Applications of New Forms of Data to Demographics

Alistair B. Leak

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Geography

Department of Security and Crime Science

University College London

December 5, 2017

I, Alistair B. Leak, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

At the outset, this thesis sets out to address limitations in conventional population data for the representation of stocks and flows of human populations. Until now, many of the data available for studying population behaviour have been static in nature, often collected on an infrequent basis or in an inconsistent manner. However, rapid expansion in the use of online technologies has led to the generation of a huge volume of data as a byproduct of individuals' online activities. This thesis sets out to exploit just one of these new data channels: raw geographically referenced messages collected by the Twitter Online Social Network.

The thesis develops a framework for the creation of functional population inventories from Twitter. Through the application of various data mining and heuristic techniques, individual Twitter users are attributed with key demographic markers including age, gender, ethnicity and place of residence. However, while these inventories possess the required data structure for analysis, little is understood about whom they represent and for what purposes they may be reliably employed. Thus a primary focus of this thesis is the assessment of Twitter-based population inventories at a range of spatial scales from the local to the global. More specifically, the assessment considers issues of age, gender, ethnicity, geographic distribution and surname composition.

The value of such rich data is demonstrated in the final chapter in which a detailed analysis of the stocks and flows of peoples within the four largest London airports is undertaken. The analysis demonstrates both the extraction of conventional insight, such as passenger statistics and new insights such as footfall and sentiment. The thesis concludes with recommendations for the ways in which social media anal-

ysis may be used in demographics to supplement the analysis of populations using conventional sources of data.

Acknowledgements

In completing this PhD, I have been fortunate of the support of a significant number of individuals and organisations. Chief amongst these being my principal supervisor, Professor Paul Longley, who has driven me to achieve both this PhD and to develop myself within academia. Without his support, patience and humour I would not have had the opportunity to complete this thesis. I also wish to thank my second supervisor, Dr James Cheshire, who provided the initial opportunity for me to pursue the PhD. It was Dr Cheshire who first supervised me within UCL and who provided me with a wealth of opportunities to develop both academically and professionally. In particular, he has been inspirational in the use of new forms of data and the effective visualisation and communication of information.

I wish to acknowledge my sponsor DSTL for the provision of financial and technical support that I received. I am grateful that they provided me with this opportunity and for the advice which I received from them. Specific thanks are given to my points of contact Mr Leo Borrett and Dr Lucy Burton. DSTL Grant number: 12/13NatPhD_61

Thanks also to my friends and colleagues within the UCL Department of Geography and Department of Security and Crime Science. In particular, I wish to highlight the luncheon group who have provided a constant respite and source of humour. Many of my best memories of this PhD stem from time spent with these people.

Lastly, I wish to thank my friends and family who have encouraged me throughout the completion of this thesis. In particular, thanks must be given to my parents and brother who have both been huge influences and motivation in my life. Unfor-

tunately not all of them are now with us, but I hope that this will make them all proud.

Thesis Outputs

Peer Reviewed Journal Publications

2014 A Geocomputational Analysis of Twitter Activity Around Different World Cities. **Geo-Spatial Information Science**. M. Adnan, A. Leak, P. Longley.

Peer Reviewed Conference Proceedings

2015 Assessing the Usefulness of Population Inventories Derived From Twitter Data in Defining Regions. **European Colloquium on Theoretical and Quantitative Geography**, Bari, Italy. A. Leak, P. Longley, J. Cheshire.

2015 Towards a Seamless Worldnames Database. **GISRUK 2015**, Leeds, UK. A. Leak, M. Adnan, P. Longley.

2014 How Representative Are Social Media Datasets of the True Population: A Case for London. **GIScience**, Leeds, UK. A. Leak, M. Adnan, P. Longley.

2013 Social Dynamics of Twitter Usage in London: Ethnicity, Gender, and Age Analysis. **European Colloquium on Theoretical and Quantitative Geography**, Leeds, UK. M. Adnan, A. Leak P. Longley.

