nineteenth century (Greenberg 1959). Such simple but effective parallel mirrored the way in which many scientists have actually attempted to disentangle the history of human evolution across the planet.

With subsequent advances in modern genetic techniques, population geneticists—a subfield of human genetics—have demonstrated the existence of such a relationship between, on the one hand, the genetic evolution and migration of humans across the world, and on the other, the evolution and geographic spread of languages. They have mapped human origins, gene frequency distribution, and the geographical spread and intermixing of population groups across the planet, and compared it with the evolution and geographical spread of language families as a proxy for the "cultural relatedness" between populations (Cavalli-Sforza 1997; Piazza et al. 1987). They have even compared the two evolutionary trees with the archaeological record (Renfrew 1987) and with historical sources to come to similar broad conclusions about the geography and timing of the major episodes in the ancient genetic and cultural history of the world's "populations".

4.2.1 Language Evolutionary Trees

Luigi Luca Cavalli-Sforza at Stanford University is considered the "father of population genetics" (Stone and Lurquin 2005) and has carried out a very successful stream of research projects in this area for over 40 years, summarised in his masterpiece *The History and Geography of Human Genes* (Cavalli-Sforza et al. 1994). Throughout his career he has primarily used a mother tongue language criterion to define populations (Cavalli-Sforza 1997). His justification for using this criterion 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). Beyond the advantage of standardisation, language classifications offer another key benefit to geneticists. They can be conceived as an evolutionary tree, just as Darwin suggested.

The vast majority of the 6,000 or so languages currently in existence in the world (Lewis 2009) can be arranged into a hierarchical taxonomy that relates each of them to one of a few major language families. The concept of a language family essentially means that there was once a common language from which all the subsequent languages in the family evolved. Such language evolution process, as originally intimated by Darwin's citation, was a result of population isolation and divergence, either as a consequence of migration and population expansions, or the appearance of different types of barriers to cultural and genetic contact and exchange. These barriers could be: physical features, such as mountain ranges, rivers or sea masses, no longer surmountable because of climate or technological changes; socio-political ones as a result of war, military and political decline (e.g. the Roman Empire), imperial expansions (e.g. the Ottoman empire cutting trade routes to the Far East), ancient norms; or cultural ones, such as dietary, customs, religious beliefs, etc. Therefore, there are obvious parallels between the evolution of an ancient common language into separate languages as a consequence of population isolation and divergence, and the genetic evolution of populations into the diverse human groups that can we can observe today. This has been a powerful analogy with which to justify the genetic findings made in this area.

According to one of the most widely cited language taxonomies, that of Greenberg and Ruhlen (Ruhlen 1987), there are six major language families from which most of contemporary languages come from: Congo-Saharan; Eurasian/American; Dene-Caucasian; Austro-Tai; and Pacific. These authors also propose that the respective *proto-languages* that these families represent were spoken in these regions between 40,000 and 20,000 years ago. According to this view, these six families form the "trunk" of a tree, branching out into 19 sub-families in the next level. From this second level tens and even hundreds of sub-branches emanate further out splitting several times until they finally link the currently surviving 6,000 languages, which form the "leaves" at the tip of the thinnest branches of the evolutionary tree. Each of those "splits" represent an event in which two or more languages drifted apart, initially as dialects of the same language, until they were mutually unintelligible (Greenberg 2005). This process happened for example when the Romance languages evolved from vulgar Latin around the fifth to ninth centuries, eventually giving birth to Italian, French, Castilian, Catalan, Portuguese and so on.

Moreover, language change is not only confined to these "linguistic fossils". It is in fact constantly taking place. Language is modified through the mutations we all introduce in our speech, reflecting the choices speakers make within certain strata, and through the way these innovations propagate unequally through social networks or over space (Coulmas 2005). If these "language mutations" are not counteracted by frequent contact with groups outside these social networks or places, first slang springs up, then dialects evolve and drift away, and with time become unintelligible languages (Croft 2000). The way these social and spatial processes of language change operate today are well documented in the field of sociolinguistics (Milroy and Gordon 2003). It is precisely this type of evolutionary behaviour of languages, through the effect of stratified community dynamics, which drew the attention of Darwin and subsequent human biologists. They have proposed analogies between speciation events represented by bifurcations in biological evolution, and these linguistic bifurcations as a consequence of "cultural evolution" and population dynamics.

Going back to language evolutionary trees, the way in which these are actually conceptualised is not from the past to the present but inversely. They are drawn following an inductive process that starts with the current 6,000 languages at the base, which are then linked together to their "closest relatives" back in time in a stepwise fashion, retracing evidences of the linguistic bifurcation events previously mentioned. In this way, linguists have reconstructed past language divergence episodes and with them our cultural history of populations over the last 30,000 years or so.

The trunk of the language tree is thus formed by the aforementioned six language families together with some isolates. Below them, at the "root of the tree", there is no evidence to suggest if these six families were actually once related to each other. That is, so far it has been impossible to establish whether a common proto-language to the whole of humanity that could have been spoken around 65,000 years ago ever existed. At this level, and even at much more recent events in the language evolutionary tree, some linguists place doubts on the evidence supporting the whole attempt to establish such ancient language families before written records were established. They draw a line in this field of language taxonomy between the likes of Greenberg and Ruhlen, as proponents of the "long-range" comparison, and others that are happy to stop establishing links between languages at prehistoric times, approximately 6,000 years ago (Ringe 1999). Their reasoning is that without precise written records it is very difficult to accurately establish how languages were actually spoken in the past. The "long-rangers" address this criticism proposing the comparative analysis of languages that are available in the historic period, in order to reconstruct the ancestral languages they are derived from, drawing upon universal and well known processes of language change. They argue that all languages are somehow related to each other, but the key is to assessing the degree of relation and the time depth and sequence of their points of departure (Greenberg 2000). With the necessary caveats, the "long-range" comparison view is the one adopted in this book, since it is directly relevant to the argument developed in this chapter.

Population geneticists have compared such language evolutionary trees with the genetic linkages between the contemporary populations that speak these languages, drawing branches and linkages at sub-continental and continental level. In doing so, they have corroborated the overall trails left by the geographical spread of such human groups, explaining any differences found between the genetic and linguistic linkages using historical or archaeological data. One of these differences is, for example, extreme cases of language substitution such as Spanish imposed to Native Americans, or Finno-Urgic language to Hungarians (Cavalli-Sforza 2001).

In one of these studies, 38 human groups (or populations) sampled across the world were organised into an evolutionary tree based on the similarities of certain genetic markers (Cavalli-Sforza 1997). Such a tree was then compared to how the languages spoken by those human groups relate to a taxonomy of 16 language families, finding that 11 of them correlated perfectly between the two trees and the rest could be explained with other archaeological or historical evidence. Cavalli-Sforza has refined this analysis over the years, summarising the main outcome in a diagram represented in Fig. 4.1. It shows two types of linkages between the 38 "populations" through two separate dendrograms; the one on the left represents the linkages (or distance) between their gene frequencies, while the one on the right shows the relationships between their languages through an evolutionary tree of language families.

These types of metaphorical analysis of cultural and biological "evolution" have been redrawn many times, typically showing similar striking parallels between the two. The reasoning behind the power of such analogy is that "the extent of genetic exchange is correlated with that of cultural (including linguistic) exchange, since geographic, ecological, and even linguistic barriers tend to act similarly on both" (Cavalli-Sforza et al. 1989: 1128). In a literal sense what these population

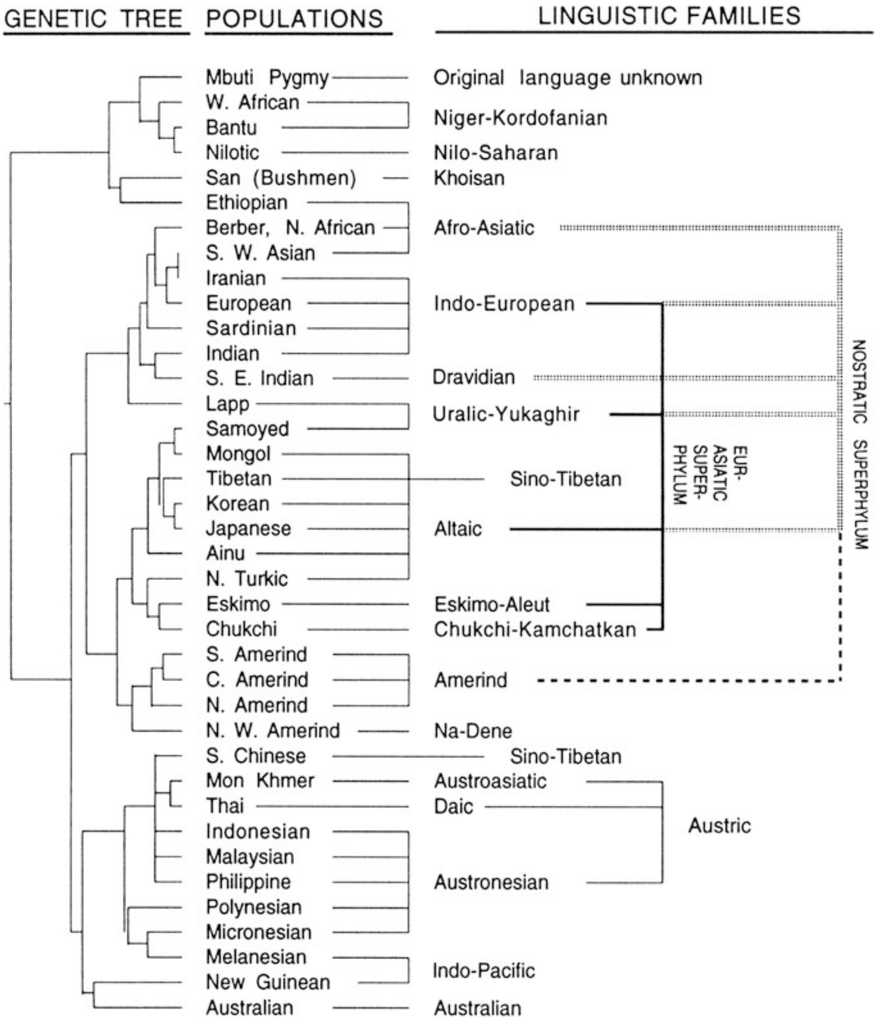


Fig. 4.1 Dendrograms showing the evolution of humans and languages. The *central column*, titled "Populations", shows linguistic or anthropological human groups. The *right column* shows the linkages on the basis of linguistic classification, while the *left column* represents a genetic tree constructed by average linkage analysis of Nei's genetic distances. *Source*: Cavalli-Sforza et al. (1994: 99) originally published in Cavalli-Sforza et al. (1988)

geneticists have done is to follow the exact path suggested by Darwin in 1859 to reconstruct "the human pedigree" based on the evolutionary tree of languages.

4.2.2 *From Evolutionary Trees to Frequency Gradients*

However, as is probably very obvious by now, these simple parallelisms do not come without problems. Although languages do evolve into isolated entities (say French and Italian), human groups do keep exchanging genes *between* language groups, except in some very extreme situations (e.g. the native population of America after migration from Northeast Asia). Only a few individual cases of gene flow between populations are required for those genes to make their way into a different population, spreading across space and diffusing over generations. Through the quest to find our "biological family history" in the age of DNA research, population geneticists have in fact unearthed an unequivocal reality no matter how our genetic makeup is analysed. This reality actually resembles much more of a continuum of human genes across continents rather than a crisp biological taxonomy with parallels in the way "speciation" is observed in the natural world (Olson 2006). Such genetic continuum found in contemporary humans across the world reflects a history characterised by a constant rate of migration and intermixing over the millennia (Sykes 1999). As is widely known, geneticists have demonstrated that we are all part of the same species; *homo sapiens,* that evolved from *homo erectus* in Africa and only then migrated to colonise the world. This view, known as the "Out of Africa" hypothesis, forms the current consensus in science, debunking alternative theories such as the multi-regional hypothesis (Wolpoff et al. 1988) that proposed separate parallel evolutions in each continent from the regional variations of *homo erectus*. They have also demonstrated that at the individual level there is much more genetic variation within population groups than between groups. Even at continental level, a selection of studies have shown that approximately 83–88 % of gene variation is found within populations while only 9–13 % between continental groups (Jobling et al. 2004: 278).

In doing so, geneticists have once and for all proved that, biologically speaking, nothing close to clear-cut human groups exist. This has been one of the greatest scientific achievements of the late twentieth century, despite the common misperceptions about biological differences between populations, namely the deeply rooted concept of *race*, that still persists today in most societies. Nonetheless, if races do not exist, and our degree of genetic similarity is so remarkable that it is now once and for all an established scientific consensus, how can population geneticists carry on insisting on their attempt to establish genetic differences between human groups? The reason is that although we are all genetically very similar, there is still a small proportion of variation in the *commonality* of certain genes between human groups. That is, what concerns these studies are variations in the *frequencies* of certain genes between populations, not the whole spectrum of genes, nor the variations between individual persons.

It must be stressed however, that the way in which genetic differences between populations are defined in the first place lies to a certain extent "in the eye of the beholder". There are two key ontological questions that concern this quest in population genetics: what actually makes a distinct human group?; and what is deemed as a significant genetic difference? Both of these key aspects need to be clearly defined by applying some sort of "thresholds" to the population continuum that unites us all through space and time. As it was argued earlier, these two types of thresholds need to be transparently and robustly established within such a genetic and population continuum. The first aspect, defining what makes a distinct human group, is very relevant to this book while the second one, conceptualising genetic differences, is definitely beyond its scope (for a review of the genetic issues see Jobling et al. 2004).

By adopting certain, albeit arbitrary, classifications with which to define populations, geneticists have demonstrated the existence of significant differences in the *relative frequencies* of certain genetic markers (usually called alleles or haplotypes) between geographical regions and human groups, sometimes very pronounced ones indeed. In other words, we might be all genetically mixed but the interest here is to discern patterns in the geographical distribution of the relative frequencies of genetic markers between populations, as opposed to just making binary decisions on present/absent genes.

Moving from individuals and particular genes to populations and overall gene frequencies introduces an important shift in this discussion. Even when distinct populations keep exchanging some genes after their languages are not mutually intelligible anymore, they are much more likely to do so with people from their own language or cultural group and, as proposed in this book, within their immediate geographical region and social networks than with more geographically and "culturally" distant groups. Following simple evolutionary dynamics, these asymmetric population exchanges in turn lead to an expansion of certain genes within a population and the retreat of others, shaping what is commonly known as a population's *gene pool* (Jobling et al. 2004). This concept in essence refers to the differential distribution of gene frequencies between groups. As such, the actual geographic variation in the frequency distribution of certain genetic markers across populations, not the individual genes themselves, form the basic material with which substantial parts of the history of human groups can be disentangled (Cavalli-Sforza et al. 2004).

As long as the previously mentioned caveats are respected, a fascinating genetic history of human evolution and migration can be partially reconstructed from these geographic variations in gene frequencies (Stone et al. 2007). Such a tale links the places and regions our ancestors inhabited, through their migration flows initially out of Africa and then back and forward between continents and regions. Population structure was then established through demographic expansions, declines and "bottlenecks", with parallels in the way cultural heritage, including language, has developed, evolved and been transmitted from one place to another and from one generation to the next (Cavalli-Sforza et al. 1982). 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 (for a review see: Cavalli-Sforza et al. 1994; Jobling et al. 2004). However, the more recent periods and micro geographical scales, below subcontinents and over the last millennia, are to a certain extent a largely uncharted and indeed risky territory in scientific discovery.

This book attempts to make a partial contribution to this literature on the most recent part of the human journey, the period that goes from the establishment of hereditary and written universal naming conventions from the Middle Ages or a few centuries ago until the present day. It also tries to disentangle regional patterns below the continental and country levels and even within territories that share the same language. This attempt should enable further insights into one of the two key ontological aspects mentioned before; what actually makes a distinct human group? It presents a set of unconventional personal names-based methods to come up with alternative definitions or dimensions of what could constitute something close to "homogenous human groups". It is proposed that through the origins of people's names, humans can be grouped in several alternative ways to study the main migration trails these have imprinted throughout most regions of the world.

4.3 In the Name of the Father: Surnames and Genetics

Moving forward in time to the more recent episodes of our human history, the evolution of languages and cultural customs produced a universal system of hereditary family naming in many countries over the last seven centuries or so, as described in Chap. 3. The fact that surnames are typically passed down systematically through generations via the male lineage (patrilinearity) makes them an ideal candidate to simulate how certain genes might have been transmitted between generations and across space. In other words, it is assumed that inheritance of surnames and biological inheritance are similar (Lasker 1985). Using surnames, human genetic diversity can be studied in more recent time periods and at much more local scales (sub-continent and sub-national level) than is currently possible with the use of genetic or linguistic information alone (Manni et al. 2005). There is a long and productive literature on the use of surnames in human genetics, with applications falling into two types of camps: the study of population genetic structure, dealing primarily with issues of endogamy and relatedness within and between populations (Jobling 2001); and the analysis of migration and cultural interaction (Cavalli-Sforza et al. 2004). For a full review of the literature on names and genetics see Lasker (1985) and Colantonio et al. (2003).

These and other authors have highlighted particular features that make surnames valuable in the genetic study of populations. We propose here to synthesise these feature in a set of *seven principles* or characteristics that summarise the value of surnames in genetics, and which have strong parallels in the biological study of non-human populations. These are dealt within each of the following seven sub-sections.

1. Surnames Originate from Particular Languages
   Surnames are a form of personal naming that were introduced within a society some time in the recent past, and were thus originated in a particular language. This may seem a somewhat obvious statement, but its implications may allow extending the language evolutionary trees, mentioned in the previous section, over the last few centuries. Since most surname systems were made hereditary between the Middle Ages and the nineteenth century in most countries, they have somehow "frozen in time" the language spoken by the family who first bore each surname, as well as the places where they lived. Hence, names in turn may allow the possibility of partially reconstructing how the people who spoke that language have expanded, migrated or disappeared from certain places. This could be done by following the different lines of ancestry of their current bearers to the places and languages of origin, even when they might not speak that language anymore, such as with millions of US, Canadian, Australian or Argentinean citizens to cite just a few examples of large scale "surname migrations" from different language homelands (see Chap. 8).
2. Most Surnames Typically Originate in a Particular Region
   Many of the surnames that currently exist are considered monophyletic, that is, they had a single origin in a particular person or family, and hence appeared at only one place. This is mainly because a large proportion of surnames in several countries are derived from placenames or topographical features that had a strong local content (Hanks 2003), alongside the need to uniquely identify an individual family at least within their area of residence. Furthermore, many other surnames have "become monogenetic" through decreasing frequencies as a consequence of the extinction of surname lineages from other areas of origin with low frequencies (Hanks 1992). Moreover, many names although polyphyletic, reflect regional dialects or even conventions in local customs and traditions, such as festivities (e.g. in East Anglia many surnames come from carnival characters, such as "Pope" or "King"). Finally, surnames, like genes, also "mutate" through new spellings, combination of surnames, and transliteration into different languages, creating in fact new unique surnames. These mutations may also be traced in many cases back to the first single individual who bore it and sometimes the place in which it occurred, introducing through them the same effects described for the original bearer (Sykes and Irven 2000). This single characteristic alone, the unique regional origin of surnames, allows for tracing the place of origin of a large proportion of surnames.
3. Couples Do Not Meet and Reproduce Randomly, and Typically Only One Surname Is Inherited by Their Offspring
   Population geneticists refer to the term *assortative mating* to "a departure from random mating or panmixia" (Mascie-Taylor 1995: 86). This phenomenon can take place in two directions: positive, when two genetically similar persons are attracted and have offspring, resulting in homogamy; or negative, when people with opposite genetic make-up procreate, resulting in heterogamy. It has been demonstrated that marital choice and reproduction tends to operate within non-random processes influenced by geographical, linguistic, religious, ethnic,

educational, and socio-economic factors. For example, studies have shown that young people tend to marry those who live nearby (Mascie-Taylor 1995). Such processes result in assortative mating having a strong role in how people mix and hence how genes are passed down the generations (Cavalli-Sforza et al. 2004). Surnames mimic well these processes of assortative mating. In most countries, out of the two surnames bore by a couple only one of them is passed on to their offspring in the next generation.[1] Through the process of assortative mating, certain surnames within social and geographical subgroups are indirectly promoted through continuous reproduction and expansion in detriment to others outside the mating circles which end up being displaced from a social group or geographical area or even in extinction. Moreover, the effect of this socio-geographic assortative mating has left a trail in marriage records where the same pairs of surnames may be found repeating themselves in an area through generations of marriages between the same families. Lasker and Kaplan (1985) formalise this effect in a summary statistic termed *repeated pairs* (RP), which is a measure of endogamy by social class. When the repeated pair is actually the same surname, the degree of consanguinity or inbreeding is even more extreme, a phenomenon measured in *isonymy* studies that will be reviewed in the next section.

4. Genetic Drift Is Reflected in Surname Frequencies

   Genetic *drift* refers "to the effect of chance on gene frequencies in successive generations" (Cavalli-Sforza et al. 2004: 18). Drift measures the change in the relative frequency with which a gene variant (also called *allele*) occurs in a population due to the random survival and reproduction probabilities of the genes present in the previous generation. Drift is a sole function of the relative frequencies of gene variants present in a population (gene pool) and the size of the population (N), and its effects can only be counter-balanced by migration. If an area does not have a large population, people do not reproduce beyond a local area, and no outsiders migrate to the area, the gene pool in the area experiences a fast rate of change attributable to drift. This situation is better known as inbreeding, which is just one particular case of the effects of drift. Carrying on with the surname analogy, changes in the surname frequencies of a population reflect processes of genetic drift, since the probability that a surname is passed on to the next generation is also a function of the surname's relative frequency, the population size and migration rates. If no new surnames enter or leave the area through migration and the population is small, the relative frequencies of common local surnames keep rising, displacing rarer surnames, and hence revealing signs of inbreeding.

[1] In the Spanish naming system the two surnames survive as both paternal and maternal surnames make up the two surnames passed onto their offspring. However, both surnames in effect reflect patrilinear heritage, since the maternal surname actually comes from one's maternal grandfather, and therefore the only difference with other naming systems is that it takes two generations to loose the maternal surname instead of simply one in the anglo-saxon system (see Mateos and Tucker 2008 for a full explanation)

5. Migration May Radically Change the "Surname Landscape" of an Area
   Migration has an important effect in the genetic structure of populations when a large number of people move between areas that have different gene frequencies. Such movements, carrying with them what is known as *gene flow*, alter not only the frequencies of the gene pool of the area of destination, but also of the area left behind (Mascie-Taylor and Lasker 1988). Therefore migration counteracts the effects of genetic drift and population isolation. When migration between two populations is substantial and sustained in both directions the result is population homogeneity. If it only happens in one direction, the process is termed *demic diffusion*, as when a large population moves into an area displacing, replacing, or intermixing with a pre-existing population (Cavalli-Sforza 1997). All of these migration processes have their parallels in surname frequencies, which are substantially affected by these events that introduce, move, deplenish, replenish, extinguish, divert and split populations bearing particular surnames and groups of surnames across space. These changes in the surname frequencies, both at origin and destination, alter the survival and reproduction probabilities of surnames. As with genes, they also face the effects of two other population genetics phenomena associated with migration—*founder effect* and *population bottlenecks*. Founder effect is "the limitation in genetic variability which can be ascribed to the small number of original members" (Lasker 1985: 142). For example, the surnames of the first European settlers in Latin America have been found to have a strong founder effect, in that they are very common today in particular populations (Bedoya et al. 2006). A population bottleneck is an evolutionary event in which a sudden decrease in a population's size drastically reduces the gene pool of the area. This typically takes place when a significant percentage of a population dies or emigrates from an area before their reproductive age is over. These types of events also reduce the surname pool in a population, substantially altering the dynamics of surname evolution in the group (i.e. rarer names might become more frequent and vice-versa). For example, in Ireland it has been found that unlike other countries, the frequency of a surname is not related to diversity of the Y-chromosomes of their bearers (McEvoy and Bradley 2006) which amongst other causes could be attributed to past epidemics or, in other words, population bottlenecks (King and Jobling 2009b). The typology of different events as reflected in surname frequencies has been summarised by Manni et al. (2005) in the diagram shown in Fig. 4.2. It presents six different hypothetical situations of a surname's frequency distribution as a result of processes of population expansion, extinction, migration and parallel evolution over time since the origin of surnames (in this case eight generations since it draws on data from The Netherlands). Certain precautions need to be taken into account when analysing the contemporary frequency distribution of surnames over space, since some of their historic or demographic processes might be hidden to us, as is clearly shown in this diagram.

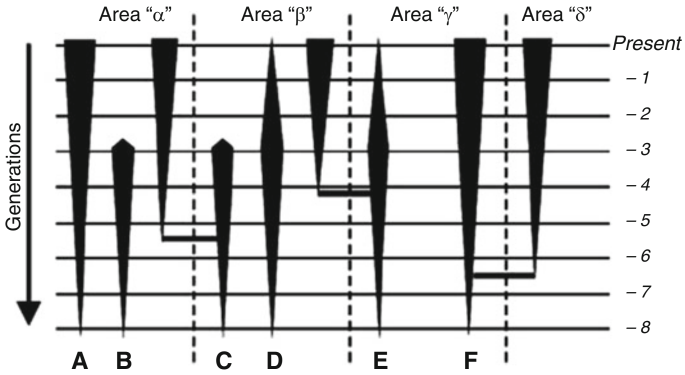


Fig. 4.2 Models of evolution of surname frequencies across space. (A) A surname geographically specific to an area, and that has experienced a population expansion in its frequency over generations. (B) A surname that became extinct in the past, and is only known through historical records. (C) A surname which has migrated to a different area where it has experienced an expansion in frequency while becoming extinct in the original area. The geographic origin cannot be inferred from its current spatial distribution. (D) A rare surname that might have gone through a population bottleneck in the past, with few individuals to accurately infer its geographic origin. (E) A combination of situations C and D, where the surname is currently more frequent in an area different from that of where it originates. (F) A surname that underwent a migration episode at a time close to the origin of surnames, and that whose present frequency distribution is balanced between two areas, what could lead to ambiguous results in unveiling its geographical origin (*Source*: Manni et al. 2005)

6. Y-Chromosomes Are Patrilinearly Inherited Just Like Surnames
   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 study the genetic linkages between the contemporary bearers of the same surnames. Those surnames that have a rare frequency, and are locally distributed, are more likely to have been "founded" by a single individual (monophyletic surnames), and hence are very prone to show close genetic linkages between their contemporary bearers.

   A good example of this type of studies is an investigation by King and Jobling (2009a) whose results are shown in Fig. 4.3. This figure plots the network of genes (*haplotypes*) of more than 100 male individuals in Britain who happen to share one of nine particular surnames (labelled *b*–*j*) compared to a control population of 110 individuals which do not have any surnames in common. The figure shows median joining networks linking the haplotypes shared within each surname, providing a powerful way to identify and visualise the genetic relationships between their bearers. The network of haplotypes among the control group (labelled *a* in Fig. 4.3) is composed overwhelmingly of singleton haplotypes, while the most common British surname, *Smith* (labelled *b*), behaves in a very similar way. However, the majority of the other surnames (labelled *c* to *f*) are very different from the control group showing one or a few haplotypes

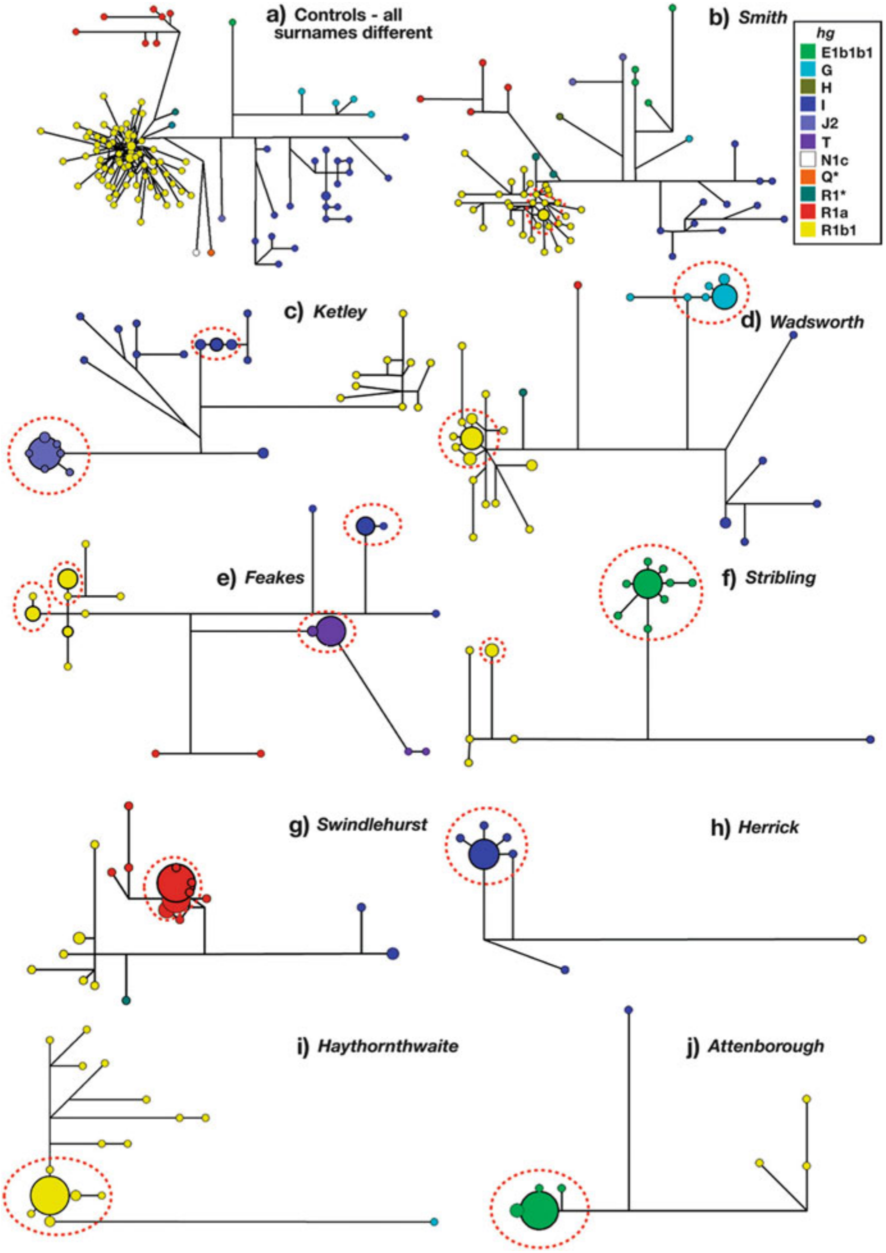


Fig. 4.3 Median-joining network of Y-chromosomes in nine surnames (labelled *b–j*) and a control group (labelled *a*), Selected median-joining networks showing Y-chromosome genetic diversity (haplogroup and Y-STR haplotype) within controls and nine surname samples. Each *circle* represents a haplotype, with areas proportional to frequency and coloured according to haplogroup as shown in the key, *top right*. *Lines between circles* represent mutational steps, with the shortest line in each network representing a single step. Boundaries of descent clusters are shown by the *dotted ellipses*. *Source*: (King and Jobling 2009a)

shared by many individuals within the surname, but not across surnames. Even more, the authors point out that people with different spellings of what originally must have been the same surname, actually share exactly the same haplotypes, pointing to the relatively recent fixation of surname spellings. For example, *Ketley* (labelled *c*) is dominated by a single haplotype circled in a dotted line, which is barely present in all other surnames, clearly pointing to a common ancestor for the 20 bearers that were sampled with this surname. These authors define a methodology with which to identify what they term "descent clusters" of haplotypes within the wider haplogroups to which these belong.

Therefore, if the necessary caveats of name change, non-paternity, 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). 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). For example, for people who do not know who their immediate ancestors were, there are now commercial services offered to search DNA databanks for their most similar genetic markers, and to obtain the common surnames associated with those markers (Sorenson Molecular Genealogy Foundation 2007). Recent research in this area has even suggested the possibility of the police being able to identify a suspect's surname by his/her DNA traces (Jha 2006), although this is a highly speculative application.

7. Cost and Availability

   Historic surname registers are not only much easier to obtain than gene frequencies from past populations, but sometimes they are the only traces left of the past inhabitants of particular regions. This makes surnames a very efficient alternative for historical, demographic and anthropological research (Smith 2002). Moreover, even for contemporary populations "surnames provide a quick, easy, cheap and crude way to study human inbreeding and migration" (Crow 1983: 383), since they are more readily available and have a much larger coverage than the alternative option of collecting DNA samples from individuals. Because of this cost advantage they have been termed "a poor man's population genetics" (Crow 1983: 383), since they cannot obviously replace genetic data but can provide very good insight into issues of population structure at a fraction of the cost of the alternative methods.

To summarise, these seven principles or key characteristics of surname analysis in genetics deal with aspects of: language and region of origin; features of assortative mating and inbreeding; processes of genetic drift and migration; relatedness through patrilinear lineages; and the cost effectiveness of the method. The rest of this chapter will draw upon the combined effect of these seven characteristics harnessed through the most frequently used technique in surnames and genetics literature; the analysis of *isonymy*.

4.4 The Isonymy Method

The "discovery" of a statistical relationship between the frequency distribution of surnames and population structure is attributed to George Darwin (1875), son of Charles Darwin. George's parents were first cousins and he was very interested in disproving the apparent deleterious effects of consanguineous marriages. In order to find out the rate of first cousin marriages, he computed the relative frequencies of same surname marriages from various records (Lasker 1985). He then made an interesting link between the relative frequency of a surname in an area and the probability of a "same surname marriage" happening at random. Any unexpected difference from randomness therefore should be interpreted as a sign of inbreeding, that is, non-random mating. He then adjusted this figure to account for the theoretical rate of same surname marriages within all first cousin marriages. As a result, he produced estimates of the degree of first cousin marriages to be 4.5 % among the aristocracy, 3.5 % among the middle classes and landed gentry, 2.25 % in the general population of rural districts and 2 % in the cities (Lasker 1985). These results demonstrated that first cousin marriages were relatively common amongst the more affluent classes of nineteenth century Victorian Britain. Moreover, to his advantage he demonstrated that the known prevalence rates for different mental disorders or hereditary diseases at the time were not related to the rates of first cousin marriage. His most interesting discovery however is in establishing an association between the relative frequency of same surname marriages with the probability of endogamy.

Since George Darwin, several researchers throughout the twentieth century have studied the relationship between surnames and population structure, building models that were then compared with the genetic evidence. Gabriel Lasker has been one of the most arduous proponents of the links between surnames and population structure. He considers that the base premise in these studies is the fact that "[s]urnames are not distributed homogeneously in different places and among different social groups" (Lasker 1985: 5). Following on from this observation, he adds that "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*, which literally means the frequency of repetition of the same surname (Lasker 2002). Isonymy can be calculated not only between marital partners but, more importantly, between places or between population groups (Smith 2002). Marital partners' isonymy is not so relevant to this book's theme, so hereafter the meaning given to isonymy will generally refer to that between places or between human groups.

4.4.1 Measuring Isonymy

The seminal paper in surname studies within human genetics is considered to be Crow and Mange's (1965) technique and theory of marital isonymy. They led the foundations for the study of inbreeding and population structure through surname frequencies (Smith 2002). Crow and Mange (1965) based their findings upon the observation that in most consanguineous marriages, there is a constant relationship between the probability of such a marriage being isonymous (where both partners had the same surname before marriage), and the inbreeding coefficient of their offspring. This constant could be expressed as the inverse of the likelihood of having the same surname multiplied by the degree of relationship. The degree of relationship is also called the *inbreeding coefficient*, which can be defined as the proportion of genes from a common ancestor inherited through each person's father and mother. For example, offspring of brothers and sisters have an inbreeding coefficient of 1/4, while they carry the same surname. However, offspring of aunts and uncles with nephews have an inbreeding coefficient of 1/8 and bear the same surname in approximately one half of cases, thus also yielding 1/4 (1/8 × 1/2), while offspring of first cousins have an inbreeding coefficient of 1/16 and share the same surname in a quarter of the cases (because of different male-female potential combinations), hence also yielding 1/4 (1/16 × 1/4). If this rationale is carried over through different combinations of inbreeding situations the value of 1/4 remains constant. In fact, what Crow and Mange demonstrated with this relationship is that a quarter of the offspring from isonymous marriages (i.e. those between partners with the same surname) are actually consanguineous descendants, while the rest are simply due to random isonymy. Hence, the *random marital isonymy* is given by

$$\mathrm{F1} = \sum_{i} \frac{\mathrm{p_i\, q_i}}{4} \tag{4.1}$$

where

$\mathrm{p_i}$ = relative frequency of the ith surname in grooms

$\mathrm{q_i}$ = relative frequency of the ith surname in brides' maiden name

Random isonymy has been subsequently used as a measure, in its own right, of kinship, relationship or genetic distance between populations (Smith 2002).

Lasker (1985) then proposes to move beyond marital isonymy to *random isonymy* to establish the relationship between two populations, instead of between two marriage partners. According to Smith (2002), this an important shift since it frees the isonymy method from marriage associations "permitting the relationship between populations to be interpreted in terms of broader determinant processes, which have mate choice or migration as outcomes, such as geography, religion, ethnicity and socio-economic factors" (Smith 2002: 119). This is indeed a key observation which lies at the core of this book's proposition; that the geographical, cultural, ethnic and linguistic relationships between populations across space, can be partially disentangled through analysing the outcomes of mating and migration decisions as reflected in people's name frequencies and geographic distribution. Moreover, random isonymy is a very efficient method, since it can be calculated cheaply and quickly from any universal list of names, and does not require access to sometimes incomplete or hard to access marriage records or historical population registers (Smith 2002).

Based on Crow and Mange's (1965) original proposition, Lasker (1977) introduced the concept of the *Coefficient of Relationship by Isonymy* (R_i) between populations defined as "the probability of members of two populations or subpopulations having genes in common by descent as estimated from sharing the same surnames" (Lasker 1985:142). It is expressed as half the proportion of isonymy between two populations:

$$\mathrm{Ri} = \sum_i \frac{p_{iA}\,p_{iB}}{2} \tag{4.2}$$

where

p_{iA} = relative frequency of the ith surname in population A

p_{iB} = relative frequency of the ith surname in population B

This formulation is based upon the observation that the degree of relationship (inbreeding coefficient) of the offspring of a consanguineous marriage is half the size of the degree of relationship between the consanguineous parents (Smith 2002). For example, first cousins have a degree of relationship of 1/8 while that of their offspring is 1/16, meaning that the degree of relationship between generations vary by 1/2 and hence the denominator of two in the Eq. (4.2) above. The value of R_i between two populations can be interpreted as the proportional correspondence between their population structures, in terms of a shared *surname pool* (Schürer 2004). This concept of the surname pool, carries over the parallel between surnames and genes mimicking the concept of a gene pool in the population genetics literature. The relative frequencies of both surnames and genes in a the population not only determine these two items' survival probabilities over generations but, when comparing two or more populations, surnames and genes also reflect their overall diversity in terms of population structure. In fact, what the concept of isonymy achieves is the comparison between the diversity of two populations' surname pools.

Whilst the Lasker's coefficient of relationship by isonymy (R_i) remains a dominant measure in surname and genetics research, it has been extended to create a distance measure between two populations or two geographical areas. This measure is widely known as the *Lasker's distance* between localities or populations (Barrai 2002; Barrai et al. 1997; Rodriguez-Larralde et al. 1998b) and is calculated as:

$$L_i = \text{-}\ln\,(2\,R_i) \tag{4.3}$$

where

L_i is the Lasker distance between two separate populations A and B

R_i is the coefficient of relationship by isonymy between the two populations as defined in Eq. (4.2)

The logarithmic transformation of twice the value of R_i is introduced here to make more prominent the differences in isonymy between populations. The Lasker's distance usually shows a strong relationship with the logarithmic transformation of geographic distance between populations, expressed for example as an Euclidean distance (Rodriguez-Larralde et al. 1998b). However, there are some interesting exceptions to such relationship primarily as a result of migration, a phenomenon that will be investigated further in this book.

4.4.2 Interpreting Isonymy

The Lasker's distance of isonymy can be thus interpreted as a measure of difference between two populations in surname space, since the greater it grows the less similar the composition of surname frequencies between two areas is. In other words, it measures the degree of difference in their surname pools.

Since the Lasker's distance of isonymy is highly correlated to geographic distance, it has also been expressed in terms of a classic human genetics phenomenon known as *isolation by distance* (Wright 1943). This refers to the divergence in the genetic differences between groups as their geographic distance increases or, its inverse process, a greater affinity the closer they are in physical space. Isolation by distance, reflects the tendency of populations to exchange genes with nearest neighbours rather than those far apart, what substantially contributes to a population's genetic drift (Cavalli-Sforza et al. 1994: 52). In Geography, the exact same property is termed *Tobler's First Law of Geography* "that everything is related to everything else, but near things are more related than distant things" (Tobler 1970: 236).

Several authors have interpreted the Lasker's distance in terms of the mentioned link between, on the one hand, genetic inheritance and, on the other, cultural, ethnic and linguistic factors influencing mating choice and migration. Such link is based upon the assumption that two populations that are closer both in "surname and genetic space" (hence more homogenous) are more likely exhibit less differences in cultural behaviour than more distant ones in both dimensions (Barrai et al. 1996; Rodriguez-Larralde et al. 1998a, b, 2000; Scapoli et al. 2007). These and other authors compare the differences in isonymy between population groups, defined across space (i.e. inhabitants of certain regions or cities), with differences in their gene frequencies searching for the identification of homogenous regions. Such homogeneity is usually expressed in terms of maximising their internal similarity while minimising that between regions. They also use isonymy to detect barriers to human interaction of gene and cultural exchange as reflected by abrupt changes in the distribution of gene and surname frequencies. In doing so they aim to uncover "geographic patterns of genetic, morphologic, and linguistic variation" (Manni et al. 2004: 173) or, in other words, the effect of the process of isolation by distance. The underlying hypothesis of this type of isonymy studies is very similar to the ideas described in the first section of this chapter at a continental or language level. That is, that areas or groups that have freely exchanged individuals between them, either through migration or mating, 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 as a direct consequence of inbreeding.

4.4.3 Assumptions in Isonymy

However, in order to sustain the genetic parallel that has been drawn in the isonymy literature, moving from surname to gene diversity, a set of core assumptions needs to be introduced (Crow 1983; Smith 2002). Firstly, surname inheritance should follow systematic rules that mimic genetic transmission over generations. Secondly, surnames should have a single original ancestor from whom all of its current bearers descend (*monophyletic surnames*). Thirdly, there should be an equal probability of migration between males and females, otherwise surnames spread geographically when males move, but not when females migrate.

As it may seem obvious at this point, a set of practices in the real world actually conspire against these assumptions. In respect to the first assumption, not all countries or societies have a method of fixed surname inheritance, as the patronymic system still present in Iceland (see Chap. 3). Likewise, processes of name change, adoption, illegitimacy, errors of spelling, etc undermine the effectiveness of the method (Crow 1983). For example, every time that a woman changes her surname to her husband's surname upon marriage, or when a surname is phonetically transcribed, translated or simplified into a different language after migration (Petersen 2001), or a child carries a surname not taken from his biological father (known as *non-paternity*), the link between genetic inheritance and surnames is broken. In spite of these situations, most people in the majority of developed countries possess a surname allocated by a system of immutable and inheritable family names that has been functioning for at least the last century or so, and that in essence resembles the general processes of genetic inheritance, at least on the male lineage (Smith 2002). The problem is thus mainly reduced to "non-paternity, child adoption, and matrilineal surname transmissions [since they] will act to introduce exogenous haplotypes into a surname" (King and Jobling 2009a: 1093). However, it has been estimated that the probability of non-paternity is rare, with rates below 5 % per generation (McEvoy and Bradley 2006; Sykes and Irven 2000) or in some cases even below 1 % (Bedoya et al. 2006). Finally, cases of matrilineal surname transmission have also been traditionally rare in most developed societies, although it must be acknowledged that they are increasing in frequency in recent decades through a rise in single-motherhood or couples exercising a more equal choice of surname now recognised in several countries (for example in Spain since 2006).

The second assumption, monophyletic surnames, is the most critical to the method and hence has generated greater concern in the literature (Rogers 1991). The key issue is that, as is well known, many surnames are polyphyletic (i.e. were introduced through different individuals), such as those derived from occupations (metonyms), for example Smith, or Taylor, which are almost ubiquitous today. However, the majority of surnames in many countries, for example in Britain (Hanks 1992) or Spain (Faure et al. 2001), are deemed to be monophyletic or have "become monogenetic" through decreasing frequencies as a consequence of surname lineage extinction (Hanks 1992). Monophyletic surnames seem to be the rule because the vast majority of surnames are derived from phenomena that present distinct local and regional variations in naming patterns, such as placenames, nicknames, topographic features or localised customs. For example, it is estimated that in Spain 58 % of the surnames are derived solely from placenames (toponyms) (Faure et al. 2001). Furthermore, several other genealogists and scholars of name studies have found time and again that a large number of surnames are believed to originate in a single origin (Hey 2000; Smith 2002).

The validity of the third assumption, an equal probability of migration between males and females, needs to be put into the context of a particular geographical scale. It is well known that migration can be gender selective in certain areas depending on local customs and the characteristics of the labour market. There are two basic types of traditions in marriages with a spouse coming from a distant area. Those localities where it is customary that men remain in their birth area are known as *virilocal settlements*, while those where it is women who remain in their area are termed *matrilocal settlements*. When men migrate to another area they "propagate" their surname to a different population, while if it is women who move the migration flow remains unnoticed in studies of isonymy. Therefore, equal probability of migration by gender between areas is required for the isonymy methodology to be effective. When comparing migration rates derived from surname frequencies between areas with actual migration statistics from the Census of Population, it has been demonstrated that in reality both are highly correlated (Piazza et al. 1987). This seems to indicate that although there might be local or temporal differences in probabilities of migration by gender, at a regional level these are not noticeable since these asymmetrical flows by gender probably end up cancelling each other out, although more work is needed in this area to measure the specific gender effect in isonymy.

Through the discussion of these three assumptions, together with the core criticisms of isonymy and some of the counter-arguments offered by the literature, it is hoped that the validity of this method for the purposes presented in this book, has been sufficiently established. Of special relevance to this book's aims is that isonymy can be borrowed from the genetics field, and conveniently applied to the analysis of population structure, read through a cultural and geographical lens.

4.5 Applications of Isonymy

The methods, data and technology available to study isonymy in entire populations were initially very limited. Human biology studies of surnames were traditionally reduced to the name structure of isolated villages or valleys (Biondi et al. 1990;

Lasker et al. 1984), provinces (De Silvestri and Guglielmino 2000), ethno-religious groups (Jorde and Morgan 1987; Tasso et al. 2004), or just a handful of surnames (Kaplan and Lasker 1983). Since the late 1960s until the early 1990s researchers were only able to manually access, digitise and process individual Parish or marriage records (Lasker 1985), or in some cases telephone books of certain cities. The exception is again a group led by Cavalli-Sforza who had access to the whole telephone directory of Italy on magnetic tapes since the late 1970s (Cavalli-Sforza et al. 2004; Piazza et al. 1987), although they also had to sample cities or surnames to be able to process millions of records with the computing resources available at the time.

4.5.1 National and International Analysis of Isonymy

In the early 1990s large population registers, such as electoral registers or telephone directories, started to become readily available in digital form, which triggered an explosion of interest in analysing the name structure of whole populations, as opposed to just local areas or specific surnames. A team of population geneticists with different linkages to the University of Ferrara in Italy has been an unquestionable pioneer and leader in this stream of studies. This loosely tied group of researchers, that we could term the *Ferrara School*, has published their findings for the whole population of the following 12 countries: Austria (Barrai et al. 2000); Switzerland (Barrai et al. 1996; Rodriguez-Larralde et al. 1998b); Germany (Rodriguez-Larralde et al. 1998a); Italy (Manni and Barrai 2001); Belgium (Barrai et al. 2003); France (Scapoli et al. 2005). the Netherlands (Manni et al. 2005); Spain (Rodriguez-Larralde et al. 2003); Albania (Mikerezi et al. 2013); Venezuela (Rodriguez-Larralde et al. 2000); Argentina (Dipierri et al. 2005); and the US (Barrai et al. 2001); Outside of this research group and the mentioned countries, most of the analyses available have focussed on a sample of surnames, rather than complete populations, in the following countries: England and Wales (Mascie-Taylor and Lasker 1985; Sokal et al. 1992); Scotland (Holloway and Sofaer 1989); Taiwan (Chen and Cavalli-Sforza 1983); China (Yuan et al. 2000); and Russia (Balanovsky et al. 2001). A complete review of the literature in this area until the early 2000s has been compiled by Colantonio et al. (2003).

The *Ferrara School* has also produced an extensive comparative review of the surname distribution of the total population of eight countries in Western Europe (Scapoli et al. 2007). They analysed a total of 26.2 million unique surnames in eight countries, whose frequencies were studied for 125 regions and 2,094 towns and cities, concluding that the present surname structure of Western Europe is intimately linked to the geographical distribution of languages. A similar conclusion was reached by a team of researchers outside genetics at University College London Department of Geography. This team, of which this book's author is a member of, used advanced spatial analysis techniques to analyse the isonymy measures for 16 European countries, derived from the surname frequency distribution of the total population (Cheshire et al. 2011). In doing so, barriers to surname exchange separating “cultural regions” were plotted, and all were found to correspond to linguistic boundaries.

Underpinning the conclusions of all the isonymy studies mentioned above is a common characteristic; at the international or sub-continental level, surname structures correlate almost perfectly with the major linguistic boundaries. A very similar type of geographic and linguistic structure between European populations has been recently found using a large survey of genes, in a widely acclaimed paper published in *Nature* (Novembre et al. 2008). This parallel between surnames and languages over space is somehow an expected finding. After all, in Western Europe both surnames and languages have been “fixed to places” for several centuries, while most nation states modern boundaries were not fixed until the nineteenth century. There are however some interesting exceptions to this association which reflect substantial migration flows or movements in national or linguistic boundaries all imprinted in surnames’ distributions. For example, Flemish migration to Wallonia (Belgium) and Northern France in the early twentieth century (Poulain et al. 2000), Italian migration to Southern France around the same period (Degioanni et al. 1996), or border movements between Germany and Denmark (Boldsen and Lasker 1996), and between Germany and France (Cheshire et al. 2011). Unfortunately, no such studies are available from Central and Eastern Europe (published in English), where political and linguistic boundaries have changed substantially over the twentieth century.

4.5.2 Sub-national Analysis of Isonymy

At the sub-national level, linguistic differences are also detected through isonymy although the boundaries are much more blurry. For example, the four different languages present today in Spain can be ascertained from the geographic distribution of surnames (Mateos and Tucker 2008), where a study of people with Basque surnames has analysed domestic migration flows and intermarriage patterns using very detailed population registers (Aranda Aznar 1998). Similarly in other multilingual countries, such as Belgium (Barrai et al. 2003) or Switzerland (Barrai et al. 1996) such internal linguistic boundaries are clearly detected using the isonymy method.

Beyond identifying boundaries between clearly distinguishable languages, the isonymy method has also been useful in detecting dialect regions within the same language. This analogy between population structure, given by isonymy, and dialect regions is well illustrated in a study conducted in France by Scapoli et al. (2005) which proves the value of isonymy analysis in broader population studies beyond genetics. Figure 4.4 shows two maps of France comparing surname clusters derived from isonymy between departments (map “a” on the left) with dialect clusters derived from dialectometric distances (map “b” on the right) between France’s *départements* (Scapoli et al. 2005). Departments are grouped in

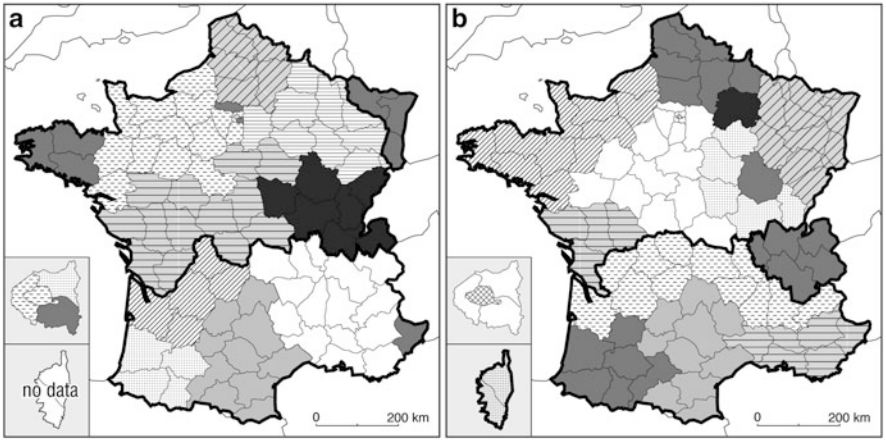


Fig. 4.4 Maps of France's (**a**) surname and (**b**) dialect distance clusters between departments. The map on the *left* (**a**) shows surname clusters derived from isonymy (Lasker distance) between departments, and the map of the *right* (**b**) represents dialect clusters derived from dialectometric distances between departments (see Scapoli et al. 2005 for full details). The *darker outer boundary* delimits the main regions with largest distances, while the *shading* within each of them groups adjacent departments together according to their closest distance in surname and dialectic space. Identical *shading* in non-contiguous clusters do not represent any type of association between clusters, but is just used here for ease of reproduction in black and white. *Source*: Redrawn based on Scapoli et al. (2005: 83–84)

each map into clusters at two levels in a regional hierarchy, the higher one dividing the country in three or four macro-regions depicted by the thickly outlined boundary, and a finer level of sub-regions identified by common shading. The similarities between surname and linguistic regions are striking, given the two independent data sources. The most important boundary is the east-to-west border that splits the country in two halves, northern and southern France, that can be interpreted as the former territories were the Langue d'Oil (north) and the Langue d'Oc or Occitan (south) were spoken. Below this macro-regional level the clusters shown capture similar areas of homogeneous populations as a result of proximal interaction, with the main differences between the two maps being interpreted as recent episodes of dialectic or surname replacement since surnames were introduced in the Middle Ages. One of these episodes is, for example, the Alsace region in the north-east which only became part of France after the First World War and that is only noticeable in the surname map (Germanic origin surnames). Another example of these recent episodes is the Franco-Provençal dialect region, in the departments of Isère, Rhone, Haute-Savoie, Savoie and Ain, where French and Provençal have mixed and evolved as the south-eastern border of France has also moved over the last five centuries.

These types of sub-national and sub-language studies comparing the surname and the dialectical distance between populations can be in fact interpreted as a "local extension" to the parallels drawn by Cavalli Sforza between genetic distance and language family evolutionary trees at continental or sub-continental level (shown in Fig. 4.1). As such, population structure over space can be expressed as a continuum in terms of genetic, linguistic/dialectic and naming similarity that, albeit with some exceptions (overlaps and folds), shows clear clines in the frequency distribution of these three dimensions. Such clines typically reveal an overall *distance decay* function in the affinity between populations, whose intergroup boundaries become more crisp as the scale of analysis moves up from the town/valley, to region, country and sub-continental levels. This thread of research into the spatial distribution aspects of surname frequencies is further expanded in Chap. 8.

4.5.3 Socioeconomic Strata in the Analysis of Isonymy

Finally, beyond the spatial patterns found in population structure, generated through the process of isolation by distance, isonymy studies can also be applied to study internal population sub-structures, for example socio-economic strata. As previously mentioned, these are derived from non-random processes of people interaction, what in population genetics is termed departures from panmixia or non-random mating. In a unique example of applications of isonymy for the historical analysis of populations, Malcolm Smith (2002), a historical anthropologist, demonstrates the value of using surnames to disentangle historical occupational structures in two English coastal villages. His analysis is based upon the surnames and occupations of the population in five historical censuses between 1841 and 1881. He calculated the coefficient of relationship by isonymy (R_i) between pairs of populations defined by the following occupations; fishermen, farmers, agricultural labourers, coastguards, mariners, ship-owners, and others. According to the degree of similarity in the surname frequencies (isonymy) between occupations over the five decennial censuses, the study finds three clear clusters of occupational sectors (a) *maritime*, including fishermen, mariners and ship-owners; (b) *agricultural*, including agricultural labourers and farmers; and (c) *coastguards*. The interpretation of these clusters is that surnames, and hence people, are primarily "exchanged" within one occupation or sector and not between them. In both villages the tightest cluster is that of the maritime trades, and within it the fishermen, reflecting a lower degree of mixing with the rest of the population and the power of a strictly hereditary profession that does not readily admit newcomers (Smith 2002). On the opposite extreme, the most "sparse" cluster is that of the coastguards, whose distribution of surnames changes rapidly from one Census to the next and is always distant, in isonymic space, from the rest of the community. This is an extremely interesting group since "the men serving as coastguards were in short-term postings [from the Admiralty] and could usually be expected to come from outside the local community" (Smith 2002: 120), what ends up clearly reflected in an sparse and distant cluster from the core of the two other "native clusters" (maritime trades and agricultural).

This example helps to illustrate the potential of name analysis to study relationships between the structure and mixing of populations, not only over space and time, but also between social strata. Moreover, this book proposes that such analogy can be successfully extended to ethnic groups.

4.6 Conclusion

The authors reviewed in this chapter have managed to compile vast evidence suggesting persistent parallels between surname frequencies, genetic structure, and geographical and socio-cultural interaction, in ways that they all seem to conspire to tell similar stories about the recent past of population interactions across space and time. Moreover, they have proven that the aforementioned opposite processes of homogenisation and divergence in surname frequencies between areas or populations, do not only resemble analogies in their variations of gene frequency distributions, but also in other cultural aspects of population interaction and affinity such as, for example, linguistic or dialectical, and even social class distance between populations.

These and other authors of isonymy studies provided the original inspiration to carry out the research presented in this book. Through unconventional multidisciplinary perspectives (in genetics, geography, linguistics, biology, and history), and a set of mundane tools of the trade (telephone directories, electoral registers, historical censuses, and rather rudimentary mapping), they pointed the way on how to compile and command the primary materials and methods for the research presented in this book. More importantly, the striking analogies provided in these accounts sparked initial investigations that attempt to unearth the links between people's surnames, geography and ethnicity and represent them in terms of *socio-spatial ties*; "our name ties". However, the large majority of these studies do not pay attention to forenames. For this reason, the next chapter moves on to analyse the value of forenames to unveil population structure, which has indeed been studied in depth, but approached from a very different set of academic disciplines than those of surnames reviewed in this chapter. The ultimate aim of this book is to link together both forename and surname analysis in the ethnicity classification of populations presented from Chap. 6 onwards.

References

Aranda Aznar J (1998) La mezcla del pueblo vasco. Empiria 1:121–177

Balanovsky OP, Buzhilova AP, Balanovskaya EV (2001) The Russian gene pool: gene geography of surnames. Russ J Genet 37(7):807

Barrai I (2002) Isonymy and isolation by distance in the Netherlands. Hum Biol 74(2):263

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. Ann Hum Biol 23:431–455

Barrai I, Scapoli C, Beretta M, Nesti C, Mamolini E et al (1997) Isolation by distance in Germany. Hum Genet 100:684

Barrai I, Rodriguez-Larralde A, Mamolini E, Manni F, Scapoli C (2000) Elements of the surname structure of Austria. Ann Hum Biol 27(6):607–622

Barrai I, Rodriguez-Larralde A, Mamolini E, Manni F, Scapoli C (2001) Isonymy structure of USA population. Am J Phys Anthropol 114:109–128

Barrai I, Rodriguez-Larralde A, Manni F, Ruggiero V, Tartari D et al (2003) Isolation by language and distance in Belgium. Ann Hum Genet 68(1):1–16

Bedoya G, Montoya P, Garcia J, Soto I, Bourgeois S et al (2006) Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proc Natl Acad Sci U S A 103(19):7234–7239

Biondi G, Raspe P, Perrotti E, Lasker GW, Mascie-Taylor CG (1990) Relationships estimated by isonymy among the Italo-Greco villages of southern Italy. Hum Biol 62(5):649–663

Boldsen J, Lasker G (1996) Relationship of people across an international border based on an isonymy analysis across the German-Danish frontier. J Biosoc Sci 28(2):177–183

Boyce AJ, Reynolds V (1995) Human populations: diversity and adaptation. Oxford University Press, Oxford

Bulmer M (1996) The ethnic group question in the 1991 census of population. In: Coleman D, Salt J (eds) Ethnicity in the 1991 census. Demographic characterisitics of the ethnic minority populations, vol 1. Office for National Statistics, HMSO, London: pp xi–xxix

Cavalli-Sforza LL (1997) Genes, peoples, and languages. Proc Natl Acad Sci U S A 94(15): 7719–7724

Cavalli-Sforza LL (2001) Genes, peoples, and languages. Penguin, London

Cavalli-Sforza LL, Cavalli-Sforza F (1995) The great human diasporas. Addison-Wesley, Reading, MA

Cavalli-Sforza LL, Feldman MW, Chen KH, Dornbusch SM (1982) Theory and observation in cultural transmission. Science 218(4567):19–27

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: bringing together genetic, archeological and linguistic data. Proc Natl Acad Sci U S A 85: 6002–6006

Cavalli-Sforza L, Piazza A, Menozzi P, Mountain J (1989) Genetic and linguistic evolution. Science 244(4909):1128–1129

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes, Abridgedthth edn. Princeton University Press, Princeton, NJ

Cavalli-Sforza LL, Moroni A, Zei G (2004) Consanguinity, inbreeding, and genetic drift in Italy. Princeton University Press, Princeton, NJ

Chen KH, Cavalli-Sforza LL (1983) Surnames in Taiwan: interpretation based on geography and history. Hum Biol 55:367

Cheshire J, Mateos P, Longley PA (2011) Delineating Europe's cultural regions: population structure and surname clustering. Hum Biol 83(5):573–598

Colantonio SE, Lasker GW, Kaplan BA, Fuster V (2003) Use of surname models in human population biology: a review of recent developments. Hum Biol 75(6):785–807

Coulmas F (2005) Sociolinguistics: the study of speakers' choices. Cambridge University Press, Cambridge

Croft W (2000) Explaining language change: an evolutionary approach. Pearson Education, Harlow

Crow JF (1983) Surnames as biological markers – discussion. Hum Biol 55(2):383

Crow JF, Mange A (1965) Measurements of inbreeding from the frequency of marriages between persons of the same surnames. Eugen Q 12:199–203

Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London

Darwin GH (1875) Marriages between first cousins in England and their effects. J Stat Soc Lond 38:153–184

De Silvestri A, Guglielmino CR (2000) Ethnicity and malaria affect surname distribution in consenza Province (Italy). Hum Biol 72(4):573–583

Degioanni A, Lisa A, Zei G, Darlu P (1996) Patronymes italiens et migration italienne en France entre 1891 et 1940. Population (French Edition) 51(6):1153–1180

Dipierri JE, Alfaro EL, Scapoli C, Mamolini E, Rodriguez-Larralde A et al (2005) Surnames in Argentina: a population study through isonymy. Am J Phys Anthropol 128:199–209

Faure R, Ribes MA, García A (2001) Diccionario de apellidos españoles. Espasa Calpe, Madrid

Greenberg JH (1959) Language and evolution. In: Meggers B (ed) Evolution and anthropology: a centennial appraisal. Anthropological Society of Washington, Washington, DC, pp 61–75

Greenberg JH (2000) The concept of proof in genetic linguistics. In: Gildea S (ed) Reconstructing grammar: comparative linguistics and grammaticalization. John Benjamins Publishing, Amsterdam, pp 161–176

Greenberg JH (2005) Genetic linguistics. Essays on theory and method. Oxford University Press, Oxford

Hanks P (1992) The present-day distribution of surnames in the British Isles. Nomina 16:79–98

Hanks P (2003) Dictionary of American family names. Oxford University Press, New York, NY

Hey D (2000) Family names and family history. Hambleton & London, London

Holloway S, Sofaer JA (1989) Coefficients of relationship by isonymy within and between the regions of Scotland. Hum Biol 61:87–97

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 Mar 2006

Jobling MA (2001) In the name of the father: surnames and genetics. Trends Genet 17(6):353–357

Jobling MA, Hurles ME, Tyler-Smith C (2004) Human evolutionary genetics: origins, peoples and disease. Garland Science Publishing, London/New York

Jorde LB, Morgan K (1987) Genetic structure of the Utah Mormons: isonymy analysis. Am J Phys Anthropol 72(3):403–412

Kaplan BA, Lasker GW (1983) The present distribution of some English surnames derived from place names. Hum Biol 55(2):243

King TE, Jobling MA (2009a) Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol 26(5):1093–1102

King TE, Jobling MA (2009b) What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet 25(8):351–360

King T, Ballereau S, Schürer KE, Jobling MA (2006) Genetic signatures of coancestry within surnames. Curr Biol 16(4):384–388

Lasker G (1977) A coefficient of relationship by isonymy: a method for estimating the genetic relationship between populations. Hum Biol 49:489–493

Lasker GW (1985) Surnames and genetic structure. Cambridge University Press, Cambridge

Lasker G (2002) Using surnames to analyse population structure. In: Postles D (ed) Naming, society and regional identity. Leopard's Head Press, Oxford, pp 3–24

Lasker GW, Kaplan BA (1985) Surnames and genetic structure: repetition of the same pairs of names of married couples, a measure of subdivision of the population. Hum Biol 57(3):431

Lasker GW, Wetherington RK, Kaplan BA, Kemper RV (1984) Isonymy between two towns in Michoacán, México. Estudios de Antropologia Biologica: 159–163

Lewis MP (2009) Ethnologue: languages of the world, 16th edn. SIL International, Dallas, TX

Manni F, Barrai I (2001) Genetic structures and linguistic boundaries in Italy: a microregional approach. Hum Biol 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. Hum Biol 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

Mascie-Taylor CGN (1995) Human assortative mating: evidence and genetic implications. In: Boyce AJ, Reynolds V (eds) Human populations. Diversity and adaptations. Oxford University Press, Oxford, pp 86–105

Mascie-Taylor CGN, Lasker GW (1985) Geographical distribution of common surnames in England and Wales. Ann Hum Biol 12(5):397–401

Mascie-Taylor CGN, Lasker GW (1988) The framework of migration studies. In: Mascie-Taylor CGN, Lasker GW (eds) Biological aspects of human migration. Cambridge University Press, Cambridge, pp 1–13

Mateos P, Tucker DK (2008) Forenames and surnames in Spain in 2004. Names 56(3):165–184

McEvoy B, Bradley DG (2006) Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. Hum Genet 119(1–2):212–219

Mikerezi I, Xhina E, Scapoli C, Barbujani G, Mamolini E et al (2013) Surnames in Albania: a study of the population of Albania through isonymy. Ann Hum Genet 77(3):232–243

Milroy L, Gordon MJ (2003) Sociolinguistics: method and interpretation. Blackwell, Oxford

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR et al (2008) Genes mirror geography within Europe. Nature 456(7218):98–101

Olson S (2006) A genetic and cultural odyssey: the life and work of L. Luca Cavalli-Sforza by L. Stone and P. F. Lurquin. Am J Hum Genet 78(1):171–172

Petersen W (2001) Surnames in US population records. Popul Dev Rev 27(2):315

Piazza A, Rendine S, Zei G, Moroni A, Cavalli-Sforza LL (1987) Migration rates of human populations from surname distribution. Nature 329:714–716

Poulain M, Foulon M, Degioanni A, Darlu P (2000) Flemish immigration in Wallonia and in France: patronyms as data. Hist Fam 5(2):227

Renfrew C (1987) Archeology and language. Cambridge University Press, Cambridge

Ringe D (1999) Language classification: scientific and unscientific methods. In: Sykes B (ed) The human inheritance: genes, language, and evolution. Oxford University Press, Oxford, pp 45–73

Rodriguez-Larralde A, Barrai I, Nesti C, Mamolini E, Scapoli C (1998a) Isonymy and isolation by distance in Germany. Hum Biol 70:1041–1056

Rodriguez-Larralde A, Scapoli C, Beretta M, Nesti C, Mamolini E et al (1998b) Isonymy and the genetic structure of Switzerland. II. Isolation by distance. Ann Hum Biol 25:533–540

Rodriguez-Larralde A, Morales J, Barrai I (2000) Surname frequency and the isonymy structure of Venezuela. Am J Hum Biol 12:352–362

Rodriguez-Larralde A, Gonzales-Martin A, Scapoli C, Barrai I (2003) The names of Spain: a study of the isonymy structure of Spain. Am J Phys Anthropol 121:280–292

Rogers AR (1991) Doubts about isonymy. Hum Biol 63:663

Ruhlen M (1987) A guide to the world's languages. Stanford University Press, Standford, CA

Scapoli C, Goebl H, Sobota S, Mamolini E, Rodriguez-Larralde A et al (2005) Surnames and dialects in France: population structure and cultural evolution. J Theor Biol 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. Theor Popul Biol 71:37–48

Schürer KE (2004) Surnames and the search for regions. Local Popul Stud 72:50–76

Shriver MD, Kittles RA (2004) Genetic ancestry and the search for personalized genetic histories. Nat Rev Genet 5(8):611–618

Smith MT (2002) Isonymy analysis. The potential for application of quantitative analysis of surname distributions to problems in historical research. In: Smith MT (ed) Human biology and history. Taylor and Francis, London, pp 112–133

Sokal RR, Harding RM, Lasker GW, Mascie-Taylor CG (1992) A spatial analysis of 100 surnames in England and Wales. Ann Hum Biol 19(5):445–476

Sorenson Molecular Genealogy Foundation (2007) Why molecular genealogy? Available at http://www.smgf.org/. Accessed 09 Mar 2007

Stone L, Lurquin PF (2005) A genetic and cultural odyssey: the life and work of L. Luca Cavalli-Sforza. Columbia University Press, New York, NY

Stone L, Lurquin PF, Cavalli-Sforza LL (2007) Genes, culture, and human evolution: a synthesis. Blackwell, Malden, MA

Sykes B (1999) Using genes to map population structure and origins. In: Sykes B (ed) The human inheritance: genes, language, and evolution. Oxford University Press, Oxford, pp 93–117

Sykes B, Irven C (2000) Surnames and the Y-chromosome. Am J Hum Genet 66(4):1417–1419

Tasso M, Lucchetti E, Pizzetti P, Caravello G (2004) Distribution of surnames and linguistic-cultural identities of the Slovenian and German minorities of northeastern Italy. Anthropol Anz 62(2):185–202

Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(2):234–241

Wolpoff MH, Spuhler JN, Smith FH, Radovcic J, Pope G et al (1988) Modern human origins. Science 241(4867):772–774

Wright S (1943) Isolation by distance. Genetics 28:114–138

Yuan YD, Zhang C, Ma QY (2000) Population genetics of Chinese surnames. (in Chinese with English abstract). Yi Chuan Xue Bao 27:471–476

Chapter 5
Forenames and Social Stratification

> *"[A name is] a repository of accumulated meanings, practices and beliefs, a powerful linguistic means of asserting identity . . .and inhabiting a social world"*
> (Rymes 2001: 160)

Abstract The apparently innocuous act of naming a newborn baby anywhere in the world belies the parents' cultural baggage of social expectations and ethno-cultural customs that have evolved over generations. In this chapter we argue that the choice of forenames that parents select for their children is by no means a random one. A multitude of cultural, religious, linguistic, geographic and socio-economic factors play an important role in the forename/s chosen for a new member of a family. When these forenaming preferences are carefully studied over time, space and social groups, evidence repeatedly shows that forename choice follows very clear "fashion waves" and changes in cultural preferences. These choices in return leave distinct cultural traits in forename distributions by ethnic, social and religious groups. The factors driving forenaming choices operate both at the individual and societal levels through internal and external influences in parents. A review of these studies in several countries is presented providing additional clues that allow us to disentangle social strata in a population, especially with respect to ethnicity in forenaming practices.

The act of forenaming, granting a forename to a new-born child, is a custom that acts as a social bond, a rite of passage or social incorporation that is present in all societies and human groups (Alford 1988). Through such social act, as argued in Chap. 3, personal names serve a dual purpose; to serve as a social marker as well as an individual identifier (Alford 1988; Finch 2008). A person's full name marks her/him as a unique individual, but at the same time it also gives some indication of her/his location in the various social worlds which she/he inhabits (Finch 2008). Within such dual framework, surnames and forenames have typically been

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_5, © Springer-Verlag Berlin Heidelberg 2014

identified with one of the two functions of naming. On the one hand, a person's surname is intimately associated with a family, typically acts a marker of kinship, signalling ancestry and continuity and hence performing well the aforementioned social identifying function of names. On the other hand, a person's forename represents a cultural label that tends to represent her/his individuality and uniqueness. However, despite acknowledging the parents' agency in their choice of a forename for a new-born baby, their decision is far from being taken within a socio-cultural vacuum. In other words, the dual purpose of naming is also found embedded within forename selection decisions, as well as in its everyday usages in particular social contexts. As a result, as it will be argued throughout this chapter, the aggregated outcomes of millions of forenaming choices made in a given society over a period of time, do reflect entrenched but subtle socio-cultural influences. The purpose of this chapter is to reveal the core characteristics driving such influences and patterns in forename choice.

A diverse range of studies about people's forenames in Psychology, Linguistics, Sociology, Economics, Education, Health and other disciplines over the last century have amply recognised the underlying patterns and consequences of forenames as socio-cultural labels. This chapter will review key studies in these areas to reveal how a number of internal and external mechanisms operate to preserve distinctive forenaming practices between social and ethnic groups. Such forenaming mechanisms and dynamics can be analysed at two different scales; (a) the individual parental level (choice) and (b) the aggregated behaviour of social groups (constraints and fashion). This chapter will first focus on the review of individual (parental) forename choice. Section 5.1 analyses how such choice is driven by the intrinsic characteristics of the properties associated to specific forenames (internal factors), while Sect. 5.2 reveals wider social influences in such parental decisions on individual forenames (external factors). The second part of the chapter analyses the outcomes of social dynamics of aggregated behaviour over time. Section 5.3 deals with a broad description of such dynamics, exposing how customs and fashion are ultimately reflected in changes in the most frequent forenames rankings in various countries and periods of time. Within the key social drivers underlying such dynamics; social class, age and ethnicity, the chapter focusses on the last of these, the underlying theme of the book. Hence, Sect. 5.4 closes the chapter with an in-depth review of how forenames might signal ethnicity and cultural identity within such social dynamics, focusing on the particular example of Black American forenames.

5.1 Individual Forename Choice

According to symbolic-interactionism theory, humans create meaningful symbols through interactions with each other and within society (Mead 1934). These symbols must be defined within the social interaction context in which they arise and are recognised and used. Names have been considered as such types of symbols

(Darden and Robinson 1976), carrying a social as well as an idiosyncratic component in their meaning. The former component is shared by most members of a given culture or social group, while the latter is unique to a few people, arising from distinctive experiences, and carrying what has been defined as "surplus meaning" (Kaplan 1964).

Hence, forenames on their own also highlight the "individuality and connectedness" dimensions of a name, signalling the dual dimensions of personhood (Finch 2008). Furthermore, the act of naming a new-born baby, in cultures that use a forename and surname, is primarily concerned with the decision of choosing a forename. This is the case in cultures where the surname element is automatically assigned (typically patrilinearly) from the family (sur)name according to the naming customs in use in a society, sometimes reflected in the civil registry regulations in force. Therefore, forename choice is a proactive decision that constitutes a key element in a person's identity within a given social context.

As Benson et al. (2006: 180) put it "[N]aming [is] a quintessentially social act [. . .] naming acts as a critical element in processes of social incorporation and the constitution of personhood". Parents as social actors choose and use particular forenames to convey social meaning in particular circumstances (Finch 2008), but these "seemingly idiosyncratic expressions of tastes in names are in general affected by underlying cultural themes" (Lieberson and Bell 1992: 511). "In selecting a name (especially for a first-born child) parents are not only determining the personhood of their child but are also taking a key step in defining their own new identity as parents" (Finch 2008: 718), identifying "what sort of child they want to be the parent of" (Zittoun 2004: 143).

Lieberson (2000) presents a convincing account of how forenames preferences reflect patterns of personal and social taste. He establishes that parents' liking or disliking this or that particular forename, especially when it comes to selecting one for their own children, has to do with a combination of internal and external influences. In his extensive study of forename practices Lieberson finds that patterns of forename selection and usage in the U.S. reflect six major factors (Lieberson 2000: 24), which are listed in Box 5.1. The first five factors have to do with individual agency in forenaming decisions, more a less constrained by social influences, while the sixth factor relates to external conditions, such as the extended family, religious rules and institutional pressures. Lieberson suggests that in the context of the U.S. there has been a decline in the role of such external pressures in return for greater individual freedom, with forename selection becoming more "a matter of taste", the actual title of his book. These six factors together encompasses a wide range of socio-psychological and cultural factors that determine its meaning to the parents and eventual behaviour (Wallace and Wolf 1999). Therefore, the purpose of this section is to untangle a range of such concealed and explicit, idiosyncratic and socio-cultural factors of forenaming practices.

Box 5.1 Major factors of forename selection and usage.

1. The imagery associated with each name
2. The notions parents have about the children's future
3. Estimates of others' responses to a name
4. The awareness and knowledge of names through the mass media and other sources
5. Parent's beliefs about what names are appropriate for people of their status
6. Institutionalized norms and pressures (religious, family and other institutional pressures)

Source: Lieberson (2000: 24)

5.1.1 Family Connection and Kinship

As stated in the previous section, names serve to identify unique individuals within family and social relationships. We will first turn our attention to how forenames are used to map family connections, despite this being a function primarily reserved to surnames.

5.1.1.1 Family Continuity

According to Tan (2004: 367) "all cultures [. . .] place great emphasis on the choice of names as they provide the necessary link to the future, in terms of the parents' (or other name givers') hopes and aspirations for the child, and to the past, in terms of the connectedness of the name to the child's ancestors or identification with a particular community". Indeed, parents have an opportunity to highlight family connections as part of "the repertoire of symbols and meanings available" when they select a child's name (Rossi 1965: 503). Through the selection of particular forenames that are closely associated with the family, parents therefore might choose to confirm and reinforce aspects of their kin relationships that are important to them (Finch 2008). The extent to which they choose to do so varies by culture, religion, social class, type of society, gender of the child and wide range of other factors. For example, Burnard (2001) argues that in English naming traditions children are portrayed less as unique individuals than as part of an on-going family and lineage. As such, the process of naming can serve to connect family members over generations, "symboli[zing] the close tie between the youngest generation and those that have gone before" (Williams 1956: 80). In another example, Lawson (1984) describes a Eskimo custom to name a new born child after a recently deceased person, so that his/her spirit can live on. Herzfeld (1991) described a rigid set of rules on naming in the island of Crete, in that the first male child takes the parental grandfather forename, the first daughter the parental grandmother's forename, and subsequent children typically the maternal grandparents' forenames. Furthermore, family connections can also be made synchronically spanning and linking an extended family (Lieberson 2000). For example, in the same Cretan study, after all of the grandparents forenames are exhausted, the godfather is given the uncontested right to choose a forename (Herzfeld 1991). When these customs of tribute and obligation to older generations or extended family are perpetuated through generations a set of family namesakes appear commonly recurring within a family. For example, in the Anglo-Saxon naming tradition namesakes are a very common choice for a child's middle name/s, and hence much less subject to fashion (Finch 2008).

5.1.1.2 Gender Differences

There are also important gender differences on the influence of family tradition and kinship in naming. Rossi (1965) noticed for the U.S., that forenaming practices for sons are more closely linked to family continuity (names of relatives are common), compared to daughters' forenames. Lieberson (2000) offers and explanation of this gender differential focusing on society's traditional function of forenames. Such functions have been driven by fairly patriarchal and sexist customs; in boys, family continuation and patrilineal heritage are reinforced through the use of family namesakes in order to signal fatherhood and stress property inheritance and patrilineal lineage, as is the case with surnames. Thus, it is suggested, a father whose paternity and family responsibilities himself or others can question, is more likely to remain attached to his offspring if his son/s carry his own forename or his father's forename (in addition to their surname). On the contrary, daughter names are selected from a much broader pool of names, or invented altogether, in order to reflect beauty and uniqueness, as opposed to family tradition. This forms part of a sort of "dowry endowment" to better her chances in the "marriage market". This tradition reveals the ephemerality of women's names, since in many cultures she loses her paternal surname at marriage, and hence a "beautiful" forename ensures portability to another family as opposed to loyalty to a patrilineal linage.

5.1.1.3 Recent Trends in Forenaming

Over the last few decades, such traditional naming conventions have been declining in Western societies as a result of greater individual freedom from institutional pressures (Lieberson 2000), and greater gender equality. Rossi's (1965) analysis in the 1960s showed that the gender difference in U.S. traditional forenaming customs had actually been declining over the first half of the twentieth century. However, even today these processes are still active in some countries. In The Netherlands, Bloothooft and Groot (2008) have found striking patterns in frequent connections between forenames that are frequently "found together" running within families, shaped by social class as well as ethnicity (as discussed later in this chapter). Using longitudinal data from large population registers, they establish that the name of an older child can still be predictive of what other forenames are used by parents for the rest of the offspring, proposing "conditional probabilities" of subsequent names. Once again this comes to show that forename choice is not random, but a decision constrained by a range of factors that vary according to type of family and group of kin.

5.1.2 Forename Desirability

Before analysing the societal factors influencing forename choice decisions, it is important to review other key intrinsic aspects that determine a particular forename's desirability in parental naming decisions. Forename desirability has to do with the first and third factor of forename selection listed by Lieberson (2000) and reproduced in Box 5.1; the imagery associated with each name, and estimates of others' responses to a name.

Various scholars from a diverse range of disciplines have directly or indirectly studied forename desirability with regards to perceptions, experiences, stereotypes and life outcomes (Copley and Brownlow 1995; Crisp et al. 1984; Erwin 1993; Garwood 1976; Harari and McDavid 1973; Lieberson and Bell 1992). By analysing the consequences of certain types of names in childhood or adulthood, as well as the subjective influences in both parental forename choice and peer responses to a forename, they establish a range of factors here termed as "for name desirability". Amongst them, a key compilation of seven factors of for name desirability are synthesised over the following paragraphs (our own terminology); aesthetic preferences, phonetic preferences, pronunciation and spelling difficulties, name recall, name commonness, gender distinctiveness, and name perception.

5.1.2.1 Aesthetic Preferences

Aesthetic preferences refer to the overall "beautiness" of a forename as a label for a person beyond its meaning or socio-cultural connotations within a group. A forename we aesthetically like is appealing in the way it sounds *(phonetic preferences)*, the way it looks in written form or when it is read by others, and the feelings those signs evocate. Those same feelings are also studied by marketeers when attempting to delineate the subjective aesthetic power of a brand. Girl names have traditionally carried a much more "ornamental" role and their selection has always primed their aesthetic properties, along the traditional factors associated with a subordinated gender role in most societies. Certain aesthetic properties are preferred by certain social groups at specific points in time or generations, such as for example particular name stems or endings for each of the sexes (Tucker 2009). This point will be expanded later in this chapter when discussing Fig. 5.7, and repetition is avoided here.

5.1.2.2 Pronunciation and Spelling

Pronunciation and spelling difficulties determine repulsiveness towards certain forenames for certain socio-cultural groups, for example linguistic minorities or less literate people. If parents perceive difficulties in determining how a name is spelt or how it is read such name is likely to be dropped from a potential shortlist (Lieberson 2000). Furthermore, in many countries there is strict civil registry legislation to prevent some of these diction problems. For example, in the UK the Deed Poll Service does not accept names "which are impossible to pronounce, which contain symbols rather than words, or which are 'vulgar or offensive'" (Finch 2008: 715).

5.1.2.3 Name Recall

Name recall is the probability that a name is retained by our interlocutor the first time s/he hears or reads the name. Such property is of course contingent on the name frequency or commonness in a given society, the name length, the difficulties in its pronunciation or spelling, and the interlocutor's characteristics, such age, education level, language ability, ethnicity, as well as memory and cognitive capabilities. Furthermore, name recall also largely depends on whether we already know someone with such forename (McCarty et al. 1997).

5.1.2.4 Name Commonness

Name commonness is a property of a name derived from its frequency or popularity in a population group at a given point in time. This property in turn carries a series of other value judgements associated with a name frequency that drives name desirability (Busse and Seraydarian 1978); its trendiness (a forename becoming fashionable at a given point in time), recognition (common names are easier to recognise and recall), originality (greater in rarer names), and attitudes towards commonness in a given group (traditional or liberal societies). To this respect, names at the two ends of the frequency spectrum—very common and very rare—tend to be avoided for two very different reasons. Common names might be considered by many as mundane and unoriginal, denoting conformity and eroding the identification power of the bearer. Meanwhile, rare names might be identified with "deviant behaviour" in a group, they indeed lower the probability of name recognition and proper pronunciation and spelling, increasing the chances of a child being discriminated against (Lawson 1971). Preferences towards name commonness also vary by gender, with male names being much more common than female ones (Busse and Seraydarian 1978; Lieberson and Bell 1992).

5.1.2.5 Gender Distinctiveness

Gender distinctiveness is a key property of a forename determining the type of gender association made with a forename. As reviewed in Chap. 3, one of the primary functions of naming in all societies is to distinguish between the sexes (Alford 1988), and hence this is an important although declining factor in selecting a forename. Generally speaking, parents will avoid forenames that are frequently used by the opposite sex of their child, or new or invented forenames that could attract a gender perception not conforming to their child's sex. For example, most languages have sounds and forename endings that signal gender identification. In English the most common vowel endings pattern being; –a, –e and –i endings frequently reserved for females while –o endings for males, with the three last letters sufficing to identify most female forenames (Barry and Harper 2000). The –a ending for girl forenames seems to be a near universal rule in most languages (Sue and Telles 2007). Some examples of the most salient spelling and sound endings favoured for girls and boys in England and Wales over the last 300 years are shown in Fig. 5.7 (this figure is further explained in Sect. 5.2.1 where the concept of "forename pool" is discussed).

The advantage of these clear gender patterns in naming is that when we encounter a new or invented name, we do very well at guessing the gender of the child, primarily because of embedded associations between gender and sound in any language (Lieberson and Mikelson 1995). However, when we encounter a name that contravenes these gender sound or name-ending rules, we are bound to make mistakes in guessing the gender of an interlocutor, at least when no other visual or voice cues are available, such as when responding to an e-mail for example. Barry and Harper (2000) found that such rules present more exceptions for female than for male forenames, given the narrower pool of male forenames and their more traditional character. Indeed since these rules vary by language and country, guessing the gender of forenames borne by international migrants or ethnic minorities becomes more challenging, although name commonness and lack of familiarity with other naming systems also play a role. To complicate things further, and contradicting one of Alford's (1988) axioms of naming, there are some names that can be interchangeably used by persons of both sexes, for example Andrea, Nicola, Frederick, Francis, Lynn, Leslie, or nicknames such as Pat and Tony (Rickel and Lynn 1981). However such gender neutral names tend to be rare, and as soon as they become more associated with one of the sexes their use in the opposite sex tends to decline rapidly.

5.1.2.6 Name Perception

Name perception relates to a broad range of subjective meanings associated with forenames that have been studied by various psychologists (Bolin 2005; Darden and Robinson 1976; Dinur et al. 1996). One example of name perception is *name* *warmth*, a subjective property relating to the degree of kindness and friendliness that a name evocates in others. Psychologists Copley and Brownlow (1995) conducted an extensive study of perceptions towards job applicants to positions that were deemed to require "warmth" (such as childcare, or customer service), versus more technical jobs. The warmness of a candidate's name seemed to ascertain some influence in the participant's decision to decide his or her appropriateness for each type of job. It is almost impossible to list all possible types of name perceptions, since these are culturally and historically contextual. However, in a classic study in the 1970s Darden and Robinson (1976) made a first attempt to map name perceptions. They measured the subjective perceptions of a group of students towards ten male English forenames. They did so measuring perception along a type of measurement scale known as "semantic differentials" (SD) (Osgood et al. 1957), through 20 pairs of bipolar adjectives aimed at capturing connotative meaning. They then conducted multidimensional scaling (MDS) analysis suggesting four dimensions in male forename perception termed as "character", "maturity", "sociability" and "virility". In Fig. 5.1 an alternative visualisation of the original data is presented in a heatmap-table, ordered through a dendogram (hierarchical clustering) along each axis of the table. This representation shows that these ten names can be grouped together into clusters according to some common subjective qualities perceived in them by the study participants. This visualisation is perhaps more useful than the 2D plots derived from the MDS analysis presented in the original paper. Abbreviated forenames (in reality "nicknames") such as Lou, Charlie or Bill, appear in the left half of the table, together with Bruce and Kevin. Within them, Lou, Bruce and Kevin are clustered in the same branch of the dendogram. These names are perceived as: friendly, youthful, common, simple, exciting, soft, sociable, hot, colourful, emotional, urban, active, and so on. On the right side of the table, we find John, Scott, William and Lance, all clustered into the same branch of the dendogram with Mathew at a further distance. These names are perceived as: passive, slow, weak, ornate (specially Lance), sophisticated, rural, colourless, cold, dull, noble, deep, and mature. These stereotypes are driven by a host of factors and are very much time and place specific, following naming fashions that will be discussed later in this chapter. However, some of these types of forename perceptions are intrinsically related to the internal influences associated with a particular forename as reviewed in these seven factors of name desirability.

In this section about name desirability we have concentrated on those intrinsic factors attached to particular names. However, an individual forename's qualities are difficult to detach from wider social preferences and stereotypes. As Lieberson (2000: 7) puts it "individual differences in preferences are the final step, but they operate within the context of broad societal and subgroup influences". Zelinsky (1970: 748) also adds that "the nature of [forename] choice is pregnant with meaning because each name charges a special field of images, significations, and emotional reverberations for the giver, bearer, and community at large". Therefore name desirability is also determined by differential preferences in social strata,

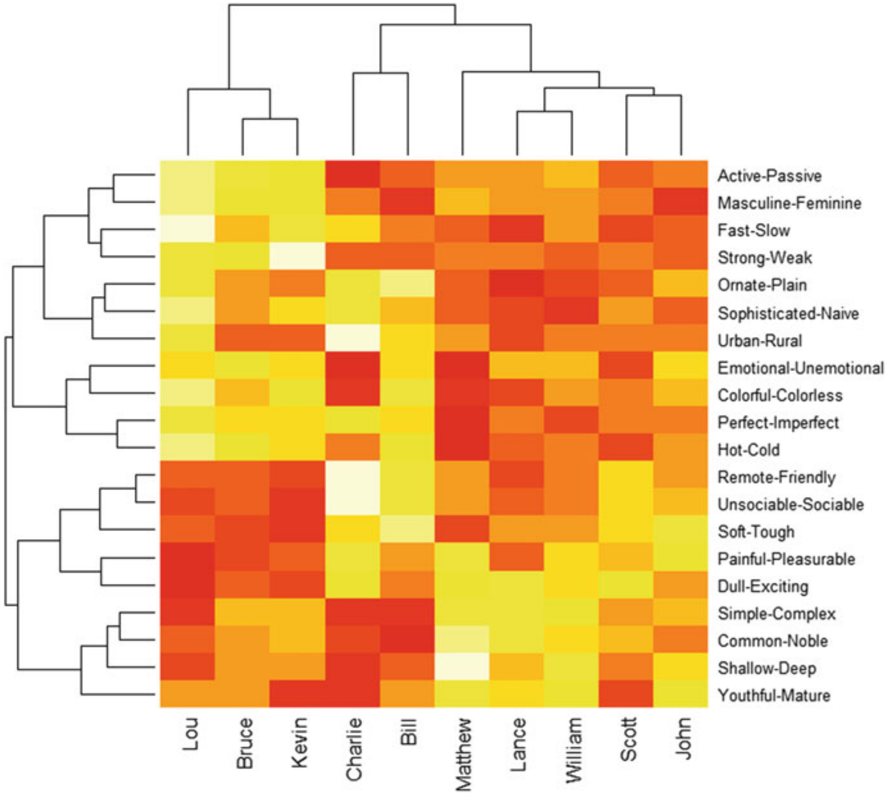


Fig. 5.1 *Heatmap* of perceptions towards 10 male forenames. This heatmap represents the mean values given to 20 bipolar semantic differentials (*rows*) for 10 male forenames (*columns*) in a study conducted by Darden and Robinson (1976). Darker (*red*) colours reflect higher mean values, while lighter (*yellow-white*) lower values. Unfortunately the authors did not provide the relationship between such values and the direction of the bipolar adjective scale. They seem to have altered the order for each pair, and we cannot ascertain higher values to a particular adjective to the *right* or the *left* of each pair. However, some informed guesses are offered in the text. *Source*: Produced by the author (heatmap in R) using data reported in Darden and Robinson (1976: 426)

such as social class, ethnicity, religious beliefs, and other group-derived preferences, all of which change over time and space through name fashion trends.

5.2 Social Influences in Individual Forenaming Practices

Beyond the close bounds of the nuclear and extended family more distant kinship and wider socio-cultural relationships are obviously also influential in forenaming decisions. Lieberson and Bell (1992: 549) state that their analysis of forenames "systematically indicate[s] how underlying cultural and structural conditions drive [forename] taste choices that on the surface appear to be idiosyncratic and haphazard". These conditions could be grouped into a "socio-cultural continuum" spanning; religious, ethnic, linguistic, social class and historical-geographical influences that bound social groups together. Because of space restrictions we cannot possibly summarise the vast range of studies published in this area, since these cover naming systems around the world, many of these socio-cultural groups, and span along various time periods (for systematic reviews see, Alford 1988; Lawson 1984, 2004; Lieberson 2000). However, a brief representative selection of societal influences in individual parental forename choice is presented here, together with some key structural characteristics about how they operate in practice.

5.2.1 Social Identity and the "Forename Pool"

Social identity can be defined as "our understanding of who we are and of who other people are" (Jenkings 1996: 5). According to the aforementioned dual quality of names—individuality and group connection—a name not only identifies the individual person but also it signals its bearer's social identity. As Finch (2008) puts it, a name has "social purchase", a power to symbolize social connection, placing the beholder within a social matrix (Alford 1988). The obvious consequence of such social nature of names is "that [fore]names may demarcate subgroups of a society along such lines as gender, race, class, ethnicity" (Lieberson and Bell 1992: 514).

How are those patterns of social demarcation in forenames produced from individual parental decisions? The most obvious social influences on forename decisions are ethno-linguistic, religious, and historical/geographical patterns and customs. These social dimensions all somehow constraint the type and number of forenames available to parents as feasible options to name their child. In other words, parents choose a forename for their child in a three-step process, even when they might not be fully aware of them. They first; (a) consider a wide "pool of available forenames", from which they; (b) select a shortlist of final options, and (c) through elimination proceed to select the final preferred forename. Of course, these steps are not apparent to most people, but in some way or another such shortlisting process from a forename pool tends to be the norm in most liberal societies (i.e. those in which choice is not largely constrained by external norms).

Central to this explanation is the concept of a *forename pool*, which can be defined as a large repository comprised of forenames that parents are aware of *and* consider appropriate for their social group, that is, are deemed suitable to signal their social identity (Rossi 1965). This section focusses on the social identity dimension of a forename pool, while later in this chapter the concept will be expanded in relation to broader societal forenaming patterns (outcomes).

The forename pool from which a particular parent, or couple of parents, draw/s their forename shortlist is first and foremost conformed by the forenames of people they know, either personally as acquaintances, or through the media, literature, or experience. Research has demonstrated that the property of name recall, or name eliciting, is directly related to those names in one's personal network of acquaintances (McCarty et al. 1997). Names in such network would normally be considered by the parents as a potential forename choice, especially persons that the parents associate with positive qualities, or those person's children's forenames. Therefore, the range of such names in the "forename pool" will be necessarily constrained by name awareness and the size of one's social network. These in turn are determined by linguistic, religious, geographic, cultural, ethnic and other social factors. Taken together, these factors not only shape parent's first-hand experience of forenames, but equally importantly they also set their appropriateness to signal their social identity (Hanks et al. 2006). On top of that "bedrock of forenames" in the pool, parents maybe more or less inclined to add to it more "innovative" name options for consideration. These are forenames that they might have not remembered or elicited quickly from the top of their minds, names that are not used by living persons anymore, or that they were unaware of altogether, for example taken from dictionaries or baby name lists from the web.

Names of religious characters, regional or national historical heroes, certain celebrities, book characters, community leaders, and all sort of "role models" are highly likely to occupy the top ranks of many parents' shortlist of potential baby names. All of these influences are likely to have a social gradient as well as geographic distance. That is, the probability of a name being selected from a forename pool, decreases with distance from the region where the forename is popular or traditional (Zelinsky 1970), or from the social class with which some names are mostly identified with (Lieberson and Bell 1992). For example, even the use of nicknames and informal names has been demonstrated to follow a geographical pattern (Callary 1997;who studied informal names used by US State male legislators c.1994).

The same "social distance" would apply for ethnic, linguistic and religious minorities, with certain names increasing in popularity within such groups and sharply decreasing outside them (Fryer and Levitt 2004). Such patterns are found regardless of actual degree of religious practice or language knowledge, since forenaming customs are perpetuated within groups and families over generations long after the initial trigger setting such name preference in motion (Lieberson 2000; Tucker 2003a). For example, once a forename is identified with a religion, such as *Isaac*, *Joseph* or *Mohammed*, it is highly likely to be avoided by people from other faiths or traditions.

Therefore, parents' forename pools are largely shaped by name preference and avoidance decisions, which are themselves driven by social identity factors. Such homophilic practices are reflected in structural differences in the forename pools from which social groups select their forename shortlists, which in turn are primarily driven by what Lieberson (2000) defines as "social taste".

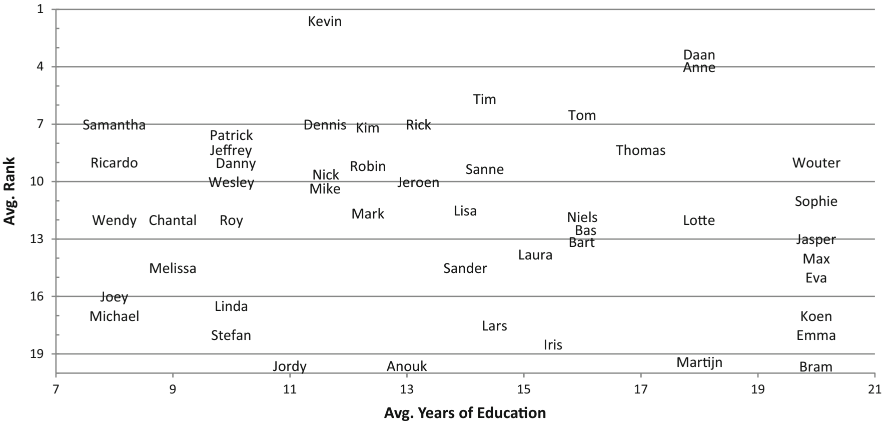


Fig. 5.2 Top forenames in The Netherlands: average rank by years of education (1982–2005). *Horizontal axis* represents average years of education of a forename's bearers, and *vertical axis* represents average rank of the forename if it appears in the top 20 forenames of each of seven educational groups in The Netherlands. Both averages were calculated by the author aggregating rank data for each of the seven educational groups reported in Bloothooft and Onland (2011). *Source*: Graph produced by the author based on table data reported in Bloothooft and Onland (2011)

5.2.2 *Social Class Influences in Forenaming*

Finding religious, ethno-linguistic and geographic traditions and preferences in forenaming practices is not at all surprising, and such associations conform to the expected characteristics of forenaming choice in any society. Of these, only ethnicity preferences will be discussed in detail (see Sect. 5.3), since it conforms the running thread of this book. However, entrenched social class differences in naming is a much more intriguing and understudied field of investigation, and a practice as common as all the other forms of social demarcation in naming (Bloothooft and Onland 2011). Disentangling the scarce evidence available on social class influences in forenames is hence the aim of the remaining of this section.

A number of studies have analysed differences in forename frequencies according to mother education, household occupation/s, child educational attainment, household income, or other proxies for socio-economic class (Bloothooft and Onland 2011; Fryer and Levitt 2004; Levitt and Dubner 2005; Lieberson and Mikelson 1995; Lieberson 2000; Stewart and Segalowitz 1991). For example, the scatter plot graph in Fig. 5.2 shows how the most popular forenames in The Netherlands present stark variations when classified by the number of years of mother's education (Bloothooft and Onland 2011). Such social class patterns in naming are related to forenames' symbolic power to denote status. Dinur et al. (1996: 192) conclude that "it is clear why, on the whole, 'our' children's names appear to be better than the names other groups give to 'their' children".

Hence forenames are denoting social stratification, a ubiquitous human characteristic widely demonstrated through other social outcomes such as residential segregation (Massey and Denton 1988).

Part of these class differences in forenaming practices have to do with semantic connotations of the names themselves, such as those reviewed earlier in this chapter, under the name desirability section. Some personality traits or perceptions of forename sounds or meaning are favoured or repelled by different social classes or households' lifestyle types (Lieberson and Bell 1992). However, most of such social class patterns in naming are produced by social dynamics in forename popularity, that in turn have to do with a belief in parental power to make a difference in a child's life from day 1 (Levitt and Dubner 2005, Chap. 6). Such dynamics will be discussed in the rest of the chapter.

5.3 Social Dynamics in Aggregated Forenaming Outcomes

So far in this chapter attention on forename choice decisions has been focused on household or family level decisions, and how these are shaped by broader community and societal level influences. Therefore, discussions up to now had to do with selecting a specific forename for a specific individual child. Attention is now turned to the aggregated result of millions of such individual level decisions across a society or country. Such aggregations show an apparent concerted pattern driven by changes in social taste. However, no one is directing or promoting these choices "from above". As Lieberson (2000: 7–8) puts it; "more is going on than merely the sum of random individual decisions. [. . .] there are orderly movements which suggest that some set of principles must be operating to drive these changes." His goal is to understand such principles, and Lieberson's book should be consulted for further details on the broader social processes summarised here.

5.3.1 From Custom to Fashion in Forenaming Practices

Although most societies have traditional naming systems that tend to be conservative (Lawson 1984), there is increasing evidence that most western societies have become less traditional and have substantially broadened the forename pool from which they select child names (Lieberson 2000; Tucker 2009). It is tempting to associate this trend towards less traditional names, with the spread of modern communication media, such as the TV, the movies, the telephone, or even the radio in recent western history. However, Lieberson (2000: 43) does not find any association between the onset of such trend towards less traditional forenames, and the spread of these media in the US around the mid-twentieth century. He actually dates the beginning of this shift in the US, back to the spread of literacy and mass urbanisation during the late nineteenth century. Such social changes facilitated

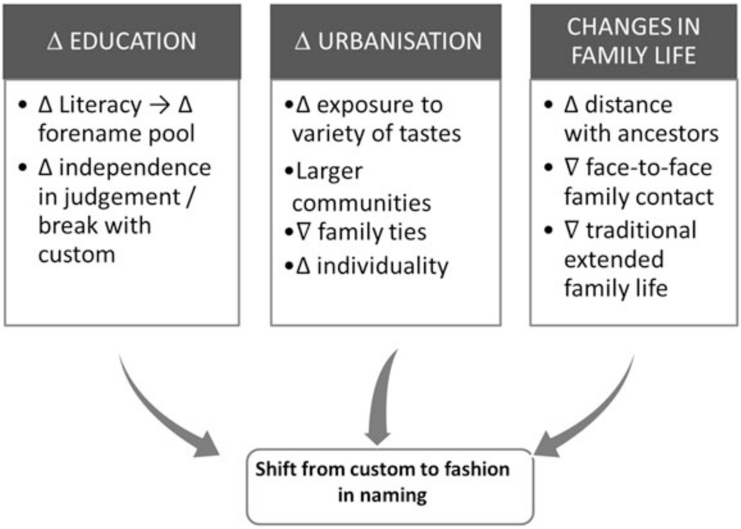

Δ EDUCATION
• Δ Literacy → Δ forename pool
• Δ independence in judgement / break with custom
Δ URBANISATION
•Δ exposure to variety of tastes
•Larger communities
•∇ family ties
•Δ individuality
CHANGES IN FAMILY LIFE
• Δ distance with ancestors
• ∇ face-to-face family contact
• ∇ traditional extended family life
Shift from custom to fashion in naming

Fig. 5.3 Changes in external influences in naming: from custom to fashion. *Key*: *triangle*, increase; *inverted triangle*, decrease. *Source*: Created by the author based on descriptions by Lieberson (2000: 42–43, 66)

greater independence in individual judgement and exposure to broader social groups, breaking away with local and traditional customs and moral pressures. Since then, youth and individuality has been primed while being named after others in older (alive) generations has been unfashionable. Thus, in the U.S., naming ceased to reflect tradition and extended family life to become a question of fashion and individuality. These changes in the external influences in naming, from custom to fashion in naming practices are summarised in a diagram shown in Fig. 5.3. Of course, in each individual country or society such shift has its own idiosyncratic characteristics and particular timing. However, Lieberson's explanation for the US and other Western countries seems plausible and generalizable to most societies (Alford 1988; Bloothooft and Onland 2011; Jayaraman 2005; Mateos and Tucker 2008; Tucker 2002, 2003b).

5.3.2 *Top Forenames' Socio-dynamics*

The expansion of forename pools, through an increase in the influence of fashion, is an outcome of constantly changing social processes in naming. These can be summarised as: *name innovation* and *propagation*, *name replacement* (expansion and decline), and n*ame preference and avoidance* practices (Bloothooft and Onland 2011; Levitt and Dubner 2005; Lieberson 2000; Tucker 2009). All of these practices operate at a group level almost inadvertently by the parents that actually take individual naming decisions.