Other Conference Proceedings

2016 Identification of Global Regions Based on a Synthesis of Traditional and New Population Inventories. **Association of American Geographers Annual Meeting**, San Francisco, USA. A. Leak, M. Adnan, P. Longley.

2015 Towards a Seamless Worldnames Database. **Association of American Geographers Annual Meeting**, Chicago, USA. A. Leak, M. Adnan, P. Longley.

2014 Towards a Seamless Worldnames Database. **Association of American Geographers Annual Meeting**, Tampa, USA. A. Leak, M. Adnan, P. Longley.

2014 Using Twitter Data As Demographic Data. **Popfest 2014**, London, UK. A. Leak, M. Adnan, P. Longley.

2013 Age, Ethnicity, and Gender Analysis of Twitter Users in Great Britain. **Royal Geographic Society Annual Conference**, London, UK. A. Leak, M. Adnan, P. Longley.

2013 Data Mining to Understand International Dimensions to Online Identity - a Classification of 2+ Billion Names and Their Linkage to Virtual Identities and Social Network Traffic. **Colloquium on Spatial Analysis**, Copenhagen, Denmark. A. Leak, M. Adnan, P. Longley.

Posters

2015 Twitter as a Demographic Data Source. **GISRUK CDRC Workshop**, Leeds, UK. A. Leak.

2014 Applications of Online Social Network Data to Demography and Security. **International Crime Science Conference**, UCL, UK. A. Leak.

Invited Talks

2014 Geography Seminar. **Geography Seminar Series**, UCL, UK. A. Leak.

2013 Geography Seminar. **Geography Seminar Series**, UCL, UK. A. Leak, K. Kempinska.

Book Reviews

2015 M Leitner (ed.), Crime Modelling and Mapping Using Geospatial Technologies. **Environment and Planning B: Planning and Design**, 43,5:960-961.

Prizes

2015 **Travel Bursary**. GISRUK 2015, Leeds, UK.

2013 **Student of the Year**. UCL ESRI Development Centre.

2012 **2nd Place** International Crime Science Conference Poster Competition.

Contents

List of Figures

List of Tables

List of Acronyms

Abbreviation	Definition
API	Application Programming Interface
BIC	Bayesian Information Criterion
CDRC	Consumer Data Research Centre
CEL	Cultural, Ethnic and Linguistic
CRM	Customer Relationship Management
GADM	Global Administrative Database
GDP	Gross Domestic Product
GIS	Geographical Information System
GPS	Global Positioning System
LBS	Location Based Services
LOAC	London Output Area Classification
LOOCV	Leave One Out Cross Validation
LpOCV	Leave p Out Cross Validation
LSOA	Lower Super Output Area
LQ	Location Quotient
NUTS	Nomenclature of Territorial Units for Statistics
MAUP	Modifiable Areal Unit Problem
OA	Output Area
OAC	Output Area Classification
OLS	Ordinary Least Squares regression
ONS	Office of National Statistics
OSN	Online Social Networks
PII	Personally Identifiable Information
PPP	Purchasing Power Parity
RMSE	Root Mean Square Error
SOCMINT	Social Media Intelligence
TOS	Terms of Service
UCL	University College London
UN	United Nations

Chapter 1

Introduction

Throughout human history, governments and rulers have sought to measure and record their populations. Motivations have varied, although in many cases, the driving force has been military conscription and the administration and collection of taxes (Dewdney, 1981). More recently, records of population, or Censuses, have become commonplace and are now employed in a broad range of applications ranging from the provision of education to the planning of major infrastructure. The first example of a modern Census was completed in 1790 in the United States and was shortly followed by the UK and France in 1801. Since then, a Census has been completed in the UK on a decennial basis with the only exception being 1941 during which time the UK was at war. While the recording period has remained consistent, the questions asked have been expanded upon to better reflect the changing needs and interests of society. Beyond the questions, the general mechanics of the Census have remained largely consistent. At set intervals, a survey is conducted of the nation, and the results of this are aggregated and presented for general consumption. In effect, a Census is a static depiction of the population at a specific point in time. It is the static nature and extended period between surveys which form their main sources of criticism. However, across the globe, various alternative means of population recording are now being employed. One example being the use of rolling surveys in the United States. Of the alternatives, there has been increasing interest in the potential applications of new forms of data. New forms of data are a range of data that may be employed beyond their original purpose for the benefit of research, and

commercial gain. Examples include smart meter data, bulk communication data and from data collected through social media. These data, which may be considered as digital exhaust, are providing a new and exciting means to understand the behaviour of populations' at a greater spatiotemporal resolution that has ever previously been possible.

In this thesis, the aim is to make a contribution in regards to how new forms of data may be employed in the study of stocks and flows of populations. The availability of data which are rich in both space and time is unprecedented and provides an entirely new means by which population insight may be generated. Rather than being constrained to static analysis, there is now the potential to investigate the population in a dynamic and evolving manner. This work builds upon existing literature which seeks to extract demographic insight from geotagged Twitter data based on the application of a range of data-mining techniques. Seeking to establish social media data as a valid alternative to conventional population data, this thesis will explore the potential of Twitter at a range of spatial scales from the local to the global. It should be noted that the integration of such new forms of data into conventional analysis is not without its challenges. Many of the qualities of conventional population data, such as detailed attribution and publicised methodologies are not readily available. Consequently, social media based analyses have often been depicted in a negative light, with major questions being raised around for whom and for what, the data are representative. Thus, a key aim of this thesis is to establish how representative Twitter data are of the observable population, and thus, to provide a framework upon which future researchers may analyse and interpret such data. The analysis employs a global database of 1.4 billion geotagged Tweets collected between December 2012 and January 2014.

It should be noted at the outset, that it is not simply the aim of this thesis to be a collection of novel applications based on the analysis of social media data. In reality, such work forms the majority of literature written in regards to social media. Too often, the focus of such analyses is the creation of unique visualisations or the extraction of particular insight within a constrained environment. Such analyses

are often written retrospectively, and thus, are liable to a publication bias in which positive outcomes are published, and negative outcomes are not. Rather, it is to take a step back and critically examine the way in which social media data could be analysed. Thus, this thesis seeks to set out a systematic approach upon which social media analysis, where possible, should be performed.

1.1 Aims

As indicated in the above, the main objective of this thesis is to develop a comprehensive understanding of for whom and for what social media data are representative. Various anecdotal evidence exists in this regard. However, such analyses are either based on survey data, are constrained in their scope, or are beset with other limitations which may adversely impact upon their outcomes. Thus, this thesis has four key aims.

1. To review the current state of geodemographics and identify potential opportunities for progression within the context of New Forms of Data.

2. To develop a methodology for the construction of functional population inventories based on the analysis of geotagged Twitter data.

3. To assess the representativeness of social media population inventories.

4. To deliver recommendations upon which Twitter, and social media more broadly, should be analysed, building upon what is delivered in aims 2 and 3.

The substantive aspects of this thesis are reported in Chapters 4 through 7. These chapters are concerned with the analysis and processing required to transform the raw data collected via Twitter into functional population records and subsequently, on their validation at a range of spatial scales. For reference, the analysis will draw on a range of additional datasets including the Worldnames Database which has been compiled by the UCL Department of Geography and also the UK Consumer Register provided by CACI Ltd. A general analysis at the global scale is conducted in Chapter 5 and a more in-depth analysis is performed at the UK scale in Chapter 6.