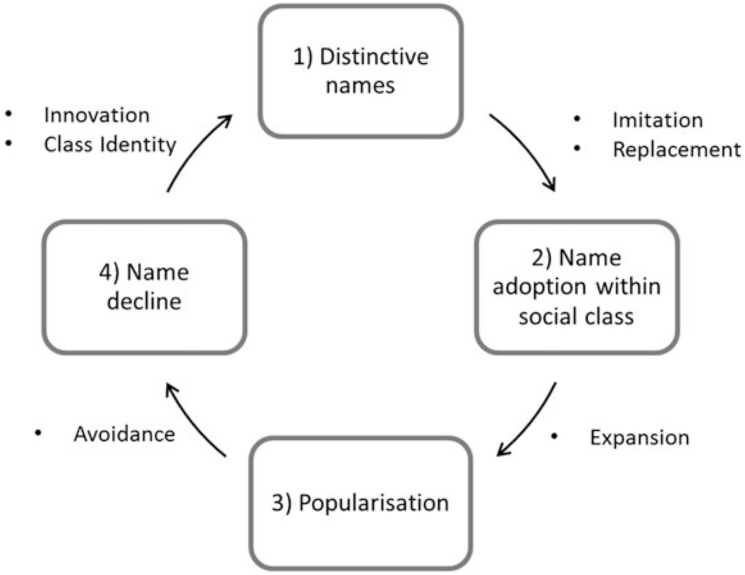


Fig. 5.4 Social influences in naming fashions. The *arrows* and text attached to them represent key processes in forenaming fashion, while the *boxes* represent their main outcomes. *Source*: Created by the author based on processes described by Lieberson (2000: 14)

In brief, such practices consist in the following sequence. Certain social groups introduce new names (or names in disuse) into the forename pool in order to distinguish themselves from "the crowd". Typically such groups are disproportionally the very affluent and the very popular at both extremes of the social class scale. Through a process of behaviour imitation and social identification these names rapidly expand in fashion waves that are socially asymmetrical. Those associated with the lower classes (such as invented names, foreign-sounding names, or movie character names), experience a sudden boom in popularity but tend to wane rapidly, while those associated with affluent groups are slowly imitated by other classes, being quickly abandoned by the upper classes after their popularisation. These processes repeat themselves in a sort of cycle as depicted in Fig. 5.4 that in turn produces temporal "fashion waves" in naming. In other words, a sequence of a name expansion trend is followed by its reversal, a pattern repeated in cyclical waves generally spanning at least one generation (Lieberson 2000).

The main outcome of such temporarily changing "orderly movements" (Lieberson and Bell 1992) is a constant restructuring of the contents and ranking within forename pools, which in turn gets reflected in the parent's preferred shortlists drawn from it. Such top choices aggregated for all children born in a given year comprise the "top hits list" of the most popular forenames preferred in that year. Therefore, the concepts of *forename pool* and *top forenames rank list* comprise the two main tools through which scholars have measured such “orderly movements” in a society or in subgroups, such as ethnic religious or linguistic groups (see next section).

5.3.2.1 Predominant Forename Sounds

One of the clearest patterns arising from such cyclical patterns of naming fashion expansion and decline is reflected in variations in the *predominant sounds* in the top forenames. Figure 5.5 shows the cyclical nature in predominant sounds and spellings in the top 50 forenames in England and Wales over a period of nearly three centuries (1700–1985) (Lieberson 2000: 100). The variety and volatility in girl forenames is striking, while boy forenames tend to be less diverse and stable. For example it is clear how boy forenames ending in “*d*” sounds experienced a boom from 1850 to c. 1925 to decline sharply into the 1960s. Such decline is paralleled by a boom of forenames ending in “*n*” sound from 1950, which peak in 1975 and is widely sustained at the end of the time series in 1985. Therefore, a fashion trend favouring “*d*” endings for boys over 75 years reversed with parents slowly moving to preference for “*n*” sound endings over the next 75 years. The chart on girl names in Fig. 5.5 is perhaps too complex to be described in the space available here, but it clearly shows several of these cyclical patterns with much shorter “boom and bust” periods than for boy names (periods of around 35–50 years from the beginning of an expansion to the end of a decline phase). Such gender difference is probably related to a greater creativity in girl’s names because of their historic ornamental function (Alford 1988), while a preference for traditional forenames for boys has produced more uniformity in sounds over time (Lieberson and Bell 1992).

5.3.2.2 Forename Popularity

In a paper titled “increased competition and reduced popularity”, Tucker (2009) has demonstrated that there is a long term trend dislike of popular (very frequent) forenames in the US, producing greater volatility (competition) in the top forenames rank list. As a result, Tucker concludes, the most frequent forenames in the US and the UK each decade represent less and less population. In Fig. 5.6 this long-term trend is shown for the last 120 years or so in the US (1880s to 2000s), showing both the top 100 and top 1,000 forename frequencies. It is striking to notice that in the 1880s the top 100 boy forenames covered 73.6 % of boys born in that decade, while the same figure is 69.3 % for girls. Today the population covered by the top 100 forenames is 50.3 % for boys and 36.3 % for girls. Such loss in popularity over 120 years reflects the shift from custom to fashion in US forename trends over the same period (Lieberson 2000), as described in the previous section. There is only a period from the 1920s to the 1950s during which this trend seems to reverse,

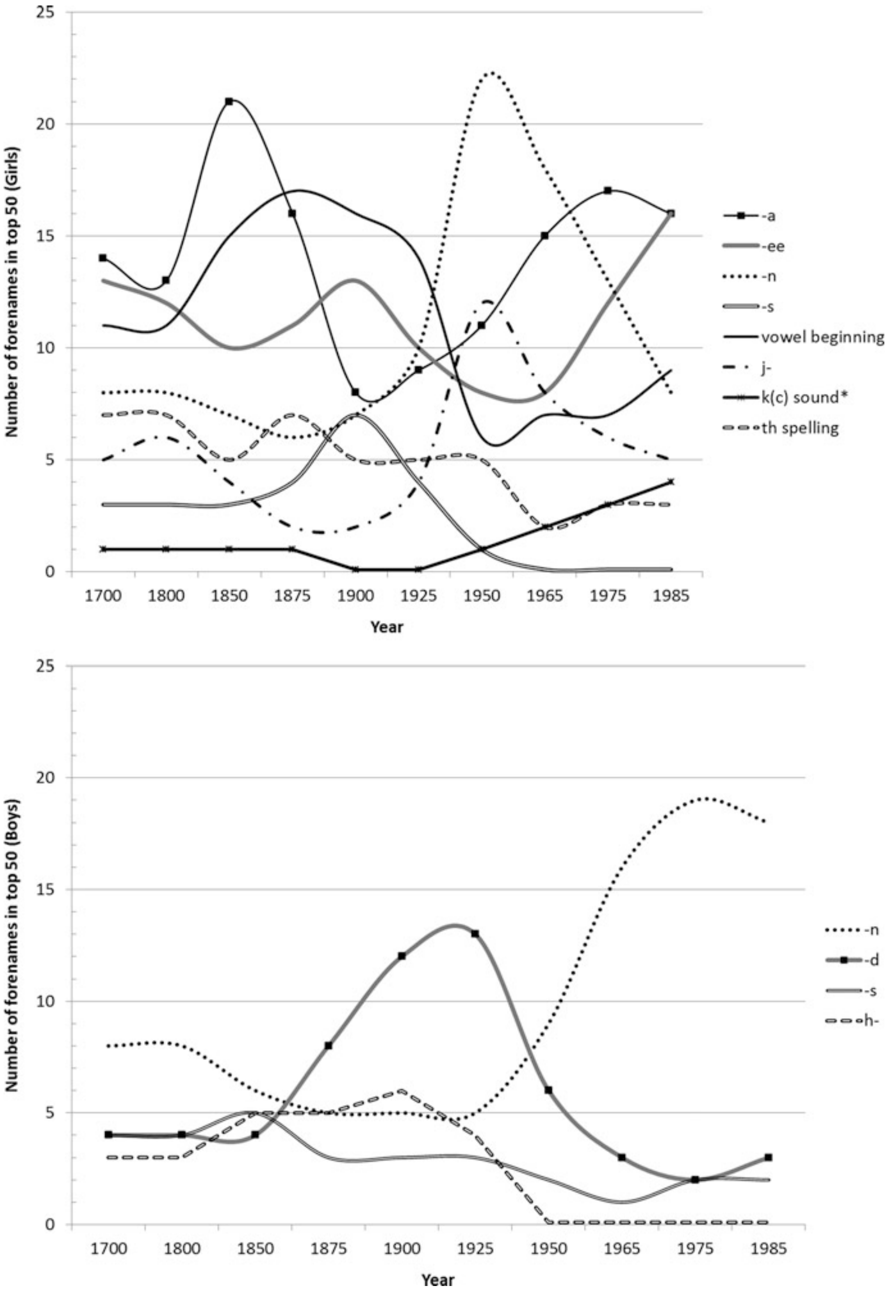


Fig. 5.5 Linguistic characteristics of top 50 forenames for children born in England and Wales (1700–1985) (Girls, *top chart*; Boys, *bottom chart*). The *top chart* shows sounds and spellings of girls forenames, while the *bottom chart* of boys' forenames. They are taken from the top 50 names given to both girls and boys in England and Wales between 1700 and 1985. Each *line* depicts a sound ending ("-" as prefix), beginning ("-" as suffix), in any position, or other spelling characteristics. *Source*: Chart produced by author from table data obtained from Lieberson (2000: 100)

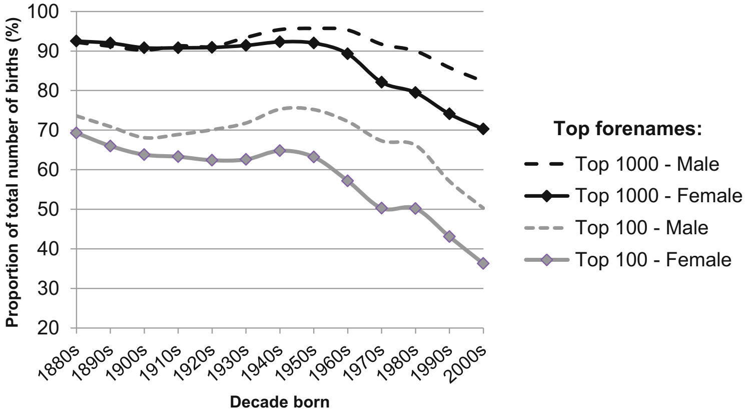


Fig. 5.6 Total population covered by the top 100 and top 1,000 forenames in the US (1880s to 2000s). *Source*: Produced by the author based on baby name frequency data provided by the US Social Security Administration (http://www.ssa.gov/OACT/babynames/)

perhaps as a result of the 1930s depression and the Second World War, a period that also coincides with an overall decline in fertility rates.

The gender differences shown in Fig. 5.6 come to reinforce the differential gender function in forenames that has been referred to at various points in this chapter. The patterns for the top 1,000 forenames parallel those of the top 100. However, these come to show the expansion in the overall size of the US forename pool. While in the 1950s, 95.7 % of boys and 92 % of girls born during these baby boom years, shared forenames appearing in the top 1,000 list for each gender, in the 2000s these rates respectively were 82.3 % and 70.3 %. This means that the forenames of a substantial proportion of children born over the last decades fall within a "long tail" outside these top 1,000. Amongst the most probable explanations are a rise in immigration and a greater exposure to and openness towards foreign, rare or invented names.

5.3.2.3 Forenames and Age Associations

Finally, as a result of such cyclical dynamics in forenaming practices, and beyond the social class associations discussed earlier in this chapter, forenames also present a certain "age association". A forename's age pattern has to do with generational association. Because of the aforementioned volatility in the top forename preferences over time, once a particular forename is popularised and subsequently "abandoned", a forename will necessarily be associated with people born in the particular generation born during its peak popularity. Once that generation reaches their thirties, those forenames will be considered "old-fashioned" and will not be commonly selected for new-born children, perhaps until such generation has died. For example, various authors have identified a recent preference for old testament biblical names which had not been used for almost a century in the US or the UK (Lieberson 2000; Tucker 2009). These age patterns in forename frequencies are starker for some names than others, for people born in particular socio-political and geographic contexts. As such, name-age associations can vary between countries, even within the same language or close cultures. For example, the forename *Brenda*, is associated with very elderly women in the UK, with middle-aged women in the US (see Fig. 5.7), and with young women in Mexico, where it became fashionable since the 1970s. As a result, a woman called Brenda will confront very different reactions in these three countries.

Figure 5.7 shows the ranking of an illustrative sample of forenames in the US over 130 years (from 1880 to 2011). To remove clutter, the figure focusses on name rankings that have made it to the top 200 names each year. In girl's names a major turnover in preferences is observed in the period between the late 1960s to the late 1970s, coinciding with a youth revolution in cultural values in Western countries. This period saw a major shift in top female forenames following more imaginative choices in a context of greater female freedom. Some particular spikes in rare name popularity in girls born around those years are observed in the data, such as for example the name *April* shown in Fig. 5.7. It is interesting to see the number of female forenames that meet half-way on the popularity rankings during this period, with names such as *Barbara* and *Brenda* quickly loosing popularity while *Amanda*, *Sarah* and *April* were on the rise (these five lines cross each other in the chart around the year 1972). *Amanda* is a particular good example of a cyclical name. It was amongst the top 200 between 1880 and 1910, it then lost popularity for almost six decades only to make a fast and short lived comeback between 1970–2005, when it made it to the top five forenames during 14 consecutive years around the 1980s. After this boom, it disappeared altogether from the top 200 in less than two decades, completing a peak cycle of about 35 years. Some of these more short lived periods of popularity can be sometimes explained after a fashion wave triggered by a particular celebrity or movie character name.

Boys names are much more stable over time and only very gradual decreases in preferences are observed in the US dataset shown in Fig. 5.7. For example the names *Richard*, *Robert* and *Edward* show a slow decline over a period of five decades. By contrast, some innovations have expanded to make it to the top 20 rank positions over shorter periods, such as *Christopher* in the 1950s and 1960s, and *Dylan* and *Noah* in the 1990s and 2000s. It is interesting to notice that these last three names had already been relatively popular in the period 1880–1900 (bottom left corner of the graph), and made a comeback around 60–90 years later. This example seems to support our hypothesis that some names become popular only after the previous generation associated with such names has passed away.

Beyond the few examples that can be reproduced here, the reader is referred to two invaluable visualisations of the US Social Security Administration dataset on forename frequencies for over 130 years. The first one is the "Rank clock" visualisation produced by CASA research centre at University College London (http://casa.oobrien.com/rankclocks/), by Oliver O'Brien following Prof. Michael Batty's innovative methodology to visualise rank-size relationship over time (Batty 2006,

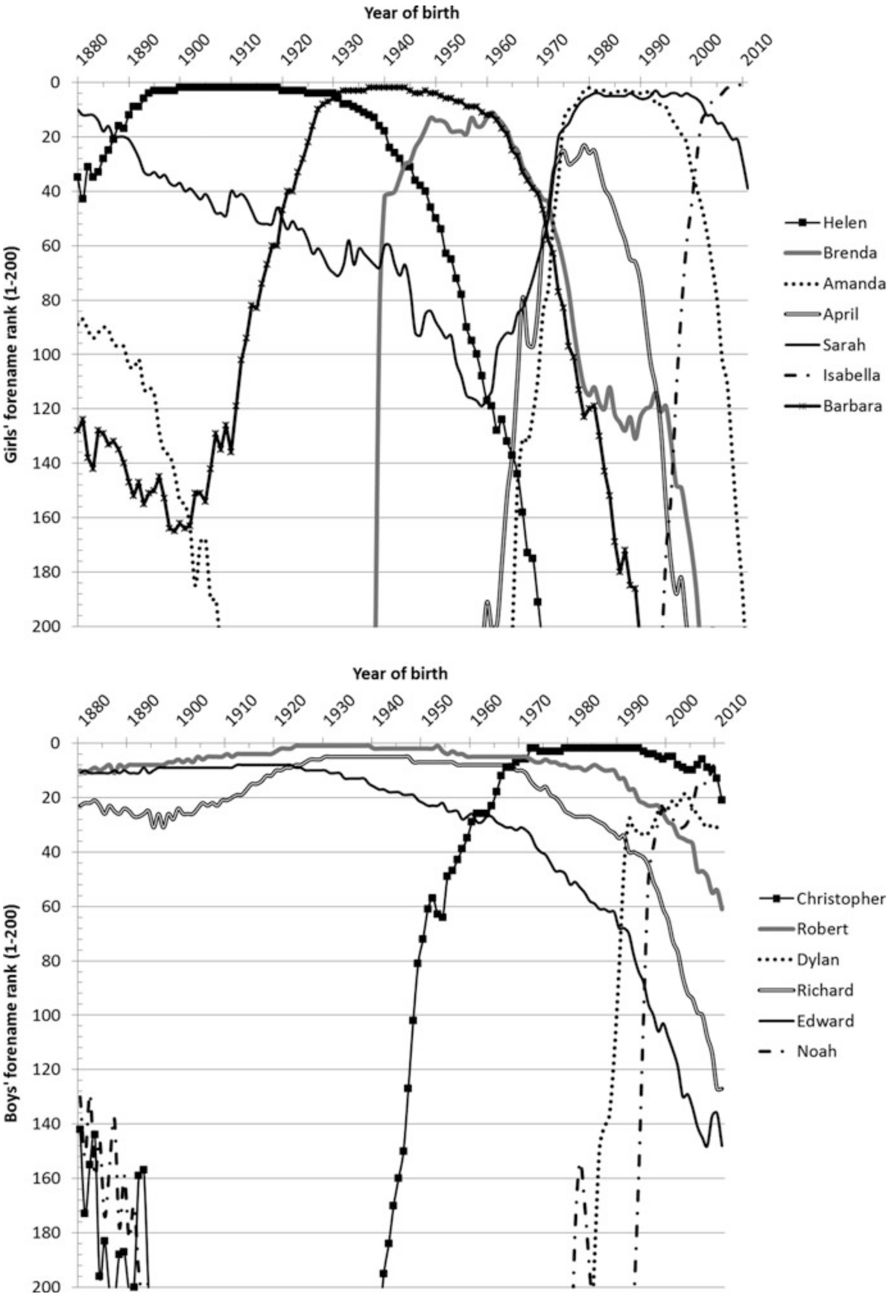


Fig. 5.7 Historic trends in the popularity of a selection of US baby names (1880–2011) [Girls: *top chart*, Boys: *bottom chart*]. *Source*: Chart and name selection produced by the author based on baby name frequency data provided by the US Social Security Administration (http://www.ssa.gov/OACT/babynames/)

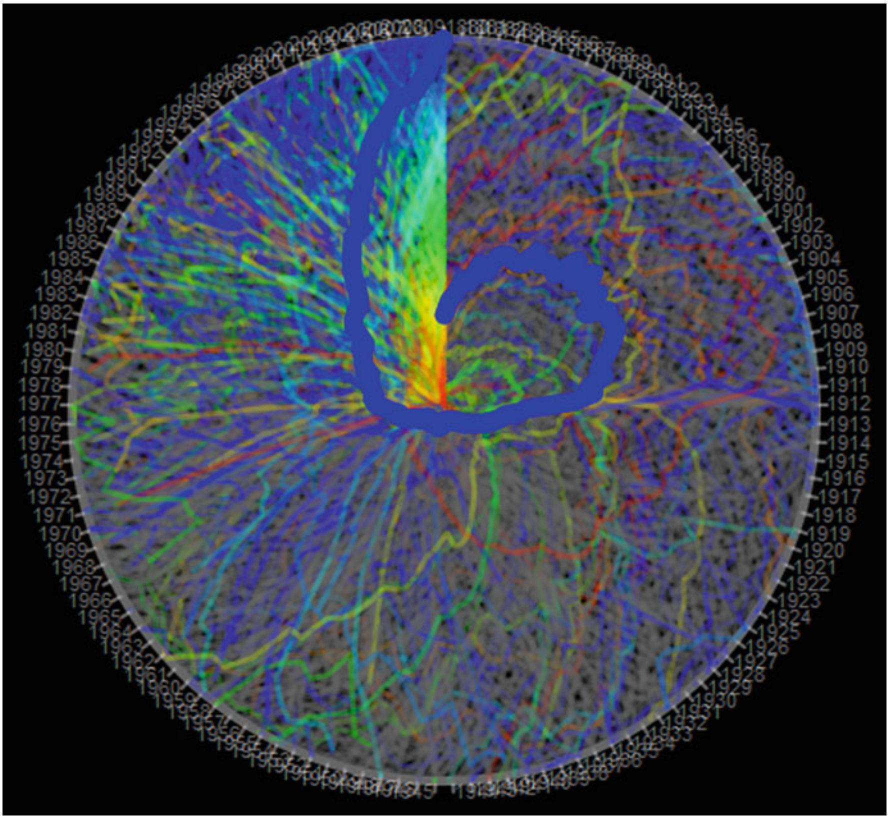


Fig. 5.8 "Rank clock" of US Baby Names 1880–2010. The external rim of the "clock" represents time in years 1880–2010 (clock-wise) while the internal *vertical axis* represents a male forename's rank 1–200 in each year. The trajectory highlighted in *blue* corresponds to the name "*Richard*" which starts around rank 50 to reach the top ranks in the 1940s and then decrease in popularity from the 1970s (see Fig. 5.7 for the ranks of this name). For further details on this type of visualisation see Batty (2006, 2010). *Source*: Oliver O'Brien, Centre for Advanced Spatial Analysis, University College London, http://casa.oobrien.com/rankclocks/ (reproduced with permission) based on US Social Security Baby Names data

2010). A screenshot of these rank clocks highlighting the sample name *Richard* is shown in Fig. 5.8. The second visualisation is the dynamic chart provided by the company "The Baby Name Wizard" (http://www.babynamewizard.com/voyager) that allows for quick and flexible interaction with such a huge dataset.

Commercial geodemographic companies have exploited such age associations in forenames, producing household-level age classifications based on the forenames of adults registered at a particular address (for example CACI's MONICA product, see: Beaumont and Inglis 1989; Birkin and Clarke 1998; Cole et al. 2005; Mitchell and McGoldrick 1994). This information is invaluable for marketers who only have a list of name and addresses in order to produce age-probabilities for each householder (Longley et al. 2006; Mateos et al. 2007). For example, in the UK a household with forenames *Fred* and *Elsie* would be most likely aged beyond 70 today, whereas one where *Sharon* and *Kevin* live will be most likely in their 30s (Birkin and Clarke 1998). Geodemographics companies such as Experian illustrate their neighbourhood classifications with sample forenames that are common in each area, mixing age patterns with social class and ethnicity stereotypes (Experian UK Ltd 2006). Although these approaches have generated some doubts in academia about the robustness and generalisation of such age associations, influential sociologists such as Mike Savage and Roger Burrows (2007) suggest that this type of pragmatic geodemographic modelling of populations still has an untapped potential in academic social science. This book aims to make a small contribution in this direction, through the introduction of various evidence and concepts drawn from a multidisciplinary literature.

5.4 Ethnicity and Forenames

Ethnicity, understood as a multidimensional concept, as discussed in Chap. 2, clearly shapes the forenames selected by most ethnic groups, following the social stratification processes discussed throughout this chapter. These preferences are of course entangled with linguistic, religious, nationalistic, geographic, racial, migratory, regional and other cultural identities influencing forename choice. Lieberson (2000) finds that forename preferences within ethnic groups in the US reflect both internal mechanisms as well as external influences. Amongst the former, old tastes in terms of religion, language, and the “pool of commonly accepted forenames” are found to influence new tastes in naming, in ways that group cultural elements are maintained even after the causal conditions may have attenuated or even disappeared: thus, for example, forenames with certain sounds that are difficult to pronounce for first generation migrants continue to be avoided by second or third generation parents who are nevertheless fluent in the majority language (English in Lieberson’s example). Amongst the external influences—or as Lieberson puts it “macrosocietal conditions”—disposition toward assimilation, nationalism, opposition toward the larger society, symbolism, recency of arrival, and group history, all determine the choice of names in an ethnic group. Through various studies we have also found similar naming patterns for different ethnic minorities and migrant groups in the U.K. (Cheshire and Longley 2011; Mateos et al. 2006, 2011; Mateos 2007).

5.4.1 Assimilation vs. Group Identity

Amongst the internal mechanisms identified by Lieberson (2000), symbolic association is the most powerful driver of tastes in name choice for certain ethnic groups. This mechanism can act in two different ways; (a) signalling integration (some would say assimilation) into the "host" society through a rapid tendency to adopt forenames used by the majority ethnic group in a society, or (b) moving away from the majority to signal independent identity, such as with Black names in the U.S. since the late 1960s (Fryer and Levitt 2004).

Many immigrant groups decide to name their children born in the "destination" country in ways that substantially differ from customs in their country of origin. Such general trend is typically explained as a question of adaptation to the new cultural medium in which their children will grow (as sustained by the social integration or even "assimilation" hypotheses). Many Asian immigrants in the U.S. name their children using Anglo-saxon naming patterns, what some authors have interpreted in terms of willingness to adapt to the "host society". For example, this is less difficult in the case of Koreans in the U.S. who are predominantly Christian (Lieberson 2000). Other authors suggest that Asian "prescriptions" might be stressing personal and financial success and hence it is their desire to mirror White names (Fryer and Levitt 2004).

Symbolic association in forename choice by ethnic group operates through a series of mechanisms that we here group in five factors.

5.4.1.1 Name Avoidance

Forenames that are difficult to pronounce or spell in the destination country main language are generally avoided. Names which might be identified with negative semantic connotations in another language are also avoided (e.g. a girl name that might sound like "lazy" in English). Mencken (1963) provides a range of examples of distinctive forenames from 20 different languages that, despite being initially common amongst foreign born immigrants in the U.S. in the late nineteenth and early twentieth century, they rapidly lost popularity within the second generation. However, to what extent are these historic trends maintained today in the age of globalisation and migrants' transnational lives? (Levitt et al. 2004). Lieberson (2000: 113) concludes that "the increasing interaction between nations and cultures, particularly through the mass media, is now leading to the emergence of an international system" of naming. An updated study of these name avoidance trends by ethnic minorities over the last decades is certainly required.

5.4.1.2 Name "Translation"

Beyond name avoidance, a common custom by many migrant and ethnic minority communities is to modify a name from a "foreign" language to make it more palatable to people with another mother language, sometimes "translating" the name altogether. For example, the biblical name *John* can be found in at least 23 languages with as many forms appearing in the Dictionary of First names by Hanks and Hodges (Hanks et al. 2006), facilitating its automatic "translation".

There is an extensive literature on the Anglicization of names, both forenames and surnames, especially in the U.S. (Fucilla 1943; Hanks and Tucker 2000; Jayaraman 2005; Tan 2004). Similar work is published in other languages for different adaptations of migrant names into French, Spanish, German, Swedish and so on (Arai and Skogman Thoursie 2009; Mateos and Tucker 2008).

These trends towards assimilation in naming customs and away from traditional names common in an ethnic group, present stark variations according to three aspects: (a) generation and recency of migration, (b) the group's social status and (c) the geo-historical origin of an ethnic group.

5.4.1.3 Generation and Recency of Migration

There is ample evidence in various facets of life, that the descendants of immigrant groups born in the destination country adopt cultural customs and behaviour that is much closer to those of the majority ethnic group than to those of their parent's. For example, residential segregation patterns have been found to be less pronounced for native born than for foreign born ethnic minorities in both the UK and the US (Iceland et al. 2011). In Fig. 5.9 we can appreciate such effect for a subset of the Hispanic population in Texas, US. The two graphs (one for each gender) compare the number of the top 20 ranked forenames given to children born in Texas between 1965 and 1995, from mothers in four "ethno-generational" groups (Lieberson 2000: 186): Mexicans foreign-born, Mexicans US-born, Non-Hispanic Whites US-born (Anglos), and Blacks US-born (using Lieberson's terminology). The overlap between the most popular forenames given by US-born Mexican mothers and those termed "Anglo", is striking; both groups share between 10 and 15 forenames in the top 20 over nearly three decades, both for boys and girls. By contrast this figure is reduced to between 2 and 7 for children of Foreign-born Mexicans. When comparing both generations of Mexican mothers (foreign and US-born), their children shared between 7–11 boy forenames and as many as 9–14 girl forenames. This gender difference is commonly found for other ethnic groups (for example in Black names; see also Fig. 5.2 and the explanation offered in the next section), where there is a greater presence of girls' forenames signalling group identity than in boys'. One possible explanation of this difference in Hispanic naming would be the patriarchal/traditional view in naming, explained earlier in this chapter, which considers girls' names mostly a matter of ornamentation, providing a more traditional role for boys. In the case of ethnic minorities, the preference in boy names would be to signal assimilation with the mainstream society in order to avoid discrimination in education and the labour market, while this is perhaps not considered as important for girls, for whom originality, ornamentation, and attractiveness to ethnic group members (homophily) are the features favoured to increase her chances (and group endogamy) in a future marriage market. However, other authors using Hispanic names from California have found the opposite pattern, with a greater frequency of English names for girls than boys (Sue and Telles 2007), claiming that traditional values are reinforced through male naming. A plausible

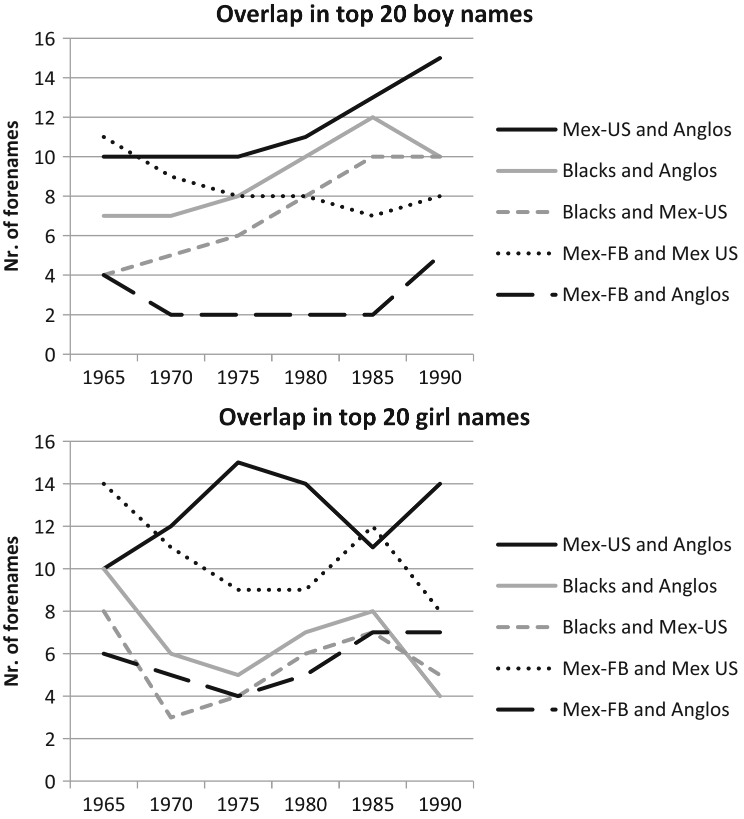


Fig. 5.9 20 most popular names in Texas: overlap between Blacks, Mexicans and Anglos in Texas (1965–1990). Number of forenames that overlap between the top 20 forenames for each of four "ethnic"/generation groups: Mex-FB—children born in Texas to Mexican born mothers; Mex-US—children born in Texas to US-born mothers reporting Mexican ancestry; Anglo—children born in Texas to US-born mothers reporting Non-Hispanic White race; Blacks—children born in Texas to US-born mothers reporting Black race. *Source*: Graph created by the author using data provided by Lieberson (2000: 186, Table 7.5)

explanation could be differential naming behaviour between Mexicans in Texas, a conservative State where Mexican ancestry has a long history and the anglo-saxon naming tradition might not necessarily be "the norm", and the much broader Hispanic population in California, a more liberal and diverse State.

Finally, Fig. 5.9 also shows how the US-born Mexican group compare with Blacks in their name patterns. Their convergence in naming has been increasing over these three decades. There is a lower overlap in children names between Blacks and US-born Mexicans than between Blacks and Anglos, meaning less integration with the White society for US-born Mexicans, although this difference seems to disappear at the end of the available period (1990).

5.4.1.4 Ethnic Group Status

Lieberson (2000) suggests that names closely associated with immigrant groups that are regarded as "inferior" by the majority of the population (i.e. possessing unattractive characteristics), tend to lose popularity within the group. In the U.S. these have traditionally been non-White or non-Christian ethnic groups, but in the context of conflict, this "fall of grace" also affects white groups, such as German names which were negatively associated during and after the Second World War. However, Lieberson suggests, as an ethnic group's social standing improves, traditional forenames re-gain popularity a few generations later. These then become "ethnic markers", providing a symbolic statement of proudness for descendants of those migrant groups. This is the case of various historically discriminated European groups such as Irish or Italians who have experienced an improvement in their status and imagery over recent decades, explaining the rise in popularity of forenames such as *Patrick* or *Frank* (associated with Franco and Francesco in the U.S.). Furthermore, Adamic (1942) provides various examples of Danish and Finnish forenames re-gaining popularity in the Midwestern States of the U.S. during the 1930s and 1940s as a result of royal family visits or in response to Second World War events. Today, a preference for typical Irish forenames amongst U.S. families identifying with such ancestry is well documented (Lieberson 2000: 178). In rare cases, names brought my migrants end up becoming popular with the majority population, such as for example *Eric* (or *Erik*), *Carl*, *Karen*, *Kelly*, *Maria* or *Dolores* in the U.S. (Lieberson 2000: 176).

5.4.1.5 Ethnic Group's Geo-historical Origin

However, not all ethnic minorities are descendants of recent immigrants. Many others are comprised of historic ethnic groups or linguistic minorities that have been traditionally discriminated in the territory they consider their "homeland". This is the case for example of Basques, Catalans and Galicians in Spain, whose traditional names were not accepted for all official purposes during General Franco's 40 years dictatorship and had to be translated to (Castilian) Spanish (Nieto 2000). Similar cases of discriminated ethnic minority groups are Welsh, Irish and Scottish in the U.K.; Chechens and other minorities in Russia; Corsicans in France; Kurds in various Middle Eastern countries, Indigenous groups in various American countries, Maori in New Zealand, and a long list of many other groups, including descendants of more historic migrant groups through complicated colonial pasts where a racial or linguistic difference exists today, such as Blacks in the U.S., or Quebecois in Canada (for an extensive coverage of these naming traditions see various issues of *Names; A Journal of Onomastics*). Most of these groups have experienced a renaissance in their identity awareness and political empowerment over the last four decades, henceforth increasing their "social status". Such renaissance has been often symbolised by a trend towards favouring distinctive forenames for children born during this period, as a marker of ethnic membership (Fryer and Levitt 2004).

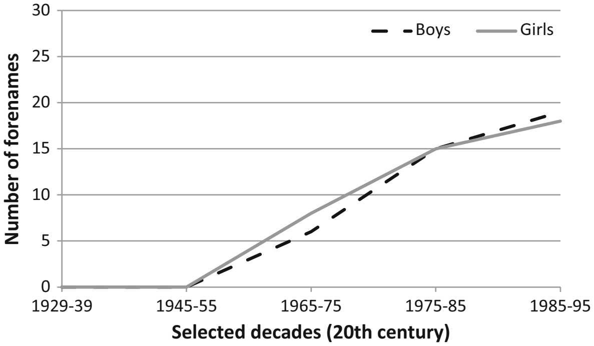


Fig. 5.10 Number of Basque forenames in top 30 rank (1929–1995). Number of distinctive Basque forenames in the top 30 forenames given to children born in the Basque country (Spain) during each of the five selected decades (twentieth century). Each *line* represent boy and girl forenames out of the top 30 used for each gender. *Source*: Elaborated by the author based on data from Nieto (2000: 155–157, Tables 1a and 2a)

An example of these trends is represented in Fig. 5.10, showing the number of distinctive Basque forenames that appear in the top 30 rank in each of five decades of the twentieth century (1929–1995) for children born within this region of contemporary Spain (Nieto 2000). It shows a clear rise in the popularity of Basque forenames since the early 1970s, coinciding with a renewed identification with Basque roots as the Franco dictatorship started to lose its grip before his death in 1975 (otherwise Basque names could not have been officially registered). While before the 1970s no distinctly Basque forenames appear in the top 30 rank, by the mid-1990s these comprise a large majority of the top forenames (63 % for boys and 60 % for girls). Furthermore, the popularity of Basque names have extended well beyond the Basque country, especially for girl names, and some are now found elsewhere in Spain and some Latin American countries.

A lack of space in this book prevents us from mentioning a few more examples of how different ethnic groups' preferences and social tastes drive idiosyncratic naming patterns across the world. Extensive descriptions of the key naming patterns of many ethno-religious groups is provided in generic works, such as Lieberson (2000 especially Chap. 7) and Hanks et al. (2006). A sample of specific group-oriented research on forenames are: Hispanic (Woods 1984), Jewish (Livingstone 2005), Turkish (Razum et al. 2001), Hindu and other South Asian names (Jayaraman 2005; Parekh and Parekh 2003), French (Besnard and Desplanques 2001; Heuvelink 2012), Dutch (Bloothooft and Groot 2008), Nordic/Scandinavian (Kotilainen 2011), or Russian (Lawson 2004) amongst many others. An exception will however be made in the next section, the highly discriminated group of Black Americans in the US, for the reasons explained below.

5.4.2 Black Forenames in the U.S.

Black Americans forenaming patterns, have dramatically changed since the 1970s, paralleling gains in political empowerment, self-esteem and group identification. Such trend is beyond the expected language or migration influences affecting the naming dynamics described for the more recently arrived ethnic groups mentioned earlier in this section. Hence, Black Americans comprise a useful example to explain how group identity can quickly re-shape social tastes in naming.

According to various authors, forenames are one of the few quantitative indicators available to measure culture (Fryer and Levitt 2004; Lieberson 2000; Zelinsky 1970). In particular, Fryer and Levitt (2004) have chosen to study Black forenames as a proxy to measure "cultural investments". They conducted an extensive study on the rise in the use of distinctive Black forenames in the US since the early 1970s, which they see as an indicator of Black "cultural investments". They find that in the 1960s Blacks and Whites used relatively similar forenames for their children. However, since the 1970s and over a short period of time the pattern changed radically with most Blacks adopting increasingly distinctive forenames. These authors found that the rise of the Black Power movement influenced how Blacks perceived their identities and this is clearly reflected in changing forenaming patterns, paralleling changes in musical tastes (Waldfogel 1999), linguistic patterns (Wolfram and Thomas 2008), and consumption choices in this ethno-racial group.

Such new pattern in Black forenaming practices has diverging intensities; those Blacks living in racially isolated neighbourhoods are much more likely to choose distinctive names, while a small subset of Blacks in more educated or upper-middle class families are actually moving toward more assimilating (White) names. Furthermore, these trends in Black naming are more accentuated for girl names than boys. As a result, Black names of people born since the 1970s provide a strong indicator of socioeconomic status, which was not previously the case (Fryer and Levitt 2004).

Part of the sharp rise in the prevalence of distinctively Black names is caused by a high frequency of *unique names*. These are defined as names that are not shared with any other child born in the same year. In a study of Californian baby names, approximately 30 % of Black girls and 20 % of Black boys had a unique name in the 1990s, while the figures for Whites were 5 % and 4 % respectively (Fryer and Levitt 2004). In a different study conducted in Illinois, the prevalence of unique Black names reached a peak of 60 % in the 1980s (Lieberson and Mikelson 1995).

Beyond the use of unique names, *distinctive names* are those that are barely used outside an ethnic group. More than 40 % of the Black girls born in California in the 1990s had distinctively Black names, meaning that we do not find a single white girl—of which 100,000 per year are born—with these names (Fryer and Levitt 2004). Even with popular names, the racial distinction of many names is striking. Over 90 % of the children with the following popular Black names born in California in the 1990s are Black, each with a frequency greater than 300: *DeShawn* (99 %), *Tyrone* (97 %), *Shanice* (97 %), *Precious* (95 %), and *Deja* (94 %). This polarisation in naming works both ways. It is not only that the rest of the population moves away from such distinctively Black names, but also Blacks moving away from other "mainstream" or White forenames. Conversely, the opposite is true for names like *Connor, Cody, Jake, Molly, Emily, Abigail, and Caitlin*, each of which has a frequency greater than 2,000 and less than 2 % of their bearers are Black (Fryer and Levitt 2004). *Molly* comprises an extreme case with only 9 Black girls out of 2,239 girls in the whole 1990s decade. Fryer and Levitt (2004) conclude that contemporary Black forename choices differ significantly more from Whites than do forenames selected by native-born Hispanics and Asians. Hence race, class and even socio-political identity are clearly driving such cultural distinctiveness and not necessarily language, religion, or migration experience.

The main consequence of the spread of unique and distinctive Black naming practices is that such names reinforce existing processes of discrimination. Black names are disproportionally associated by the general population with certain disadvantaged socioeconomic traits such as; single mothers, racially isolated neighbourhoods, lower education level, or poorer households (Fryer and Levitt 2004). Several studies have demonstrated that the general population approaching these distinctive Black names are prone to associate such stereotypes with their bearers, who in turn are discriminated in the job and housing markets (Bertrand et al. 2004; Bolin 2005; Cotton et al. 2008; Guéguen and Pascual 2011; Hogan and Berry 2011; Kalist and Lee 2009; Riach and Rich 2002). Henceforth, "giving one's child a minority name may impose important economic costs on the child" (Fryer and Levitt 2004: 771). However, these authors did not found any association between Black names and worst life outcomes, after controlling for a wide range of socioeconomic and demographic factors. Hence they conclude that Black names are a consequence and not a cause of discrimination *per se*. They might affect the chances of getting a job interview through a CV shortlisting process, but a name is not relevant beyond the interview stage. Black parents choosing these names might be aware of their disadvantages, as one might do when incorporating other socially stigmatizing symbols such as a tattoo in a visible area. However, they choose to bestow a Black name to their child to signal a sense of proudness in group and class identity. Other ethnic groups such as many Asian groups in the US, actually do not want to highlight group distinctiveness but the contrary, assimilation, at least in the first two generations.

5.5 Conclusion

This chapter has presented a broad range of key evidence to explain how internal and external mechanisms in forenaming practices operate to preserve distinctive naming patterns in social and ethnic groups. Starting from parental forename choice decisions at the level of the individual, constrained by family custom and small community traditions, we have explored how the rise of individuality and fashion over the twentieth century has resulted in greater freedom and a broader forename pool. Broader social influences in individual decisions have also been analysed, such us for example, regional, social class, age, or ethnicity influences. As a result of millions of individual forenaming decisions, the aggregated behaviour of parents shows clear patterns delimiting social and cultural groups within the framework of the unfolding of a future global system of naming. The chapter has unbundled some of the key socio-cultural dynamics behind such aggregated behaviour over time. It has defined important concepts such as forename pool, forename frequency, top forename rank list, name avoidance, forename desirability, and various other concepts operating at different scales. Through these explanations, the workings of a delicate balance between individual parental choice and broader societal influences in forenaming patterns has been established throughout the chapter. Within this, attention has been brought to the ethnicity implications of forenames, especially for ethnic minorities. As it will be unfolded in the next chapters, forenames hence provide the missing link that will allow the triangulation between surnames and ethnicity in order to establish a proxy for 'population structure', hence closing a loop and conforming the trilogy which gives title to this book; names, ethnicity and populations.

References

Adamic L (1942) What's your name? Harper & Brothers, New York

Alford R (1988) Naming and identity: a cross-cultural study of personal naming practices. Hraf Press, New Haven, CT

Arai M, Skogman Thoursie P (2009) Renouncing personal names: an empirical examination of surname change and earnings. J Labor Econ 27(1):127–147

Barry H, Harper A (2000) Three last letters identify most female first names. Psychol Rep 87(1): 48–54

Batty M (2006) Rank clocks. Nature 444(7119):592–596. doi:10.1038/nature05302

Batty M (2010) Visualizing space–time dynamics in scaling systems. Complexity 16(2):51–63. doi:10.1068/a210587

Beaumont J, Inglis K (1989) Geodemographics in practice: developments in Britain and Europe. Environ Plann A 21(5):587–604

Benson S, Bruck G, Bodenhorn B (2006) Injurious names: naming, disavowal, and recuperation in contexts of slavery and emancipation. In: Vom Bruck G, Bodenhorn B (eds) An anthropology of names and naming. Cambridge University Press, Cambridge, pp 178–194

Bertrand M, Mullainathan S, Ullainathan SEM (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Am Econ Rev 94(4): 991–1013

Besnard P, Desplanques G (2001) Temporal stratification of taste: the social diffusion of first names. Revue francaise de sociologie 42:65–77

Birkin M, Clarke G (1998) GIS, geodemographics, and spatial modeling in the UK financial service industry. J Hous Res 9(1):87–111

Bloothooft G, Groot L (2008) Name clustering on the basis of parental preferences. Names 56(3): 111–163. doi:10.1179/175622708X332851

Bloothooft G, Onland D (2011) Socioeconomic determinants of first names. Names 59(1):25–41. doi:10.1179/002777311X12942225544679

Bolin AU (2005) The effects of first name stereotypes on ratings of job. Am J Psychol Res 1(1): 11–20

Burnard T (2001) Slave naming patterns: onomastics and the taxonomy of race in eighteenth-century Jamaica. J Interdiscip Hist 31(3):325–346

Busse TV, Seraydarian L (1978) Frequency and desirability of first names. J Soc Psychol 104(1): 143–144. doi:10.1080/00224545.1978.9924050

Callary E (1997) The geography of personal name forms. Prof Geogr 49:494

Cheshire J, Longley P (2011) Spatial concentrations of surnames in Great Britain. Procedia Soc Behav Sci 21:279–286. doi:10.1016/j.sbspro.2011.07.047

Cole K, Dingle R, Bhayani R (2005) Pledger modelling: help the aged case study. Int J Nonprofit

Copley JE, Brownlow S (1995) The interactive effects of facial maturity and name warmth on perceptions of job candidates. Basic Appl Soc Psychol 16(1–2):251–265. doi:10.1080/01973533.1995.9646112

Cotton J, O'Neill B, Griffin A (2008) The "name game": affective and hiring reactions to first names. J Manag Psychol 23(1):18–39. doi:10.1108/02683940810849648

Crisp DR, Apostal RA, Luessenheide HD (1984) The relationship of frequency and social desirability of first names with academic and sex role variables. J Soc Psychol 123(1): 143–144. doi:10.1080/00224545.1984.9924527

Darden DK, Robinson IRAE (1976) Multidimensional scaling of men's first names: a sociolinguistic approach. Sociometry 39(4):422–431

Dinur R, Beit-Hallahmi B, Hofman JE (1996) First names as identity stereotypes. J Soc Psychol 136(2):191–200. doi:10.1080/00224545.1996.9713993

Erwin PG (1993) First names and perceptions of physical attractiveness. J Psychol 127(6):625–631

Experian UK Ltd. (2006) Mosaic consumer classification for the UK. Experian UK Ltd., Nottingham, Nov 2006. Retrieved from http://www.business-strategies.co.uk/upload/downloads/mosaic uk brochure.pdf

Finch J (2008) Naming names: kinship, individuality and personal names. Sociology 42(4): 709–725. doi:10.1177/0038038508091624

Fryer RG, Levitt SD (2004) The causes and consequences of distinctively black names. Quart J Econ 119(3):767–805

Fucilla JG (1943) The anglicization of Italian surnames in the United States. Am Speech 18(1): 26–32

Garwood S (1976) First-name stereotypes as a factor in self-concept and school achievement. J Educ Psychol 68(4):482–487

Guéguen N, Pascual A (2011) Are people with attractive names more employable? An evaluation in a field setting. Eur J Econ 38:164–166

Hanks P, Tucker DK (2000) A diagnostic database of American personal names. Names 48(1): 59–69

Hanks P, Hardcastle K, Hodges F (2006) Oxford dictionary of first names. Oxford University Press, Oxford

Harari H, McDavid JW (1973) Name stereotypes and teachers' expectations. J Educ Psychol 65(2):222–225

Herzfeld M (1991) A place in history: social and monumental time in a Cretan town. Princeton University Press, Princeton, NJ

Heuvelink C (2012) Forming impressions from English and French first names: is there an in-group effect in Québec? Psychol Rep 110:166–172. doi:10.2466/07.17.28.PR0.110.1.166-172

Hogan B, Berry B (2011) Racial and ethnic biases in rental housing: an audit study of online apartment listings. City Commun 10(4):351–372. doi:10/1111/j.1540-6040.2011.01376.x

Iceland J, Mateos P, Sharp G (2011) Ethnic residential segregation by nativity in Great Britain and the United States. J Urban Aff 33(4):409–429. doi:10.1111/j.1467-9906.2011.00555.x

Jayaraman RDA (2005) Personal identity in a globalized world: cultural roots of Hindu personal names and surnames. J Popular Cult 38(3):476–490

Jenkings R (1996) Social identity. Routledge, London

Kalist D, Lee D (2009) First names and crime: does unpopularity spell trouble? Soc Sci Quart 90(1):39–49. doi:10.1111/j.1540-6237.2009.00601.x

Kaplan A (1964) The conduct of inquiry: methodology for behavioral science. Chandler, San Francisco

Kotilainen S (2011) The genealogy of personal names: towards a more productive method in historical onomastics. Scand J Hist 36(1):44–64, Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21542200

Lawson ED (1971) Semantic differential analysis of men's first names. J Psychol 78:229–240

Lawson ED (1984) Personal names: 100 years of social science contributions. Names a Journal of Onomastics 32(1):45–74

Lawson ED (2004) Religious, patriotic, and ethnic factors involved with names and naming in Russia, Latvia, Lithuania, and Azerbaijan. In: Boullón Agrelo AI (ed) Novi te ex nomine: estudos filolóxicos. Fundación Pedro Barrié de la Maza, A Coruña, Spain, pp 1–10

Levitt SD, Dubner SJ (2005) Freakonomics: a rogue economist explores the hidden side of everything. Penguin Books, London

Levitt P, Glick-Schiller N (2004) Conceptualizing simultaneity: a transnational perspective on society. Int Migrat Rev 38(3):1002–1039. doi:10.1111/j.1747-7379.2004.tb00227.x

Lieberson S (2000) A matter of taste: how names, fashions, and culture change. Yale University Press, New Haven, CT

Lieberson S, Bell EO (1992) Children's first names: an empirical study of social taste. Am J Sociol 98(3):511–554

Lieberson S, Mikelson KS (1995) Distinctive African American names: an experimental, historical, and linguistic analysis of innovation. Am Sociol Rev 60(6):928–946

Livingstone RDA-A (2005) Some aspects of German-Jewish names. Ger Life Lett 58(2):164–181

Longley P, Webber R, Li C (2006) The UK geography of the e-society: a national classification. CASA Working Papers nr. 111. UCL London. Retrieved from http://discovery.ucl.ac.uk/3343/1/3343.pdf

Massey DS, Denton NA (1988) The dimensions of residential segregation. Soc Forces 67:281–315

Mateos P (2007) An ontology of ethnicity based upon personal names. Implications for neighbourhood profiling. Unpublished PhD Thesis. Department of Geography, University College London, London. Retrieved from http://eprints.ucl.ac.uk/16145/

Mateos P, Tucker DK (2008) Forenames and surnames in Spain in 2004. Names 56(3):165–184. doi:10.1179/175622708X332860

Mateos P, Webber R, Longley PA (2006) How segregated are name origins? A new method of measuring ethnic residential segregation. In: Priestnall G, Aplin P (eds) GIS Research UK 14th Annual Conference (GISRUK). University of Nottingham, pp 285–291

Mateos P, Webber R, Longley PA (2007) The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. (C. for A. S. Analysis, Ed.) CASA Working Paper 116. University College London, London. Retrieved from http://www.bartlett.ucl.ac.uk/casa/publications/working-paper-116

Mateos P, Longley PA, O'Sullivan D (2011) Ethnicity and population structure in personal naming networks. PLoS One 6(9):e22943. doi:10.1371/journal.pone.0022943

McCarty C, Bernard HR, Killworth PD, Shelley GA, Johnsen EC (1997) Eliciting representative samples of personal networks. Soc Networks 19(4):303–323

Mead G (1934) Mind, self, and society from the standpoint of a behaviorist. University of Chicago Press, Chicago

Mencken HL (1963) The American language: a preliminary inquiry into the development of English in the United States. Alfred A. Knopf, New York

Mitchell V, McGoldrick P (1994) The role of geodemographics in segmenting and targeting consumer markets: a Delphi study. Eur J Market 28(5):54–72

Nieto MG (2000) Evolución del nombre de pila en el País Vasco peninsular. Fontes linguae vasconum: Studia et documenta 32(83):151–168

Osgood CE, Suci GJ, Tannenbaum PH (1957) The measurement of meaning, vol 49. University of Illinois Press, Urbana-Campaign, IL

Parekh S, Parekh S (2003) Asian babies' names from the hindu, muslim and sikh traditions. Elliot Right Way Books, Tadworth, Surrey

Razum O, Zeeb H, Akgun SDA (2001) How useful is a name-based algorithm in health research among Turkish migrants in Germany? Trop Med Int Health 6(8):654–661

Riach PA, Rich J (2002) Field experiments of discrimination in the market place. Econ J 112(483): F480–F518. doi:10.1111/1468-0297.00080

Rickel AU, Lynn RA (1981) Name ambiguity and androgyny. Sex Roles 7(10):1057–1066

Rossi AS (1965) Naming children in middle class families. Am Sociol Rev 30:499–513

Rymes B (2001) Names. In: Duranti A (ed) Key terms in language and culture. Blackwell, Oxford, pp 158–161

Savage M, Burrows R (2007) The coming crisis of empirical sociology. Sociology 41(5):885–899. doi:10.1177/0038038507080443

Stewart J-AL, Segalowitz SJ (1991) Differences in the given names of good and poor readers. Can J Educ/Revue 16(1):103–105

Sue CA, Telles EE (2007) Assimilation and gender in naming. Am J Sociol 112(5):1383–1415

Tan PKW (2004) Evolving naming patterns: anthroponymics within a theory of the dynamics of non Anglo Englishes. World Englishes 23(3):367–384

Tucker DK (2002) Distribution of forenames, surnames, and forename-surname pairs in Canada. Names 50(2):105–132

Tucker DK (2003a) Surnames, forenames and correlations. In: Hanks P (ed) Dictionary of American family names. Oxford University Press, New York, pp xxiii–xxvii

Tucker DK (2003b) An analysis of the forenames and surnames of England and Wales listed in the UK Census data. Onoma 38:181–216

Tucker DK (2009) Increased competition and reduced popularity: US given name trends of the twentieth and early twenty-first centuries. Names 57(1):52–62. doi:10.1179/175622709X402663

Waldfogel J (1999) Preference externalities: an empirical study of who benefits whom in differentiated product markets. NBER Working Paper No. 7391, National Bureau of Economic Research, Washington. Retrieved from http://www.nber.org/papers/w7391

Wallace R, Wolf A (1999) Contemporary sociological theory. Prentice Hall, New Jersey

Williams W (1956) Gosforth: the sociology of an English village. Free Press, Glencoe, IL

Wolfram W, Thomas E (2008) Development of African American English. Blackwell, Oxford

Woods RD (1984) Hispanic first names: a comprehensive dictionary of 250 years of Mexican-American usage. Greenwood, Westport, CN

Zelinsky W (1970) Cultural variation in personal name patterns in the eastern United States. Ann Assoc Am Geogr 60(4):743–769

Zittoun T (2004) Symbolic competencies for developmental transitions: the case of the choice of first names. Cult Psychol 10(2):131–161

Part II
Methods: Name-Based Ethnicity Classifications

Chapter 6
Classifying Ethnicity Through People's Names

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

Abstract Several approaches have been proposed to classify populations into ethnic groups using people's names, as an alternative to ethnicity self-identification information when this is not available. These methodologies have been developed, primarily in the public health literature in different countries, in isolation from and with little participation from demographers or social scientists. This chapter brings together these isolated efforts and provides a coherent comparison, a common methodology and terminology. A systematic review of the most representative studies that develop new name-based ethnicity classifications has been conducted, extracting methodological commonalities, achievements and shortcomings. Their current limitations are mainly due to a restricted number of names and a partial spatio-temporal coverage of the reference population datasets used to produce name reference lists. The chapter concludes with a review of unconventional computational approaches that set the baseline for the development of an innovative name classification methodology in the next chapter (Chap. 7).