1.2 Thesis Structure

1.2.1 Chapter 2: Geodemographics, Identity and Personal names

Chapter 2 is concerned with establishing the foundation upon which this thesis is constructed with a critical evaluation of current practices in geodemographics and population data. Having identified a series of limitations in the use of such data, it is suggested that New Forms of Data, specifically those that are collected via Online Social Media may provide a plausible alternative. The chapter goes on to establish the link between conventional population data and social media data based on the novel analysis of personal names.

1.2.2 Chapter 3: Social Media and Geodemographics Applications

Having established a linkage between geodemographics and social network data, Chapter 3 is concerned with establishing a research context within social media from which this thesis can draw information and further build upon. Here, the focus is two-fold. First, concerning the various ways in which social media are employed. Second, regarding the limitations observed in the various analyses. The chapter goes on to introduce Twitter, the social media platform employed in the completion of this thesis.

1.2.3 Chapter 4: Database Creation, Linkage and Validation

Chapter 4 is concerned with establishing and implementing a framework by which the raw Twitter data collected via the Streaming API can be transformed into functional population inventories. Using a collection of 1.4 billion geotagged Tweets, the aim was to assign each person to a single location at a range of spatial scales. Subsequently, heuristics are applied to individuals' display names with the purpose of extracting their probable forenames and surnames. These, as will be discussed, provide the means by which key identities may be inferred. These analyses were conducted in two countries, the UK and Spain. These countries were chosen given

that each represents a different major language group and because comprehensive names data were available via the UCL Worldnames Database for reference.

1.2.4 Chapter 5: Creation of a Seamless Worldnames Database

Chapter 5 is concerned with applying the inventory creation framework at the global scale. The aim of the chapter is two-fold. First, it provides an opportunity to supplement the UCL Worldnames Database with data for countries where no existing data are held. Second, it provides a means by which the global geography of Twitter may be examined. Seeking to model the probable representativeness of the Twitter inventories on a global scale, a series of analyses are performed seeking to identify what factors were associated with the Twitter inventory performance. Applying this model, it was possible to gain an improved understanding of the global geography of Twitter. The analysis is useful for three main reasons: it may be used to inform the research regarding where analysis of Twitter is likely to be effective; when investigating data related to nationality it provides a means for standardisation, and lastly, it provides a means to identify those individuals who are not resident in the country or area of study.

1.2.5 Chapter 6 : Twitter in the UK: A Basis for Analysis

Chapter 6 is a UK specific assessment of to what degree Twitter data are representative of the observable population in regards to a series of key demographic attributes: age, gender, ethnicity and geographic distribution. These attributes are inferred through the use of a series of heuristics based on individuals' personal names and their long-term tweeting behaviour. The aim of the chapter is to ascertain what proportion of the observable population the Twitter-derived inventories are representative. The results confirm many anecdotal beliefs regarding age and gender bias within the Twitter cohort. The chapter concludes with a discussion regarding how knowledge of the Twitter users' demographic structure may be incorporated into the analysis of population stocks and flows. This is supported by examples in which knowledge of the demographic outcome may have been advantageous.

1.2.6 Chapter 7 : Social Media Demographics

Chapter 7 is an applied chapter in which the potential of demographically attributed Twitter data is showcased through an analysis of London's four largest airports. The analysis is delivered in two parts. In the first, conventional forms of analysis including demographic profiling and airport catchment analysis are performed. In the second, various novel insights are generated drawing on the Twitter data. These include footfall modelling and textual data mining. The significance of the analysis is that it is completed in the absence of direct observation and that the methods employed are easily transferable. The chapter concludes with a discussion of potential strengths, weaknesses and opportunities for the methods demonstrated.

1.2.7 Chapter 8: Contributions and Future Work

Chapter 8 consolidates the main finding from the preceding chapters with the aim of highlighting the key methodological contributions in regards to the application of new forms of data to demographics. In particular, the chapter reiterates the various recommendations that have been put forward in regards to the effective use of social media in academia and industry. The chapter concludes with a discussion of potential future work building on what has been achieved over the course of this thesis.

1.3 Notes on Population Data

Throughout the course of this thesis numerous reference is made to the use of population data. It should be noted that this definition encapsulates two distinct forms which the data may take. In the first, to be referred to as 'individual level population inventories', each individual is represented as a single entity. This entity may contain multiple additional attributes such as age, gender, ethnicity or place of residence. In the second, referred to as 'aggregate population inventories' individuals are grouped based on a common value, such as family name and one or more spatial parameters. Where not explicitly mentioned, the unit of aggregation is the country for which the individual is believed to be resident.

Figure 1.1 provides a useful reference by which the two forms of population

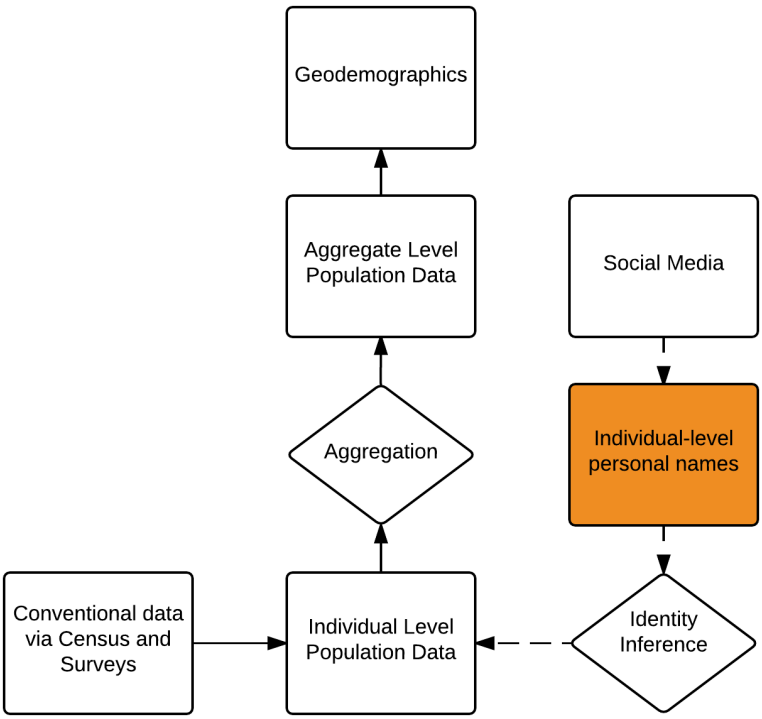


Figure 1.1: Framework by which geodemographics and social media data may be linked via the novel analysis of personal names.

should be understood. At the base level, all population data are constructed using individual level population data. Traditionally such data may be collected via surveys or through national censuses of population. In effect these may be considered the raw form of population data. For the purpose of application, such data are typically aggregated to show the typical characteristics of specific groups of individuals. Typically the grouping of individual level population data is informed by a specific geographic boundary such as national boundaries. A key feature of this thesis is the construction of proxy individual level population data based on the systematic analysis of data harvested from the Twitter online social network. More detail is provided on this process in Chapters 4 and 5. It should be reiterated that the goal is to not to replace the existing Worldnames Database population inventories. Rather, it is the goal to assesses the performance of the Twitter Inventories for these countries and potentially employ the Twitter inventories to fill gaps in the database at the national scale.

1.4 Notes on Software and Data

In this thesis, a range of software and data have been utilised. In regards to software, the majority of analysis has been performed using R, an open-source, cross-platform statistical programming language. The majority of data storage and major handling operations were achieved using PostgreSQL, an open-source relational database management system. Regarding data, the main datasets have been: a corpus of 1.4 billion geotagged Tweets and the UCL Worldnames Database, both collected as part of the Uncertainty of Identity project; The GADM 2.0 global administrative dataset; and a selection of administrative geographies published by the Office for National Statistics. In the case of both software and data, more information is provided as is deemed appropriate.