The value of using personal names to delineate ethno-cultural groups has been established through the arguments and evidence presented in the preceding five chapters. The basic hypothesis of this methodology is that the classification of surnames and forenames into ancestral groups of origin provides a viable

This chapter is partly based on material previously published in: 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. Part of the text, tables and figures from this article is reproduced here with permission of the journal publisher.

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_6, © Springer-Verlag Berlin Heidelberg 2014

alternative to create subdivisions of populations into groups of common origin or classifications of neighbourhoods according to population diversity. As discussed in Chap. 2, this is of particular importance when ethnicity, linguistic or religious data are not explicitly available at appropriate temporal, spatial or nominal (number of categories) resolutions. In this chapter the different theoretical and methodological approaches that have been independently developed in the fields of public health/ epidemiology, population genetics, linguistics, and statistics are reviewed. The purpose is to bring together these isolated efforts from very different research angles, so far reduced to the study of a small number of ethnic groups in a few migration destination countries. This is achieved through a systematic comparative analysis of the existing literature, proposing a common terminology in order to foster new research and applications. The chapter is organised in four sections. Section 6.1 reviews historical efforts in the mid-twentieth century to classify populations into some ethnic groups using name origins. As shown at the end of Chap. 3, the US government historically has been a key player in the use of this approach to population classification, and in this chapter its influence in more recent time periods are reviewed. Section 6.2 introduces the literature review conducted to select relevant academic studies which have built name-based ethnicity classifications and have evaluated their accuracy using an independent source. The key common aspects of the 13 selected studies are established in this section, while Sect. 6.3 focusses on how these approaches have been evaluated, comparing their resulting accuracy through indicators that reflect several research validation dimensions. Finally, Sect. 6.4 introduces the reader to other—non-academic, computational and marketing—approaches developed in the commercial sector. This view is presented here in order to prove the broad interest in name analysis applications and to justify the need for innovative but transparent methodologies that are open to scrutiny in order to advance research in this field. This section, together with the conclusion, serves to justify the need for a new approach to name-to-ethnicity classification, such as the proposed in this book, which is then described in detail in Chap. 7.

6.1 A Recent History of Name-Based Ethnicity Classifications in the US

Chapter 3 ended presenting a case study featuring a well-known historical use of surnames to classify ethnicity using the individual responses to the US 1790 Census. As discussed then, this analysis was carried out in several phases during the first decades of the twentieth century commissioned by the US Congress in order to inform new migration policy. That precedent ended in 1932 with the presentation to Congress of the final study by the American Council of Learned Societies. Not long after, a separate name classification effort was developed by the

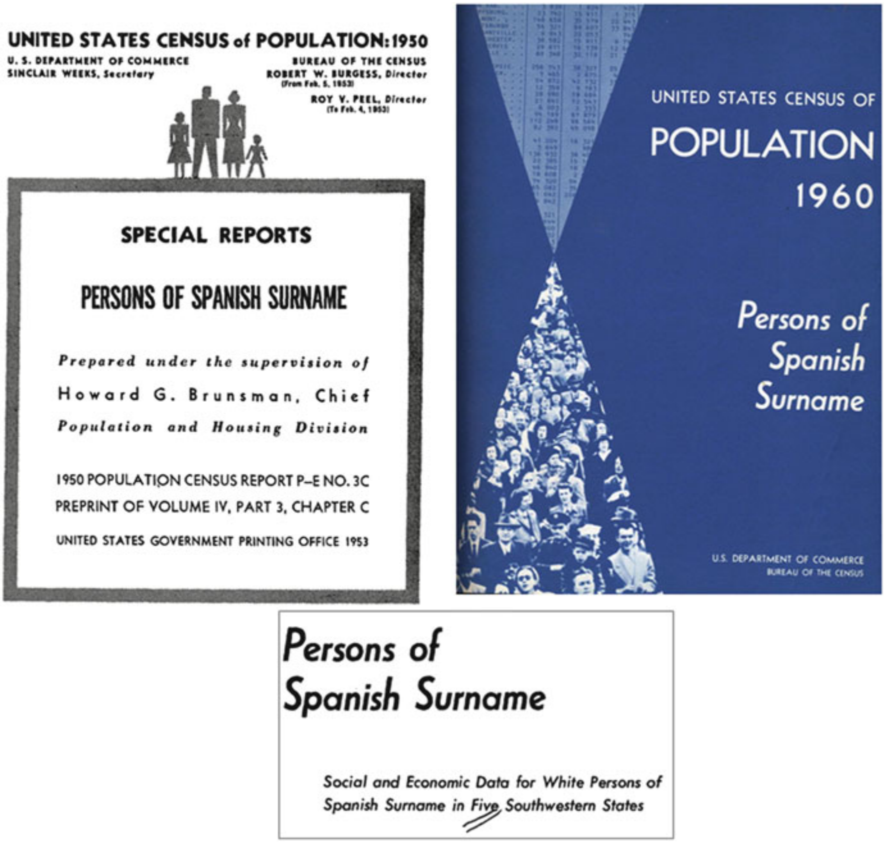
US Census Bureau, albeit this time the aim was to classify contemporary surnames as to ascribe ethnicity in the resident population, as opposed to historic migration.

Following the Second World War, there were large population movements in Europe and in the US, political boundaries were re-drawn, new nationality and citizenship rules were applied to migrants, large volumes of war refugees moved countries and ethnic minorities started to be recognised within national state systems. In this context, name origin analysis began to be used to ascribe ethnicity in the fields of demography and public health, especially with respect to Hispanic populations in the US (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 over several decades, lending robust official support to the use of these methods and their derived statistical results. Examples of the front cover of two of these Census outputs using the "Spanish surname list" in 1950 and 1960 Censuses, are shown in Fig. 6.1. These reports clearly represent the "official halo" involving the use of surname analysis, and the census validation of this approach.

A key figure in this effort was Robert Buechley, an epidemiologist who conducted studies on the health of Mexican migrants using Spanish surnames in the 1950s and 1960s (Buechley et al. 1957). He realised that the country of birth cross-tabulations commonly used by the Census Bureau, were insufficient to account for certain health inequalities prevalent in populations with Mexican ancestry. Some of these populations, and their ancestors, had actually never moved internationally. Their families had been living in the area now occupied by the US states of Texas, New Mexico, Arizona, California, Colorado and parts of Utah and Nevada, and which was annexed by the US in the 1848 war with Mexico. For the rest of the nineteenth century the new border was not actually enforced in preventing population movements from Mexico. Local movements that only seemed natural in this area continued for decades, as it had been happening for centuries under nomadic indigenous groups as well as during the Spanish rule in North America. Therefore, in the 1940s only 3–4 generations had passed since the US-Mexico border had moved south, and hence Mexicans in the border states of the Southwest comprised a long-term ethnic minority, rather than a newly arrived migrant group as such. Hence the usefulness of the surname method to capture that Hispanic heritage during this epoch in the US.

Buechley 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 "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 50 years (1950–2000 Censuses) 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



UNITED STATES CENSUS of POPULATION: 1950

U. S. DEPARTMENT OF COMMERCE
SINCLAIR WEEKS, Secretary

BUREAU OF THE CENSUS
ROBERT W. BURGESS, Director
(From Feb. 5, 1953)
ROY V. PEEL, Director
(To Feb. 4, 1953)

SPECIAL REPORTS

PERSONS OF SPANISH SURNAME

Prepared under the supervision of
Howard G. Brunsman, Chief
Population and Housing Division

1950 POPULATION CENSUS REPORT P-E NO. 3C
PREPRINT OF VOLUME IV, PART 3, CHAPTER C

UNITED STATES GOVERNMENT PRINTING OFFICE 1953

UNITED STATES CENSUS OF
POPULATION
1960

Persons of
Spanish
Surname

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

Persons of
Spanish Surname

Social and Economic Data for White Persons of
Spanish Surname in Five Southwestern States

Fig. 6.1 Covers of U.S. Census publications on persons of Spanish surname (1950 and 1960). Reports summarising data from US Census of 1950 (*left*) and 1960 (*right*) for persons with Spanish surnames. An inset from the second page of the 1960 report is shown in the *bottom part*, noting the underlying intention to study persons of "white race" who happen to have Spanish surnames. This is one of the first examples of the U.S. Census Bureau struggling to accommodate the concept of ethnicity as distinct from race. These reports ended with the creation of the Hispanic "race" in the 1970 Census and as a separate "ethno-cultural" category from 1980 onwards. *Source*: US Bureau of the Census (1963, 1953)

et al. 1978). This work resulted in the widely used *Word-Passel* Spanish surname list in the 1980s (Passel and Word 1980; US Bureau of the Census 1980) and the *Word-Perkins* surname list in the 1990s (Perkins 1993; Word and Perkins 1996). These lists have been distributed as official statistics by the US Census Bureau for several decades (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) (Abrahamse et al. 1994). The innovativeness of this survey is that it classified both forenames and surnames, an exception in this field which still tends to concentrate just on surnames to ascribe ethnicity. Despite these efforts, no official Asian surnames list has been published or documented by the U.S. Census in the same way as the Spanish surnames list has, although there has been some attempts to produce alternative Asian and Middle Eastern name lists by various researchers (Elliott et al. 2008; Lauderdale and Kestenbaum 2000; Lauderdale 2006). Instead, the US Census Bureau has distributed frequency statistics of the most common surnames cross-tabulated by the race or Hispanic questions. Researchers can thus use these frequency lists to tailor them to a particular classification application.

Many other researchers have developed and applied these types of name-based ethnicity classification techniques. Indeed, academic interest in this area has grown very rapidly over approximately the last three decades, 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. The rest of this chapter will present a systematic review and evaluation of contemporary efforts to develop such name-based ethnicity classifications.

6.2 Name-Based Ethnicity Analysis: Building the Classifications

Given the level of interest in name-based techniques to ascribe ethnicity, 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 (Mateos 2007). There is a vast range of studies that developed, evaluated or applied these techniques, and therefore the purpose of this and the next section 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. The great majority of these studies have been carried out in public health/epidemiology taking a population studies (demographic) approach.

The objective of this and the next section (Sects. 6.2 and 6.3) 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 book. Through this review, the methodological commonalities, achievements and shortcomings of the selected studies have been extracted. This section (Sect. 6.2) presents a summary of this review, compares the main characteristics of the studies evaluated, and how they built the name to ethnicity classification, while the following section (Sect. 6.3) separately analyses their evaluation and the results of such comparison.

6.2.1 Literature Review

A literature search and systematic review was carried out to identify the most representative research papers that comply with three requirements:

1. Specifically deal with the problem of classifying lists of names of individuals into ethnic groups
2. Do so through the development of new methods, rather than applying those name lists developed by others
3. And that provide a full evaluation of their accuracy using an independent data source where reported ethnicity is known for a number of individuals

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).

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 6.1 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 6.1 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.

6.2.2 Structure of the Selected Studies

The 13 selected papers aimed to demonstrate a satisfactory accuracy rate in separating individuals of one—or just a few—ethnic minority group/s, 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

Table 6.1 Summary of the general characteristics of the 13 studies reviewed

	Geographical area of study		Name to ethnicity assignment	
Paper reference	Country (region)	Ethnic minorities (E.M.) classified	Method	Name components
Choi et al. (1993)	Canada (Ontario)	Chinese	A	S
Coldman et al. (1988)	Canada (British Columbia)	Chinese	A	F, S, M
Lauderdale and Kestenbaum (2000)	US (National)	Chinese, Japanese, Filipino, Korean, Indian and Vietnamese	A	S
Razum et al. (2001)	Germany (Rhineland-Palatinate and Saarland)	Turkish	A	F, S
Word and Perkins (1996)/ Stewart et al. (1999)	US (National)	Hispanic	A	S
Harding et al. (1999)	UK (Bradford and Coventry)	South Asian + Hindu, Muslim and Sikh	A	F, S
Cummins et al. (1999)	UK (Thames, Trent, W. Midlands and Yorkshire)	South Asian	A	F, S
Nanchahal et al. (2001)	UK (London, W. Midlands, Glasgow)	South Asian	A	F, S, M
Sheth et al. (1999)	Canada (National)	South Asian and Chinese	A/M	S
Martineau and 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 and Moll (2003)	Netherlands (Rotterdam; 1 Hospital)	Turkish, Moroccan, Surinamese	M	F, S
Nicoll et al. (1986)	UK (Selected areas)	South Asian	M	F, S
Harland et al. (1997)	UK (Newcastle)	Chinese	M	F, S

Method of name to ethnicity assignment: *A* automatic, *M* manual. Name components used in the classification: *S* surname, *F* forename, *M* middle name

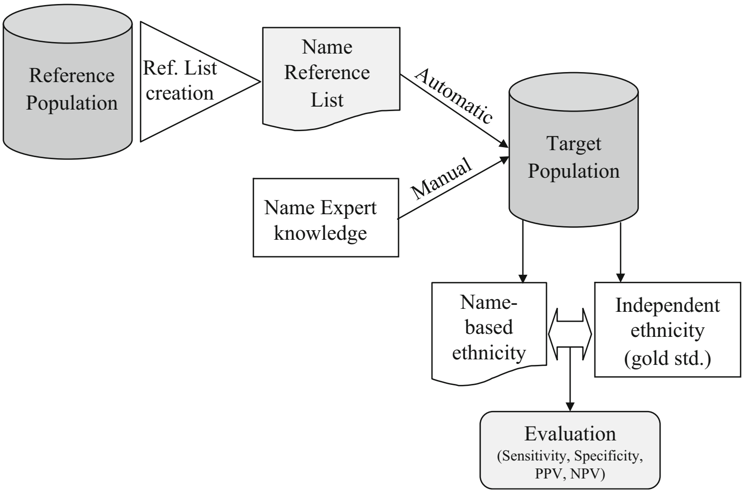


Fig. 6.2 Common structure and processes of name classifications

population. These common structures and processes are summarised as a flow chart in Fig. 6.2.

6.2.3 Source Data, Reference and Target Populations

The primary source materials 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 generally 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.

6.2.4 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 6.2. 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 6.2). Exceptions to this are Choi et al. (1993) and Coldman et al. (1988), with important consequences for their results, as will be discussed below.

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.

Table 6.2 Characteristics of reference populations and reference lists in the automatic methods

	Reference population					Reference list		
Paper reference	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 et al. (1988)	203,354	5,430	2.7	Death registrations	1950–1964	Ethnicity (family)	544	16 (Chinese) 1.7 (other)
Lauderdale and 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 et al. (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)	2,782,00 (estimated)	N/K	N/K	Canadian mortality data base (CMBD)	1979–1993	Country of birth (deceased and parents)	4,271	N/K
Word and 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 et al. (1999)	List of 2,995 surnames in *Nam Pehchan* program			*Nam Pechan* 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

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)

E.M. ethnic minority, *N/K* not known, *N/A* not available

6.2.5 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:

$$n \geq \frac{\log(1-x)}{\log y} \tag{6.1}$$

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*; $n = 13.4 \times 4 = 53.6$ people per surname and the results are presented in Table 6.3. 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.

6.2.6 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 6.4 ("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

Table 6.3 Comparison of actual reference population sizes used in five studies with the minimum reference population size criterion established by Cook et al. (1972)

	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 (13.4 × 4)	Actual size as % of Cook expanded criterion
Paper reference	*a*	*b*	$c = a \times 13.4$	*b*/*c*	$d = a \times 13.4 \times 4$	*b*/*c*
Choi et al. (1993)	427	1,899	5,722	33	22,887	8
Coldman et al. (1988)	544	5,430	7,290	74	29,158	19
Lauderdale and Kestenbaum (2000)	27,000	1,609,679	**361,800**	**445**	**1,447,200**	**111**
Razum et al. (2001)	12,188	108,500	163,319	66	653,277	17
Word and Perkins (1996)/ Stewart et al. (1999)	25,276	597,533	**338,698**	**176**	1,354,794	44

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; $n = 13.4 \times 4 = 53.6$ people per surname. Only two studies satisfy the first "Cook first criterion" and only one satisfies the "Cook expanded criterion" (highlighted in bold)

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 6.1).

Table 6.4 Summary of target population characteristics and results of the evaluation of classification accuracy in the 13 papers reviewed

		Target population						Method evaluation (single value or a range)			
Paper reference	Division of reference and target population	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 et al. (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 and 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 et al. (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 and 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 et al. (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) and (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 and 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 and White (1998)	N/A	137	107	78.1	Family health service authority register (FHSA)	Born Oct 1993–Sep 1994	Ethnicity (third party reported)	0.87–0.98 (outlier 0.5)	0.60–0.97	N/K	N/K

Bouwhuis and Moll (2003)	N/A	335	99	29.6	Hospital internal survey to parents of children	Sep–Dec 1999	Parents' country of birth (COB)	0.40–0.95	0.80–0.99	0.61–0.86	N/K
Nicoll et al. (1986)	N/A	846	348	41.1	(a) Child register, (b) school survey, (c) stillbirth certificate	N/K	Ethnicity [(3^{M} pty. (a), parents (b)]; Mother COB (c)	0.67–1.00	0.92–1.00	0.72–1.00	0.96–1.00
Harland et al. (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

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

E.M. ethnic minorities, *COB* country of birth, *N/K* not known, *PPV* positive predictive value, *NPV* negative predictive value

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 6.1 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).

6.3 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.

6.3.1 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 6.5 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").

The results for these four variables in the 13 studies are given in Table 6.4 ("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 6.4) 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.

Table 6.5 Explanation of measures of classification accuracy: sensitivity, specificity, PPV and NPV

Classification (predicted ethnicity)	Gold standard ("true" 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)

6.3.2 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 (Razum et al. 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 6.2: "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 2007).

(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)–(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).

6.3.3 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 Sect. 6.2.1), 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).

6.4 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 fill these three general gaps.

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 "computational and marketing approaches" will be reviewed in this section 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. Another innovative approach to the ethnicity classification of names was developed by Ken Tucker and Patrick Hanks (Hanks and Tucker 2000; Tucker 2003), in the field of Onomastics. Its contribution opened up a whole new research sub-field, and set the seeds upon which the name classification developed in this book is based (termed Onomap). Therefore, Tucker and Hanks approach is reviewed in the next chapter (Chap. 7) as to set the scene for our own proposed approach.

6.4.1 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. Another exception is Humpert and Schneiderheinze who built a German company around an international onomastics classification and published their approach (in German) (Humpert and Schneiderheinze 2000).

However, the market of global name classifications is dominated by US companies. 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, most of their public papers disappeared from their website (Dance 2007, Personal communication). 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 Management" 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 11 world (The Economist 2007).

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)
- Experian Mosaic Origins (http://www.experian.co.uk/business-strategies/mosaic-origins.html)
- Humpert & Schneiderheinze GbR (http://www.stichproben.de/)

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 do not reveal how they have tackled the wide range of issues associated with ascertaining the geographic, cultural and linguistic origin of names. They are indeed "black-boxes" and hence not useful to scientific research, which require elements of investigation that are fully open to scrutiny to be validated by third parties.

6.5 Conclusion

This chapter has reviewed a large number of research projects which developed name-based ethnicity classification techniques, in disciplines as diverse public health/epidemiology, demography, linguistics, computer science, and marketing. Research in this area has grown very rapidly over the past 30 years, following increasing interest in international migration and ethnicity, technological improvements in computing power, and the availability of large digital name datasets for whole countries 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 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. In a scale from 0 (no accuracy) to 1 (full accuracy), 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 in which ethnicity has not been collected.

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 fill the research gaps identified in this chapter. However, they lack the required transparency to be able to fully judge their merits.

To conclude, the chapter has demonstrated that formal name-based ethnicity classification methods make it possible to subdivide populations to a sufficient degree of accuracy when ethnicity information is not available, especially at the aggregate ethnic group level. The approaches evaluated here have produced relatively accurate total figures and order of magnitude estimates when compared to the little cost and effort involved. Therefore, 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.

Furthermore, 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 public directories, population registers and a growing realm of genealogical internet resources. In addition of this raw material, two other ingredients are required: a typology or taxonomy of cultural, ethnic and linguistic (CEL) groups into which to classify the names; and a set of techniques to perform the classification that places each name into a slot of the CEL taxonomy assigning a probability level to such attribution.

These are no trivial tasks and together comprise a research agenda that this book aims to promote in the future. The next chapter (Chap. 7) takes on this challenge to propose a new automatically generated names-to-ethnicity classification, overcoming some of the issues mentioned in this chapter. The rest of the book makes some additional research contributions towards filling some of these research gaps, justifying through selected applications the validity of this approach.

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. J Public Health Med 21(1):46–54

Abrahamse AF, Morrison PA, Bolton NM (1994) Surname analysis for estimating local concentration of Hispanics and Asians. Popul Res Policy Rev 13:383–398

Adebayo C, Mitchell P (2005) Patient profiling. Presented at GEONom, London, 25 May. Available at http://www.casa.ucl.ac.uk/geonom/Initial_meeting. Accessed 12 May 2006

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 Nov 2005

Bonaventura P, Gori M, Maggini M, Scarselli F, Sheng J (2003) A hybrid model for the prediction of the linguistic origin of surnames. IEEE Trans Knowl Data Eng 15(3):760–763

Bouwhuis CB, Moll HA (2003) Determination of ethnicity in children in the Netherlands: two methods compared. Eur J Epidemiol 18(5):385–388

Buechley RW (1961) A reproducible method of counting persons of Spanish surname. J Am Stat Assoc 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 10:63–66

Chaudhry S, Fink A, Gelberg L, Brook R (2003) Utilization of papanicolaou smears by South Asian women living in the United States. J Gen Intern Med 18(5):377–384

Choi BCK, Hanley AJ, Holowaty EJ, Dale D (1993) Use of surnames to identify individuals of Chinese ancestry. Am J Epidemiol 138:723–734

Coldman AJ, Braun T, Gallagher RP (1988) The classification of ethnic status using name information. J Epidemiol Community Health 42(4):390–395

Cook D, Hewitt D, Milner J (1972) Uses of the surname in epidemiologic research. Am J Epidemiol 95:38–45

Coronado GD, Koepsell TD, Thompson B, Schwartz SM, Wharton RS et al (2002) Assessing cervical cancer risk in Hispanics. Cancer Epidemiol Biomark Prev 11(10 Pt 1):979–984

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. J Public Health Med 2(4):401–406

Elliott MN, Fremont A, Morrison PA, Pantoja P, Lurie N (2008) A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. Health Serv Res 43(5):1722–1736

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

Fucilla JG (1943) The anglicization of Italian surnames in the United States. Am Speech 18(1):26–32

Hage BH, Oliver RG, Powles JW, Wahlqvist ML (1990) Telephone directory listings of presumptive Chinese surnames: an appropriate sampling frame for a dispersed population with characteristic surnames. Epidemiology 1(5):405–408

Hanks P, 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. Popul 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

Himmelfarb HS, Loar RM, Mott SH (1983) Sampling by ethnic surnames: the case of American Jews. Public Opin Q 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. J Nerv Ment Dis 186(11):677–683

Hofstetter CR, Hovell MF, Lee J, Zakarian J, Park H et al (2004) Tobacco use and acculturation among Californians of Korean descent: a behavioral epidemiological analysis. Nicotine Tob Res 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 Jun 2006

Humpert A, Schneiderheinze K (2000) Stichprobenziehung für telefonische zuwandererumfragen. Einsatzmöglichkeiten der namenforschung. ZUMA-Nachrichten 24(47):36–64

Jobling MA (2001) In the name of the father: surnames and genetics. Trends Genet 17(6):353–357

Kimmerle MM (1942) Norwegian-American surnames in transition. Am Speech 17(3):158–165

Kitano HH, Lubben JE, Chi I (1988) Predicting Japanese American drinking behavior. Int J Addict 23(4):417–428

Kolehmainen JI (1939) Finnish surnames in America. Am Speech 14(1):33–38

Lai DW (2004) Impact of culture on depressive symptoms of elderly Chinese immigrants. Can J Psychiatry 49(12):820–827

Lasker GW (1985) Surnames and genetic structure. Cambridge University Press, Cambridge

Lasker G (1997) Census versus sample data in isonymy studies: relationship at short distances. Hum Biol 69(5):733–738

Lauderdale DS (2006) Birth outcomes for Arabic-named women in California before and after September 11. Demography 43(1):185–201

Lauderdale D, Kestenbaum B (2000) Asian American ethnic identification by surname. Popul Res Policy Rev 19(3):283–300

Linguistic Minorities Project (1985) The other languages of England. Routledge & Kegan Paul, London

Longley PA, Maguire DJ, Goodchild MF, Rhind D (2005) Geographic information systems and science. Wiley, Chichester

Lyra F (1966) Polish surnames in the United States. Am Speech 41(1):39–44

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

Mateos P (2007) A review of name-based ethnicity classification methods and their potential in population studies. Popul Space Place 13(4):243–263

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. J Public Health Med 23(4):278–285

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. J Epidemiol Community Health 40(4):364–368

Passel JS, 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, CO, 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, CA, April 1982

Peach C, 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 Aug 2006

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 May 2005

Petersen W (2001) Surnames in US population records. Popul Dev Rev 27(2):315

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 Tob Res 7(1):103–109

Razum O, Zeeb H, Akgun S (2001) How useful is a name-based algorithm in health research among Turkish migrants in Germany? Trop Med Int Health 6(8):654–661

Rissel C, Ward JE, Jorm L (1999) Estimates of smoking and related behaviour in an immigrant Lebanese community: does survey method matter? Aust N Z J Public Health 23(5):534–537

Senior PA, Bhopal R (1994) Ethnicity as a variable in epidemiological research. Br Med J 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. Can Med Assoc J 161:132–138

Smith-Bannister S (1997) Names and naming patterns in England 1538–1700. Oxford University Press, Clarendon, PA

Stewart SL, Swallen KC, Glaser SL, Horn-Ross PL, West DW (1999) Comparison of methods for classifying Hispanic ethnicity in a population-based cancer registry. Am J Epidemiol 149(11): 1063–1071

The Economist (2007) What's in a name? *The Economist*, Technology Quarterly Survey, 10 March: 27

Tu SP, Yasui Y, Kuniyuki A, Schwartz SM, Jackson JC et al (2002) Breast cancer screening: stages of adoption among Cambodian American women. Cancer Detect Prev 26(1):33–41

Tucker DK (2003) Surnames, forenames and correlations. In: Hanks P (ed) Dictionary of American family names. Oxford University Press, New York, pp xxiii–xxvii

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. US Government Printing Office, Washington, DC

US Bureau of the Census (1963) U.S. Census of Population: 1960. Subject reports, Persons of Spanish surname. U.S. Government Printing Office, Washington, DC

US Bureau of the Census (1980) 1980 census of population and housing: Spanish list technical documentation. Data User Services Division, Washington, DC

US Census Bureau (2006) US Census Bureau geneaology resources. Available at http://www.census.gov/genealogy/www/. Accessed 12 May 2006

Williams K (2007) US Patent application: name classifier algorithm. Available at http://www.uspto.gov (search for patent number '20070005597'). Accessed 19 Mar 2007

Williams K, 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 26 May 2006

Winnie WW Jr (1960) The Spanish surname criterion for identifying Hispanos in the southwestern United States: a preliminary evaluation. Soc Forces 38(4):363–366

Word DL, 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 May 2005

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, TX, October

Yavari P, Hislop TG, Abanto Z (2005) Methodology to identify Iranian immigrants for epidemiological studies. Asian Pac J Cancer Prev 6(4):455–457

Chapter 7
Naming Networks and Clustering

Abstract Most of the literature on names and ethnicity reviewed in the previous chapters is typically only concerned with one of the two elements in a person's name, either forenames or surnames, but not both in conjunction. This is rather striking, since the previous chapters have demonstrated that different socio-cultural interactions result on uneven name frequencies between population groups and across space. These distinct naming practices simultaneously interplay with both surnames and forenames weaving distinct networks of naming connections between the two. This chapter first reviews one previous, yet rather limited, approach attempting to exploit these connections, to then propose an innovative network representation of such linkages. In doing so it establishes a remarkable relationship between cliques of highly dense connections of forename-surname pairs in social networks and cultural proximity of ethnic groups using network clustering techniques. The existence of these naming communities can be demonstrated without any prior knowledge about a name's origins. The resulting new name-based ethnicity classification, termed *Onomap*, conforms an innovative method of community assignment to reveal the degree of isolation, social integration or overlap between population groups in our rapidly globalising world.

The studies reviewed in the previous chapter have developed systems to classify populations primarily using surnames, with only a few of them also including forenames to their analysis. All of these studies required a reference list of names and their ethno-cultural origin, a sort of dictionary against which to classify a target list of specific individual people. Most of these studies built their reference lists by aggregating pre-existing databases of immigrant names by country of birth, hence

Parts of this chapter are based on previously published material as Mateos, P., Longley, P. A., & O'Sullivan, D. (2011). Ethnicity and Population Structure in Personal Naming Networks. PLoS ONE, 6(9), e22943. doi:10.1371/journal.pone.0022943. Reproduced here with permission. The author is very grateful to his co-authors on that paper for contributions to this chapter, specially to David O'Sullivan for key support to scale up the clustering techniques described.

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_7, © Springer-Verlag Berlin Heidelberg 2014

assuming these were representative of the ethnic group at large, at least in a particular destination country. Some of the studies also combined this approach with the procurement of names lists by origin, by recurring to published dictionaries, amateur genealogical interest groups or loose lists of common surnames by country, religion or ethnic group of origin published by a host of sources of various quality. In general, approaches to build such reference lists are *ad hoc* and hence difficult to reproduce and validate in different contexts.

The previous chapter, drawing from a thorough literature review, thus concluded by setting specific recommendations on how to improve the name reference lists compiled in this field. Specifically, it recommended including all ethnic groups present in a society (not just a few), expanding the pool of countries and spatio-temporal coverage from which name frequency data are drawn, and introducing name scores to measure the probability of a name being associated with a particular ethnic group. This chapter tackles these challenges and proposes a new approach to build a name-based ethnicity classification, in a way that is replicable, and most importantly does not require an extensive name reference list as a starting point. It does so by developing an innovative proposition; to represent forename-surname relationships as a network in order to apply network clustering techniques to identify naming communities within its structure. This approach, derived from the author's own research, draws upon a range of concepts selected from the linguistic, sociological and network science literatures. Key to this method is analysing the combined relationships between the ethno-cultural patterns in surnaming and forenaming practices, having separately reviewed its supporting literature in Chaps. 4 and 5. This combined forename-surname approach to ethnicity classification sets apart this book from the rest of the studies reviewed in previous chapters, which have surprisingly analysed issues of population structure focusing only in one of the two types of names.

Section 7.1 first reviews the only proposal found in the literature that has exploited the combined effects of surnames together with forenames as indicators of cultural or ethnic groupings. Section 7.2 proposes an innovative approach to build name reference lists using a naming network approach to automatic forename-surname clustering, justifying it theoretically in the context of social network theory. Section 7.3 describes the materials and methods used to demonstrate the validity of such naming network clustering approach. Section 7.4 presents the results of such validation using global name frequency data from 17 countries, mostly in Europe, as well as a single city in New Zealand, demonstrating that the concept of community structure in naming networks applies to both spatial scales. Section 7.5 finally introduces how this naming network clustering technique has been applied to build a new name-based ethnicity classification termed *Onomap*. The chapter closes with a set of conclusions on the implications of this approach, as a preamble to the presentation of a range of applications in the last two chapters of the book.

7.1 Previous Research in Forename-Surname Clustering: The 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 an experienced name researcher would have a productivity of assigning 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 that were of particular interest to linguists and historians (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.

7.1.1 The CELG Technique

Faced with the need to publish a large "Oxford Dictionary of American Family Names" (DAFN) (Hanks 2003), Tucker and Hanks developed a semi-automatic means of classifying a large list of surnames into ethno-cultural origins (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 DAFN. 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 at the end of the 1990s, from which he computed forename and surname frequencies and established relationships between the two sets of names.

This entailed a number of stages. First, a set of "diagnostic forenames" (deemed to be good predictors of a CELG) 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 would be the case with surnames, since there is a much smaller number of forenames than surnames in any given population. As explained in Chap. 5, a short number of forenames make the top list of a forename rank, most of which are typically distinctive of a CELG and hence easier to be intuitively classified by someone with a general language knowledge, than is the case with surnames. In this way, Hanks and Tucker (2000) took those forenames with a frequency greater than 9 in the US, totalling 85,000 unique forenames, and classified them by CELG creating a reference list, which for simplicity will be termed the “F_list”. For each forename in the 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 a CELG or not
- Gender: (Female, Male, Both, Unknown)
- Cultural Ethnic Linguistic Group (CELG): 1 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 to pick up “more forenames like them”, as will be explained in the next paragraphs.

Second, the F_list was linked to the forenames of the individuals in the US telephone directory. Third, a new surname reference list was produced, comprising all surnames in the telephone directory (1.75 million unique surname instances) and which will be termed the “S_list”, which was also linked to the telephone directory through the surnames of individuals. The structure of Tucker and Hank’s 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:

1. Forenames considered as “diagnostic” in the F_list are given double weight when computing the counts, i.e. counts are multiplied by a value of 2
2. Female forenames are weighted down to 80 % of their count values, i.e. counts are multiplied by a value of 0.8 (Tucker 2007, Personal communication). 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 in the US naming system.

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 as discussed in Chap. 5, 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).

Tucker's 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 × 2 = 20
12 female diagnostic forenames; 12 × 2 × 0.8 = 19.2
80 male non-diagnostic forenames; 80 × 1 = 80
60 female non-diagnostic forenames; 60 × 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 by Tucker 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.

7.1.2 A Small Trick; Forenames Are Much More Frequent Than Surnames

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 self-reported ancestry (Lieberson 2000). This important contribution has only been possible through the recent availability of digital registers containing almost entire populations at the level of the individual, including full forenames and surnames. Moreover, this method is very efficient because it leverages the differential skewness of the name frequency distribution between forenames (extremely positively skewed) and surnames (largely positively skewed). Let's explain how this works.

In most populations it has been found that there is a smaller number of forenames than surnames (the exceptions being some Asian language communities) hence presenting a clear difference in their frequency distributions (Tucker 2007). 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, as discussed in Chap. 5. This pattern is in contrast with the fixed nature of surnames, a proportion of which disappear due to a process of "natural selection" when there are no male descendants in a surname linage (Manni et al. 2005). This feature of names has been noted in different countries, for example in the U.S. (Tucker 2003), Spain (Mateos 2007b; Mateos and Tucker 2008), the UK (Mateos et al. 2007), and Canada (Tucker 2002).

Figure 7.1 illustrates this difference in the frequency distribution of forenames and surnames for the Great Britain's electoral register in 2004. The graph depicts the rank-size distribution of forenames and of surnames, each represented with a different line. The horizontal axis depicts the rank of names ordered from left (1 = most frequent) to right (1 million = least frequent) represented in a logarithmic scale (base 10). If the logarithmic scale is not used both curves are so highly positively skewed that no difference between the two is appreciated. The vertical

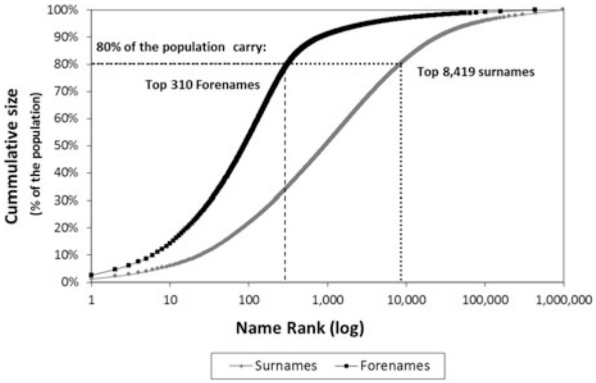


Fig. 7.1 Number of surnames and forenames (log scale) against proportion of the total population (Great Britain 2004). The graph depicts the rank-size distribution of forenames and of surnames (the two *lines*), in the 2004 Electoral Register of Great Britain. The *horizontal axis* depicts the rank of names ordered from *left* (1 = most frequent) to *right* (1 million = least frequent) represented in a logarithmic scale (base 10). The *vertical axis* represents the cumulative population size (proportion of the total population) at each point of the name rank. Hence as highlighted by the *dotted lines*, to reach 80 % of the population only 310 top forenames are required while for surnames this requires the 8,419 top surnames, showing the difference in the skewness of their name frequency distributions. *Source*: Adapted from Mateos et al. (2007)

axis represents the cumulative size (proportion of the total population) represented at each point of the name rank. Hence, as highlighted by the dotted lines, to reach 80 % of the population only the 310 top forenames are required, while for surnames this requires the 8,419 top surnames, showing the difference in the skewness of their name frequency distributions. The area between the two curves actually represents the number of people whose forename is included in a given population size threshold (e.g. 80 %) but whose 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 is the key to understand Hanks and Tucker Forname-Surname pairing technique, since it actually permitted the classification of thousands of names in a relatively effortless way. Their technique leverages on this differential frequency distribution in order to classify the ethnicity of surnames bore by a large proportion of the population using just a limited number of forenames. Thus Hanks and Tucker (2000) report that the initial diagnostic forenames list used in the DAFN had 85,000 unique forenames with a frequency greater than 9 in the US. To illustrate this property again with a US 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 the forenames' ethnicity of 95 % of the US population, and hence their corresponding surname ethnicity (Tucker 2001). Furthermore, by applying the CELG technique in several iterations this population coverage can be increased to nearly 100 %, while improving the overall accuracy of the names classified. This is further eased by building robust "diagnostic forenames" lists using etymological dictionaries of forename origins (sometimes also called books on "baby names"), which are much more common than surname dictionaries for most ethnic groups.

Using such knowledge of diagnostic forenames and the "correlations"—as they term them—between surnames and forenames in the telephone directory (Tucker 2003), Hanks and Tucker (2000) have established—albeit implicitly—the close links between the distinctive ethno-cultural naming processes of forenames and surnames that had only been separately identified in previous studies. Important though this discovery is, their work stops short of attempting to explain the socio-cultural linkage mechanisms behind these "correlations", it requires prior expert knowledge of good diagnostic forenames, and their analysis is limited to first order linkages (one surname to its bearer's forenames), rather than a broader vicinity of multiple surname-forename-surname configurations, that is, a broader *naming network* structure. In the rest of this chapter an innovative network analysis approach to forename-surname clustering will be developed in detail.

7.2 Naming Networks

7.2.1 Representing Forename-Surname Relationships as Networks

If indeed, as Hanks puts it, "*a person's name is a badge of cultural identity*" (Hanks 1990: vii), the aim of this chapter is to establish how they can be empirically classified using a representation of naming networks. This is done by building upon Hanks and Tucker's work to reveal the connections between forenames and surnames that lie beneath distinct cultural naming practices. Here, we go beyond this work to interpret the sum of these connections as "social networks" that emerge from the overlapping structure of millions of links between individually paired forenames and surnames. In such an interpretative framework, unique names can be arranged as nodes, connected through common bearers forming a network with distinctive ethno-cultural clusters each separated by various degrees of "social distance" in their naming practices.

The analysis presented in this chapter utilizes the pairings of *surnames*, which normally correspond to the components of a person's name inherited from his or her family (Hanks 2003), and *forenames*, which refer to the proper name given to a person, usually at birth. This technique necessarily only applies to societies that use both types of personal names.

The key contribution of this book is to conceptualise the ethno-cultural relationships between people as a network representation of personal names (vertices or nodes) connected by weighted forename-surnames pairs (links or edges). Such networks are derived from complete population registers such as telephone directories or electoral registers. Here, our main empirical analysis entails unsupervised classification of the topological structure of a naming network to detect ethno-cultural clusters using population registers from 17 countries across three continents. Surname networks are then extracted from the full network and weighted using relative frequencies of occurrence of shared forenames. Hence we here demonstrate that they have distinctive structure, which can be related to cultural, ethnic, and linguistic groups, and that they can reveal details of socio-cultural structure that are hard to identify by other methods.

7.2.2 Social Network Theory Applied to Naming Communities

The underlying hypothesis of this chapter is that naming networks structure mirrors socio-cultural structures in populations, as separately described in Chaps. 4 and 5. Drawing a parallel with *amazon.com*'s recommendation service; "people who bought this book also bought..." we could say that "people who bear this surname often choose these forenames". Pursuing this analogy, just like book titles at amazon.com are automatically clustered into genres using purchasing behaviour in a network representation (Newman 2006), we propose to cluster surnames into cultural, ethnic and linguistic groups of forenaming preference in a similar fashion using population registers. Let's first review a couple of concepts in social network theory that are useful to justify this approach.

Granovetter's (1973) "The Strength of the Weak Ties" presented a convincing argument that shifted the attention in social network research from the study of dense relations of a person's strong ties (one's close friends and family), to focussing on the importance of the casual or weak ties (one's acquaintances) in the diffusion of information. This distinction between the sections of social networks with low-density relationships (areas in which many of the possible relational links are absent) and the densely knit parts of the network (where many of the possible links are present) (Granovetter 1983), is very relevant to the research on name networks presented in this chapter.

The success of Granovetter's thesis was to shift attention away from redundant (dense) connections in a network and focus on the rare connections. Because densely knit relations provoke redundancy in the flow of information in the network, they have a sort of "provincialism" effect in their members, where local news are much more likely to travel than distant ones. On the contrary weak ties, play an essential role in the diffusion of information and innovation across the entire network, and as such they are often termed "bridges" in the social network literature. As Granovetter puts it "[t]he contention here is that removal of the average weak tie would do more 'damage' to transmission probabilities than would that of the average strong one" (1973: 1366). In fact, carrying on with the parallelism between social and naming networks, such "damaging effect" is actually what we should be aiming for in attempting to cluster naming networks. The idea is that by removing the weak links or bridges between name clusters we will end up isolating names into islands of self-contained clusters. This will have the same effect as removing the most cosmopolitan people or cross-cultural names from the names network (highly frequent names, names resulting from intermarriage, international families, or second generation migrants), leaving just the local connections that perfectly reproduce the cultural traits identified by the naming literature (see Chaps. 4 and 5). Therefore, the reverse of Granovetter's theory will be applied here; "the strength of the strong ties" in names networks, but the established measures for identifying weak ties or bridges will be used in order to "remove them".

Another parallel with the sociometry literature can be drawn in terms of "social distance" (Bogardus 1925) and "naming distance". In name networks, names are topologically close to one another as a result of intense shared naming (Lieberson 2000). This closeness would correspond with social and cultural groups that are close to each other in social space. In Bogardus seven levels "social distance scale" such situation would correspond to level one; "close kinship by marriage" (shared surnames and forenames), and level two; "would have as regular friends" (common forenames). The more segregated two groups are in social and physical space, the further their social distance in Bogardus scale, and the less intense the cross-linkages between their naming practices would appear in a naming network. At the same time, some names within any given cluster might present a peripheral position with respect to the "cluster core", and hence only related to other names within the cluster through intermediate links. Conversely, other names will be more central and highly connected to a large number of similar names.

Therefore, the social network framework of analysis of people's names presents a key advantage over Hanks and Tucker (2000) forename-surname pairing approach, as described in the previous section. The network approach proposed in this chapter allows to measure the connectivity of each name with all the rest up to several orders of adjacency, and not just to its immediate neighbours (close social distance). As such, to our knowledge this is the first study to propose and test this type of empirical approach to detect the ethnicity structure of whole populations using people's names.

7.3 Name Clustering Methodology

7.3.1 Building Naming Networks

The key idea underpinning the naming networks modelling approach presented here is that cultural-ethnic-linguistic (hereinafter "CEL") affiliations and practices are revealed as topological structures in a network in which unique forenames or surnames are considered as nodes, linked via common bearers. For any large population, network structure will manifest CEL communities (Hanks 2003) separated by the "social distance" of distinctive naming practices (Bogardus 1925). Figure 7.2 presents an illustrative two-mode (bipartite) network based upon

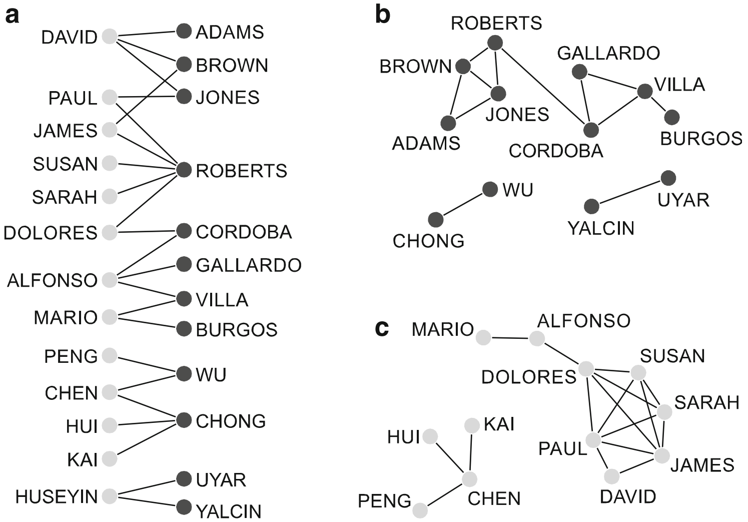


Fig. 7.2 Simple naming networks derived from a population of 23 people. (**a**) shows a two-mode network of 23 people, comprised of 13 unique forenames (*grey nodes*) and 12 unique surnames (*black nodes*) connected by 23 links each representing one person. (**b**) and (**c**) are one-mode transformations from network (**a**). (**b**) shows a one-mode network of the 12 surnames linked by common forenames, while (**c**) shows a one-mode network of 13 forenames linked by common surnames. Four CEL clusters emerge in (**b**); Anglo-Saxon, Spanish, Chinese and Turkish. Notice that the first two CELs networks are joined together by a cross-CEL name ("Dolores Roberts")

forename and surname (F-S) associations of 23 people (Fig. 7.2a), along with two derived one-mode associations based upon surnames (S-S) (Fig. 7.2b) and forenames (F-F) (Fig. 7.2c) alone. CEL cluster strength is reinforced by using one-mode networks, because of the multiplicative effect of combining the non-randomness of F-S and S-F links into a one mode (S-S or F-F) network. Here we will use only one-mode networks, defined by the preponderance of common cross-occurrences of (fore- or sur-) names within CEL communities, and their relative absence between communities.

We define *naming proximity* by the unequal probabilities with which forename–surname pairs are selected from the pool of all of those used in a society. The frequency of people connecting the two name components (nodes f and s) in the network (n_{fs}) becomes the weight of each link (fs). This weighted network is filtered using threshold values of "naming proximity".

Our fundamental premise is that the number of occurrences of a particular forename–surname pair n_{fs} will substantially exceed a naïve expectation of its rate of occurrence were forenames randomly selected within a population. Thus

$$n_{fs} > \left\lceil \frac{k\, n_f n_s}{N} \right\rceil \tag{7.1}$$

where k is the rate ($k \gg 1$) by which we require the observed number of cases of the forename-surname pair n_{fs} to exceed a naïve expectation, given n_f occurrences of the forename and n_s occurrences of the surname in the total population of N people. In practice, observed name associations are retained if the observed frequency exceeds expectations by a threshold k. Raising this threshold value focuses attention on the most strongly over-represented *fs* name-pair combinations, identifying the most tightly knit naming communities. The expected value of k is rounded up to the nearest integer count. This has the effect of removing from consideration name-pairs which occur only once (in practice a large number of pairs) which might otherwise be considered important because even one instance is many times more than the naive (random) expectation would suggest.

7.3.2 One-Mode Naming Networks

So far our analysis deals with a two-mode (bipartite) network, which is conveniently represented as a (sparse) n_s-by-n_f coincidence matrix $\mathbf{W}$ where non-zero entries w_{fs} represent the existence of the forename-surname combination *fs*. An important consideration here is the assignment of values to the coincidence matrix entries w_{fs}, representing the weights of the *fs* links in the two-mode network. In this chapter we are primarily interested in identifying surnames strongly connected by shared forenames, and therefore we define an *fs* weight as:

$$w_{fs} = n_{fs} / \sqrt{\frac{n_f (n_f - 1)}{2}} \tag{7.2}$$

The weight w_{fs} reflects the importance to forename f of the *fs* link with surname s. This approach is asymmetric in that if the interest is in clustering forenames strongly connected by shared surnames, it would be necessary to render the denominator of Eq. (7.2) dependent on the frequency of occurrence of the linked surname (n_s). A variety of formulations for this weighting were investigated, and it was found that provided that the weights increase with the number of instances of the *fs* name pair and decrease with the frequency of the forename in the population, the final outcome was not much affected. This approach reduces the importance of very common names that bridge CEL clusters (weak ties), in the one-mode network, and is desirable because such "cosmopolitan" names tend to obscure the distinctiveness of naming communities.

The coincidence matrix $\mathbf{W}$ is readily transformed into the edge-weights or adjacency matrix of either of the one-mode surname or forename graphs discussed above (Fig. 7.2b, c) by multiplication of $\mathbf{W}$ by its transpose:

$$\mathbf{D_s} = \mathbf{W^T W} \tag{7.3}$$

$$\mathbf{D_f} = \mathbf{W W^T} \tag{7.4}$$

where $\mathbf{D_s}$ and $\mathbf{D_f}$ are the distance (adjacency) matrices of the one-mode surname and forename networks respectively. This matrix multiplication is in effect a two-mode to one-mode network transformation, where the final strength of connection w_{ss} between two surnames in matrix **Ds** is given by the sum of products of the multiple w_{fs} connections to a set of common shared forenames. We describe this as the *naming proximity* (NP) between each pair of surnames x and y. Using Eq. (7.3), this can be expressed as

$$\mathrm{NP}_{xy} = \sum_{f} w_{fx} w_{fy} \tag{7.5}$$

Substituting Eq. (7.2) in Eq. (7.5) we formally define *naming proximity* (NP) between distinct surnames x and y as:

$$\mathrm{NP}_{xy} = \sum_{f} \frac{2 n_{fx} n_{fy}}{n_f \left(n_f - 1 \right)} \tag{7.6}$$

where x and y are distinct surnames, summation is over all shared forenames f, n_{fx} and n_{fy} denote the frequency of occurrence of forename-surname combinations f-x and f-y and n_f is the overall frequency of occurrence of forename f. In this chapter we cluster only surname networks linked via forenames, but the same procedure could in principle also be applied to forename networks (see Fig. 7.7 for a visual example of a forename type of network).

7.3.3 Input Data

One of the key strengths of the approach presented in this chapter lies in the ease of access to population register data to build a global naming network, as well as the availability of published work on the CEL origins of many names. Our analysis consisted of two stages. First, we developed a preliminary clustering analysis of the ethnically diverse population of Auckland, New Zealand, to demonstrate the existence of population structure in naming networks without any prior knowledge of CEL groups. Second, we extended this network clustering analysis using a global synthetic network covering 17 countries in four continents, using a dictionary of name origins to ascertain the CEL provenance of each cluster and to assess the accuracy of our automatic classification procedure.

Data used for this analysis derive from a very extensive database of 300 million people's names from 26 countries in four continents, assembled from publicly available telephone directories and electoral registers for a project developed at University College London (see http://www.worldnames.publicprofiler.org/). This database has been used, *inter alia* to build maps of population ethnic origins (Gibin et al. 2008; Williams 2011), to measure residential segregation (Mateos 2011) and to classify populations in public health registers (Lakha et al. 2011; Petersen et al. 2011).

Table 7.1 Description of the global names dataset with 17 WorldNames countries

Country name	Year	Country's population	Individuals in WorldNames	Forename-surname pairs
Austria	1997	8,316,487	2,516,864	1,707,653
Belgium	2007	10,511,382	3,378,147	2,504,949
Denmark	2006	5,457,415	3,075,509	1,153,183
Ex-Yugoslavia (a)	2006	10,159,046	1,704,633	757,355
France	2006	64,102,140	20,257,382	11,077,105
Great Britain	2006	60,587,300	45,688,172	11,454,381
Hungary	2006	10,064,000	281,305	162,683
India (4 city-regions b)	2004	*n/a*	321,662	250,818
Italy	2006	59,131,282	15,907,519	8,438,659
Luxemburg	2006	480,222	112,434	107,198
Norway	2006	4,770,000	3,581,614	2,071,687
Poland	2007	38,518,241	8,015,669	3,244,993
Romania (Bucharest)	2006	*n/a*	333,545	234,812
Slovenia	2007	2,019,245	344,709	277,934
Spain	2004	45,116,894	10,397,093	2,769,590
Sweden	2004	9,142,817	792,421	570,357
Switzerland	2006	7,508,700	1,559,532	1,204,039
Total			118,268,209	47,987,396

Summary of key characteristics from the global names dataset from 17 countries extracted from *WorldNames*. The year refers to the publication date of the telephone directory (Electoral Register in Great Britain), and the country's population refer to the closest available year

[a] Ex-Yugoslavia in 2006 includes current day Serbia, Montenegro & Kosovo

[b] The four city-regions in India are Delhi, Mumbai, Chennai, Hyderabad metropolitan areas

The first subset extracted from the dataset is the 887,021 electors resident in Auckland and recorded in the 2008 New Zealand Electoral Register (hereinafter Auckland dataset). This was chosen as a good example of a small yet ethnically diverse population of a single city, about which very little information is available in the naming literature. This subset comprised 79,855 unique surnames and 88,760 unique forenames, constituted in a two-mode network with 711,807 unique forename-surname pairs (links or edges).

The second subset of this database was created comprising records from 17 countries in Europe and the Indian subcontinent (see Table 7.1 for a full list of countries and name frequencies), in order to exclude imported naming systems in countries settled by colonisation—in which intermarriage between ancestral ethnic groups is likely to be greater. The extracted dataset comprised 118.3 million individuals in 17 countries, organised in a forename-surname network with 4.6 million unique surnames and 1.5 million unique forenames (hence 6.1 million nodes), and 46.3 million unique forename-surname pairs (links or edges: an average of 2.55 people per F-S pair).

Additionally, a reference list of "diagnostic" surnames whose cultural provenance is known was compiled from the academic literature and official statistical sources, in order to validate the results of network clustering (a full list of bibliographic and data sources is listed in Table 7.2). This comprised 30,479 surnames, each identified with 1 of 40 cultural ethnic and linguistic groups (CELs) which was used here as the "gold standard" (see Tables 7.1 and 7.3 for full details).

7.3.4 Network Clustering Analysis

The two datasets used in this analysis (Auckland's and the global 17-counties), are simply large registers of people's names, listing each person's forename and surname. These raw records were aggregated into forename-surname pairs along with their frequencies. They were initially represented as a two-mode (bipartite) network of forenames and surnames as nodes linked by forename-surname pairs as edges in a similar fashion to Fig. 7.2a. This two-mode network was subsequently transformed into a one-mode surname-to-surname (*s*-*s*) network (as in Fig.7.2b) and the unexpectedness rate (k) and naming proximity (NP) weights calculated for all links as specified in the previous section.

After finalisation of each weighted *s*-*s* one-mode network, standard network clustering algorithms were applied to detect its community structure (Girvan and Newman 2002). We have tested three different algorithms to find communities in very large networks following the criteria that they are able to handle very large weighted networks (up to 10,000 nodes and around a million edges) and that the chosen algorithm be implemented in some form of software capable of running within hours using a powerful desktop computer. The three candidate algorithms were *Fastcommunity* (Clauset et al. 2004), *Walktrap* (Pons and Latapy 2006) and *Label propagation* (Raghavan et al. 2007) which were all tested for their suitability in finding communities in very large naming networks. Clustering performance was measured using *modularity* (Q), defined as the quotient of the number of edges that fall within clusters to the number outside the clusters (Girvan and Newman 2002). *Walktrap* and *Label propagation* repeatedly came up with identical results, which were always outperformed by *Fastcommunity* in terms of higher modularity (Q) values. For ease of interpretation and conciseness only reports results based on the *Fastcommunity* clustering algorithm are reported here.

7.4 Analysis of Two Sets of Naming Networks

7.4.1 Auckland's Naming Network

The case study of Auckland, New Zealand, was chosen as a good example of a small yet ethnically diverse population of a single city, which has hitherto received very little attention in the naming literature. The naming network of Auckland's

Table 7.2 List of sources used to build the diagnostic surnames list

CEL group	Source	Peer reviewed or nat'l. stats	Number of surnames	Reference citation	Comments
Armenian	Federation of East European Family History Societies	No	1,818	Kazerian (1997)	
Belgium	Statbel	Yes	178	Statbel (2006)	
British	Mascie-Taylor and Lasker (1985)	Yes	85	Mascie-Taylor and Lasker (1985)	
Cambodian	Tu et al. (2002)	Yes	84	Tu et al. (2002)	
Chinese	Quan et al. (2006)	Yes	1,186	Quan et al. (2006)	
Czech	Kysilka (2009)	No	99	Kysilka (2009)	
Danish	Danmarks Statistik (Statistics Denmark)	Yes	20	Danmarks Statistik (2009)	
Finnish	Finnish Population Registration Centre	Yes	10	Finnish Population Register Center (2009)	
French	Darlu et al. (1997)	Yes	100	Darlu et al. (1997)	
German	Kunze (1999)	No	57	Kunze (1999)	
Greek	Dimitrios (2009)	No	405	Dimitrios (2009)	
Hungarian	Hungary's Ministry of Interior	Yes	100	Hungary's Ministry of Interior (2006)	
Iranian	Yavari et al. (2005)	Yes	25	Yavari et al. (2005)	
Irish	Tucker (2006)	Yes	13	Tucker (2006)	
Italian	Alfemminile.com	No	148	Alfemminile.com (2009)	Only the top 150 were taken
Jewish	Himmelfarb et al. (1983)	Yes	35	Himmelfarb et al. (1983)	
Norwegian	Statistics Norway	Yes	100	Statistics Norway (2008)	
Polish	Poland Ministry of Interior	Yes	33	Polish Ministry of Interior and Administration (2009)	
Russian	Balanovsky et al. (2001)	Yes	89	Balanovsky et al. (2001)	
Slovenian	Statistical Office of the Republic of Slovenia	Yes	100	Statistical Office of the Republic of Slovenia (2009)	

Spanish	Word and Perkins (1996)	Yes	890	Word and Perkins (1996)	Only the top 890 surnames were taken (heavily Hispanic)
Multi-CEL group					
Various Asian	Lauderdale and Kestenbaum (2000)	Yes	10,925	Lauderdale and Kestenbaum (2000)	
Various Middle East	Lauderdale & Morrison	Yes	22,362	Lauderdale (2006)	
Polish and various Slavic	Worldnames	No	400	Worldnames (2009)	
Several nationalities	Spanish National Statistics Institute (INE)	Yes	600	Instituto Nacional de Estadistica (2008)	

Notes: See the reference list at the end of this document for details of each individual citation. The sum total of the number of surnames column is 34,287 but this includes some duplicate entries of surnames that arrived to the reference list from different sources in the Multi-CEL group (the last four rows in the table). The total number of unique surnames is 30,479 as reported in the text

Table 7.3 List of CEL groups and name frequencies extracted from the global dataset

CEL code	CEL name	Number of unique surnames	Number of forename-surname pairs	Total number of people
afg	Afghan	255	1,525	1,907
afr	African	73	16,788	37,089
ara	Arabic	2,747	62,181	134,183
arm	Armenian	25	76	90
bri	British	80	308,143	8,455,394
bul	Bulgarian	17	2,428	4,057
cam	Cambodian	67	3,514	4,305
chi	Chinese	974	171,843	346,654
czk	Czech & Slovak	88	17,941	31,948
dan	Danish	20	78,877	1,558,343
dut	Dutch	115	90,344	335,331
fin	Finnish	10	1,596	5,899
fre	French	149	200,825	2,021,921
ger	German	62	98,722	489,983
gre	Greek	223	9,719	21,001
hun	Hungarian	92	38,521	137,040
ind	Indian	901	139,698	376,322
iri	Irish	26	42,422	682,850
ita	Italian	147	250,527	1,445,061
jap	Japanese	1,851	16,917	19,808
jew	Jewish	35	18,342	45,682
kor	Korean	82	3,990	5,623
lit	Lithuanian	20	241	261
nig	Nigerian	14	1,496	1,971
nor	Norwegian	83	112,860	759,288
pak	Pakistani	597	79,395	241,847
per	Persian	4,775	34,744	39,123
pol	Polish	196	87,723	1,202,623
por	Portuguese	20	35,478	162,787
rom	Romanian	37	11,717	27,862
rus	Russian	17	306	381
sla	Slavic	9	5,212	18,440
slo	Slovenian	96	26,493	53,440
spa	Spanish	880	667,778	5,477,346
ssl	South Slavic	199	118,456	776,359
sud	Sudanese	135	616	653
swe	Swedish	18	67,752	219,181

(continued)

Table 7.3 (continued)

CEL code	CEL name	Number of unique surnames	Number of forename-surname pairs	Total number of people
tur	Turkish	2,174	67,049	92,013
ukr	Ukrainian	18	1,087	6,221
vie	Vietnamese	84	16,397	38,078
Total		17,411	2,909,739	25,278,365

Definition of CELs and abbreviations adapted from Hanks and Tucker (Hanks and Tucker 2000; Tucker 2005). Out of the 30,479 unique surnames collected in the reference list (see text under Sect. 7.3.3) only 17,411 were present in the global names dataset (17 countries selected from WorldNames). This table lists the surname frequency distribution per CEL and the number of forename-surname pairs in which they are involved in the global names dataset used in this book
CEL cultural ethnic and linguistic groups

887,021 registered electors is shown in Fig. 7.3 was transformed into a surnames network and filtered at $k > 100$, $NP \geq 0.0$ (i.e. no NP filtering). We believe that this is the first naming network ever drawn of a complete city's population. The graph shows the highly structured outcome of naming practices in a city with high rates of immigration from all over the world, in which tightly knit clusters are strongly suggestive of CEL communities. In the centre of the graph, one giant connected component reflects the "majority of the population" whose surnames are connected with the largest number of other surnames through shared forenames. Visually, we can easily distinguish three distinct sub-components within this giant component, but its structure becomes much clearer after applying a community detection algorithm. Such network clustering techniques necessarily only work on a single connected component in a network, since the presence of any other isolated components already reflects membership of different communities (i.e. no clustering required). Therefore, we applied the *fastcommunity* algorithm to the giant component at the centre of Fig. 7.3.

We classified all of the surnames into 22 clusters, depicted using different colours in the colour version of this graph. One of the three sub-components is magnified in order to expose its surnames and structure (Fig. 7.3a), in this case names of South Asian origin, with the three node colours assigned by the cluster analysis indicating likely internal sub-structure (orange denotes Sikh, and green and blue different regions of India—see Fig. 7.3 caption for an on-line colour version of this figure). We have noticed that this giant component includes the most common names that are also the most likely to be found in other countries and also in the literature that traces each name's ethno-linguistic origins. However, if we turn our focus to the rest of the components in the graph, disconnected from the giant component, we find very interesting unique CEL communities that are particular to New Zealand. Three of these smaller components are magnified to show the tightly knit internal structure of their CEL communities, which from local knowledge we know are; Tongan (Fig. 7.3b), Samoan and other Pacific Islanders (Fig. 7.3c), and Eastern European (particularly Dalmatian, a late nineteenth century

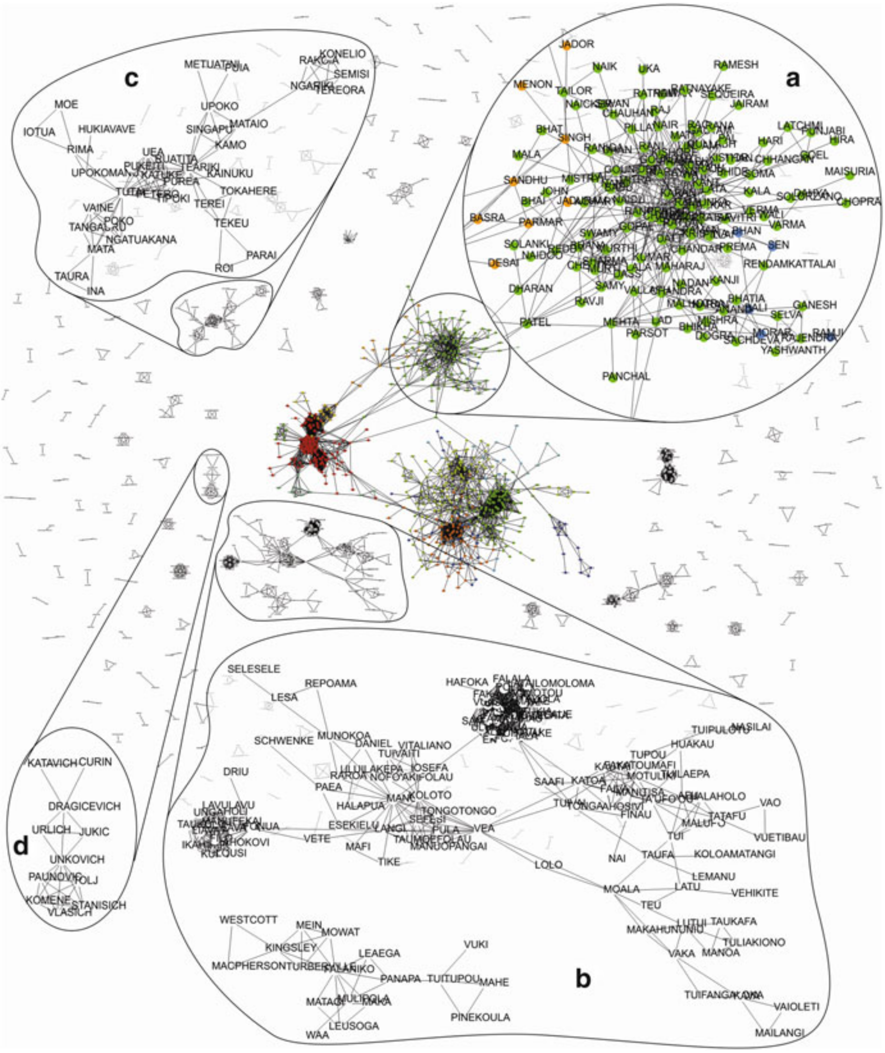


Fig. 7.3 Naming network of the city of Auckland New Zealand. The Auckland surnames network filtered at $k > 100$, $NP \geq 0.0$. The graph shows the highly structured outcome of naming practices in a city with high rates of immigration from all over the world. The giant component in the centre of the graph has been classified with *fastcommunity* algorithm into 22 clusters, each depicted by a different node *colour*. Four subgraphs are magnified to show the tightly knit internal structure of some CEL communities. One (**a**) is classified as part of the giant component (and is South Asian/Indian), the others are Tongan (**b**), Samoan and other Pacific Islanders (**c**), and Eastern European (particularly Dalmatian: **d**). The last three are disconnected from the network giant component. A colour version of this figure can be found at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0022943

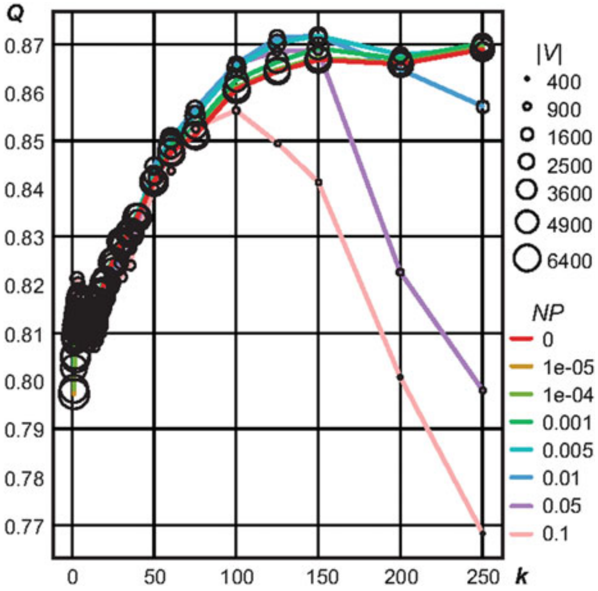


Fig. 7.4 Modularity results for different values of *k* and *NP* thresholds. Each point (*circle*) in the graph shows the modularity results (Q, *y*-axis) of running the *fastcommunity* algorithm (Clauset et al. 2004) on one-mode surname networks filtered using different values of *k* (*x*-axis), and naming proximity (*NP* as *line colours*), with the sizes of the *circles* (|V|) depicting the number of surnames (nodes) in the filtered network. A colour version of this figure can be found at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0022943

immigrant group: Fig. 7.3d). Other much smaller components are scattered around the periphery of this “constellation of naming galaxies”. The obvious tightly knit and geometrically compact topologies clearly show the outcome of the exclusive nature of naming practices, as predicted by the literature reviewed in Chaps. 4 and 5. It is striking that such clear ethno-cultural structure within a single city automatically emerges from the naming network representation proposed here, even without previous knowledge on the origins of these names or the existence of such communities in Auckland.

7.4.2 *Global Naming Network*

After demonstrating the existence of such clear structure in naming networks for a single city, we proceeded to undertake an analysis of the much larger 17 country “global dataset”. The diagnostic list of 30,479 surnames for which origins are asserted in published sources were linked to the matching surnames in the extracted global dataset (see Tables 7.1 and 7.3). The resulting two-mode network had 17,411 surnames linked to 243,135 forenames through 2,909,739 unique forename-surname pairs (see Table 7.3 for a breakdown by CEL group). We experimented with threshold values of *k* (Eq. 7.1) and *NP* (Eq. 7.6) when transforming this two-mode network into a one-mode surname network measuring the performance of *fastcommunity* in terms of modularity values (Q) and the final number of surnames (nodes(|V|) in the filtered network. Some results of this experimentation are shown in Fig. 7.4 and demonstrate that over-representation of a forename with

Table 7.4 Validation of clustering results: Percentage of surnames in cluster by reference CEL group

CEL group (ref. list)	Cluster ID (largest 20)																				Nr. of surnames
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
afr	**52**																				55
ara	13	**75**			1				1		3	3			9	7				1	469
bri			**73**																		13
iri			20											1							5
chi	10	2		**86**	40		5	7		3	2	1	1		1		3	10		1	466
vie				1	32													2			44
dut					2	**100**	**75**	**53**													46
fre	3		7		20		20	40						1			2	7			63
gre									**68**												58
ind	4	2		1							**83**	**87**			8	1		2			541
jap	2			2							1	1	**98**	1	1		3				343
nor														**55**							64
dan														17							20
swe														15							17
pak	2	2			1						3	4			**50**	1					291
per	6	17		1	1				4	4	4	3		1	24	**89**				2	778
spa	3			1	1				1					2			**72**	44			629
por																		32			20
ssl									4	6	1			1					**100**		103
tur		1							4	2		1			2	1				**93**	1,042
czk									1	16											46
slo				1						24											66
hun										30											78
Other CELs	5	1		7	2				17	15	3		1	6	5	1	20	3		3	**297**

Grand Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	5,554
Nr. of unique surnames	104	444	15	384	126	20	20	15	78	249	371	189	302	117	507	573	814	59	80	1,087	5,554
Most probable cluster-CEL	afr	ara	bri	chi	chi-vie	dut	dut	dut-fre	gre	hun-slo-czk	ind	ind	jap	nor-dan-swe	pak-per	per	spa	spa-por	ssl	tur	

The table shows clustering results on the global network filtered at $k = 150$ and NP $= 0$. The columns represent the largest 20 clusters and the rows the CEL groups in the diagnostic list, while the rows are a selection of 23 CELs with higher values in the table. The cell values are the percentages of unique surnames within each cluster that matches a particular CEL group in the reference list ($\geq$50 highlighted in bold). Percentages are rounded to the nearest integer and zero values are not shown. The largest 20 clusters shown here account for 5,554 surnames out of a total of 5,787 surnames assigned to 82 clusters. The last row lists the most probable CEL allocation (or CEL combination) to each cluster based on the highest percentages. For example, cluster 4 is 86 % Chinese, while cluster 9 is 68 % Greek, while cluster 6 is 100 % Dutch (see Table 7.3 for a description of CEL codes)

Table 7.5 Binary classification results of 14 families of CEL groups

Families of CEL groups	afr	ara	bri	chi-vie	dut	gre	hun-slo-czk	ind	jap	nor-dan-swe	pak-per	spa-por-ita	ssl	tur
Amalgamated cluster ID (Table 7.4)	1	2	3	4;5	6;7;8	9	10	11;12	13	14	15;16	17;18	19	20
Nr. of surnames	104	444	15	510	55	78	249	560	302	117	1,080	873	80	1,087
True positives	54	333	11	422	43	53	173	473	295	101	890	769	80	1,013
False positives	50	111	4	88	12	25	76	87	7	16	190	104	0	74
False negatives	1	136	2	88	3	5	17	68	48	0	179	19	23	29
True negatives	5,449	4,974	5,537	4,956	5,496	5,471	5,288	4,926	5,204	5,437	4,295	4,662	5,451	4,438
Classification accuracy														
Sensitivity	0.98	0.71	0.85	0.83	0.93	0.91	0.91	0.87	0.86	1.00	0.83	0.98	0.78	0.97
Specificity	0.99	0.98	1.00	0.98	1.00	1.00	0.99	0.98	1.00	1.00	0.96	0.98	1.00	0.98
PPV	0.52	0.75	0.73	0.83	0.78	0.68	0.69	0.84	0.98	0.86	0.82	0.88	1.00	0.93
NPV	1.00	0.97	1.00	0.98	1.00	1.00	1.00	0.99	0.99	1.00	0.96	1.00	1.00	0.99

This table summarises the binary classification results of an amalgamation of the 20 clusters shown in Table 7.4 into 14 amalgamated clusters that correspond to CEL families of one, two or three closely related CEL groups (as specified in the second row). The top half of the table shows the raw counts of surnames correctly or incorrectly classified according to the reference list, while the bottom half reports results of measures of classification accuracy (Sensitivity, Specificity, PPV = Positive Predictive Value, NPV = Negative Predictive Value) respect to a surname (k) drives the success of the clustering results, rather than the naming proximity metric (NP).

After filtering this global surname network at $k \geq 150$ and $NP \geq 0$, a giant component comprised of 5,787 nodes (surnames) was classified into 82 clusters using *fastcommunity*. The breakdown of surnames in each of the largest 20 clusters belonging to each CEL in the reference list is summarised in Table 7.4. For example cluster 4 is 86 % Chinese while cluster 9 is 68 % Greek and cluster 13 is 98 % Japanese. The great majority of these surnames (77 %) were assigned to clusters with a single CEL allocation in the reference list. The remainder presented a mix of multi-origin names or culturally close CEL groups, such as different Romance, Slavic, Germanic or Nordic languages, or Muslim names that cannot be attributed to a single CEL group. To accommodate some of these overlaps, pairs or triads of the largest 20 clusters were amalgamated into 14 clusters if they contained the same CEL or culturally similar CELs (see Table 7.5). Addition of these clusters increased the percentage of surnames "correctly" classified to 85 %. Measures of binary classification success were calculated for the 14 clusters, with very satisfactory results as shown in Table 7.5 (Sensitivity: 0.71–1; Specificity: 0.96–1; Positive Predictive Value: 0.52–1; Negative Predictive Value: 0.96–1; with ranges denoting extreme values for different CEL groups).

In order to produce a graph that can be clearly visualised, the global surname network was filtered with values of $k \geq 150$ and $NP \geq 0.01$, as shown in Fig. 7.5. The network's giant component of 2,232 surnames was classified into 53 distinct clusters (node colours), and cluster assignments remained consistent with those from the CEL reference list (shown with bounding boxes). The layout of sub-clusters within the graph, which places nodes in proximity to their directly connected nodes, clearly shows a geographical proximity arrangement of CELs, with frequent overlaps between some groups (e.g. between Spanish, Italian and Portuguese or between Chinese, Vietnamese, Cambodian and Korean names). CELs that are proximal in ethno-religious space, rather than in a geographical sense, also appear to share naming practices (e.g. Turkish, Arab, Persian and Pakistani names), or those close geographically but distant in ethno-religious space are distinctly clustered yet separated (e.g. Indian and Pakistani names or Chinese and Japanese names). Furthermore, it is striking to notice that although the global data are drawn principally from European countries, it is non-European CEL groups which show up clearly in the network analysis community structure. As we have argued, this is again proof of the distinctiveness of naming practices that are preserved after migration.

Finally, we illustrate the internal network structure of one particular CEL group in this dataset. Figure 7.6 presents a graph of the network formed by those surnames assigned as "Turkish" in the diagnostic list. This is an example of a highly coherent community as identified by the network clustering algorithm, where almost all Turkish surnames in the diagnostic list end up automatically assigned to the same cluster (orange) despite Turkey not being part of the 17 countries in the global dataset. Furthermore, the graph shows clear internal sub-structure within the Turkish surnames network, for example around a few "pivotal" surnames, which could

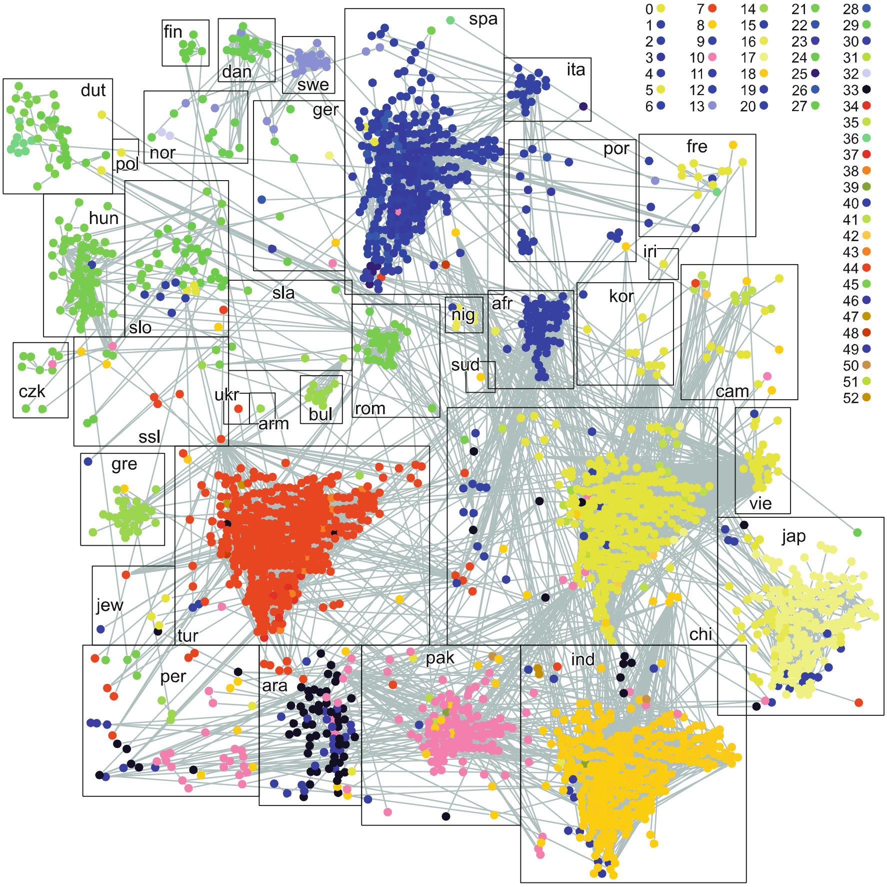


Fig. 7.5 Cultural clusters in the global surname network. Global surname network from 17 countries with 2,232 nodes (surnames) and 7,515 edges (shared forenames between each surname pair). Each node is *coloured* according to the cluster assigned by *Fastcommunity* ($k > 150$ $NP \geq 0.01$ producing 53 clusters), while the rectangles group surnames assigned to the same CEL group in the reference list (see Table 7.3 for CEL abbreviations) (see on-line version at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0022943.g004/largerimage)

be investigated further to understand relationships between communities of Turkish migrants in Europe.

The same reasoning could be used to cluster forename networks. Figure 7.7 shows a network of 328 forenames in the UK 2004 Electoral Register with origins in four ethnic groups; Nigerian, Vietnamese, Spanish, and Turkish CELGs. Each node in this network represents a unique forename while the links represent those forename pairs that share one or more surnames through common bearers. The cluster in the centre represents Nigerian forenames, at the top Vietnamese forenames, to the right Spanish forenames, and to the left Turkish forenames. Notice the strength of links between the Nigerian and Turkish clusters, because of some

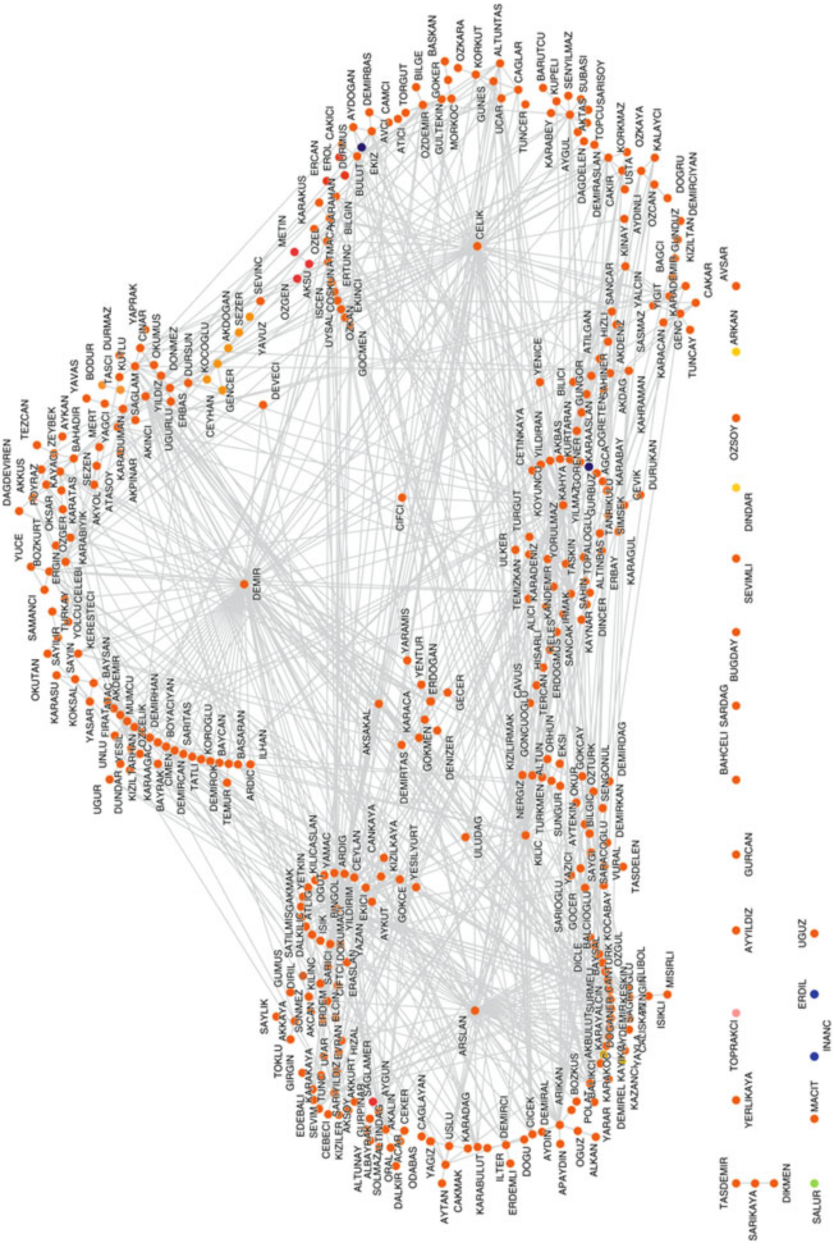


Fig. 7.6 Sub-network of Turkish CEL surnames. Nodes represented in the graph are those surnames assigned as "Turkish" in the reference list. The nodes are *shaded* by the clustered assigned by Fastcommunity. This is an example of a highly coherent cluster, where almost all Turkish surnames in the diagnostic list end up assigned to the same cluster (*orange*). Notice the highly structured nature of the network around some pivotal surnames that act as hubs connecting other surnames (e.g. *Celik*, *Demir*, and *Arslan*)

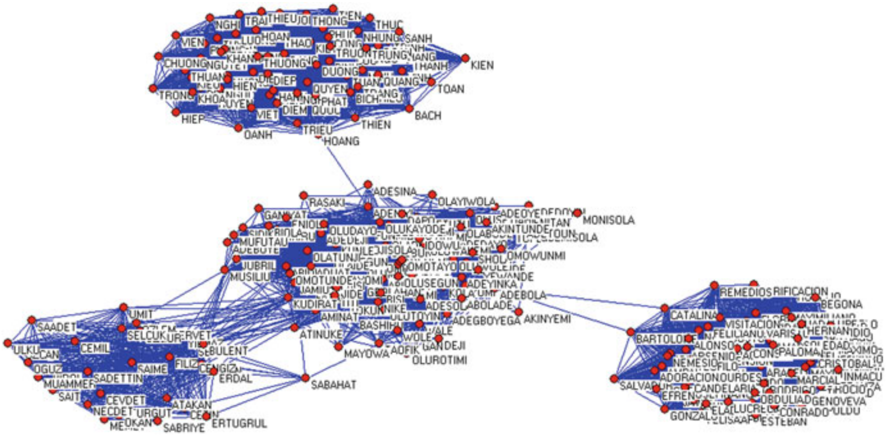


Fig. 7.7 A forename network of Nigerian, Vietnamese, Spanish, and Turkish forenames in the UK. The network includes a sample of forenames from four different ethnic groups. Each node (*red points*) represents a unique forename while the links (*blue lines*) represent those forename pairs that share one or more surnames through common bearers. The cluster in the *centre* represents Nigerian forenames, at the *top* Vietnamese forenames, to the *right* Spanish forenames, and to the *left* Turkish forenames. Notice the strength of links between the Nigerian and Turkish clusters, because of some common general Muslim surnames, and the sparseness of links between the other three clusters

common general Muslim surnames, and the sparseness of links between the other three clusters. This network of 328 forenames was derived from an initial sample of 401 connected forenames that was filtered using threshold values of *k* and *NP*, as explained in the previous sections, in order to eliminate potentially non-diagnostic forenames.

7.4.3 Discussion of Naming Networks Clustering Technique's Results

The naming network model proposed here demonstrates the existence of clear cultural naming practices based on much more complex attachments than geographic origins alone, and indicates that socio-cultural practices are sustained for generations after migration. Naming networks thus reveal the links that bind us together in communities of cultural practice, and provide a useful framework for classifying populations into cultural ethnic and linguistic communities.

This methodology is valuable for detecting the emergence of new naming communities, as well as revealing the ancestral hierarchies of cultural, ethnic, linguistic and religious attachment that underpin existing ones. Furthermore, the patterns that we have identified have been detected independently of geographic location. Additionally, sensitivity analysis allows investigation of overlaps and apparent exceptions when defining communities. In the context of millions of individuals across 17 diverse countries, the forcefulness of the evidence presented here is overwhelming.

7.5 The Onomap Classification

This final section describes how the naming networks clustering technique presented in the previous sections, was further expanded in order to build a global classification of forenames and surnames into categories of cultural, ethnic or linguistic (CEL) origin. Such classification was later termed *Onomap* and extensive evaluations and examples of applications have been published elsewhere (Booth et al. 2012; Dancygier 2010; Lakha et al. 2011; Wood et al. 2011). The aim here is to explain how the Onomap list of CEL categories was derived and provide a brief summary of how the classification was built.

7.5.1 Classifications of Human Groups

Using Harré's (1981) terms, what starts as a taxonomic collective—which only exists in the mind of the classifier—with time becomes a relational collective—whose members end up having real relationships with each other. With regards to ethnicity classifications, this is exactly what happened with various groups such as "Hispanics" in the U.S. This category was only introduced as an ethno-cultural label in the 1970 Census form. Since then, the concept of "Hispanic" as a true homogenous collective quickly took root, and in 2010 it was used for self-identification by 50.5 million people (16 % of the total population) conforming the largest ethnic minority in that country. However, the Hispanic etnicity has no meaning outside the U.S.

This is an illustrative example demonstrating that the practice of classification and measurement cannot be neatly detached from the actual phenomenon under consideration, in this case the concept of ethnicity as a socially constructed characteristic of human beings worth investigating. Various other ontological implications of this concept have been discussed in Chap. 2, and here reference will be only made to the task of constructing a workable and meaningful ethnicity classification of people's names.

In order to create an ethnicity classification of names, a purpose-built list of ethnicity categories must be conceived as a preliminary step. As reviewed in Chap. 6 the name classifications that predominate in the literature typically classify a particular geographical region or ethno/cultural group, such as for example; South Asian (Harding et al. 1999; Nanchahal et al. 2001); Asian American (Abrahamse et al. 1994; Lauderdale and Kestenbaum 2000), Hispanic (Morgan et al. 2004; Passel and Word 1980; Word and Perkins 1996), and Arab names (Lauderdale 2006). The effort presented here is aimed at creating a "universal" names classification, hence covering a global set of ethnic groups as much as possible.

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, linguists have established a genealogical tree of how current languages fit into language families with a common proto-language in the distant past. They do so by studying similar observable characteristics of contemporary 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) (see Chap. 4). This human taxonomy work from these three academic collectives has spanned almost two centuries, generally surrounded by a great deal of debate, controversy and speculation between different schools and political misuses of this research. Part of these efforts has been reviewed in Chap. 4, when discussing language evolutionary trees and their use in surnames and genetics. Let's link here the evidence between surnames, forenames and language trees.

As we have discussed in Chaps. 3 and 4, surnames derive from the languages in which they were originally created, largely between two to ten centuries ago. Since their coinage, they have been passed down to us in written form (sometimes in modified form) following specific morphologies, linguistic conventions and religious and civil identification rules, through those same origin languages and other foreign ones, when those holding particular surnames emigrated. Forenames follow similar linguistic rules but, as discussed in Chap. 5, are voluntarily selected and almost freely modified by parents following a set of cultural, religious, linguistic, social interaction and identity conventions. Forenames thus are more flexibly transmitted vertically through generations, and horizontally through uneven propagation across a specific spatio-temporal and social medium.

Therefore, in both cases of surnames and forenames, language represents the primary factor in the processes of creation, modification, transmission and geographic 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 (Darwin 1859) in the *Origin of Species* quoted at the beginning of Chap. 4.

7.5.2 Language Classification

There are several classifications of the world's languages, which aim to list the languages currently spoken and organise them into linguistic families. One of the language family classifications most widely accepted is that of Greenberg and Ruhlen (Cavalli-Sforza 2001), 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). Although there are many debates around these classifications, a body of literature share many elements of Greenberg and Ruhlen's classification (see Fig. 4.1).

The application of such macro-language families to particular languages is the field of linguistic cataloguists. One of the few standardised and updated catalogues of existing contemporary languages is the *Ethnologue* system. It comprises a language coding system organised hierarchically into a language taxonomy in combination with the international standard for language codes termed 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). The version of this standard used here is ISO 639-3, the third version of the international coding of languages, released in 2007, which contained 7,618 languages. *Ethnologue* 15th edition (Gordon 2005) was used as a basis to produce the Onomap Taxonomy described in the next section. It contained 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 http://www.ethnologue.com for the complete list and hierarchy). However, most of these 108 language families are considered language isolates, and in relative terms the great majority of the population worldwide is assigned to languages that fall within Ruhlen's 20 linguistic families.

The tandem *Ethnologue*—ISO 639-3 language classification forms the basis for the taxonomy of ethnicity based upon personal names developed to build the Onomap names classification. As such, this taxonomy initially distinguishes ethno-linguistic groups through the names currently present in the 26 countries included in *WorldNames* (http://www.publicprofiler.org/worldnames). This taxonomy is modified by cultural, religious and geographic criteria where required, in order to reflect the uniqueness of the group's names as available in the 26 countries represented in *WorldNames*.

7.5.3 The Onomap "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 book 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 most common languages spoken today in the 26 countries present in the WorldNames database, and modified by cultural, religious and geographic classifications when it was considered appropriate.

In this book the CEL concept is used as a basis for classifying both forenames and surnames currently present in the *WorldNames* database, defined as those names of residents with a frequency of three or more occurrences per surname or forename in a country. 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, simultaneously, 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), as described in previous chapters. The assumption underlying this book 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, frequencies, and, most importantly, pairings of the forenames or surnames that are assigned 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 (Mateos et al. 2007).

The criterion used to create the CEL taxonomy, both in DAFN and in Onomap, is primarily an onomastic one, that is, a list of human groups based on name origins. The taxonomy created in this research is based on a "bottom-up" approach, through 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, Catalan 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 Onomap 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 Onomap taxonomy presented here is optimised for the names present in the contemporary *WorldNames* data population, and currently includes 185 CEL categories of which 7 describe different aspects of "void or unclassified names" and 178 "true" CELs (see Table 7.6 for the complete list). The resulting Onomap 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 multilevel 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.

Table 7.6 The Onomap taxonomy (Onomap Types and their assignments into Onomap Groups)

Onomap Group	Onomap 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

The process by which the Onomap 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 process 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 "Onomap

Types", which are grouped into 15 coarser categories according to onomastic criteria and termed here "Onomap Groups" as detailed in Table 7.6.

7.5.4 Building the Onomap Classification

The core name classification methodology used to build Onomap is the naming networks and clustering method described earlier in this chapter. Starting with the initial list of "diagnostic" names identified in the literature (see Table 7.2), these were further expanded through the identification of communities (clusters) of names in either forename-forename or surname-surname one-mode networks (see examples in Figs. 7.3 and 7.5). As long as the great majority of the known "diagnostic" names within a cluster, pointed unequivocally to the same CEL Type, all the names in the cluster were provisionally classified under such category (see examples in Tables 7.4 and 7.5). Repeated iterations of the network clustering technique, applied to alternative subdivisions of the networks into smaller chunks of connected sub-networks were performed. After a three clustering iterations, if a name repeatedly fell within the same cluster of names and most of them (applying various thresholds) had been provisionally labelled with the same CEL Type in three iterations, that particular name was confirmed to belong to such Onomap Type. Hence the final decision was not only based on the overlap of the same Type in the three iterations, but also on the same recurring pattern happening of the "neighbouring names" in the same cluster. In turn, that name was added to the "diagnostic" set of names for inclusion in future iterations of the clustering algorithm. The analysis moved from the "core", highly connected components of the network (see for example Fig. 7.5), to smaller, less connected name networks, some of which comprised a single Onomap Type cluster in themselves. Examples of these smaller disconnected networks are visible at the margins of Fig. 7.3, which correspond to closely knit Maori, Samoan, and Fiji names in Auckland, New Zealand, and also Fig. 7.6, depicting Turkish regional sub-communities or even close families present in Western Europe.

Finally, various other methodologies were used to expand the pool of diagnostic names and to classify highly disconnected rare names, which are summarised elsewhere (Mateos et al. 2007; Mateos 2007b). Furthermore, extensive validations and applications of Onomap have been conducted in various fields, such as Public Health, Economics, Political Science, and Computer Science, and the reader is encouraged to consult these publications for further detail (Booth et al. 2012; Dancygier 2010; Lakha et al. 2011; Lewis et al. 2009; Nathan 2011, 2014; Petersen et al. 2011; Wood et al. 2011). In the various evaluations, the classification accuracy of Onomap ranges from 0.7 to 0.9 in the various measures of sensitivity, specificity, positive predictive value (PPV), and negative predicted value (NPV) (defined in Chap. 6). Furthermore, a range of applications of Onomap that relate to the geographical analysis of population diversity will be presented in detail in the next two chapters (Chaps. 8 and 9).

7.6 Conclusion

This chapter has developed an innovative proposition; to represent forename-surname relationships within people's names as a naming network in order to identify closely knit naming communities within its topological structure. This approach permits to develop a name reference list for the purpose of building name-to-ethnicity classifications. However, differing from the large majority of studies in this area, the naming network approach presented here does not require a large pre-existing name reference list by ethnic origin and derived from other sources. Drawing upon a range of concepts selected from the linguistic, sociological and network science literatures, the approach proposed in this chapter identifies community structure within naming networks. It does so by exploiting the combined relationships between surnaming and forenaming practices in ethno-cultural groups and the significant difference between their rank-size frequency distributions. This technique is then applied to a large database of nearly-complete populations of individual residents in 17 countries, in order to build a completely new name-to-ethnicity classification termed *Onomap*, including its own typology (termed Onomap taxonomy) of cultural, ethnic and linguistic groups (CEL).

The findings stemming from this chapter suggest that the net effects of human migration over the last several centuries has been to spawn new "naming communities", and that names remain important pointers to community membership—or the lack of it. As reviewed in the previous chapters of this book, naming practices provide enduring tokens of cultural affiliation in the era of globalisation. Conversely, the transience of naming conventions renders them important indicators of population composition over space and time, as well as to track the scale and pace of ethnic affinity and cultural change. In this chapter, we have linked disparate concepts from a range of diverse fields around such naming practices into an empirical investigation of naming, proposing a network representation of forename to surname relationships. Through the formulation of this approach, which lies at the core of this book, we hope to have moved the research frontier in name analysis breaking new grounds in different directions. Some of them will become clearer in the last two chapters of the book, since they present some applications of the Onomap classification at different geographical scales. As it will be shown, inherently vague concepts such as "social integration" or "spatial segregation" of minority groups may be monitored using this name classification approach. A consequence of this work may thus be supplementation of static mapping of fixed cultural and ethnic classifications in national Censuses with a more dynamic understanding of the human Diaspora in the broadest sense.

References

Abrahamse AF, Morrison PA, Bolton NM (1994) Surname analysis for estimating local concentration of Hispanics and Asians. Popul Res Policy Rev 13(4):383–398

Alfemminile.com. 2009. Classification of the most common surnames (translated from Italian). Available at http://cognome.alfemminile.com/w/cognomi/cognomi-piu-diffusi-in-italia.html. Accessed 29 July 2009

Balanovsky OP, Buzhilova AP, Balanovskaya EV (2001) The Russian gene pool: gene geography of surnames. Russ J Genet 37(7):807

Bogardus E (1925) Measuring social distances. J Appl Sociol 9:299–308

Booth AL, Leigh A, Varganova E (2012) Does ethnic discrimination vary across minority groups? Evidence from a field experiment*. Oxf Bull Econ Stat 74(4):547–573. doi:10.1111/j.1468-0084.2011.00664.x

Cavalli-Sforza LL (1997) Genes, peoples, and languages. Proc Natl Acad Sci 94(15):7719–7724

Cavalli-Sforza LL (2001) Genes, peoples, and languages. Penguin Books, London

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys 70(6):066111

Dancygier RM (2010) Immigration and conflict in Europe. Cambridge University Press, p 352

Danmarks Statistik (2009) Most popular forenames and surnames in the Danish population (translated from Danish). Statistics Denmark, Copenhagen. Available at http://www.dst.dk/Statistik/Navne/pop.aspx. Accessed 23 July 2009

Darlu P, Degioanni A, Ruffie J (1997) Some statistics on the distribution of surnames in France (translated from French). Population 52(3):607–634

Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London

Dimitrios J (2009) A database of Greek surnames and their ancestral origins. Available at http://www.dimitri.8m.com/surnames.html. Accessed 25 July 2009

Eriksen TH (2002) Ethnicity and nationalism, 2nd edn. Pluto Press, London

Finnish Population Register Center (2009) Most common surnames 2009. Available at http://192.49.222.187/Nimipalvelu/default.asp?L=3. Accessed 28 July 2009

Gibin M, Singleton A, Milton R, Mateos P, Longley P (2008) An exploratory cartographic visualisation of London through the Google Maps API. Appl Spat Anal Policy 1:85–97

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826

Gordon RG Jr (ed) (2005) Ethnologue: languages of the world. SIL International, Dallas, TX. Available at http://www.ethnologue.com/. Accessed 21 Nov 2006

Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380

Granovetter MS (1983) The strength of weak ties: a network theory revisited. Sociol Theory 1:201–233

Hanks P (1990) A dictionary of first names. Oxford University Press, Oxford

Hanks P (2003) Dictionary of American family names. Oxford University Press, New York

Hanks P, Tucker DK (2000) A diagnostic database of American personal names. Names 48(1):59–69

Harding S, Dews H, Simpson SL (1999) The potential to identify South Asians using a computerised algorithm to classify names. Popul Trends 97:46–50

Harré R (1981) Philosophical aspects of the macro–micro problem. In: Cicourel A, Knorr-Cetina K (eds) Advances in social theory and methodology. Toward an integration of micro- and macro-sociologies. Routledge, London, pp 139–160

Himmelfarb HS, Loar RM, Mott SH (1983) Sampling by ethnic surnames: the case of American Jews. Public Opin Q 47:247–260

Hungary's Ministry of Interior (2006) Most common surnames with a minimum of 10 000 people (translated from Hungarian). Central Personal Data Processing Office. Available at http://www.nyilvantarto.hu/kekkh/fixhtml/nepessegfuzet/2006/nepesseg2006/nepesseg_fuzet_2006_fuggelek_014_f-csalad.html. Accessed 29 July 2009

Instituto Nacional de Estadistica (2008) Most frequent first surname by nationality (translated from Spanish). INE, Madrid. Available at http://www.ine.es/daco/daco42/nombyapel/apellidos_por_nacionalidad.xls. Accessed 23 July 2009

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 5 Mar 2007

Kazerian N (1997) Armenian surnames. The Federation of East European Family History Societies. Available at http://www.feefhs.org/links/armenia/am-sur.html

Kunze K (1999) DTV-atlas of names. First and last names in the German language (translated from German). Deutscher Taschenbuch, München

Kysilka K (2009) Czech Surnames. Available at http://zlimpkk.tripod.com/Genealogy/czechsurnames.html. Accessed 25 July 2009

Lakha F, Gorman DR, Mateos P (2011) Name analysis to classify populations by ethnicity in public health: validation of Onomap in Scotland. Public health 125(10):688–696. doi:10.1016/j.puhe.2011.05.003

Lauderdale DS (2006) Birth outcomes for Arabic-named women in California before and after September 11. Demography 43(1):185–201

Lauderdale D, Kestenbaum B (2000) Asian American ethnic identification by surname. Popul Res Policy Rev 19(3):283–300

Lewis D, Mateos P, Longley P (2009) Choice and the composition of general practice patient registers. London

Lieberson S (2000) A matter of taste: how names, fashions, and culture change. Yale University Press, New Haven, CT

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

Mascie-Taylor CGN, Lasker GW (1985) Geographical distribution of common surnames in England and Wales. Ann Hum Biol 12(5):397–401

Mateos P (2007a) A review of name-based ethnicity classification methods and their potential in population studies. Popul Space Place 13(4):243–263

Mateos P (2007b) An ontology of ethnicity based upon personal names. Implications for neighbourhood profiling. Unpublished PhD Thesis, Department of Geography, University College London, London. Available at http://eprints.ucl.ac.uk/16145/

Mateos P (2011) Uncertain segregation: the challenge of defining and measuring ethnicity in segregation studies. Built Environ 37(2):226–238

Mateos P, Tucker DK (2008) Forenames and surnames in Spain in 2004. Names 56(3):165–184

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.bartlett.ucl.ac.uk/casa/publications/working-paper-116. Accessed 5 Mar 2007

Morgan RO, Wei II, Virnig BA (2004) Improving identification of hispanic males in medicare – use of surname matching. Med Care 42(8):810–816

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. J Public Health Med 23(4):278–285

Nathan M (2011) The economics of super-diversity: findings from British cities, 2001–2006. London School of Economics, London

Nathan M (2014) Same Difference? Ethnic inventors, diversity and innovation in the UK. Journal of Economic Geography (in press)

Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103(23):8577–8582

Passel JS, Word DL (1980) Constructing the list of Spanish surnames for the 1980 Census an application of Bayes theorem. Annual meeting of the population association of America, Denver, CO, April 1980

Petersen J, Longley P, Gibin M, Mateos P, Atkinson P (2011) Names-based classification of accident and emergency department users. Health place 17(5):1162–1169. doi:10.1016/j.healthplace.2010.09.010

Polish Ministry of Interior and Administration (2009) Statistics of the 20 most popular surnames in Poland (translated from Polish). Available at http://www.mswia.gov.pl/download.php?s=1&id=7972. Accessed 3 Aug 2009

Pons P, Latapy M (2006) Computing communities in large networks using random walks. J Graph Algorithm Appl 10(2):191–218

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

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E Stat Nonlin Soft Matter Phys 76:036106

Reaney PH (1958) A dictionary of British surnames. Routledge and Kegan Paul, London

Ruhlen M (1987) A guide to the world's languages. Stanford University Press, Standford, CA

Ruhlen M (1994) On the origin of languages. Studies in linguistic taxonomy. Stanford University Press, Standford, CA

Statbel (2006) Most frequent family names – Belgium and regions (translated from French). Statistics Belgium, Brussels. Available at http://statbel.fgov.be/figures/d21a_fr.asp

Statistical Office of the Republic of Slovenia (2009) Most frequent family names. Available at http://www.stat.si/eng/imena_top_priimki.asp?r=True. Accessed 3 Aug 2009

Statistics Norway (2008) Last names used by 200 or more. Available at http://www.ssb.no/english/subjects/00/navn_en/. Accessed 21 July 2009

Tu SP, Yasui Y, Kuniyuki A, Schwartz SM, Jackson JC et al (2002) Breast cancer screening: stages of adoption among Cambodian American women. Cancer Detect Prev 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: Hanks P (ed) Dictionary of American family names. Oxford University Press, New York, pp xxiii–xxvii

Tucker DK (2005) The cultural-ethnic-language group technique as used in the Dictionary of American Family Names (DAFN). Onomastica Canadiana 87(2):71–84

Tucker DK (2006) A comparison of Irish surnames in the United States with those of Eire. Names 54(1):55–75

Tucker DK (2007) Surname distribution prints from the UK 1998 Electoral Roll compared with those from other distributions. Nomina 30:5–22

Williams AR (2011) What's in a surname? *National Geographic*, 22–23 Feb

Wood J, Badawood D, Dykes J, Slingsby A (2011) BallotMaps: detecting name bias in alphabetically ordered ballot papers. IEEE Trans Vis Comput Graph 17(12):2384–2391. doi:10.1109/TVCG.2011.174

Word DL, 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 May 2005

Worldnames (2009) Worldnames database. University College London. Available at http://worldnames.publicprofiler.org/. Accessed 25 July 2009

Yavari P, Hislop TG, Abanto Z (2005) Methodology to identify Iranian immigrants for epidemiological studies. Asian Pac J Cancer Prev 6(4):455–457

Part III
Applications: Mapping Names and Ethnicity

Chapter 8
The Geography and Ethnicity of People's Names

Abstract Research on the spatial mobility of the population in Europe has demonstrated that most people don't move too far away from where they were born, and tend to marry with people born in the same area. This has the obvious effect of maintaining family names "in-situ". This chapter uses the current geographical distribution of people's names in Europe and America to uncover historic and contemporary migration flows as well as regions of cultural interaction. The validity of different methods of spatial analysis using geocoded surname frequencies as raw material is justified. The chapter reviews a gallery of examples gathered from different countries and a range of scales, from the continent to a city's neighbourhood. These maps show that once settled at their destinations, cliques of forenames and surnames have continued to operate following the same socio-cultural patterns described in the previous chapters. As a result, analysing a "destination" country's contemporary name register over space we can identify not only the settlement pattern of current migrants, but also of historic migrant populations even several generations after they died.