Chapter 2

Geodemographics, Identity and Personal Names

2.1 Introduction

In developing applications for new forms of data to geodemographics, there is first a need to establish a means by which they may be linked. Thus, this chapter initiates with an introduction to geodemographics, before exploring the means by which the two may be connected based on the novel analysis of personal names. It will be demonstrated how personal names, mined from online social media, may be used in the inference of the key demographic identifiers. Having established this relationship, Chapter 3 will be concerned more broadly with the analysis of online social media and the exploration of potential applications.

2.2 Geodemographics

2.2.1 Introduction

The ability to analyse data within the context of space has facilitated the extraction of new insight from aggregate and individual level population data, enabling the discovery of previously hidden geographic phenomenon (Singleton and Longley, 2009). Such a capability has supported the emergence of geodemographics. An extension of demographics and sociology, geodemographics is concerned with the characterisation of small areas through the classification of regions based on social,

economic and demographic data (Singleton and Longley, 2009).

Figure 2.1: Extract of Charles Booth's poverty map showing the region in the immediate vicinity of University College London (source: https://booth.lse.ac.uk/map/).

One of the earliest examples of geodemographics is attributed to Charles Booth, a social reformer who assembled a detailed map of poverty in London in 1889. Motivated by a desire to disprove statistics alleging that more than 25% of Londoners lived in poverty, his final assessment was, in fact, 30.7% (Harris et al., 2005). Illustrated in Figure 2.1, Booth's map of poverty clearly delimits the streets around Camden, highlighting areas of wealth and poverty. Over a period of years, Booth compiled surveys across the city designed to assess the 'general condition of its inhabitants.' These surveys were subsequently employed in the authorship of the map which classified streets on an ordinal scale of poverty. Aside from Booth, much of the early work in geodemographics may be attributed to the Chicago School of Sociology. The Chicago School was one of the first to place the impetus on the use of empirical analysis to study urban and spatial phenomena (Poston Jr and Micklin, 2006).

Moving forwards in time, the general approach to area classification has shifted towards identifying areas based on their general characteristics. The approach em-

ployed in their creation is largely consistent (Harris et al., 2005): First, a series of variables, which in combination are considered to fulfil the mandate of the classification, are identified. Second, the variables are clustered such that distinct groups/partitions are identified. And, finally, through quantitative and qualitative analysis of the identified groups, names are assigned to each group designed to encapsulate each group's character. Examples of such names, drawn from the 2011 Output Area Classification (OAC) include 'English and Welsh Countryside', 'London Cosmopolitan' and 'Mining Heritage and Manufacturing' (Gale, 2014). The objective of the clustering is to partition areas such that homogeneity within groups and heterogeneity between groups is optimised (Singleton and Longley, 2009).

A fundamental factor in the development of new geodemographics has been the rise of Geographical Information Systems (GIS). GIS being: "An integrated collection of computer software and data used to view and manage information about geographic places, analyse spatial relationships, and model spatial processes. A GIS provides a framework for gathering and organising spatial data and related information so that it can be displayed and analysed" (Wade and Sommer, 2006). It is important to differentiate between GIS which is concerned with the application of methods and GIScience which is the development of new methods and technologies.

From an academic perspective, GIS have several definitions for which Maguire (1991) presents an overview. He suggests three interrelated concepts: the map, the database and the spatial analysis. The map component is a reference to the creation of cartographic products based on existing geographic data. The database component concerns the efficient storage and querying of geographic data. The spatial analysis component is centred on the analysis and modelling of geographic data and is referred to as a spatial information science rather than technology. Maguire (1991) highlights that in many cases, these three components are used in parallel to achieve a specific goal and that the differences in definitions are a consequence of their origin or application.

2.2.2 Applications

When considering the strengths and weaknesses of geodemographics, it is important to consider both their construction and subsequent application. Through an understanding of how these data may be used and by whom the data are employed we can more easily identify potential opportunities for developing new applications. In the subsequent text we investigate the use of geodemographics across a range of applications including customer segmentation, crime analysis and health.