Previous chapters in this book have amply demonstrated the value of exploiting various properties of names to decipher population communal origins. Chapter 7 in particular introduced a methodology to build our own name classification by cultural, ethnic and linguistic groups; termed Onomap. In this chapter, a review of the relevant literature on the spatial analysis of people's names is presented, with an emphasis on ethnicity and on applications that have used the Onomap classification or that have been developed by the author's research team at University College London.

Specifically, this chapter reviews research on the geographical analysis of people's names to study underlying socio-spatial processes, primarily migration, ethnicity and population mixing over space and time. It does so through three well identified sections, and an abundant selection of maps and illustrative material that bring innovative visualisations to the socio-historical processes described in previous chapters. Section 8.1 reviews a range of methodological approaches to the

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_8, © Springer-Verlag Berlin Heidelberg 2014

spatial analysis of population structure, migration and ethnicity in people's names. Such review is internally structured in four core methods that have all attempted to contribute to the question of how best to systematically unveil the ethno-cultural patterns in naming practices over space. The ultimate aim is to detect and delineate diverse strata in population diversity over space. Section 8.2 reviews examples of investigations that have used simpler geographical techniques to reveal historical migration patterns, within or over national state borders. Section 8.3 in turn reviews research into the mapping of contemporary migrations, closely linked to the aforementioned research team at University College London, particular an Atlas of ethnicity through names in London, and a World Atlas in 26 countries termed WorldNames. Finally, the chapter closes with some conclusions and recommendations for future research in developing applications of name analysis to understand spatial distributions of populations.

8.1 Approaches to the Spatial Analysis of Ethnicity and Migration in Names

Most of the published studies on surname distribution across space and time primarily seek a single purpose: finding surprising population geographic distributions and movement patterns with which to delineate cultural regions and/or migration episodes. They do so by focusing on the analysis of regions within a single country or language community, as opposed to attempting international comparisons, except for some border regions spanning neighbouring countries. Temporally these studies tend to focus on a time period encompassing the last 200 years, for which reliable population registration records and surname frequencies at small area level exist. The number of studies of this kind is not very large, but they represent a wide range of research fields, primarily; history, linguistics, demography, geography, and population genetics. The great majority of these spatial studies concentrate a single type of name; surnames and not forenames or very rarely the two of them combined (an exception being Zelinsky 1970, who focuses just on forenames). A few examples of this type of spatial analysis of surnames will be mentioned here, taken from France, Belgium, Great Britain, Spain, Italy and Canada. This review will be structured according to four different methodological approaches to the analysis of the geography of surnames for the purpose of disentangling ancient population identity and diversity over space. The four methods are individually described and discussed in the next four subsections.

8.1.1 Method 1: Surnames' Geography and the Search for Cultural Regions

A wide range of studies on the geographical patterns of names focus on the study of spatial distributions of surname frequencies. At the base of the great majority of these studies lies a single type of data source: a contemporary or historical register of surname frequencies broken down over space. These frequencies are then compared between spatial units using different measures of surname homogeneity between neighbouring populations. One of them is isonymy, the indicator fully described in Chap. 4. Such measures are used to determine the degree of "isolation by distance", or the existence of barriers to population exchange, between neighbouring areas (Manni et al. 2004). In the great majority of cases stark barriers do not exist within modern countries or contemporary languages, since hereditary surnames in Europe are 200–800 years old. Instead, the literature focuses on identifying zones, or "clines", of "transition" or "gradual change", between areas featuring distinct dynamics of surname reproduction. The reverse pattern of such barriers, borders, or transition zones to surname transmission is hypothesised to represent some sort of "cultural regions".

This type of studies are hence concerned with identifying somewhat homogenous surname areas in what could be conceived as the classic problem of regionalisation in Geography. Regions have been defined as "spatial compartments" of formal, functional, or perceptual significance (Murphy 1991), or just as a distinct area on the Earth's surface (Massey 1984). Empirically, regions have been typically classified as areas that have more interaction within each other than with other, outside areas (Brown and Holmes 1971). Therefore a good proxy candidate for such cultural interaction is the surname exchange between areas, through migration flows and social and biological reproduction.

An early study in the use of surname frequencies for regionalisation purposes is Guppy's (1890) book "*Homes of family names in Great Britain*". He computed the frequencies of approximately 8,000 surnames from land-owning farmers sourced from *Kelly's Post Office Directory*, sampling their locations and proposing a division of the country into seven surname regions (see the map in Fig. 8.1). He was concerned with the extent to which politically delineated regions, such as the Wales-England border or parliamentary constituencies, actually resembled meaningful and homogenous historical communities. Guppy's study opened up a stream of research primarily driven by historians, geneaologists and linguists concerned with the identification of geographical patterns in surname distributions that could be used as proxies for underlying socio-cultural processes over space. However, most of the large scale (national level) studies on surnames have actually come from population geneticists and geographers. Various examples of this type of regionalisation studies were reviewed in Chap. 4 when discussing applications of using isonymy as a methodology in population genetics (see Sect. 4.4) and others will be commented here.

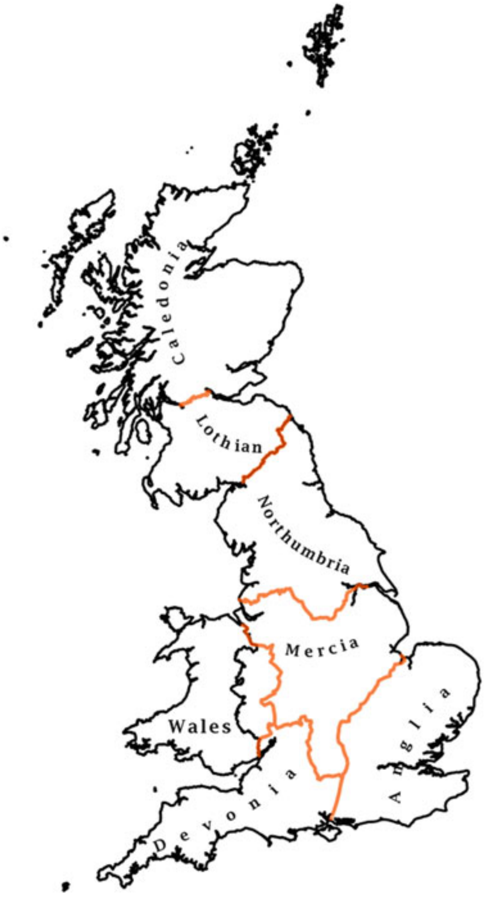


Fig. 8.1 Guppy's map of surname regions in Great Britain (1890). *Source*: Prepared by James Cheshire, UCL, reproduced with permission

The indisputable leader in this area has been a group of scholars here refered to as the "Ferrara school" of population genetics in Italy, led by Barrai, Dipierri, Scapoli, Rodriguez-Larralde and others. This group has undertaken a multitude of studies of European and some American countries spanning over the last three decades (citations to this school appear in Sect. 4.4). Except for Scapoli et al. (2007), which study the combined space of eight European countries, all of these studies are concerned with identifying surname regions within individual countries. Examples from other authors leveraging on the study of isonymy or similar surname distance measures are cited in Chap. 4 (Sect. 4.4) and will not be repeated here. For a full review of studies in this area the following three publications should be consulted: (Lasker 1985; Colantonio et al. 2002; Darlu et al. 2012).

Another distinct thread of studies have used various spatial analysis and clustering techniques to empirically identify such surnames regions. They all use a matrix of isonymy-derived distance measures between pairs of areal units as their primary input data. One prolific author in this area is Franz Manni, at the France's National Museum of Natural History (Musée de l'Homme) who has used Self-Organising-Maps (SOMs) (Manni et al. 2004) and Monmonier's barrier algorithm (Manni and Gue 2004) to identify surname regions in France and The Netherlands. He and his co-authors have compared the resulting regionalisations to those derived from alternative cultural indicators, such as the distribution of languages and dialects, demonstrating a close correlation save for some well documented exceptions. Furthermore, our team at University College London (UCL) Department of Geography led by Paul Longley, James Cheshire and myself, have proposed various measures of population similarity over space for the purposes of creating cultural regionalisations through surnames. We have used various classification techniques, namely: hierarchical clustering, multidimensional scaling (Cheshire et al. 2009a, b), k-means clustering (Longley et al. 2011), and network analysis (Mateos et al. 2011). Our work at UCL Geography has primarily focussed on applications in the U.K. (Longley et al. 2011; Winney et al. 2012; Cheshire and Longley 2012), Czech Republic (Novotný and Cheshire 2012), Spain (Mateos and Tucker 2008; Mateos 2006), the European Union (Cheshire et al. 2011), Mexico (Mateos 2010), Japan (Cheshire et al. 2013) and New Zealand (Mateos et al. 2011).

An example of a map with a regionalisation of surnames in Great Britain using multidimensional scaling (MDS) clustering is shown in Fig. 8.2. The map shows local authority districts classified with a continuous colour code according to the degree of similarity of their surname distributions (isonymy) based on the 2001 Electoral Register. A multidimensional scaling (MDS) clustering along three axes is applied (3-D MDS) and each of these dimensions is assigned a value from 0 to 255 corresponding to one of Red, Green or Blue value in the RGB colour scheme (Cheshire et al. 2009a, b). The resulting map clearly shows five distinct surname regions; Northern Ireland, Scotland, Wales, Northern England and Southern England. Each of these regions appears dominated by a primary colour while the areas in between them show a mix in the colour palette depicting a transition between these regional origins. Furthermore, using a surname dataset containing 16 European countries derived from the Worldnames project (see Sect. 8.3.2), MDS, Ward's hierarchical clustering, and k-means were applied to the task of delineating Europe's surname regions (Cheshire et al. 2011). One of these analyses is presented in Fig. 8.3 showing the spatial distribution of each of the three multidimensional scaling (MDS) dimensions across Europe. Each dimension is displayed separately in the black and white maps and all combined in the colour one. The three MDS dimensions tend to show an axis more concentrated in the British Isles, a second one in Germany and the Nordic countries and a third one in Spain and Southern Europe. These examples prove the value of using surnames to uncover cultural regions whose boundaries can be delineated using different criteria and are best represented as a continuous transition between different surname regimes, such as those shown in the MDS maps (Figs. 8.2 and 8.3).

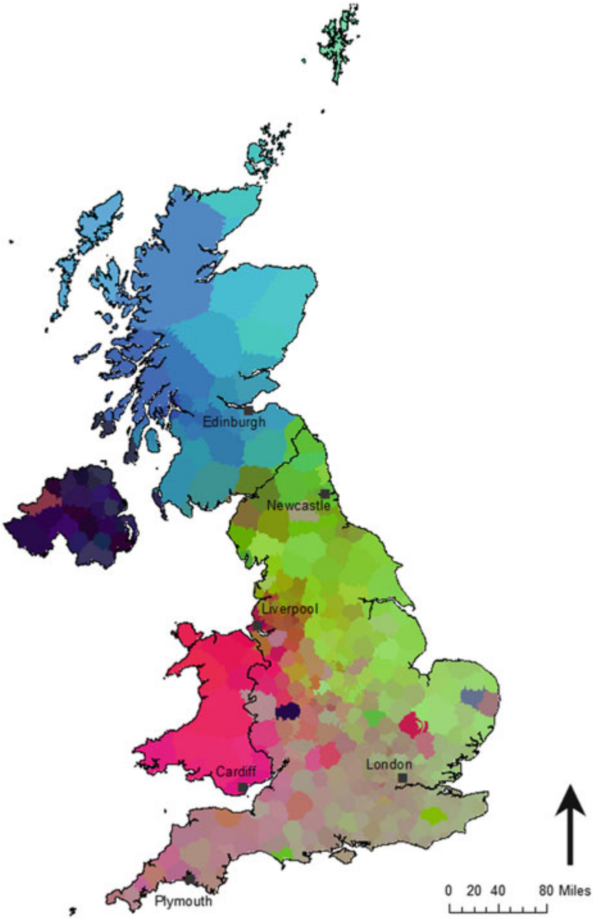


Fig. 8.2 Map of the UK's surname regions. The maps shows local authority districts classified with *colours* according to the degree of similarity of their surname distributions (isonymy) based on the 2001 Electoral Register. The clustering method used is multidimensional scaling (MDS) along three axis depicted by a *Red-Green-Blue* colour scheme. *Source*: Cheshire et al. (2009a)

8.1.2 Method 2: The Geography of Surnames' Morphological Patterns

The great majority of the studies on cultural regionalisation and surnames cited so far, make no presumption on the cultural or geographical origin of a specific name. Thus, they use the spatial dimension of individual surname frequencies as the single raw material with which to attempt to establish historical episodes in populations such as migration, intermarriage, language replacement and so on. In some cases, such as Lasker (1985) and Rogers (1995) the analysis is reduced to displaying sequences of maps of individual names, perhaps grouped by those concentrated in particular areas of a country. That is, both in studies of isonymy and in these two publications, no aggregation of individual names is performed.

Why would we be interested in aggregating names? It seems logical that if we are after unveiling geographical and historical processes behind name distribution

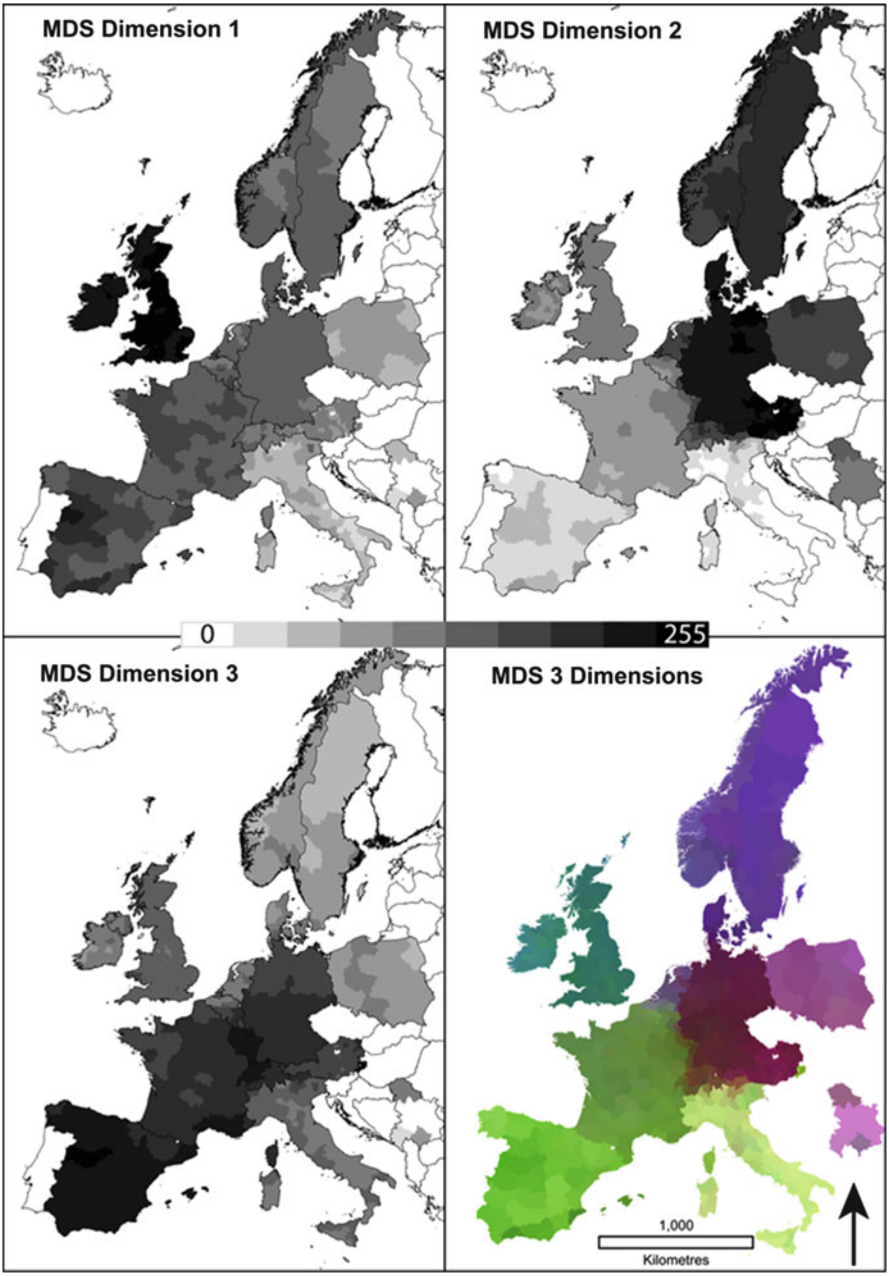


Fig. 8.3 Europe's surname regions. Maps showing the spatial distribution of each of the three multidimensional scaling (MDS) dimensions. Each dimension has been rescaled to a value of between 0 and 255 to facilitate the creation of RGB *colours*, and displayed separately in the *black* and *white maps* (*top maps* and *bottom left*) and all combined in the *colour* one (*bottom right*). *Source*: Cheshire et al. (2011)

patterns, having some prior knowledge on the probable linguistic, regional, or religious origin of a name, and perhaps the type and meaning of names would be a tremendous help in accomplishing such regionalisation exercise.

A first attempt into such aggregations of names by origin, when no other knowledge is available, is to group together names with particular morphological patterns, such as endings, prefixes, common morphemes or phonetic similarities that can be identified in their spelling. Some of these patterns derive from ancient languages or dialects, regional accents, cultural preferences for certain spellings or sounds, and historical episodes such as migration and conquest deriving in language or religious replacement. Such groupings of names sharing a common pattern can then be mapped together as a single cultural feature over space.

In Britain Kevin Schürer (2004) has been interested in unveiling geo-historical patterns in surname morphologies. He has developed searches for particular name morphologies to account for receding languages (Cornish, Scots, Welsh), dialectal variations in spelling, or certain historic episodes such as various "invasions" (Roman, Viking, Norman, etc), all which have left a clear trail in the contemporary geographical distribution of surnames in the British Isles. Malcom Smith (2002) has also studied certain surname patterns, and even meanings and occupations of the bearers, grouping certain occupations to search for socio-spatial structure in historical records.

Furthermore, many of the types of surnames described in Chap. 3 (Sect. 3.2), such as patronyms, locatives, metronyms, or diminutive names also present common morphological patterns that have a regional explanation. An example of one of such regional patterns in surnames' morphology is shown in Fig. 3.1. It represents a map of Great Britain with the 1998 frequency distribution of occupational surnames ending in "-man", most of them present in the southeast of the country. Most of the work in this area has dealt with patronyms, surnames derived from the father's forename, such as *Johnson*. For example, prefixes such as "Mc-"/"Mac-" in Scotland and Ireland, or endings, such as "-son" in English or "-ich" "-ova" in Russian can be easily searched for in a surname database and mapped. This type of approach, when applied to the search for the geographic origin names, and hence of migration, has been termed "the patronymic method" (Darlu and Degioanni 2007: 259).

Poulain and Darlu (Poulain et al. 2000) used this approach in a historic demography study that proposed the use of patronyms to measure historical migration flows of Flemish people to Wallonia (Belgium) and Northern France in the early twentieth century. This method worked well as a proxy to map this migration episode since contemporary frequencies of Flemish patronyms in Belgium and France, such as those starting with "*Van*" or "*Ver*", will clearly indicate the major destination areas of Flemish migrants. Similar work by these authors have revealed Italian migration flows to France between 1891 and 1940 (Degioanni et al. 1996).

These examples demonstrate the value of analysing the spatial distribution of "classified names" into groups of common origin, as opposed to individual frequencies or isonymic distances. Furthermore, this geographical exercise can be of great assistance to determine the cultural ethnic and linguistic origins of a name, the theme of this book. The morphological and patronymic methods described in this section could be used to identify the geographical origin of a surname within a country or neighbouring countries through a heuristic or "trial and error" approach. This would involve two steps; first looking for probable morphological patterns in naming, and then testing for uneven spatial distributions pointing to certain regions as the probable geographical origin of a name (Mateos et al. 2007). However, this method is rather unproductive, requires prior knowledge of meaningful patterns, some names originated at various places simultaneously (polyphyletic) and the geographical origin of many names, that lack salient morphological patterns, would simply remain undetected through morphological naming patterns alone.

8.1.3 Method 3: Comparing Geographical Distributions over Time

A more feasible alternative to establish the probable geographic origin of a name is to compare surname geographic distributions over two time periods, especially if a historical population register is available, in order to remove the migratory effects of mass migration and urbanisation over the last century. Such spatio-temporal comparisons can be useful to account for recent migration flows, especially to cities, as well as to provide two time snapshots with which to corroborate or discard the probable geographic origin of surnames or at least indicate plausible lines of direction in episodes of name expansion or retreat.

Amongst such approaches, for example Degioanni and Darlu (2001) propose the application of a Bayesian method to estimate the probable geographical origin of migrants using surname frequencies measured over two time periods. By using birth registers in France in 1891–1915 and 1916–1940, a period of mass urbanisation, they demonstrate the validity of this method to estimate the probable region of origin of migrants between these two time periods when no other information on surname origin is available (i.e. beyond the "patronymic" or morphological method). Other similar studies using surname frequencies from historical records to reconstruct past population structure have typically been limited to specific areas or towns, because of restrictions in the availability of administrative or religious population registers. Worth mentioning is the work of French historians and population geneticists Pierre Darlu, Pascal Chareille, and Guy Brunet, who, have produced interesting insights to historic distributions of populations all the way back to the Middle Ages in France (including the Savoy and neighbouring countries), Italy, and Belgium (beyond the work already cited; Darlu et al. 2011, 2012; Chareille 2002; Brunet and Bideau 2000; Brunet et al. 2001).

In Britain, the 1881 Census of Great Britain was fully digitized (including approximately 30 million people) by the Church of Jesus Christ of Latter-day Saints (better known as the Mormon Church). This database was geocoded at the level of the Parish and made available to researchers in digital form (Schürer 2002, 2004). Such unprecedented resource led to a series of studies that exploited spatio-temporal comparisons between historic and contemporary registers. These studies unveiled cultural regions, domestic and international migration, emigration, urbanisation, and even genetic patterns (Schürer 2004; Longley et al. 2007, 2011; Webber 2004; Mateos et al. 2007; Winney et al. 2012).

For example, Cheshire et al. 2009a compared the 1881 Census with the 2006 electoral register, and used hierarchical clustering to classify Great Britain into surname regions. The city of Corby in the English Midlands was classified as part of the Scottish cluster in 2006 but not in 1881. Looking at the surnames frequencies in Corby they found them very similar in composition to Scottish areas. The socio-historical literature revealed that in 1932 there was an episode of mass migration from Scotland, after the establishment of a new iron and steel works company in Corby that recruited workers in Scotland for several decades (Grieco 1985). In the 1970s, 57 % of Corby's population reported Scottish origin and a range of Scottish customs and festivities are still very popular today (Grieco 1985). Therefore, Scottish surnames do still resemble Scottish origins in this area of the English Midlands, hence pointing to the usefulness of name analysis to the study of identity over space, the central theme of this book.

Using this 1881 Census database the UCL Geography team has proposed alternative methods to determine the "geographical core" area of origin of a name (Winney et al. 2012; Longley et al. 2011; Cheshire and Longley 2012) implementing a spatial analysis method known as kernel density estimation (KDE) (De Smith et al. 2007). This "point density" method allows the delineation of an adaptative contour on a map, enclosing the areas with the highest relative concentrations of a surname. In this case, this method produced a map of the areas with the highest population density for each surname in the 1881 Census, independent of any political boundaries. As such, this analysis brings us much closer to the original area where the surname was probably originated in the Middle Ages, since we "remove" over 130 years of population mobility, or between five and six generations of population mixing over space. Examples of these KDE maps for a few surnames in Great Britain are showed in Fig. 8.4. These individual surname "core area" maps can be overlapped to start delineating common borders that acted as barriers to the transmission of surnames over space. A derivation of this work has been used to determine if each surname from DNA donors in a population genetics study in Britain (http://www.peopleofthebritishisles.org/) was probably originated in the same area where the four grandparents of the donor where born (Winney et al. 2012). This fine-grained analysis was based on four geocoded grandparental birthplaces (three generations ago) and the surname geographical concentration in the late nineteenth century (five generations ago). Such rich dataset has allowed population geneticists to apply different distance thresholds to their donor data to determine the specificity of a DNA sample to a geographic area, in ways that were unthinkable just a few years ago. As Tyler-Smith and Xue put it, this "microcosmic survey of genetic variation in a set of small islands off the western coast of the Eurasian continent is revealing the level of differentiation that builds up over millennia via events well documented by archaeology and history, so these alternative data sets can be compared to address questions about the initial peopling of

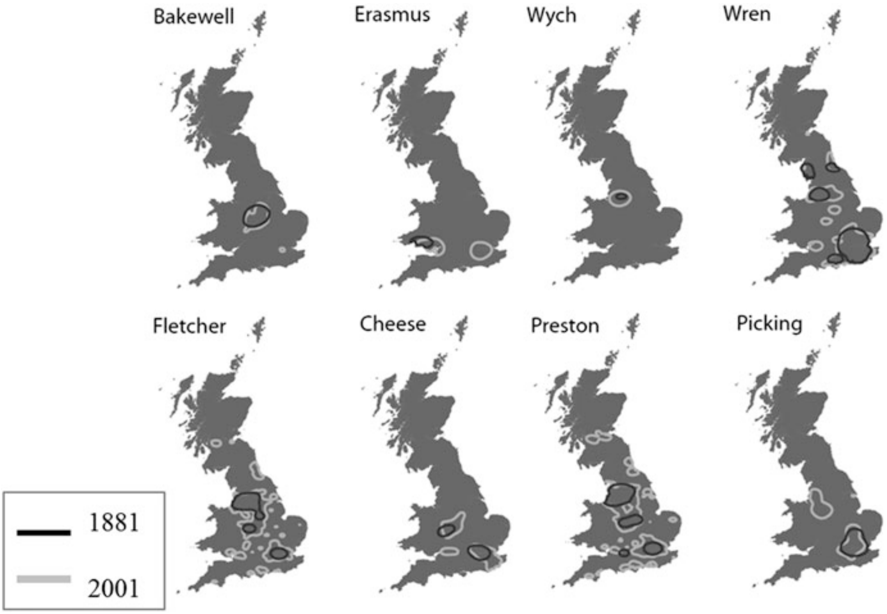


Fig. 8.4 Maps of some surname "core" areas in Great Britain in 1881 and 2001. Each map correspond to an individual surname. The areas delineated correspond to the "core" area within the Kernel Density Estimation (KDE) surface defined as those areas that contain at least 50 % of bearers of that surname. Two datasets are mapped, 1881 Census and 2001 Electoral Register, respectively depicted by *black* and *grey* contours on the map. *Source*: James Cheshire, UCL, reproduced with permission of the author

the area, and its subsequent reshaping by internal and external forces" (Tyler-Smith and Xue 2012: 130). In other words, the transdisciplinary work described here has been put to use to delineate migration episodes that takes us back to the origin of surnames in the Middle Ages and even to prior centuries and millennia through DNA analysis. However, the method described in this section requires two temporally distant population registers in order to ascribe names to a particular region of origin.

8.1.4 Method 4: Mapping Pre-classified Names by Origins

Despite the usefulness of the three methods of spatial analysis of names origins described so far, the clear challenge is to map names pre-classified into groups of common origin. This challenge becomes a unique opportunity in this book, since its underlying theme precisely focuses on proposing such classification of names into a taxonomy of probable cultural ethnic and linguistic origins. We will now draw upon the developments presented in Chaps. 6 and 7, to review examples of this type of geographical analysis.

The core interest of this method four is to develop a productive method to study the geographical origin and current distribution of names. This involves the pre-classification of names according to external knowledge (i.e. not spatio-temporal) about their cultural, ethnic, linguistic or religious origin. Such knowledge, as discussed in Chap. 6, must be procured from one of two types of sources: a) name dictionaries or list of name origins provided by reputable sources, or b) international name frequencies geographically disaggregated by residence, place of birth, or nationality.

The first type of external data source, dictionaries and lists of names' origins, has been traditionally produced by linguists, historians, and genealogists (see for example (Hanks et al. 2006; Hanks 2003). However, as we have seen in Chaps. 6 and 7, most of these dictionaries refer to forenames, while the coverage of surnames is patchy and sometimes they present multiple origins. Therefore, a host of researchers from other disciplines have contributed to this task over the last decades, producing lists of names by origin derived from local knowledge and administrative records, focusing on specific collectives; East Asian, South Asian, Hispanic, Arab, and specific languages and countries, as reviewed in Chap. 6 (Lauderdale and Kestenbaum 2000; Razum et al. 2001; Cummins et al. 1999; Nanchahal et al. 2001; Word and Perkins 1996).

The latter examples point to the direction of the second type of external data source, the use of international name frequencies as a primary material to provide the most probable area or country of origin. These are derived from the places where a name presents a higher concentration. This method works by aggregating the number of people with the same name (surname or forename) in a large administrative database where geographical information on either current residence, place of birth, parents' place of birth, or nationality/ies is available (Mateos et al. 2007). Name frequencies by area are then computed and probabilities of a name being representative of a certain area are then computed. Names that exceed a set probability threshold are added to a name origin list. This second method is obviously less reliable than the first one, in the sense that name to origin attribution has not been validated by experts or other authors, but could work well for less frequent names for which no existing knowledge is available. Furthermore, many national or regional statistical agencies already provide such name frequencies by country of birth and hence can be used for this purpose. An example of some of these agencies and sources is provided in Table 7.2, that lists 25 different sources of countries or languages used to build the so called "diagnostic surnames list" for the purposes of network clustering.

Using either method to build a reference list (Mateos 2007), a population register with names and locations would then be classified by their cultural, ethnic, linguistic or religious origin (CEL) according to a specific categorization of 'identity' devised for a particular purpose (see Chap. 7). After this, population frequencies by these CEL categories and location would be calculated. It is important to notice that this method involves the use of aggregated population counts by each single CEL, and hence differs substantially from the use of individual names (this chapter, Sect. 8.1.1 or 8.1.3) or common naming patterns (Sect. 8.1.2). Hence the interest here is to understand a whole "population" (from the title of this book). Even when a proportion of name's bearers might not ascribe themselves to the particular CEL category with which it has been classified, the aggregated results are very likely to represent the majority of ancestral origins at population level, thus revealing interesting population diversity patterns over space and time. In the next sections, some examples from the literature that use this method four will be reviewed before reporting our own research in this area.

8.2 Mapping Historic Migrations

Based on method number four described in the previous section, the use of a pre-classified list of names by origins, a few researchers have revealed interesting historical migration patterns within specific regions. Three representative studies from Great Britain, Spain and Canada will be summarised here for illustrative purposes.

From a geodemographic perspective, Longley et al. (2007) focus on Cornish surnames in Britain, as an alternative method to study historic migrations and social mobility at very fine geographical scales. They did so through the classification of the 1998 UK electoral register at postcode level, using a list of Cornish names previously derived from the spatial concentrations of surnames in Cornwall in the 1881 Census. They mapped this data and found a clear "distance-decay" function away from Cornwall. Figure 8.5 shows a scatterplot with the spatial concentration of Cornish surnames by postal area in 1998 against distance from Cornwall, depicting such distance-decay pattern with a power function curve. The dot at the top left hand corner of the scatter plot corresponds to Truro postal area, in the south-western tip of the country, with the highest concentration of Cornish surnames, while descending through the trend line we find neighbouring postal areas Plymouth first, and then Torquay and Exeter, followed by the rest of the country with much lower levels of Cornish surnames concentrations. It was through the analysis of this plot that Longley et al. (2007) found an outlier in Middlesbrough in Northern England, which presents a disproportionally high concentration of Cornish surnames in 1998 given its distance to Cornwall (the dots in the scatterplot around coordinates $x = 23.5$, $y = 8.5$, corresponding to Cleveland and Harrogate postcode areas). Such outliers were explained by historical migration evidence from the literature. A mass migration episode from Cornwall took place in the 1850s at a time of a decline in Cornish tin and copper mines while miners' skilled labour was demanded for iron ore in the Middlesbrough area. The poor economic conditions in this destination area since the late nineteenth century to the present, have meant that descendants of Cornish migrants have remained in place, sometimes in very specific types of neighbourhoods, and thus preserving the concentration of Cornish surnames (Longley et al. 2007). This analysis exemplifies the need to conduct

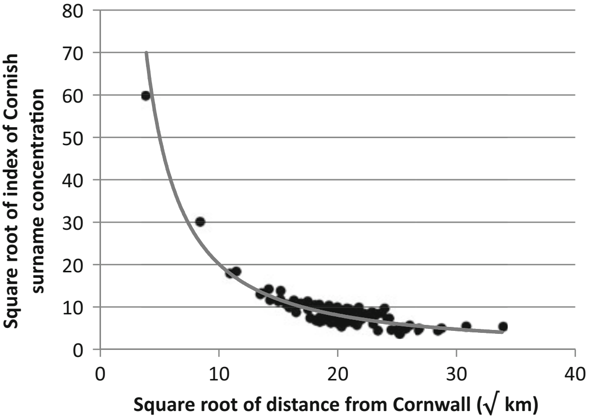


Fig. 8.5 Distance decay of Cornish surnames in Great Britain with distance from Cornwall (1998). Each *dot* in the scatter plot represents a postal area of Great Britain. The *vertical axis* shows the square root of a location quotient depicting the concentration of Cornish surnames compared to the national average (=100). The *horizontal axis* shows the square root of the distance from Cornwall, calculated between the geographic centroid of a given postal area and that of *Truro* postal area in South-western Cornwall. The *line* represents a power trend-line over the *dot* distribution. *Source*: Produced by the author based on data supplied by Richard Webber

spatial and statistical analysis techniques on the names data rather than just look for visual hints on the map.

In Spain, Aranda Aznar (1998) studied Basque surnames using an official dictionary to classify the Spanish population register. He detected migration patterns to and from the Basque Country and the rest of Spain, analysing the degree of intermarriage between Basque and non-Basque surnamed populations in the 1990s. He used very detailed population registers taking the advantage of his post as the director of the Spanish National Statistics Institute at the time, what reveals the critical need of securing access to very detailed digitized and georeferenced population registers in order to conduct this type of studies.

My own work on Spanish surnames has demonstrated how the spatial analysis of names can reveal historical as well as contemporary migration episodes of population flow and settlement. For example the Christian "re-conquest" of the Iberian Peninsula from the Arabs in the eleventh to fifteenth century is revealed in its current surname distribution over space, after classifying the 2004 telephone directory into Basque, Catalan, Portuguese-Galician and Castilian names using clustering techniques (Mateos 2006; Mateos and Tucker 2008). Such historic population settlement process is clearly discernible in the maps shown in Fig. 8.6.

Furthermore, this work has revealed the uneven expansion of Spanish surnames in Latin America through the analysis of surname frequencies in Mexico, Argentina, Venezuela, and Hispanic surnames in the U.S. (Mateos 2010; Mateos et al. 2006). Finally, the migration story comes back full circle with mass migration to Spain over the last 15 years. This is revealed by the rapid expansion of certain

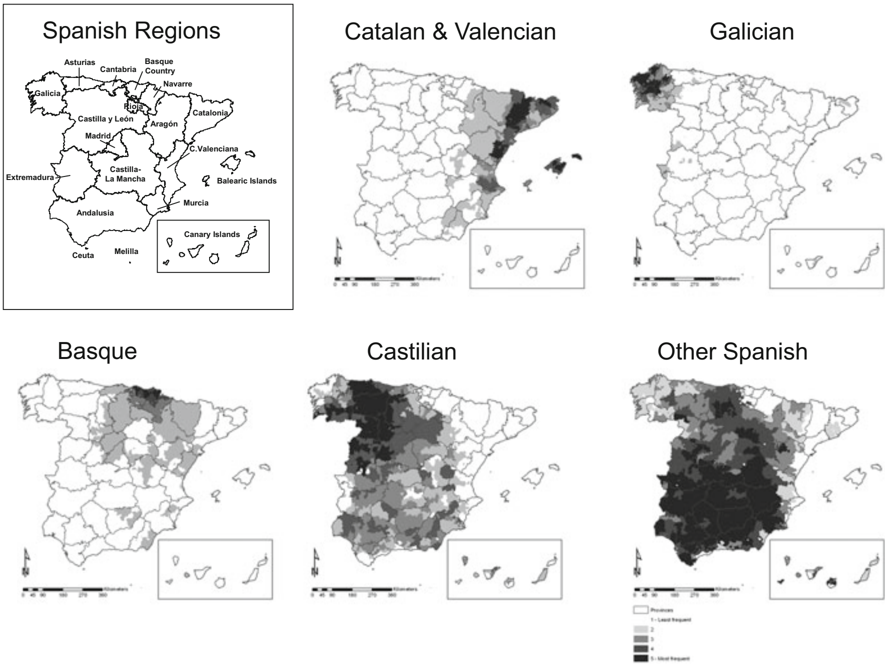


Fig. 8.6 The geography of Spanish surnames by linguistic or regional origin. The distribution of Basque, Catalan and Valencian, Galician, Castilian, and "other Spanish" surnames in Spain by postal area, according to the 2004 telephone directory. *Source*: Mateos and Tucker (2008: 180)

rare surnames many of which had disappeared altogether in Spain while being preserved in Latin America, or had been coined there after Spanish colonization (Mateos 2006).

As shown in the examples from Belgium, Great Britain and Spain cited so far, there seems to be a special interest on the spatial analysis of ethnicity in names in multilingual countries. Another example is Canada, which has been studied in historical demography to analyse the spatial patterns of French families in Quebec (Desjardins et al. 2000), and more recent migration episodes, such as Chiarelli (1992) who used surnames to identify the regional origins of Italians who emigrated to Toronto over the last two centuries.

These three studies (using method 4 described Sect. 8.1.4) differ from the rest of studies previously mentioned in this chapter (methods 1–3) in that rather than studying the geographical distribution of individual surnames or some types (e.g. patronyms, other morphologies, or historical migrants), all surnames present in a population are grouped and mapped together according to their language or regional culture of origin (e.g. Flemish, Cornish, Scottish, Basque, Quebecois, Italian, etc.). This is an important step for reasons that will become clear in the rest of this book, in which the use of name classifications into groups according to origin will be further developed to analyse geographical patterns of contemporary migration.

8.3 Mapping Contemporary Migrations

Most of the examples cited so far in this chapter, have used surname spatial analysis to study domestic migrations and historical processes of population settlement within one country (Scottish, Cornish, Catalan, Galician, Basque, Quebecois, French regions), neighbouring border regions (Flemish in Wallonia, Savoy & Italian-France borders), or homogeneous language areas (Latin America, France-Belgium). However, despite the breadth of the literature available to study contemporary international migration through name analysis, as reviewed in the previous chapters (Chaps. 2–7), there seems to be a substantial void in the spatial analysis of such processes. In other words, researchers interested in the spatial analysis of name distributions have only focused their attention in local populations (up to the Nation State) and on historical peopling episodes up to the early twentieth century. Meanwhile, scholars using names to identify ethnicity in studies of population diversity in contemporary cities and regions (in public health, demography, sociology, or economics, as reviewed in Chap. 6) do not seem to have been interested in the spatial analysis of such populations. Amongst the key possible reasons for this void in the literature are; a lack of accessibility to geographically disaggregated and geo-referenced datasets, issues of data privacy and risk of disclosure combined with a probable lack of spatial analysis skills.

This and the next chapter conform the application sections of this book, which are primarily devoted to make a contribution towards filling this research gap. Here the approach is to review representative examples from the literature and close collaborators, combined with new unpublished research from the author.

8.3.1 London's Diverse Population: "The World in One City"

It is often said that London is a small sample of the world's population (Benedictus 2005). With over 322 languages spoken at London's schools and 40 % of school children not speaking English at home, London cannot be considered a standard European capital city (Von Ahn et al. 2010). London and New York are probably the most ethnically diverse cities in the world, in terms of the number and relative sizes of ethnic groups residing in its neighbourhoods, and the diversity in their geographical and cultural origins across the world.

London is in itself an "ethnic minority city". In 2011 55 % of London's population of 8.17 million was not "White British" (19.5 % in England and Wales) rising from 40.2 % in 2001 (Office for National Statistics 2012). This share of the population is comprised of many different ethnic groups, and for a full description of London's diverse population and its future prospects see Mateos (2013a). Although "Non-White" groups tend to get all the media and academic attention, actually 15 % of London's population belong to White groups different to

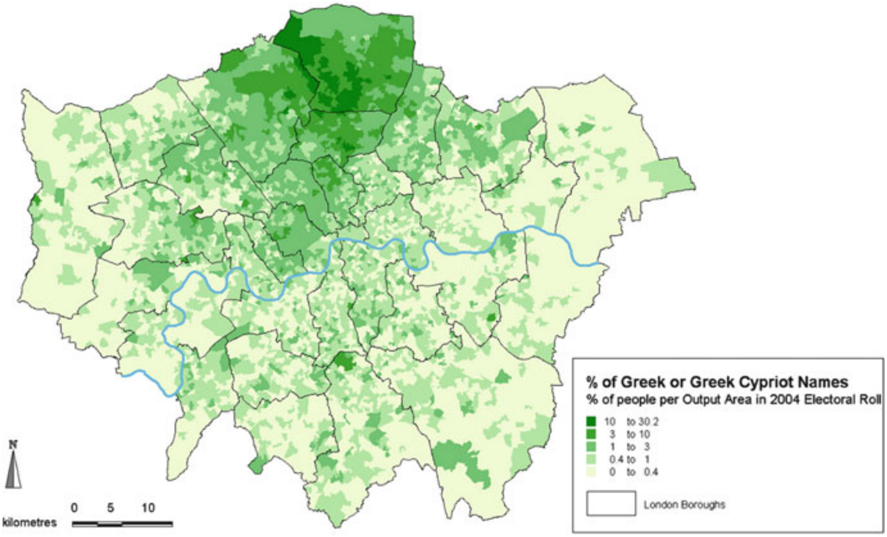

% of Greek or Greek Cypriot Names
% of people per Output Area in 2004 Electoral Roll
10 to 30.2
3 to 10
1 to 3
0.4 to 1
0 to 0.4
London Boroughs
N
0 5 10
kilometres

Fig. 8.7 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 five intervals, using the 2004 Electoral register. *Source*: Mateos et al. (2007)

White British, encompassing a variety of origins in Europe, North Africa, the Middle East, the whole American continent as well as Oceania. These, more subtle, aspects of ethnicity play testimony to the need of a multidimensional concept of ethnicity, as discussed in Chap. 2.

Despite this rich level of population diversity, the traditional statistical tools available to understand population composition have failed to capture the rapidly changing, small scale geographical processes that relate particular neighbourhoods with regions all around the world. As discussed in Chap. 2, in the UK, the Census of population ethnicity classification is too coarse and rigid to reflect the fine-grained reality of London's population changes over the last 20 years.

It is precisely to fill this gap that we embarked on the 9-year research project whose overall efforts are partly reflected in this book. Throughout this journey, there was a single city-laboratory in which ideas, data and maps were tested one and again against neighbourhoods we were familiar with: "London, the world in one city". We devote an introductory example to this city in this chapter but full details are available Chap. 9.

Figure 8.7 shows one of the early maps produced from an early version of the Onomap name classification in 2005. It shows the distribution of people in London with Greek or Greek Cypriot surnames using an Onomap-classified version of the UK 2004 electoral register. The spatial resolution of this map is the Census Output Area, a standard administrative unit at the neighbourhood level representing an average population of approximately 250 people. This map shows the distribution of the absolute number of people, that is a population headcount taken from the

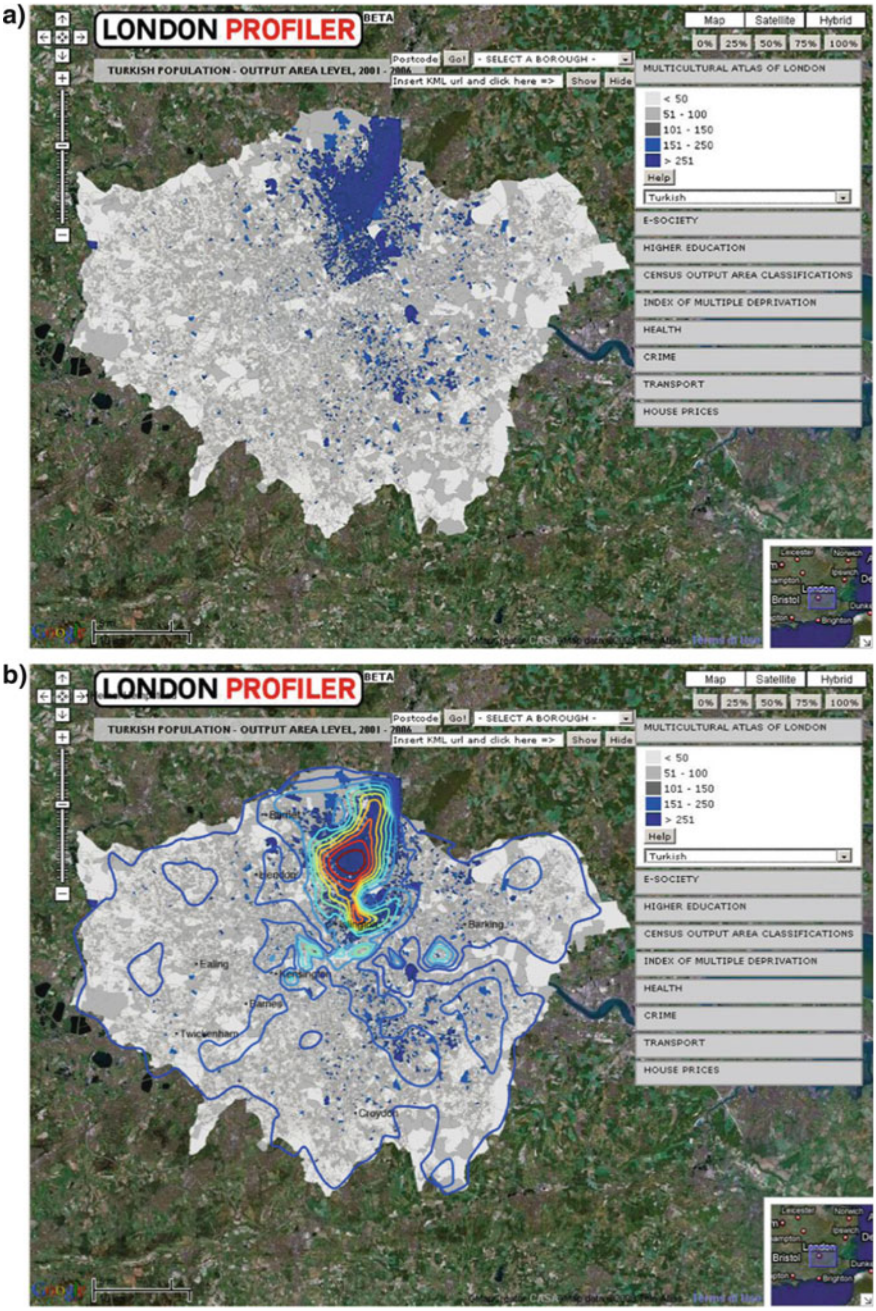


Fig. 8.8 (continued)

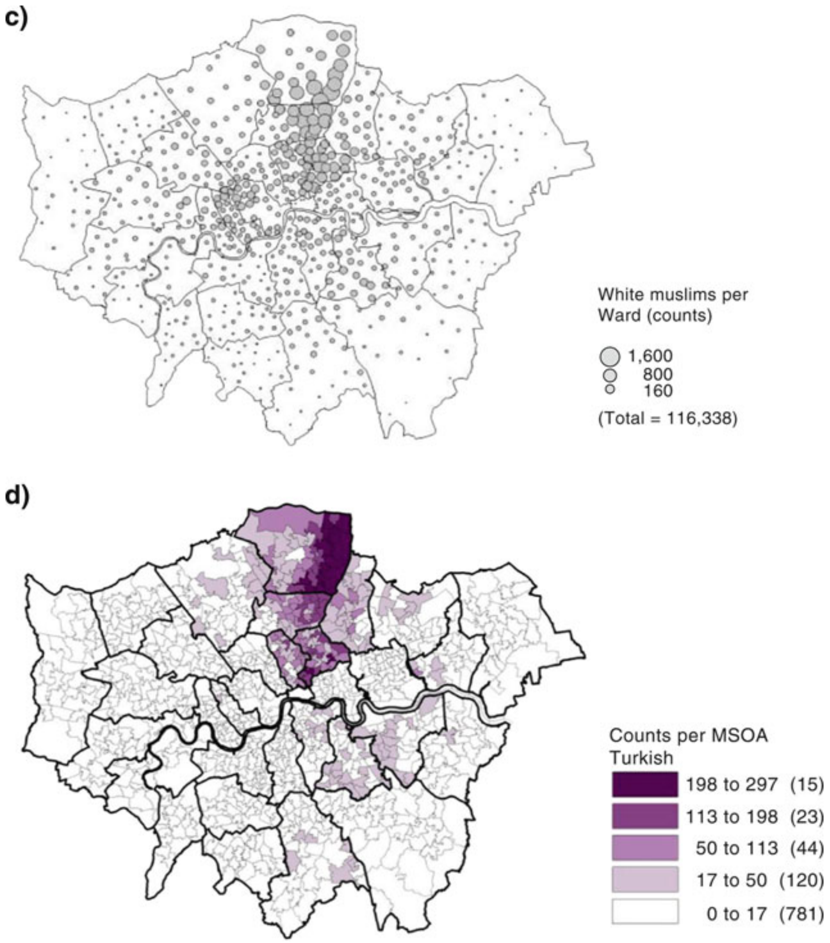


Fig. 8.8 Four maps of the Turkish population in London (2001–2008). Each of the four panels contain a self-standing map representing: (**a**) Distribution of Turkish surnames by Output Area in the 2002–2006 Electoral Register (*Source*: http://www.londonprofiler.org); (**b**) Countour map with the distribution of phone calls to Turkey in 2008, provided by a telecom operator (*Source*: Jon Reades, CASA, University College London); (**c**) Distribution of white Muslim population in the 2001 Census (*Source*: Peach 2006); (**d**) Distribution of schoolchildren who speak Turkish at home, extracted from the Pupil Level Annual School Census 2008 (*Source*: Von Ahn et al. 2010)

electoral register, which represents all adults entitled to vote at one of the various types of elections (including British nationals, EU nationals and most Commonwealth countries' nationals). However, even when the Census Output Areas are designed to contain a similar number of people in each of them, they do present stark variations in population sizes across London. Because of this problem, the map in Fig. 8.7 is not an ideal type of cartographic representation in the field of population geography, since population counts do not account for differences in area sizes, hence creating different population densities over space (Kraak and Ormeling 1996). Furthermore, maps of absolute counts for different ethnic groups cannot be compared on equal terms, since ethnic group size present stark variations across population groups.

Because of these limitations, a relative measure of population distribution was selected; a *location quotient* (LQ), a ratio of two rates. It compares the relative size of a population group in a geographical unit compared to the relative size of that group in the whole study area (Plane and Rogerson 1993). It is calculated as follows:

$$LQ\,x_i = \frac{{}^{x_i}/_{P_i}}{{}^{X^*}/_{P^*} \cdot 100} \tag{8.1}$$

The Location Quotient (LQ) of Group *X* in Area *i*, where: X_i = Population of Group X in Area *i*, P_i = Total Population of Area *i*, X* = Total Population of Group X in All Areas, P* = Total Population of All Areas (i.e. a City, Region, Country, etc).

LQ values are distributed around 100 (or 1 if the 100 factor is not applied) according to the concentration of the population of group X in an area relative to its average presence in the whole city/country:

>0 and <100; less concentration than average
=100; the same concentration than average
>100; more concentration than average

Using the Onomap name classification applied to the 2002–2006 UK Electoral Register, the geographical distribution of 18 of the most symbolic ethnic groups in London were mapped at Output Area level. These 18 ethnic groups were (in alphabetical order): Bangladeshi, Chinese, English, Greek, Indian, Irish, Italian, Jewish, Nigerian and Ghanaian, "Other Muslim", Pakistani, Polish, Portuguese, Russian, Sikh, Sri Lankan, Turkish, and Vietnamese. Most of these "ethnic" groups are not collected by the UK Census of population or by standard government surveys. These datasets represent an innovative approach to measuring cultural diversity in London. Moreover, the UCL Geography research team made them easily accessible to analysts and the general public through an innovative geo-visualisation technique using Google Maps, through a project termed "London Profiler" (http://www.londonprofiler.org). A full description of these visualisations is offered in three publications (Gibin et al. 2008, 2009; Mateos 2013b), and an example of one of these maps is included in Fig. 8.8a showing the concentration of Turkish names in London. Panel (a) shows the concentration of Turkish names in London shown against three other alternative data sources: the distribution of; landline phone calls to Turkey (panel "b"), the "White Muslim" ethno-religious group in the 2001 Census (panel "c"), and school children who speak Turkish at home (panel "d"). The fact that these four maps (names, phone calls, ethno-religion, and language) present similar patterns of geographical concentration in North London is striking, visually validating the name classification methodology proposed in this book. Furthermore, these maps helps us to present a very rich picture of the multifaceted sides of cultural diversity in a city like London, which cannot be reflected by a single ethnicity question from the Census.

8.3.2 WorldNames: A World Map of Names

Moving beyond the UK and even beyond Europe, at the UCL Geography team we developed a wider research project including detailed name data for 26 countries in four continents. This project is termed Worldnames (http://worldnames.publicprofiler.org/), and its basic characteristics were introduced in Chap. 7. The full list of countries, ordered by continent, is provided in Table 8.1. For each country the project's website contains surname (and sometimes forename) frequency data broken down by two levels of administrative geographical units. For ease of reference these two levels are termed "Regions" (higher level) and "Localities" (lower level). Table 8.1 provides the precise name for each type of area per country, in the local language and its English approximate translation. For example, *Worldnames*' two levels of geographical disaggregation in France correspond to Regions (*régions*) and Departments (*départements*), while in the U.S. to States and Counties. A similar two-level scale is offered for most of the rest of the countries, while for a few of them only one level of geographical disaggregation is available.

The source data for name frequencies consisted in publicly available telephone directories, except for the UK, Ireland, Argentina, Japan and New Zealand, where population registers such as the electoral registers were used. These administrative registers were publicly available at various levels of geographical and name resolution (i.e. from lists of individual persons—very fine level—to people headcounts per name and region—coarser level). For the rest of the countries, telephone directories were used from which circa of 200 million records (telephone lines) were cleaned to eliminate duplicate households, non-residential names, standardise the use of initials and other name forms, and transform special characters into English spellings. These records were then geocoded at a postcode level through a pair of coordinates using placename gazetteers (Geonet Names Server—GNS http://earth-info.nga.mil/gns/html/, and Geonames http://www.geonames.org/). Maps of these coordinates were produced and placed within boundary maps of administrative regions using a Geographic Information System, at the two level hierarchy established in Table 8.1. Finally, population counts per name and administrative unit were computed within an Oracle database.

The website version of *Worldnames* only includes surname data, since forenames in telephone directories tend to be biased towards male forenames. A relative indicator of population concentration per name was calculated as a rate per million people (termed "frequency per million"—FPM—in the website). Finally, a location quotient was calculated (see Sect. 8.3.1 for its calculation) per name, taking as a base the whole population of a country, continent or the World, depending on the geographical extent at which the map is displayed.

Table 8.1 Countries and geographical granularity available in the Worldnames website

Continent	Country	Region (original language)	Region (English)	Locality (original language)	Locality (English)
America	US	States	States	Counties	Counties
America	Canada	Provinces	Provinces	Regional municipalities	Regional municipalities
America	Argentina	Regiones	Regions	Provincias	Provinces
Asia	India	States and Union territories	States and Union territories		
Asia	Japan	Regions	Regions	Todofuken	Prefectures, metropoly, territory and urban prefecture
Europe	Italy	Regioni	Region	Province	Province
Europe	Spain	Comunidad Autonoma	Autonomous Region	Provincia	Province
Europe	Austria	Bundesland	States	Bezirke and Statutarstädte	Political districts and statutory cities
Europe	Belgium	Région	Region	Provincies	Provinces
Europe	Denmark	Counties	Counties		
Europe	France	Region	Region	Dе́partements	Department
Europe	Germany	Lander	States	Regierungsbezirke	Administrative districts
Europe	Hungary	Megyék, Föváros and Megyei jogu város	Counties, Capital City and Urban Counties		
Europe	Ireland	Regional Authorities	Regional Authorities	Counties	Counties
Europe	Luxemburg	Districts	Districts	Cantons	Cantons
Europe	Netherlands	Provincies	Provinces		
Europe	Norway	Fylker	Counties		
Europe	Poland	Wojewodztwa	Voivodships	Powiaty, Miasta na prawach powiatu	Counties, cities of county right
Europe	Serbia	Regions	Regions		
Europe	Slovenia	Opčine	Communes (pre 1998)		
Europe	Sweden	Län	Counties	Kommuner	Municipalities
Europe	Swtizerland	Regions	Regions	Cantons	Cantons
Europe	UK	Regions	Regions	Counties	Counties
Oceania	Australia	States and Territories	States and Territories		

(continued)

Table 8.1 (continued)

Continent	Country	Region (original language)	Region (English)	Locality (original language)	Locality (English)
Oceania	New Zealand	Regional Councils (pre 1992)	Regional Councils (pre 1992)		

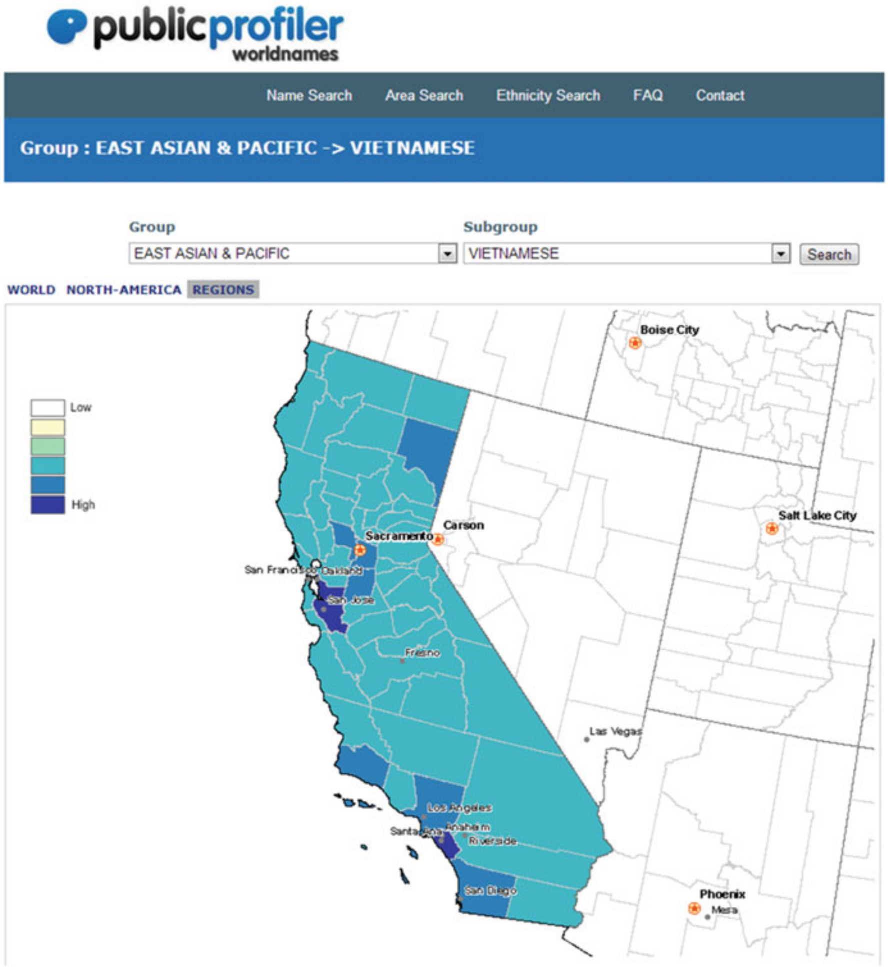
Each row in the table represents a country in the Worldnames database. There are two levels of geographical granularity for which surname data is available in the website, here termed: "Region" (higher level) and "Locality" (lower level). For some countries only the Region level is available. Columns 3–6 show the names to describe the type of "Regions" and "Localities" given to the administrative areas used in each country, providing both the English translation and local language term when this was available

To our knowledge the *Worldnames* database constitutes the first published effort to compute international name frequency data for a large number of countries (26) in various continents (4) and at very fine levels of geographical disaggregation (at municipality-equivalent level). For example, country-level statistics from the Worldnames database have been used by geneticists King and Jobling (2009) to compare the median number of bearers per name in various countries, in a study of population dynamics establishing a parallel between surnames and genetics.

Using the methods described in Sect. 8.1, the *Worldnames* database can be used to uncover various traces of historical as well as contemporary international migrations throughout the world. Only a few representative examples will be mentioned here, but the reader is encouraged to navigate the website to, instead of searching for individual surnames, search for the distribution of surnames ascribed to a particular cultural, ethnic or linguistic origin through the main menu item termed "ethnicity search".

For example, Fig. 8.9 shows a map of the distribution of Vietnamese surnames in California in 2002 at county level. Although the map shows a high concentration of Vietnamese surnames across the whole State of California compared to the national average in the US (a location quotient), a few counties present very high concentrations, specially Santa Ana and San Jose counties (respectively in the Los Angeles and San Francisco Bay metropolitan regions). As the header of the website shown in Fig. 8.9 shows, this map is arrived at by searching for, in the first instance; "East Asian & Pacific" Onomap group, and then selecting "Vietnamese" in the Onomap Subgroup dropdown menu. The taxonomy of "ethnic" groups used here is derived from the aforementioned Onomap taxonomy discussed in Chap. 7. Hundreds of maps of this type can be visualised using this facility, following the major ethnic groups and destination areas amply discussed in the contemporary migration literature (Castles and Miller 2003).

Furthermore, this detailed database is also very useful in uncovering historic migration episodes whose results have been preserved in the current geography of surnames around the world. For example, Fig. 8.10 shows the distribution of French surnames in North America (U.S. and Canada). As expected, they are highly



Group : EAST ASIAN & PACIFIC -> VIETNAMESE

Group: EAST ASIAN & PACIFIC

Subgroup: VIETNAMESE

Fig. 8.9 Map of Vietnamese surnames in California. An example of “Worldnames Ethnicity Search” (third option in the *top* menu) through which the largest Onomap Groups and Onomap Subgroups can be queried in the 26 countries. (*Source*: http://worldnames.publicprofiler.org/EthnicityResult.aspx?country_code=STATE-CA)

concentrated in Canada, especially in the Canadian province of Quebec, and along various border States in the U.S. (New England and the Great Lakes region). However, the cluster in the State of Louisiana in the South, is testimony of over 300 years of French presence in the South of this State, as shown in the map’s inset.

Figure 8.11 shows the distribution of Welsh names across the world, showing a distance decay from Wales within the UK and Ireland, and a strong presence in Anglo-Saxon destination countries (apart from the UK and Ireland, in Australia, New Zealand, U.S. and Canada). However the European map shows a mild presence of Welsh surnames in Brittany, France, because of historical contact

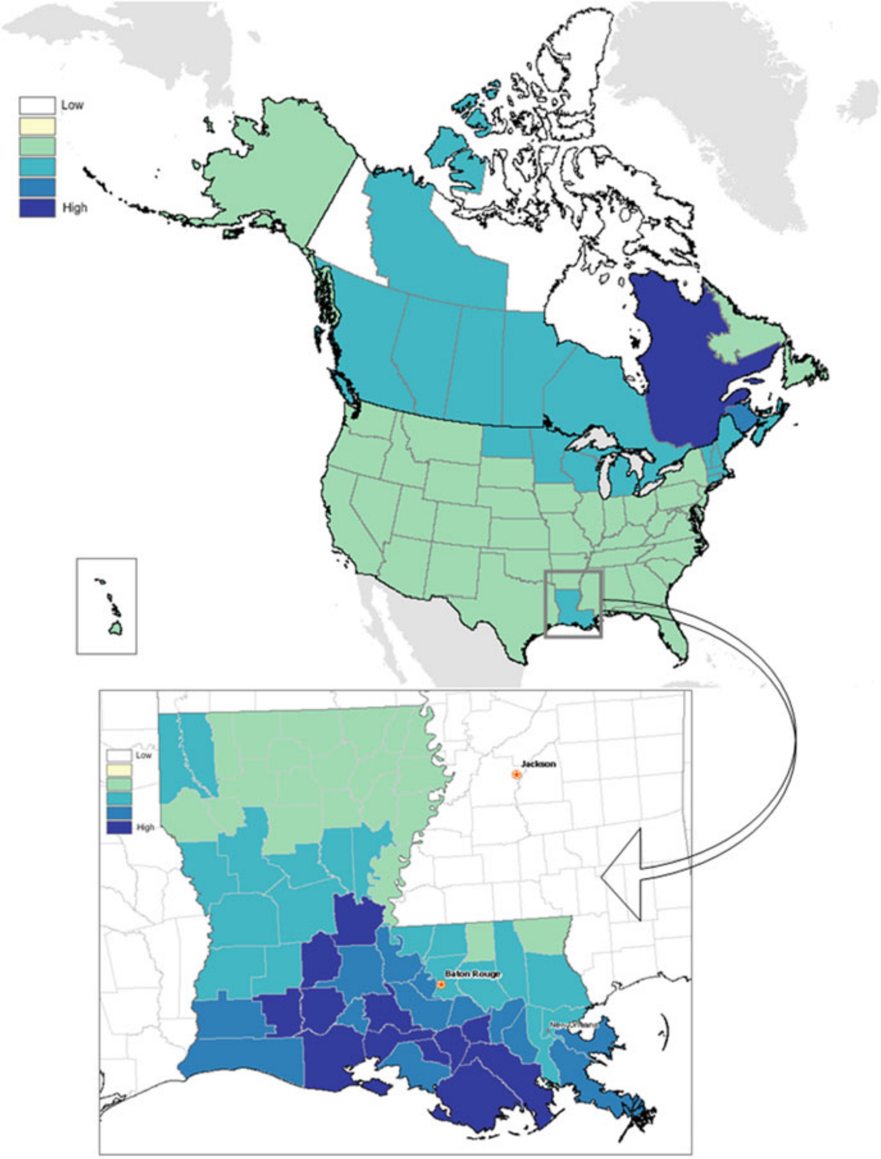


Fig. 8.10 French surnames in North America (2002). Relative concentration of French surnames by US States and Canadian Provinces, with an inset of the State of Louisiana showing the concentration by county. (*Source*: Worldnames http://worldnames.publicprofiler.org/)

across the English Channel, and in the Balearic Islands, Spain, explained by contemporary retirees. More striking is the important cluster of Welsh surnames in the province of Chubut in the Patagonia Region of Argentina. This outlier is explained by an important episode of Welsh emigration to this region from 1865

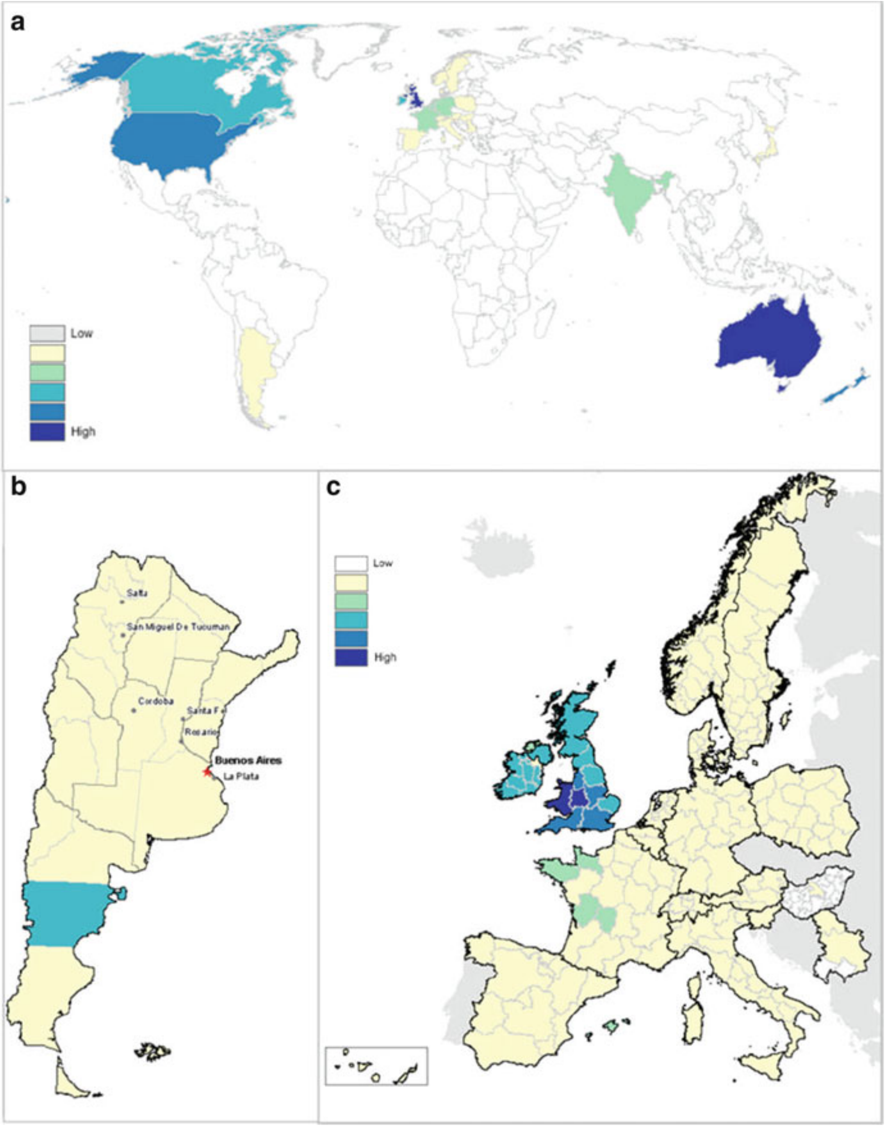


Fig. 8.11 World map of contemporary Welsh surnames diaspora, with insets in Europe and Argentina (2000–2006). (**a**) Distribution of Welsh surnames in Worldnames 26 countries, with Anglo-Saxon countries highlighted as the major destinations. (**b**) Distribution of Welsh surnames in Argentina, depicting a cluster of Welsh surnames in the province of Chubut in the Patagonia Region, following an episode of mass migration in the nineteenth century. (**c**) Distribution of Welsh surnames in Europe. In the UK and Ireland, see the distance decay from Wales, while some significant presence is noted in Brittany, France, because of historical contact across the English Channel, and in the Balearic Islands, Spain, explained by contemporary retirees

THE TIMES
No. 68793 ■ THURSDAY AUGUST 31 2006 ■ NEWSPAPER OF THE YEAR
60p
The name game
HOW TO TRACE YOUR SURNAME ACROSS THE GLOBE NEWS PAGE 5
THE TOP 500 BRITISH SURNAMES
LEARN YOUR FAMILY HISTORY
A FASCINATING GUIDE TO YOUR PAST
FREE SUPPLEMENT INSIDE
NEWSPAPER OF THE YEAR
The Observer
Sunday 15 April 2007
THE TOP 500 3
What's in your name?
bbc.co.uk
Home
TV
Radio
Talk
Where I Live
A-Z
UK version
International version
About the versions
BBC NEWS
WATCH
The News in 2 minutes
News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Education
Science/Nature
Technology
Entertainment
E-mail this to a friend
Printable version

Last Updated: Thursday, 31 August 2006, 10:13 GMT 11:13 UK

Most adventurous surnames mapped

The most adventurous Britons over the decades have been identified by academics analysing people's surnames.

Geographers at the Centre for Advanced Spatial Analysis, at University College London, looked at the spread of 500,000 names across the world.

There are 50% more Blairs now than in 1881

Fig. 8.12 News coverage of the *Worldnames* website 2006–2007. *Source*: From *top* to *bottom*; (**a**) *The Times*, 31 August 2006, cover page, UK edition; (**b**) *The Observer*, 15 April 2007, cover page and special supplement (The Sunday newspaper of The Guardian); (**c**) *BBC News* online, 31 August 2006)

because of religious and linguistic hostilities in Great Britain during the nineteenth century (Williams 1975). Today, it is estimated that 50,000 people in Chubut claim Welsh ancestry and 5,000 speak the Welsh language (The Welsh Assembly 2012).

These examples are sufficient to show the immense research possibilities offered by world-wide, geographically disaggregated surname frequency data classified by ethno-linguistic origin, such as those provided in *Worldnames*. Rather than as a resource for the more typical use of researching individual family history, aggregating surnames together in groups "of common descent" (Weber 1997) permits uncovering hundreds of migration stories such as the ones mentioned in this section (Vietnamese in California, French in Louisiana, Welsh in Argentina). In fact, the value of the *Worldnames* website as a research tool accessible to the general public has been publicised by the news media in various countries, as exemplified by various British newspaper covers shown in Fig. 8.12.

8.4 Conclusion

This is the first of two chapters on spatial applications of the name ethnicity classification approach proposed in this book. It has presented the results of a thorough review of the spatial analysis of names with the purpose of studying population socio-cultural structure from various angles. In particular it has summarised four key methodological approaches common to such analyses, drawing from research in various academic disciplines; 1) the regionalisation method to delimit homogeneous "cultural regions"; 2) the geographical analysis of surname's morphological patterns, in order to search for linguistic or dialectical differences across space as proxies of migration or ethnic minorities; 3) the comparison of geographical distributions of names over time, since abrupt temporal changes are likely a consequence of migration events in the intervening period; and 4) mapping pre-classified names by origins, or in other words, applying name classifications such as Onomap to populations over space. In each of these four general methodological approaches detailed measures and techniques have been introduced, discussing their strengths and limitations for particular contexts of application. The rest of the chapter reviewed representative examples of applications of these methods to the study of historic as well as contemporary migration over space. It particularly drew upon method four; mapping pre-classified names by origin, specially through the application of the Onomap classification discussed in Chap. 7 to various parts of the world. The range of current and potential spatial applications of name-to-ethnicity classifications of populations is astounding. This Chap. 6 has literally taken "a trip around the Western world" and over time to illustrate the value of this approach to the study of both historic and contemporary migrations, in the hope that others will be inspired to take this approach much further.

The fundamental premise underpinning this chapter's proposal, through the mapping of name frequencies, is that the majority of the individuals that share a name do not move, and that by applying various spatial analysis techniques "it possible to identify the shared heartlands of combinations of location-specific surnames that date back 700 or more years" (Longley et al. 2011: 9). Moreover, the reverse is also true, "new arrivals" can be identified in these maps and the migration and population diversity processes of cities and regions can partly untangled and traced over space and time. In this chapter such processes have been studied primarily at the national scale and looking at historical migration flows that took place various generations back in time. The next and final chapter takes a much more "microscopic" and contemporary approach, studying the spatial analysis at very fine grained levels of spatial granularity within a city; London, and monitoring recently arrived groups over the last 50 years or so. In this way the analysis of population diversity through names that this book proposes, goes all the way down to the neighbourhood level. In doing so it connects with the urban segregation literature in sociology, providing a set of new tools of analysis for social scientists that are fully justified in the next and last chapter of this book.

References

Aranda Aznar J (1998) La Mezcla Del Pueblo Vasco. Empiria 1:121–177

Benedictus L (2005) London: the world in one city. The Guardian G2 special supplement, 21 January 2005

Brown LA, Holmes J (1971) The delimitation of functional regions, nodal regions, and hierarchies by functional distance approaches. J Reg Sci 11(1):57–72

Brunet G, Bideau A (2000) Surnames: history of the family and history of populations. Hist Fam 5 (2):153

Brunet G, Darlu P, Zei G (2001) Le Patronyme: Histoire, Anthropologie, Société. CNRS Editions

Castles S, Miller MJ (2003) The age of migration, 3rd edn. Palgrave Macmillan, Basingstoke

Chareille P (2002) Methodological problems in a quantitative approach to changes in naming. In: Beech GT, Bourin M, Chareille P (eds) Personal names studies of Medieval Europe: social identity and familial structures. Western Michigan University, Kalamazoo, MI, pp 15–27

Cheshire JA, Longley PA (2012) Identifying spatial concentrations of surnames. Int J Geogr Inf Sci 26(2):309–325. doi:10.1080/13658816.2011.591291

Cheshire J, Mateos P, Longley PA (2009) Family names as indicators of Britain's changing regional geography. CASA Working Paper 109. London

Cheshire JA, Longley PA, Mateos P (2009) Combining historic interpretations of the Great Britain population with contemporary spatial analysis: the case of surnames. 2009 5th IEEE international conference on e-science workshops (art. no. 5407971), pp 167–170

Cheshire J, Mateos P, Longley PA (2011) Delineating Europe's cultural regions: population structure and surname clustering. Hum Biol 83(5):573–598

Cheshire JA, Longley PA, Yano K, Nakaya T (2013) Japanese surname regions. Papers in Regional Science (in press). doi:10.1111/pirs.12002

Chiarelli B (1992) The use of family names in the study of human migration during the last two centuries. Mankind Quart 33(1):69–77

Colantonio SE, Fuster V, Marcellino AJ (2002) Interpopulation relationship by isonymy: application to ethnosocial groups and illegitimacy. Hum Biol 74(6):871–878

Cummins C, Winter H, Cheng KK, Maric R, Silcocks P, Varghese C (1999) An assessment of the Nam Pehchan computer program for the identification of names of south Asian ethnic origin. J Public Health Med 21(4):401–406

Darlu P, Degioanni A (2007) L'origine Géographique Des Migrants Par La Méthode Patronymique. Espace Géographique 36(3):251–265

Darlu P, Brunet G, Barbero D (2011) Spatial and temporal analyses of surname distributions to estimate mobility and changes in historical demography: the example of savoy (France) from the XVIIIth to XXth century. In: Gutmann MP, Deane GD, Merchant ER, Sylvester KM (eds) Navigating time and space in population studies, Chapter 5. Springer, Berlin

Darlu P, Bloothooft G, Boattini A, Brouwer L, Brouwer M, Brunet G, Chareille P et al (2012) The family name as socio-cultural feature and genetic metaphor: from concepts to methods. Hum Biol 84:169–214

De Smith MJ, Goodchild MF, Longley P (2007) Geospatial analysis: a comprehensive guide to principles, techniques and software tools. Troubador, Leicester

Degioanni A, Darlu P (2001) A Bayesian approach to infer geographical origins of migrants through surnames. Ann Hum Biol 28(5):537–545

Degioanni A, Lisa A, Zei G, Darlu P (1996) Patronymes Italiens Et Migration Italienne En France Entre 1891 Et 1940. Population (French Edition) 51(6):1153–1180

Desjardins B, Bideau A, Brunet G, Charbonneau H, Légaré J (2000) From France to New France: Quebec family names, past and present. Hist Fam 5(2):215–226

Gibin M, Singleton A, Milton R, Mateos P, Longley P (2008) An exploratory cartographic visualisation of London through the Google Maps API. Appl Spat Anal Policy 1:85–97

Gibin M, Mateos P, Petersen J, Atkinson P (2009) Google maps mashups for local public health service planning. In: Geertman S, Stillwell J (eds) Planning support systems best practice and new methods. Springer, pp 227–242. doi:10.1007/978-1-4020-8952-7

Grieco MS (1985) Corby: new town planning and imbalanced development. Reg Stud 19(1):9–18. doi:10.1080/09595238500185021

Guppy HB (1890) Homes of family names in Great Britain. Harrison and Sons, London

Hanks P (2003) Dictionary of American family names. Oxford University Press, New York

Hanks P, Hardcastle K, Hodges F (2006) Oxford dictionary of first names. Oxford University Press, Oxford

King TE, Jobling MA (2009) What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet 25(8):351–360

Kraak MJ, Ormeling FJ (1996) Cartography: visualisation of spatial data. Longman, Harlow

Lasker GW (1985) Surnames and genetic structure. Cambridge University Press, Cambridge

Lauderdale DS, Kestenbaum B (2000) Asian American ethnic identification by surname. Popul Res Policy Rev 19(3):283–300

Longley PA, Webber R, Lloyd D (2007) The quantitative analysis of family names: historic migration and the present day neighbourhood structure of Middlesbrough, United Kingdom. Ann Assoc Am Geogr 97(1):31–48

Longley PA, Cheshire JA, Mateos P (2011) Creating a regional geography of Britain through the spatial analysis of surnames. Geoforum 42(4):506–516. doi:10.1016/j.geoforum.2011.02.001

Manni F, Gue E (2004) Variation: how barriers can be detected by using Monmonier's algorithm. Gene Geogr 2:173–190

Manni F, Guérard E, Heyer E (2004) Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. Hum Biol 76(2):173–190

Massey D (1984) Spatial divisions of labor. Routledge, London

Mateos P (2006) Segregación residencial de minorías étnicas y el análisis geográfico del origen de nombres y apellidos. Cuadernos Geográficos 39(2):83–101

Mateos P (2007) A review of name-based ethnicity classification methods and their potential in population studies. Popul Space Place 13(4):243–263. doi:10.1002/psp

Mateos P (2010) El analisis geodemografico de apellidos en Mexico [Geodemographic analysis of surnames in Mexico]. Papeles de Población 65:73–103

Mateos P (2013a) London's population. In: Bell S, Paskings J (eds) Imagining the future city: London 2062. Ubiquity Press/University College London, London, pp 7–21, http://dx.doi.org/10.5334/bag.a

Mateos P (2013b) Geovisualización de la población: Nuevas tendencias en la web social [Geovisualisation of populations: new trends in the social web]. Investigaciones Geográficas 60:87–100

Mateos P, Tucker DK (2008) Forenames and surnames in Spain in 2004. Names 56(3):165–184. doi:10.1179/175622708X332860

Mateos P, Longley P, Webber R (2006) El Estudio De Migraciones En Latinoamérica a Través Del Análisis Geográfico De Apellidos. In: *II Congreso De La Asociación Latinoamericana De Población*. Guadalajara, Mex 3–5 Sept. Asociacion Latinoamericana de Poblacion. Available at: http://www.alapop.org/Congreso06/DOCSFINAIS_PDF/ALAP_2006_pos_17.pdf

Mateos P, Webber R, Longley PA (2007) The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. Centre for Advanced Spatial Analysis. CASA Working Paper 116. University College London, London. http://www.bartlett.ucl.ac.uk/casa/publications/working-paper-116

Mateos P, Longley PA, O'Sullivan D (2011) Ethnicity and population structure in personal naming networks. PLoS One 6(9):e22943. doi:10.1371/journal.pone.0022943

Murphy AB (1991) Regions as social constructs: the gap between theory and practice. Prog Hum Geogr 15(1):23–35. doi:10.1177/030913259101500102

Nanchahal K, Mangtani P, Alston M, dos Santos Silva I (2001) Development and validation of a computerised South Asian names and group recognition algorithm (SANGRA) for use in British health-related studies. J Public Health Med 23(4):278–285

Novotný J, Cheshire JA (2012) The surname space of the Czech Republic: examining population structure by network analysis of spatial co-occurrence of surnames. PLoS One 7(10):e48568. doi:10.1371/journal.pone.0048568

Office for National Statistics (2012) 2011 census, key statistics for local authorities in England and Wales. http://www.ons.gov.uk/ons/publications/re-reference-tables.html?newquery=*&newoffset=25&pageSize=25&edition=tcm:77–286262

Peach C (2006) Islam, ethnicity and South Asian religions in the London 2001 census. Trans Inst Br Geogr 31(3):353–370

Plane DA, Rogerson PA (1993) The geographical analysis of population. Wiley, New York

Poulain M, Foulon M, Degioanni A, Darlu P (2000) Flemish immigration in Wallonia and in France: patronyms as data. Hist Fam 5(2):227–241

Razum O, Zeeb H, Akgun S (2001) How useful is a name-based algorithm in health research among Turkish migrants in Germany? Trop Med Int Health 6(8):654–661

Rogers C (1995) The surname detective. investigating surname distributions in England 1086-present day. Manchester University Press, Manchester

Scapoli C, Mamolini E, Carrieri A, Rodriguez-Larralde A, Barrai I (2007) Surnames in Western Europe: a comparison of the subcontinental populations through isonymy. Theor Popul Biol 71:37–48

Schürer KE (2002) Regional identity and populations in the past. In: Postles D (ed) Naming, society and regional identity. Leopard's Head, Oxford

Schürer KE (2004) Surnames and the search for regions. Local Popul Stud 72:50–76

Smith MT (2002) Isonymy analysis. The potential for application of quantitative analysis of surname distributions to problems in historical research. In: Smith MT (ed) Human biology and history. Taylor & Francis, London, pp 112–133

The Welsh Assembly (2012) Wales and Argentina. The Official Gateway to Wales. http://www.wales.com/en/content/cms/English/International_Links/Wales_and_the_World/wales_argentina/wales_argentina.aspx

Tyler-Smith C, Xue Y (2012) A British approach to sampling. Eur J Hum Genet 20(2):129–130. doi:10.1038/ejhg.2011.153

Von Ahn M, Lupton R, Greenwood C, Wiggins D (2010) Languages, ethnicity, and education in London. DoQSS Working Paper No. 10–12, Institute of Education, University of London. Available at: http://repec.ioe.ac.uk/REPEc/pdf/qsswp1012.pdf

Webber R (2004) Neighbourhood segregation and social mobility among the descendants of Middlesbrough's 19th century celtic immigrants. CASA. Working Paper 88. London

Weber M (1997) What is an ethnic group. In: Guibernau M, Rex J (eds) The ethnicity reader. Nationalism, multiculturalism and migration. Polity, Cambridge, pp 15–32

Williams G (1975) The desert and the dream: a study of Welsh colonization in Chubut, 1865–1915. University of Wales Press, Cardiff

Winney B, Boumertit A, Day T, Davison D, Echeta C, Evseeva I, Hutnik K et al (2012) People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. Eur J Hum Genet 20(2):203–210. doi:10.1038/ejhg.2011.127

Word DL, Perkins RC (1996) Building a Spanish surname list for the 1990s a new approach to an old problem. Technical Working Paper 13. vol 13p. US Census Bureau, Population Division, Washington, DC. http://www.census.gov/population/documentation/twpno13.pdf

Zelinsky W (1970) Cultural variation in personal name patterns in the Eastern United States. Ann Assoc Am Geogr 60(4):743–769

Chapter 9
How Segregated Are People's Names in London?

> What could be more inherently geographical than segregation?
> (Brown and Chung 2006: 125)

Abstract This final chapter illustrates an example of the potential applications of the *Onomap* classification developed in this book. It consists in a detailed geographical application of the at the small area scale, particularly in ethnicity profiling of neighbourhoods. The chapter introduces the context of measuring ethnicity in London's diverse population, justifying the analysis of residential segregation using people's names. Traditional dimensions and indicators of residential segregation, drawn from the sociological and geographical literature, are implemented comparing both the Census of Population with the name-based approach. Results are discussed identifying key findings and research challenges; in particular the implications of scale effects and the overall complex population dynamics of London.

The literature on name-based ethnicity classifications reviewed in Chap. 6 comprised 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 as opposed to theory or methods. The studies analysed in Chap. 6 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

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_9, © Springer-Verlag Berlin Heidelberg 2014

field, probably because of the majority of them have been developed by strongly empirically-led health and genetics researchers.

The primary aim of the research presented in this book is a methodological one; to develop a new ethnicity classification of personal names covering whole populations 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 chapters. However, the area of applications finally selected to illustrate the value of the Onomap classification presented in the Chap. 7 is primarily one of a geographical nature, as implied by the name of this Springer book series, and intimated by the contents of Chap. 8.

As a geographer, the author believes 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, actually takes place and acquires an interpretable meaning in everyday practices and encounters (Amin 2002, Mateos 2011). Moreover, there is a recognised need to differentiate the identity of neighbourhoods in the delivery of public services. In this public policy context, as Longley puts it, "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 final chapter will illustrate one thread of many potential fields of applications of the Onomap name-to-ethnicity methodology developed in this book. It is included here as an illustrative example of a profound geographical application of the methodology at the small area scale, particularly in ethnicity profiling of neighbourhoods. To our knowledge, this is the first such application of name analysis in the study of residential segregation.

Other examples of actual applications of name-based ethnicity classifications, specifically Onomap have been offered throughout this book, specially in Chaps. 4–8. The case studies mentioned in those chapters do not purport to provide a comprehensive account of the specific applications for the Onomap classification, but rather present a reasonably representative range of examples of potential implementations of the methodology to common problems identified in the ethnic inequalities literature, in particular with a geographical view in mind. Therefore they should be interpreted in conjunction with the case study developed in this chapter, in order to illustrate avenues for future applied research in this area.

Section 9.1 introduces the context of measuring ethnicity in London's diverse population. Section 9.2 presents the justification for the analysis of residential segregation in London and introduces the methods and data used in the chapter. Section 9.3 reviews in detail the concepts behind four of the five traditional dimensions of residential segregation, drawn from the sociological literature, expanding them with additional dimensions and approaches from a geographical perspective. Section 9.4 presents the results of the analysis of segregation using the selected indices, discussing their implications of in terms of scale effects and the overall population dynamics of London. Finally, a conclusion wraps up key findings and discusses the issues and research challenges identified.

9.1 London’s Ethnicity: Measuring “the World in One City”

As discussed in Chap. 8, London is a unique “laboratory” to study ethnicity across the world (see Sect. 8.3.1 for an introduction to London’s population diversity). However, ethnic group categories in the UK Census of population are sometimes too broad to understand the causes for residential segregation in a diverse and rapidly changing city such as London (Mateos 2013).

This chapter uses data from the 2001 Census since the 2011 was not available at the time of performing the analysis. Furthermore, name databases in the form of the UK electoral register was used for 2004 and 2006, and hence closer to the 2001 Census for comparison purposes. Preliminary 2011 Census data available presents a very similar outline, and initial exploration of these data suggest that all of the trends presented in this chapter still hold for the 2011 Census.

Table 9.1 shows the population of each ethnic group as a share of the total population of London in 2001, 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 Onomap classification 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 Onomap name classification to issues surrounding neighbourhood profiling and residential segregation debates. As exposed in the literature review presented in Sect. 2.2, these issues are of most relevance in public policy debate in Britain, and in the developed world in general.

9.2 Residential Segregation in London. Introduction and Methods

The main application presented in this chapter seeks to illustrate the relevance of the Onomap classification to the issues identified in the literature review and highlighted in Chap. 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 Chap. 2, the reality of

Table 9.1 Proportion of the population by ethnic group; London vs. UK (2001 UK Census)

	UK (%)	London (%)
White		
British	87.5	59.8
Irish	1.2	3.1
Other White	*2.6*	*8.3*
Mixed		
White and Black Caribbean	0.5	1.0
White and Black African	0.2	0.5
White and 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*

"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)

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 the UK;—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 Onomap 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 Onomap Subgroups, in order to analyse the level of segregation of ethnic groups and neighbourhoods at very fine scales (Onomap 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.

9.2.1 Data Preparation and Methods

The dataset used in this analysis is the "Onomap-classified" 2004 Electoral Register for Greater London, which contained 5 million electors, individually classified into 66 Onomap Subgroups as per the process described in Sect. 7.5. As a result, 99.79 % of the individuals could be allocated with an Onomap Subgroup, what constitutes a remarkable achievement in terms of population coverage. A summary table of the sizes of each of these Onomap Subgroups is listed in Table 9.2. Individuals were then aggregated into the 131,721 unit postcodes of London, computing counts of people per Onomap Subgroup and postcode unit. Finally this table was further aggregated into Census 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 2006). 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 Geographic Information System (GIS) using Ordnance Survey CodePoint geographical 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. (2008), was used to compute the residential segregation indices for all of the Onomap 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 at http://geoseganalyzer.ucs.inrs.ca/.

However, because of computer memory limitations the segregation indices at postcode unit level for London (n = 131,721) could not be calculated using the

Table 9.2 List of the 66 Onomap Subgroups and their total and relative population sizes in London (2004)

Onomap Subgroup	Total pop.	%	Onomap Subgroup	Total pop.	%	Onomap Subgroup	Total pop.	%
English	2,876,980	57.47	Somalian	20,376	0.41	Muslim North African	2,044	0.04
Irish	414,038	8.27	Hindi not Indian	12,643	0.25	Albania	1,908	0.04
Scottish	323,847	6.47	Black Caribbean	11,554	0.23	Czech &		
			Slovakian	1,660	0.03			
Welsh	222,429	4.44	Muslim South Asian	11,380	0.23	Ukranian	1,629	0.03
Hindi Indian	156,269	3.12	European Other	9,091	0.18	Lebanese	1,404	0.03
Pakistani	140,548	2.81	Balkan	9,035	0.18	Nordic	1,174	0.02
Sikh	83,968	1.68	Chinese	8,874	0.18	Muslim Stans	1,155	0.02
Bangladeshi	72,829	1.45	South Asian Other	8,484	0.17	Korean	1,139	0.02
Italian	71,967	1.44	Vietnam	8,415	0.17	Romanian	1,085	0.02
Nigerian	68,596	1.37	International	6,214	0.12	Baltic	1,061	0.02
Greek	61,296	1.22	Russian	5,539	0.11	Eritrean	1,053	0.02
Muslim Middle East	48,114	0.96	Dutch	5,477	0.11	Ethiopian	918	0.02
Portuguese	44,780	0.89	Swedish	5,155	0.10	Malaysia	891	0.02
Spanish	44,679	0.89	African	4,879	0.10	Ugandan	812	0.02
French	40,264	0.80	Iranian	4,761	0.10	Congolese	598	0.01
Sri Lankan	39,269	0.78	Danish	4,592	0.09			
Jewish	35,984	0.72	Sierra Leonian	3,854	0.08			
Hong Kongese	35,609	0.71	Japanese	3,469	0.07	Unknown name	10,546	0.21
Ghanaian	35,255	0.70	Afrikaans	3,036	0.06	Void name	90,715	1.81
Turkish	34,359	0.69	East Asian	2,645	0.05			
Polish	33,270	0.66	Hungarian	2,603	0.05			
German	33,264	0.66	Armenian	2,436	0.05	Grand total	5,006,490	100.00
Pakistani Kashmir	32,061	0.64	Muslim	2,335	0.05			
India North	31,888	0.64	Black South Africa	2,161	0.04			
Norwegian	24,927	0.50	Finnish	2,099	0.04			

The table is ordered by decreasing population size. The category "unknown name" has been added, although it does not constitute a Onomap Subgroup per se. See text for explanation of "International", "Void" and "Unknown Names" categories.

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.

9.3 The Traditional Dimensions of Residential Segregation

9.3.1 Selection of Segregation Indices

Drawing upon Massey and Denton's (1988) seminal "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 et al. 2003), and 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 in this case study, and therefore simpler indices were preferred. As a result of this selection process, four indices were finally adopted for further analysis:

1. **Evenness**

   ID; Index of Dissimilarity (Duncan and Duncan 1955)

$$ID = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{x_i}{X} - \frac{t_i - x_i}{T - X}\right| \tag{9.1}$$

2. **Exposure**

   ${}_xP_x{}^*$; Isolation Index (Bell 1954; Lieberson 1981)

$$\mathrm{xPx}* = \sum_{i=1}^{n}\left[\frac{x_i}{X}\right]\cdot\left[\frac{x_i}{t_i}\right] \tag{9.2}$$

3. **Concentration**

   ACO; Relative Concentration Index (Massey and Denton 1988)

$$ACO = 1 - \left\{ \frac{\left[\sum_{i=1}^{n} \left(\frac{x_i A_i}{X} \right) - \sum_{i=1}^{n} \left(\frac{t_i A_i}{T_1} \right) \right]}{\left[\sum_{i=n2}^{n} \left(\frac{t_i A_i}{T_2} \right) - \sum_{i=1}^{n1} \left(\frac{t_i A_i}{T_1} \right) \right]} \right\} \tag{9.3}$$

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

4. **Clustering**

   ACL; Absolute Clustering Index (Massey and Denton 1988)

$$ACL = \left[\frac{\sum_{i=1}^{n} \left(\frac{x_i}{X} \right)}{\sum_{j=1}^{n} c_{ij} x_j} - \frac{X}{n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \right] \Big/ \left[\frac{\sum_{i=1}^{n} \left(\frac{x_i}{X} \right)}{\sum_{j=1}^{n} c_{ij} t_j} - \frac{X}{n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \right] \tag{9.4}$$

Key to the equations:

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
$\mathrm{t_i}$ = Total population in spatial unit *i*
$\mathrm{t_j}$ = Total population in spatial unit *j*
$\mathrm{T_1}$ = The sum of all $\mathrm{t_i}$ in areal unit 1 to areal unit $\mathrm{n_1}$
$\mathrm{T_2}$ = The sum of all $\mathrm{t_i}$ in areal unit $\mathrm{n_2}$ to areal unit n
c_{ij} = cell value of the binary connectivity matrix (1 where *i* and *j* are contiguous and 0 otherwise)
$\mathrm{A_i}$ = Area of spatial $\mathrm{unit_i}$
${}_xP_x{}^*$ = 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. (2008).

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 Onomap 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 Onomap Subgroups in Greater London. Those individuals who could not be classified by Onomap 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 Onomap 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 Onomap 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 Onomap subgroups and are deemed to be of an "international" nature, as opposed to any particular Onomap subgroups. "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 Onomap Subgroups, for reasons of ease of representation and discussion. These correspond to the Onomap Subgroups with a total population size in London greater than 3,000 people, which in the list shown in Table 9.2 corresponds to the 46 subgroups that are more numerous than the "East Asian" category.

9.3.1.1 *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) (Duncan and Duncan 1955).

$$ID = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{x_i}{X} - \frac{t_i - x_i}{T - X}\right| \qquad (9.1)$$

(See Sect. 9.3.1 for explanation of variables)

The ID index was calculated for the Onomap Subgroups in London, and the results for the most frequent 46 Onomap Subgroups are listed in Table 9.3. In this table the Onomap Subgroups are ordered by descending index of dissimilarity (ID), noted by the rank, alongside their absolute population size in London. The most segregated Onomap 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 Onomap Subgroups). This is an

Table 9.3 Index of dissimilarity (ID) by Onomap Subgroups in London at Output Area level

Rank	Onomap Subgroup	Total pop.	ID	Rank	Onomap 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

ID index of dissimilarity (Duncan and Duncan 1955); Rank, rank ordered by ID in descending order. The table only lists the most frequent 46 Onomap 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 Onomap Subgroup in London is also listed. See text for explanation of "International", "Void" and "Unknown Names" categories

interesting result, since these are not precisely the groups that come up at the top on the segregation literature on London (Johnston et al. 2002; Peach 1996, 1999). This demonstrates the value of the Onomap classification in uncovering the residential patterns of carefully defined disaggregate ethnic groups. The least segregated Onomap Subgroups (out of the most frequent 46 Onomap 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 9.3 suggests that there is a relationship between the size of the Onomap Subgroup and the level of the segregation index. In order to corroborate this, Fig. 9.1 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

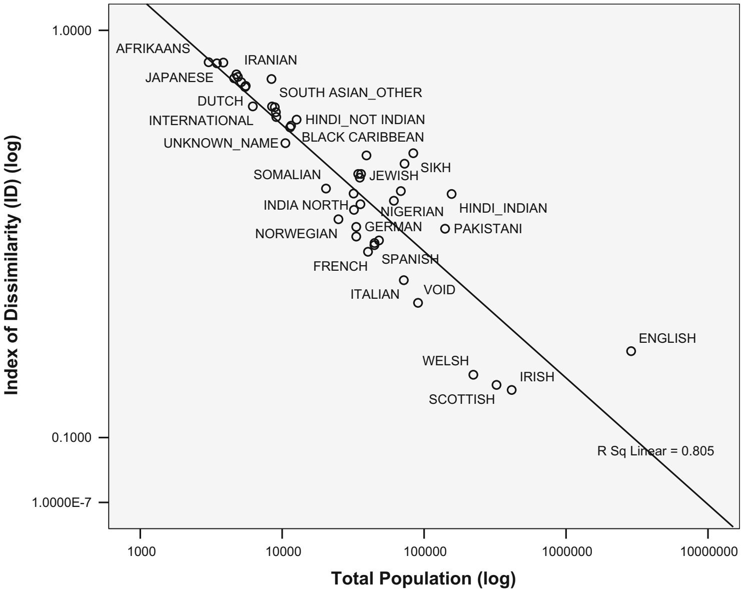


Fig. 9.1 Scatterplot of Onomap Subgroups index of dissimilarity (ID) at Output Area level vs. their total population size in London. This scatterplot only includes the most frequent 46 Onomap Subgroups with a total population size in London greater than 3,000 people. The ID (Duncan and Duncan 1955) calculated at Output Area level is represented on the *vertical axis* and the total population size of the Onomap 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

46 Onomap 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.

Nevertheless, this finding is at odds with the consensus in the literature stating that ID index is independent of the group's 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 1996: 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.

In any case, it is interesting to look at deviations from this relationship between group's size and the dissimilarity index in Fig. 9.1, which are also readily apparent in Table 9.3. It is striking to notice the position of the English Onomap 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 subgroups (Irish, Scottish and Welsh) are less segregated than the English group, as expected by their

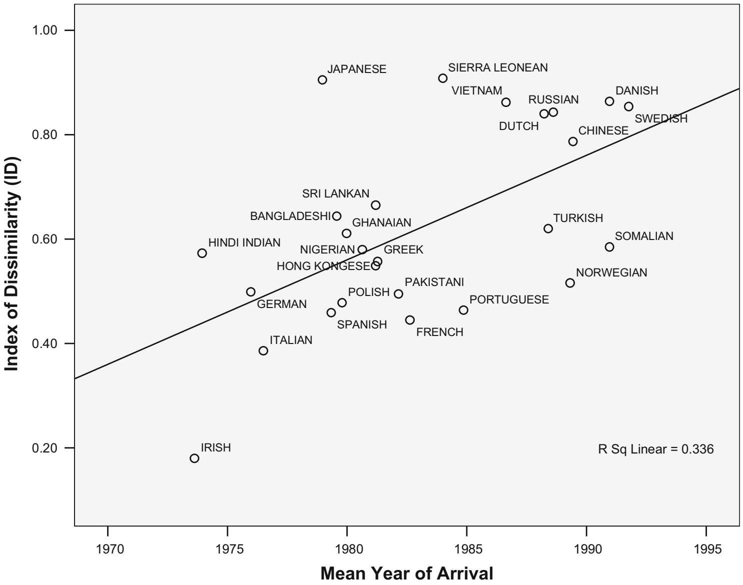


Fig. 9.2 Index of dissimilarity vs. mean year of arrival in Britain. This scatterplot shows on the *vertical axis* the index of dissimilarity of 26 Onomap Subgroups in London, against the average year of arrival in Britain on the *horizontal axis*. The year of arrival information corresponds only to current residents in the London Borough of Camden who have been born abroad. Country of birth has been matched to their associated Onomap Subgroup. *Source*: General Practice register, Camden Primary Care Trust

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 the UK are likely to have lower residential segregation. To test this point, Fig. 9.2 shows a scatterplot of the index of dissimilarity (ID) of 26 Onomap Subgroups in London, against the mean year of arrival in Britain of people born in countries associated with those Onomap 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 partner of the early work in this area led by the UCL Geography team (Mateos 2007). 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, Fig. 9.2 shows that there is a positive relation between the mean 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 Onomap Subgroups measured by the ID index.

9.3.1.2 *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 Shevky 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 (Bell 1954; Lieberson 1981).

$$\mathrm{xPx}* = \sum_{i=1}^{n} \left[\frac{x_i}{X}\right] \cdot \left[\frac{x_i}{t_i}\right] \tag{9.2}$$

(See Sect. 9.3.1 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 9.4 following the same layout as described in Table 9.3.

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, using 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 book.

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 Fig. 9.3 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 Onomap 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 Onomap Subgroups. On the other hand, Onomap 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.

Table 9.4 Index of isolation (P*) by Onomap Subgroups in London at Output Area level

Rank	Onomap Subgroup	Total pop.	P*	Rank	Onomap 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

P*, index of isolation (Lieberson 1981); Rank, rank ordered by the P* index in descending order. The table only lists the most frequent 46 Onomap 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 Onomap Subgroup in London is also listed. See text for explanation of "International", "Void" and "Unknown Names" categories

9.3.1.3 *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 Sect. 9.2), 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 (Massey and Denton 1988).

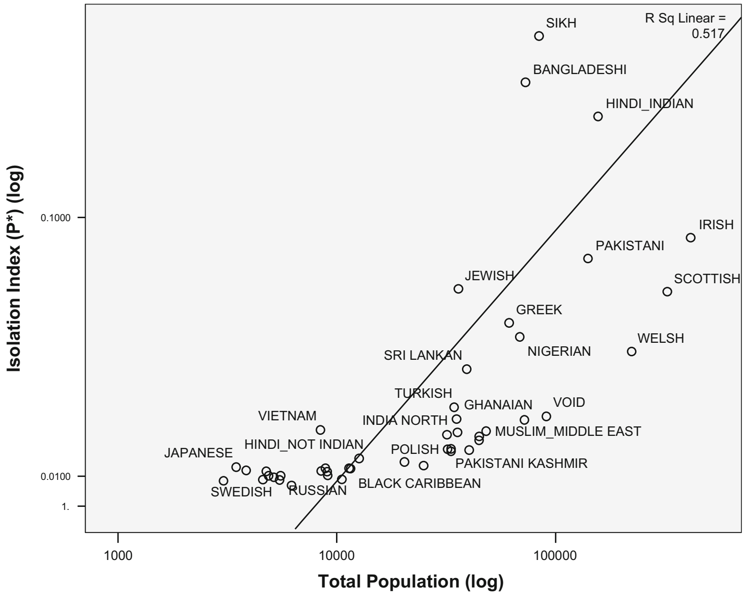


Fig. 9.3 Scatterplot of Onomap Subgroups index of isolation (P*) at Output Area level vs. their total population size in London. This scatterplot only includes the most frequent 46 Onomap Subgroups, each with a total population size in London greater than 3,000 people. The “English” Onomap Subgroup is 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 Onomap 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

$$ACO = 1 - \left\{ \frac{\left[\sum_{i=1}^{n} \left(\frac{x_i A_i}{X} \right) - \sum_{i=1}^{n} \left(\frac{t_i A_i}{T_1} \right) \right]}{\left[\sum_{i=n2}^{n} \left(\frac{t_i A_i}{T_2} \right) - \sum_{i=1}^{n1} \left(\frac{t_i A_i}{T_1} \right) \right]} \right\} \tag{9.3}$$

(Spatial units are sorted by area size in ascending order. See Sect. 9.3.1 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 Onomap Subgroups obtain very similar and high values of ACO, except for the British Isles ones (English, Welsh, Scottish, Irish). If these four Onomap 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 Onomap subgroups 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.

9.3.1.4 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).

$$ACL = \left[\frac{\sum_{i=1}^{n} \left(\frac{x_i}{X} \right)}{\sum_{j=1}^{n} c_{ij} x_j} - \frac{X}{n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \right] \Big/ \left[\frac{\sum_{i=1}^{n} \left(\frac{x_i}{X} \right)}{\sum_{j=1}^{n} c_{ij} t_j} - \frac{X}{n^2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \right] \tag{9.4}$$

(See Sect. 9.3.1 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 Sect. 9.2). It varies from a minimum of 0 (low clustering) to a maximum that approaches but never equals 1 (high clustering).

The results for the calculation of the ACL index in London are shown in Table 9.5, which only lists the most frequent 46 Onomap Subgroups. The most clustered is again the English Onomap Subgroup, since it has most neighbours of its own subgroup, followed by the Sikh, Hindi Indian, and Bangladeshi subgroups. The spatial clustering of these three groups has been persistently identified in the segregation literature on London (Brimicombe 2007; Peach 2006). Following

Table 9.5 Absolute clustering index (ACL) by Onomap Subgroups in London at Output Area level

Rank	Onomap Subgroup	Total pop.	ACL	Rank	Onomap 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

ACL, absolute clustering index (Massey and Denton 1988); Rank, rank ordered by ACL in descending order. The table only lists the most frequent 46 Onomap Subgroups (those with a total population size in London greater than 3,000 people) ranked by ACL. The total population size of each Onomap Subgroup in London is also listed. See text for explanation of "International", "Void" and "Unknown Names" categories

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 (see also maps in Fig. 8.7). Amongst the less

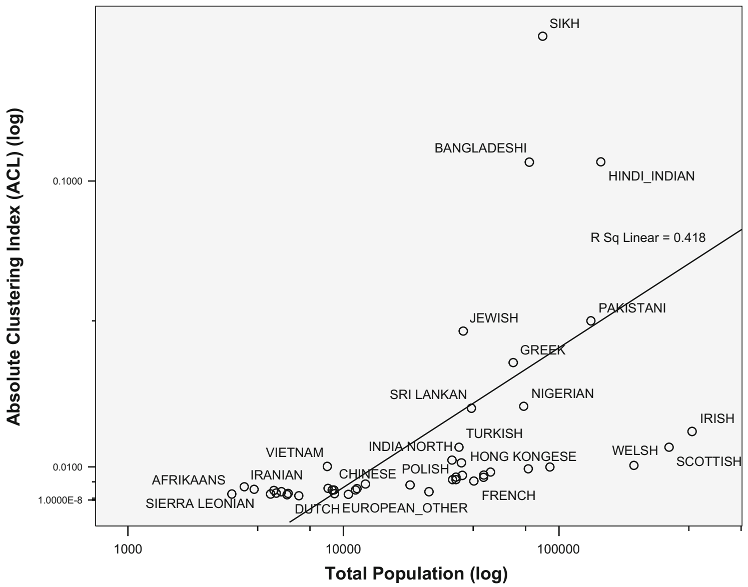


Fig. 9.4 Scatterplot of Onomap Subgroups' absolute clustering index (ACL) at Output Area level vs. their total population size in London. This scatterplot only includes the most frequent 46 Onomap Subgroups with a total population size in London greater than 3,000 people. The "English" Onomap Subgroup 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 Onomap 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

clustered groups, there are several Nordic Onomap Subgroups (Norwegian, Swedish, and Danish), some European small subgroups (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 are 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 Fig. 9.4. 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 Onomap Subgroups with a total population size above about 60,000 people, while below this size, the ACL index barely grows with population size (bottom left part of Fig. 9.4). 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.

9.3.2 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 Onomap-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 segregation indices already presented, since they represent aspects not adequately reflected by the previous measures. More precisely, these measures are not globally/spatially invariant across the study area, and they focus on the over all ethnicity composition of each neighbourhood rather than separately on each particular ethnic group.

9.3.2.1 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 (Anselin 1995):

$$I_i = z_i \sum_j w_{ij}\, z_j \tag{9.5}$$

the observations z_i and z_j are given in standard deviations from the mean $\left[z_i = \left(X_i - \overline{X}\right); z_j = \left(X_j - \overline{X}\right)\right]$, 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 Onomap 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 an area's 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 Onomap Subgroup. While the value of Moran's I varies between -1 and 1, indicating the range from strong negative autocorrelation to strong positive (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's I statistic were represented in a choropleth map for the most significant Onomap Subgroups (p values <0.05) classifying all output areas into five types of spatial correlation, following (Anselin and Regents of the University of Illinois 2004):

- *High-high*; output areas with high proportions of people from the Onomap Subgroup next to areas with similar values.
- *Low-low*; output areas with low proportions of people from the Onomap Subgroup next to areas with similar values.
- *High-low*; output areas with high proportions of people from the Onomap Subgroup next to areas with low values.
- *Low-high*; output areas with low proportions of people from the Onomap 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 Onomap Subgroups were selected representing the third with a larger number of highly clustered Output Areas in London. The 22 maps of the five types of local clustering of LISA are shown in Figs. 9.5, 9.6, 9.7, and 9.8. These maps use the following colour scheme in the colour version of this book: 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. In the black and white version of the book, two shades can be appreciated; dark grey for high-high clustering association, and pale grey for low-low.

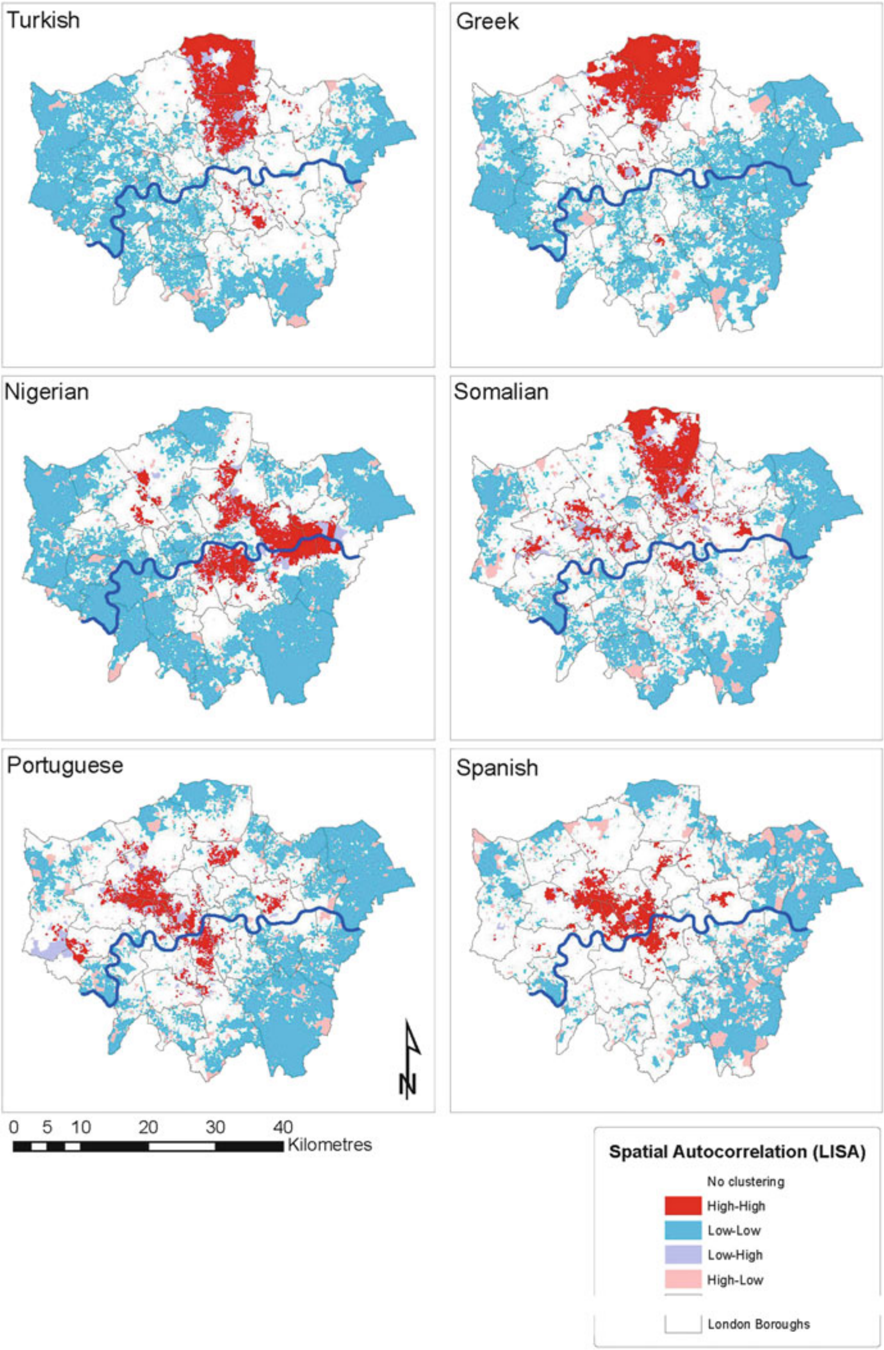


Fig. 9.5 Maps of local indicators of spatial autocorrelation (LISA): Turkish, Greek, Nigerian, Somali, Portuguese and Spanish Onomap Subgroups

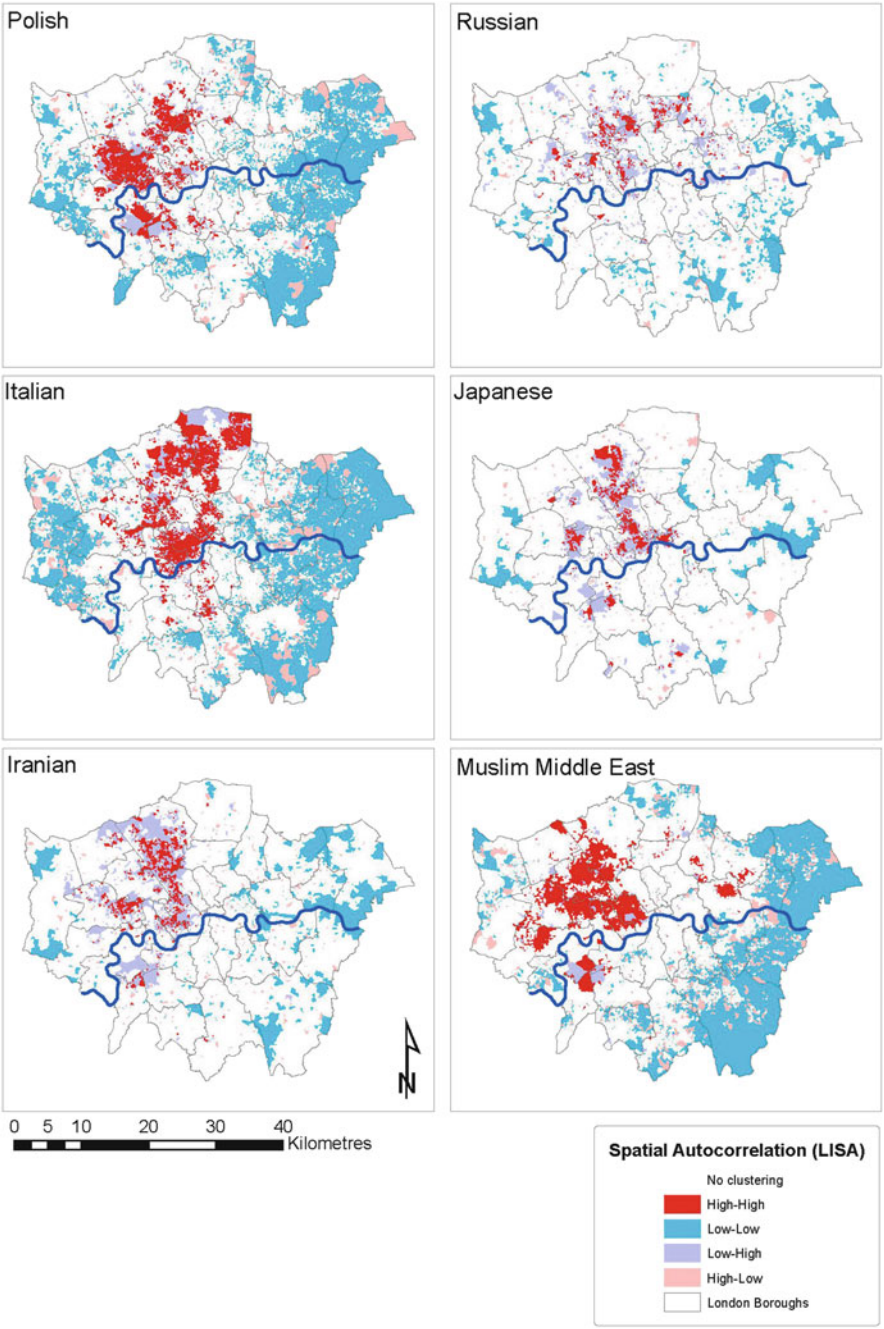


Fig. 9.6 Maps of local indicators of spatial autocorrelation (LISA): Polish, Russian, Italian, Japanese, Iranian and Muslim Middle East Onomap Subgroups

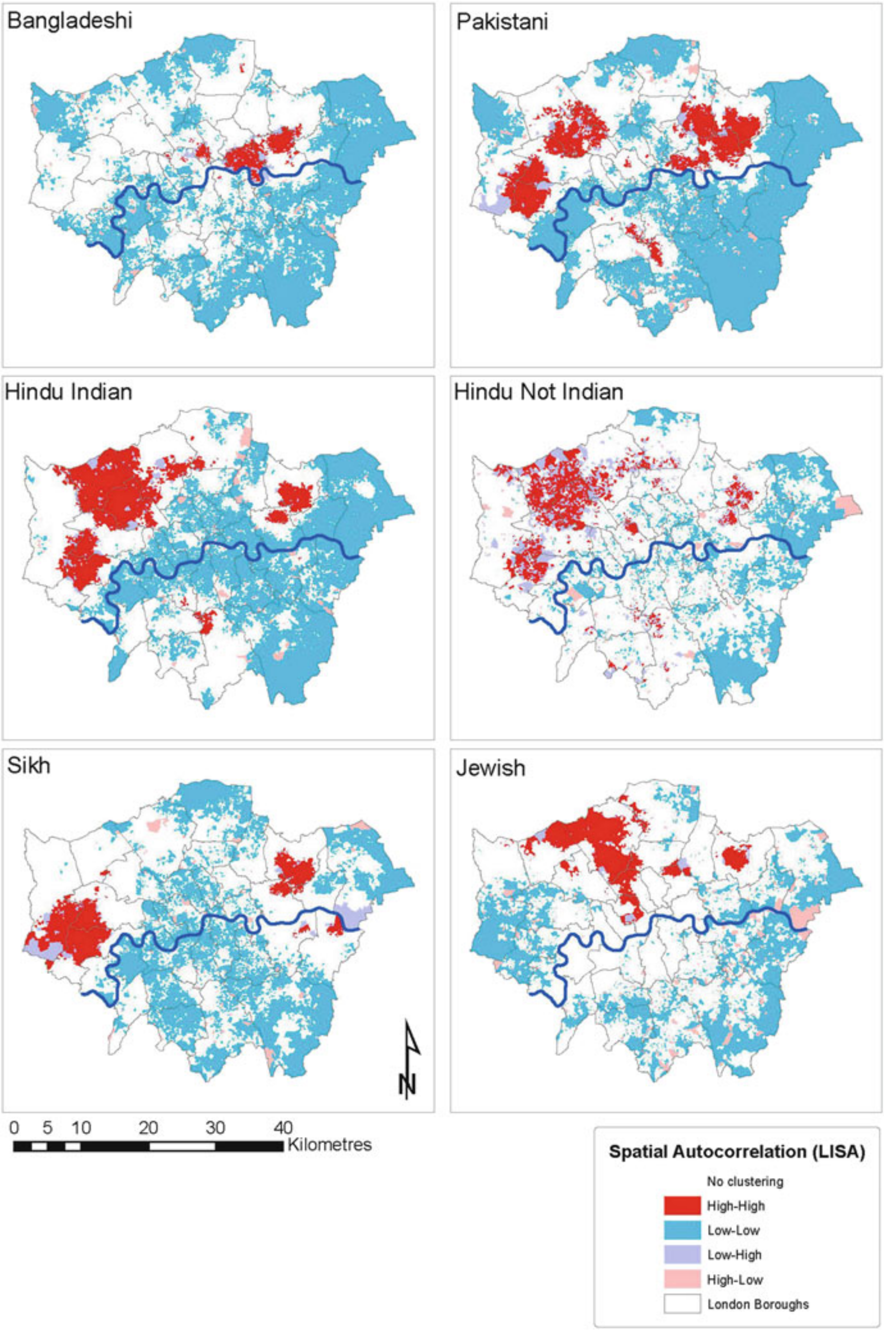


Fig. 9.7 Maps of local indicators of spatial autocorrelation (LISA): Bangladeshi, Pakistani, Hindu Indian, Hindu Not Indian, Sikh and Jewish Onomap Subgroups

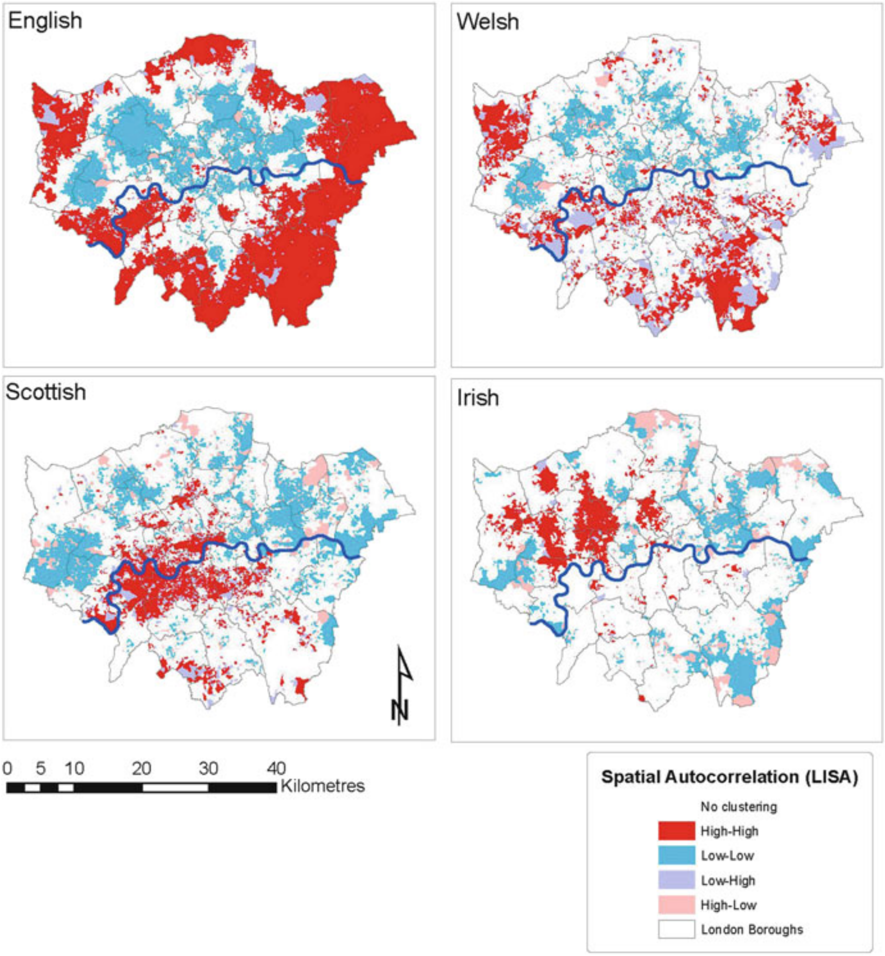


Fig. 9.8 Maps of local indicators of spatial autocorrelation (LISA): English, Welsh, Scottish and Irish Onomap Subgroups

These 22 maps show the unique geographical distribution patterns of these Onomap 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.

Figures 9.5 and 9.6 show 12 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 1999). These maps show the unique spatial clustering patterns of each Onomap Subgroups, in which each subgroup seems to occupy a distinct set of areas within the city. However, 11 of these 12 groups, appear to cluster in an area comprising approximately a third of London's total area, in what constitutes the Northwest quarter 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 Onomap subgroup which is predominantly clustered in the East of London on both sides of the river, following historic settlement areas of Black Africans in London.

Starting with Fig. 9.5 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 Onomap 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 Onomap Subgroup (Fig. 9.6) 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 several years old at the time of finishing writing this book. However, it is known that the Polish ethnic group has been one of the fastest growing in Britain since various Eastern European countries joined the EU in May 2004. 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 Fig. 9.7 represent the Onomap Subgroups associated with the most commonly reported ethnic minorities in the literature; Bangladeshi, Pakistani, Hindu Indian and Hindu Not Indian, Sikh, and Jewish Onomap 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. 2002; 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, when compared with the Onomap classification both methodologies tell a very similar story in London.

Figure 9.8 includes the LISA maps of the British Isles Onomap Subgroups English, Welsh, Scottish and Irish, whose degree of segregation is rarely analysed by the literature. The areas of high clustering of the English subgroup 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 since the UK national origin information is not usually 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 established migrants from Ireland in London. However, in a validation exercise described in Mateos (2007), 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 by a difference in the perceptions of Irish identity between different 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 (names Vs. Census) to study the differentials between their demographic and migration profiles. Furthermore, using the new question about national identity in the 2011 Census, a similar type of future analysis for the rest of the British Isles Onomap Subgroups will be very illustrative of collective identity formation processes at local level.

9.3.2.2 *Diversity*

Beyond the five dimensions of segregation analysed in Sect. 9.3.1, 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):

$$H = -\sum_{i=1}^{n} \frac{\left(\frac{P_{ij}}{P_j}\right) \ln\left(\frac{P_{ij}}{P_j}\right)}{\ln n} \tag{9.6}$$

where:

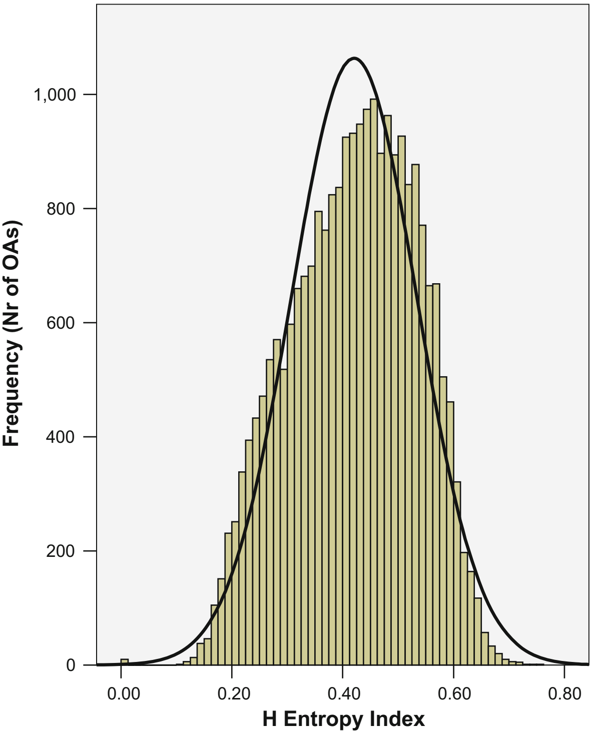
n = number of groups

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 a 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 Fig. 9.9 that shows a near-to-normal shape of its frequency distribution, which is slightly negatively skewed (its skewness is −0.208). However, when the same results are mapped, as shown in Fig. 9.10, systematic differences in OA diversity become very apparent. The map in Fig. 9.10 confirms the aggregated results of the clustering processes unveiled by the previous maps for each individual Onomap 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.

Fig. 9.9 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



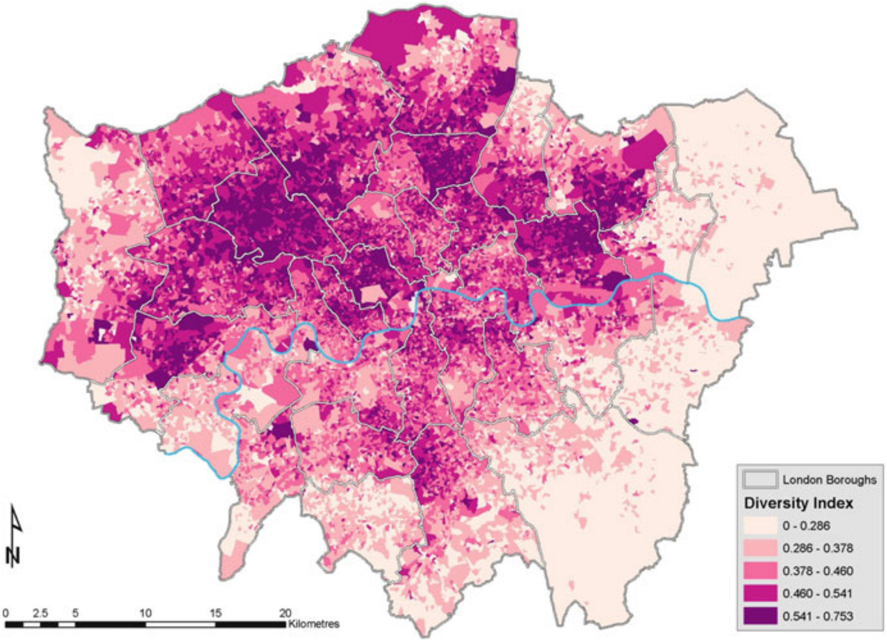


Fig. 9.10 Map of ethnic diversity in London at Output Area level, measured by the Multigroup Entropy Index (*H*)

Table 9.6 Summary of geographic units' characteristics OA: Ouput Area, LSOA: Lower Super Output Area

	Spatial scale			
	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

9.4 Discussion of Residential Segregation Results

9.4.1 Scale Effects

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 1996), 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 Onomap 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 9.6.

Firstly, the 66 Onomap Subgroups were aggregated into a set of 17 aggregations of "Onomap 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 Onomap results more comparable with the Census dataset. These Onomap 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 subgroups), Greek or Greek Cypriot, Jewish, Chinese, Japanese, Bangladeshi, Pakistani, Hindu (all Hindu subgroups), Sri Lankan, Sikh, and Other Muslim (Muslim subgroups not included in the rest). Calculations of the index of dissimilarity (ID) were made for each of these 17 Onomap 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

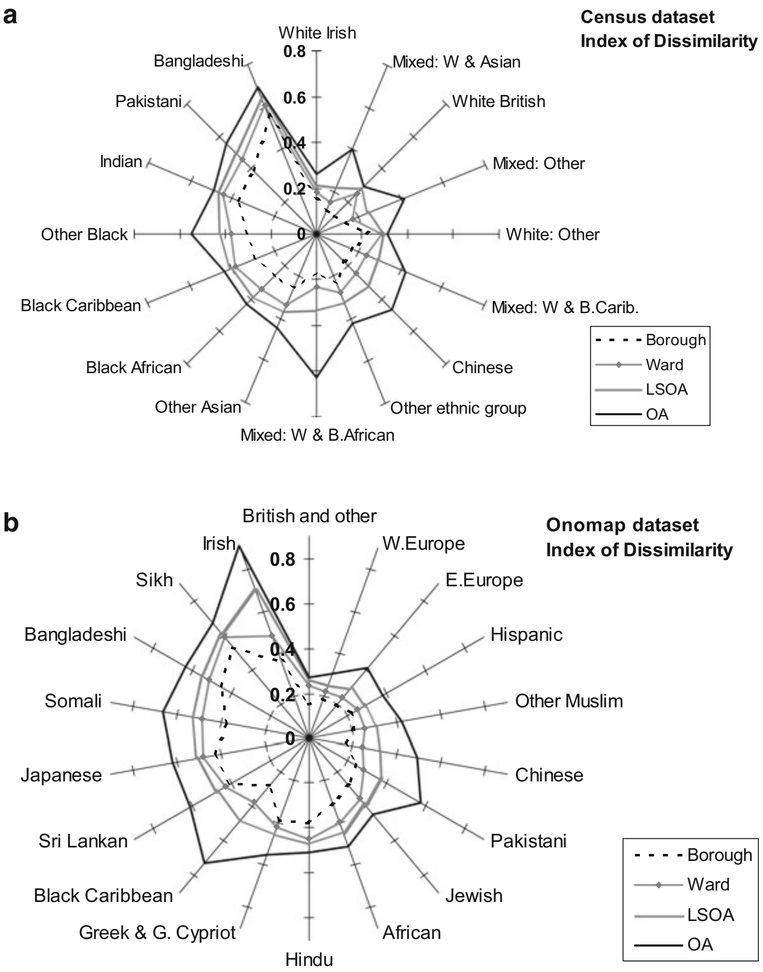


Fig. 9.11 Index of dissimilarity of the Census (**a**) and Onomap (**b**) datasets at four different geographical scales. Both figures represent the index of dissimilarity ID (Duncan and Duncan 1955) calculated for the Census (**a**) and Onomap (**b**) datasets at four different geographical scales. The ethnic categories are ordered by their average ID value, showing increasing segregation in a clockwise direction from "12 o'clock"

(comprising 7,158,904 Census respondents and 24,100 OAs), and higher geographies (LSOA, Ward and Borough). 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 Onomap classification in order to highlight the advantages of the methodology presented in this book.

The "radar" charts shown in Figs. 9.11 and 9.12 represent a graphical comparison of the ID index for both the Census and the Onomap datasets at each of the four geographical scales; OA, LSOA, Ward and Borough. As expected, the level of

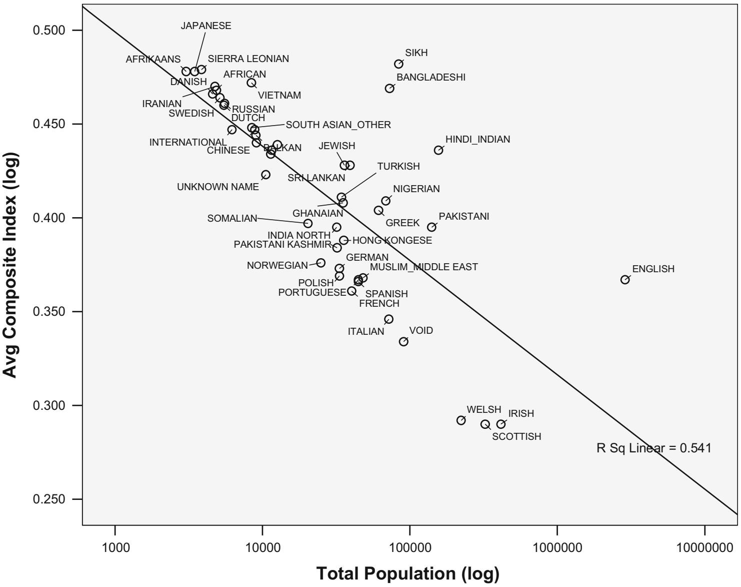


Fig. 9.12 Scatterplot of average composite index vs. total population size

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 Figs. 9.11 and 9.12 would look like parallel concentric rings. However, in the Census-based Fig. 9.11, all the "Mixed" ethnic groups are much more segregated at OA than at LSOA level. A similar difference is noticeable in the Onomap based Fig. 9.12, for the Eastern European, Pakistani, Black Caribbean, and Irish Onomap 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. 2002; Peach 2006; Simpson 2005).

Moreover, the advantage of the much finer Onomap categories is apparent in Fig. 9.12, 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 Onomap 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.

Table 9.7 Effect of MAUP and MEUP on Black African and Somali index of dissimilarity in London

		Borough	Ward	LSOA	OA
Black African (census)	ID	0.26	0.35	0.39	0.43
	MAUP index	60	80	90	100
Somali (Onomap)	ID	0.37	0.48	0.53	0.66
	MAUP index	56	73	80	100
MEUP index		141	139	136	153

ID index of dissimilarity, *MAUP* modifiable areal unit problem, *MEUP* modifiable ethnic unit problem

The table shows the effect on the ID of changing between ontologies of ethnicity (MEUP); Census based "Black African" and Onomap 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 (ID Somali/ID Black African x 100). 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"

Furthermore, changes in the ontology of ethnicity can have a significant effect in segregation levels. In Fig. 9.12 the newly created Onomap 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 Fig. 9.11 presents an ID index of 0.43 at OA level, the Onomap based Somali group in Fig. 9.12 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 detailed Onomap group is as much or even more important than having greater spatial detail. This is illustrated with an example in Table 9.7, that shows the effect on the Index of Dissimilarity (ID) of changing between ontologies of ethnicity (MEUP); Census based "Black African" and Onomap 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 Figs. 9.11, 9.12 and Table 9.7. It is envisaged that the name-based methodology developed in this book will allow other 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 segregation indices.

9.4.2 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 Onomap Subgroup and the four dimensions of evenness, exposure, concentration, and clustering, are summarised in Table 9.8. In order to rank all of the 46 Subgroups evaluated here from high to low overall segregation an average composite index has been created rather crudely 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 0–1 scale, for ease of overall interpretation. However, the “English” subgroup is an outlier in most indices and this could have an impact in the final result. Furthermore, the mean value of each of the four indices vary substantially, making it difficult to compare them through an unweighted average. We acknowledge the issues with this composite index so it should be interpreted with caution.

Table 9.8 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.

Table 9.8 Summary of the four dimensions of segregation and composite index

			Evenness		Isolation		Concentration		Clustering		Avg.
	Onomap	Total									composite
Rank	Subgroup	pop.	ID	Rank	P	Rank	ACO	Rank	ACL	Rank	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	0.810	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	Internaional	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 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 East	48,114	0.469	37	0.026	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	0.001	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

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

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 Onomap Subgroups 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 subgroups whose names are well established in London 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.

The scatterplot in Fig. 9.12 presents 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.

Underlying the relationship exposed by Fig. 9.12 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 Onomap 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 equation is designed to have a majority group and one or just very few ethnic minority groups all with substantial population size.

9.5 Conclusion

While the rest of this book has focused on different developmental aspects of the name-to-ethnicity classification methodology, this chapter has presented one in-depth application of this methodology to the study of contemporary population diversity in cities, in particular residential segregation. This application was selected given its high prominence in the media and relevance to current political debates in contemporary European cities: namely, the study of ethnic residential segregation, and in particular in London.

The application of the Onomap classification to this purpose has opened up new opportunities for much finer analysis of several dimensions of residential segregation in terms of the size and number of the geographical units as well as ethnic group boundaries, and the frequency of update beyond the decennial censuses. 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 minority 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 as Greek and Turkish names show a higher level of segregation than expected by their total population sizes. On the other hand, the three “Celtic” Onomap groups; Welsh, Scottish and Irish, show a very low level of segregation across all dimensions, even less so than the English names majority.

The number of geographical units considered, the group’s population size and the average length of residence of each Onomap 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 Onomap

Subgroups population sizes, there are some subgroups 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 (LISA). 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 London'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, these results should be put in context with the examples of other potential applications of the name-to-ethnicity classification methodology mentioned across the book. Together with the detailed case study presented in this chapter, these other examples constitute a small gallery of applications in order to illustrate the very wide potential applicability of the methodology proposed in the book.

References

Amin A (2002) Ethnicity and the multicultural city: living with diversity. Environ Plann A 34(6): 959–980

Anselin L (1995) Local indicators of spatial association – LISA. Geogr Anal 27:93–115

Anselin L, Regents of the University of Illinois (2004) GeoDa Analysis Software (version 0.9.5-i). University of Illinois. Available at https://www.geoda.uiuc.edu/. Accessed 31 Mar 2007

Apparicio P, Petkevitch V, Charron M (2008) Segregation Analyzer: a C#.Net application for calculating residential segregation indices. *Cybergeo* 414

Bell W (1954) A probability model for the measurement of ecological segregation. Soc Forces 32(4):357–364

Brimicombe A (2007) Ethnicity, religion and residential segregation in London: evidence from a computational typology of minority communities. Environ Plann B 34(5):884–904

Brown LA, Chung S-Y (2006) Spatial segregation, segregation indices and the geographical perspective. Popul Space Place 12(2):125–143

Connolly H, 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 Jan 2006

Duncan OD, Duncan B (1955) A methodological analysis of segregation indexes. Am Sociol Rev 20(2):210–217

Fotheringham SA, Brunsdon C, Charlton M (2000) Quantitative geography. Sage, London

Geary RC (1954) The contiguity ratio and statistical mapping. Incorporated Statistician 5:115–141

Goodchild MF (1986) Spatial autocorrelation. CATMOG 47. Geo Books, Norwich

Johnston R, Forrest J, Poulsen M (2002) Are there ethnic enclaves/ghettos in English cities? Urban Stud 39:591

Johnston R, Voas D, Poulsen M (2003) Measuring spatial concentration: the use of threshold profiles. Environ Plann B 30(1):3–14

Lieberson S (1981) An asymmetrical approach to segregation. In: Peach C, Robinson V, Smith S (eds) Ethnic segregation in cities. University of Georgia Press, Athens, GA, pp 61–82

Longley P (2003) Geographical information systems: developments in socio-economic data infrastructures. Prog Hum Geogr 27(1):114–121

Massey DS, Denton NA (1988) The dimensions of residential segregation. Soc Forces 67:281–315

Mateos P (2007) An ontology of ethnicity based upon personal names. Implications for neighbourhood profiling. Unpublished PhD Thesis, Department of Geography, University College London, London. Available at http://eprints.ucl.ac.uk/16145/

Mateos P (2011) Uncertain segregation: the challenge of defining and measuring ethnicity in segregation studies. Built Environ 37(2):226–238

Mateos P (2013) London's population. In: Bell S, Paskings J (eds) Imagining the future city: London 2062. Ubiquity press/University College London, London, pp 7–21, http://dx.doi.org/10.5334/bag.a

Office for National Statistics (2006) National Statistics Postcode Directory (NSPD) user guide. Office for National Statistics, London. Available at http://www.statistics.gov.uk/geography/downloads/NSPDUserGuide.pdf. Accessed 23 Nov 2006

Openshaw S (1984) The modifiable areal unit problem. Geo Books, Norwich

Owen D (2006) Spatial analysis of segregation in the UK. Presented at royal geographic society with institute of British geographers annual conference, London, 30 Aug 2006

Peach C (1996) Does Britain have ghettos? Trans Inst Br Geogr 21:216–235

Peach C (1998) South Asian and Caribbean ethnic minority housing choice in Britain. Urban Stud 35(10):1657–1680

Peach C (1999) London and New York: contrasts in British and American models of segregation with a comment by Nathan Glazer. Int J Popul Geogr 5(5):319–347

Peach C (2006) Islam, ethnicity and South Asian religions in the London 2001 census. Trans Inst Br Geogr 31(3):353–370

Peach C, 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 Aug 2006

Peach C, Robinson V, Smith S (1981) Ethnic segregation in cities. University of Georgia Press, Athens, GA

Phillips D (1998) Black minority ethnic concentration, segregation and dispersal in Britain. Urban Stud 35(10):1681–1703

Shevky E, Williams M (1949) The social areas of Los Angeles, analysis and typology. University of California Press, Berkeley, CA

Simpson EH (1949) Measurement of diversity. Nature 163:688

Simpson L (2004) Statistics of racial segregation: measures, evidence and policy. Urban Stud 41: 661–681

Simpson L (2005) Measuring residential segregation. Presented at census: present and future. ESRC/JISC Census Programme, Leicester, 16 Nov 2005. Available at http://www.ccsr.ac.uk/research/migseg.htm. Accessed 24 Apr 2006

Simpson L (2007) Ghettos of the mind: the empirical behaviour of indices of segregation and diversity. J R Stat Soc A Stat Soc 170(2):405–424

Thiel H, Finezza AJ (1971) A note on the measurement of racial integration of schools by means of informational concepts. J Math Sociol 1:187–194

Voas D, Williamson P (2000) The scale of dissimilarity: concepts, measurement and an application to socio-economic variation across England and Wales. Trans Inst Br Geogr 25(4):465–481

Wong DWS (2003) Spatial decomposition of segregation indices: a framework toward measuring segregation at multiple levels. Geogr Anal 35(3):179–194

Wong DWS (2004) Comparing traditional and spatial segregation measures: a spatial scale perspective. Urban Geogr 25(1):66–82

Chapter 10
Conclusion

Identity, though complex, can be encoded in a name
(Seeman, 1980: 129)

Abstract This book has presented a wealth of evidence of how naming ties are formed, disseminated and preserved over generations and across space through socio-cultural networks. Through a series of innovative methods and multi-disciplinary linkages this book has shown why disentangling these naming ties can be very useful in population studies of diversity between human groups, both in historic and contemporary contexts. Space is the geographical key to link those temporal, social and ethno-cultural processes, providing the backdrop onto which uncover past and current migrations, and thus letting the reader to trace identity in space, through names, ethnicity and populations.

Interest in the concept of ethnicity has surged over the last two decades in academic circles. This has followed an explosion of attention in issues of migration, race, linguistic difference, nationalism, and religion, around a renewed preoccupation with the question of defining and asserting collective identities in an increasingly globalised world. But ethnicity is a socially constructed, multidimensional concept in a constant state of flux, and hence extremely difficult to capture in static labels and categorisations. However, the definitions and measurements proposed for this complex concept have fallen short of the requirements expressed by a wide range of ethnic groups. Many of these groups do not recognise themselves, their particular perception of collective identity, in official ethnicity classifications. Therefore, the wide breadth of subsequent academic analyses on ethnic inequalities, based on these official classifications, doesn't make justice to their claims. This book has proposed an alternative methodology to bypass some of these problems: the analysis of the cultural ethnic and linguistic origin of personal names. This technique offers a technique to provide a more nuanced and flexible means to classify

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4_10, © Springer-Verlag Berlin Heidelberg 2014

populations according to various ontologies of ethnicity, which can be then fine-tuned to specific spatio-temporal and thematic contexts of application.

Research linking people's names to such identity issues has been partially addressed in the fields of genetics, epidemiology (public health), linguistics, economics, geography, demography, sociology, psychology, anthropology, history and genealogy. However, most of the publications reviewed in the first part of the book, have been developed in isolation from each other, typically focusing on a single purpose application within their disciplinary boundaries and a particular country or language. As such, most researchers in this area have failed to appreciate the fascinating global and cross-disciplinary linkages between forenames, surnames, and ethnicity over space and time. This book has summarised a thorough review of the socio-linguistic, geographical, historical and ethno-cultural aspects shaping naming practices in various countries, weaving together the disparate existing evidence into, what we hope is, a compelling common narrative. As such, it has disentangled most of the key factors driving how surnames are transmitted over generations and across space, forenames assigned to new-born children in all societies, and the combination of the two naming processes forming unique population structures.

Based on this finding, the second part of the book justified the use of name-based classifications to understand population diversity. It identified the need to develop new name-based ethnicity classifications that are comprehensive in their coverage of ethnic groups, easily expandable and reproducible by other researchers. The book then proposed an alternative view of the linkages between forenames and surnames conceived in terms of naming networks, which can be clustered to search for alternative delineations of ethno-cultural affinity. In doing so it has presented an innovative contribution that, to the best of our knowledge, the existing disparate literature had so far failed to identify.

The book's empirical findings presented in the latter chapters are grounded in the results of an ambitious academic research project on the quantitative analysis of names at the Department of Geography in University College London (UK). Some 300 million names were assembled from 26 countries in Europe, North America, and Australasia, and each was made available for study and mapping at finely detailed geographical levels. The similarities and commonalties amongst the names were then crystallised in a classification of 185 distinctive cultural, ethnic and linguistic groups worldwide termed *Onomap*.

Finally, in order to illustrate the validity of this proposition, the third part of the book has presented a gallery of applications in the spatial analysis of ethnicity and names across various countries. At the core of such contribution is the use of quantitative and spatial analysis techniques on name frequency distributions classified by *Onomap*, as a new research methodology to unveil our common pasts. Moreover, it has further expanded on an in-depth case study with a key geographical application of such ethnicity classifications; the study of contemporary residential segregation in cities (London in particular). Through these two final Chaps. 8 and 9, the core argument of the book moves away from naming practices to the ultimate aim of representing subtle variations of ethnicity constructs at population level and across space. In doing so, it engages back with the book's opening, the discussion on the definition and measurements of ethnicity presented in Chaps. 1 and 2. The book's argument then has come full circle, justifying the use of names as a—rather imperfect—alternative methodology to understand population diversity. Although names are at the forefront of this book, they are only instrumentally used here as an alternative methodological toolset, which, taking the necessary precautions, can provide a useful proxy to reach a much higher aim; the study of important contemporary debates on population diversity and ethnic inequalities, such as for example challenging entrenched stereotypes in residential segregation in cities. However, it is acknowledged that this is just one of the various alternative methods in which this task can be approached.

It is hoped that this book reaches a wide audience that breaks out of academic silos and national literatures, and that it inspires other scholars to develop further avenues of research in this fascinating journey; gaining transdisciplinary understandings about our collective identities and shared pasts in studies of populations, “tracing identity in space”.

Reference

Seeman MV (1980) Name and identity. Can J Psychiatry 25(2):129–137

List of Abbreviations

ACL	Absolute clustering index
ACLS	American Council of Learned Societies
ACO	Absolute concentration index
AD	Anno Domini
avg	Average
BC	Before Christ
CA	Canada
CASA	Centre for Advanced Spatial Analysis
CEL	Cultural, Ethnic and Linguistic Group
CELG	Cultural, Ethnic and Linguistic Group (technique)
Chap.	Chapter
COB	Country of birth
CV	Curriculum vitae
DAFN	Dictionary of American family names
DNA	Deoxyribo-nucleic acid
EM	Ethnic minority
EPSRC	Engineering and Physical Sciences Research Council
ESDA	Exploratory spatial data analysis
ESRC	Economic and Social Research Council
EU	European Union
F	Forename
F_List	Forename list
FB	Foreign born
FHSA	Family Health Service Authority Register
Fig.	Figure
FPM	Frequency per million
GB	Great Britain
GIS	Geographic information system
HM Government	Her Majesty's Government (UK)

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science,
DOI 10.1007/978-3-642-45413-4, © Springer-Verlag Berlin Heidelberg 2014

ID	Index of dissimilarity
ISO	International Standards Organisation
KDE	Kernel density estimation
km	Kilometres
LAS	Language analysis systems
LISA	Local indicators of spatial autocorrelation
Ln	Natural logarithm
log	Logarithm
LQ	Location quotient
LSOA	Lower level super output area
MAUP	Modifiable areal unit problem
MDS	Multidimensional scaling
MEUP	Modifiable ethnic unit problem
Mex	Mexican
N	Nickname
N/A	Not available
N/K	Not known
NHS	National Health Service
NP	Naming proximity
NPV	Negative predictive value
NSPD	National Statistics Postcode Directory
NZ	New Zealand
OA	Output area
OECD	Organisation for Economic Co-operation and Development
ONS	Office for National Statistics
PPV	Positive predictive value
(Q)	Modularity
R	R-project statistical computing and graphics package
RGB	Red, green and blue
S	Surname
S_List	Surname list
SANGRA	South Asian Name and Group Recognition Algorithm
Sect.	Section
SMOBE	Survey of minority-owned business enterprises
SNI	Sistema Nacional de Investigadores (Mexico)
SOA	Super output area
SOMs	Self-organising maps
UCL	University College London
UK	United Kingdom of Great Britain and Northern Ireland
US	United States of America
Y-STR	Short tandem repeat on the Y-chromosome

Index

P. Mateos, *Names, Ethnicity and Populations*, Advances in Spatial Science, DOI 10.1007/978-3-642-45413-4, © Springer-Verlag Berlin Heidelberg 2014