Retail and customer segmentation represent the lions share of geodemographics use. The ability to understand buying behaviour of consumers based on the linkage of customer data and geodemographics offers an effective means by which their buying behaviour may be profiled and understood (O'Malley et al., 1997). Such information may be used for a range of purposes including the optimisation of store placement to the delivery of bespoke marketing. In exploiting the potential of geodemographics for retail one of the most valuable data assets are store loyalty cards. Loyalty cards provide a tangible link between individuals' places of residence and their store choices providing a potential wealth of insight into the geography of purchasing behaviour. While much of the existing research has centred on traditional retail, increasing interest has emerged on the use of geodemographics for the analysis of online behaviour. For example, the Internet Users Classification which uses a range of Census and other survey data to understand individuals' online behaviours.

In terms of crime, Williamson (2008) highlight three key themes in the use of geodemographics including the profiling of specific Wards and police beats, the profiling of operational crime data and the attribution of crime survey data. It is highlighted how awareness of trends based on specific geodemographic types can be used to better understand crime and deliver more effective policing. Bowers (1999) explores the use of GIS and geodemographics in conjunction with crime analysis software to examine patterns of crime within specific segments of the population. Being able to cross-reference both victim and offender locations with specific demographic characteristics enables a greater level of insight than simply knowing the location of the offence. Further, through the application of such techniques to large collections of

crime data, a greater understanding may be obtained of the association with specific crime types and geodemographic types and geography. Consequently, such knowledge may be employed to better target policing resources such that the expenditure of time, money and resources is optimised. The use of geodemographics may thus be considered as a key tool in terms of Intelligence-Led-Policing. However, it should be recognised that these classifications may be misleading due to their depiction of the population at its place of residence. The availability of real-time and more temporally frequent data could prove transformative in such applications.

Health trends and the provision of services is a key area in terms of geodemographic applications and is built on the principle that individuals health is at some level a function of the place and situation within which they exist. Though the linkage of patient records and geodemographic classifications it is possible to examine and identify the prevalence of specific health concerns, at-risk groups and health inequalities (Abbas et al., 2009). Further, through identification of specific groups, it become possible to deliver better preventative care to those groups found to be at the greatest risk.

As may be observed in the preceding text, geodemographics provide an effective means by which the population may be aggregated into distinct homogeneous groups reducing the complexity of populations to such an extent that actionable insight may easily be generated. In each of the three example, geodemographic provide a means to supplement existing data to better understand general behaviours and subsequently to extrapolate the findings based on a broader knowledge of where specific population groups are resident. From this, it is clear that geodemographics are a highly transferable tool. However, as may be increasingly evident, the use of such data for the attribution of individuals and events has a number of potential limitations. These pitfalls and potential remedies are discussed in the subsequent section.

2.2.3 Limitations

While the mass and continued adoption of geodemographics is a testament to their success, some significant limitations persist. Common critiques include: the regular use of a ‘one size fits all’ methodology; the dependence upon residential night-time

data; the reliance on infrequently published data; and the 'black box' nature of many geodemographic classifications in terms of the data used and methods employed (Singleton and Longley, 2009).

The issue of 'one size fits all' is exemplified in the case of the 2001 and 2011 UK Output Area Classifications. In both cases, a fixed set of variables, drawn from the corresponding Census of Population are incorporated into national scale general-purpose classifications. As a consequence, many nuanced aspects of the population are hidden (Singleton and Longley, 2015). In the case of the 2011 OAC, the issue necessitated the creation of a London-specific classification known as LOAC (London OAC) which was better able to capture the diversity within the city (Longley and Singleton, 2014).

It may be argued that the greatest progress has been made in the commercial sector. CACI Ltd (see: http://www.caci.co.uk) for example, maximises the inclusion of data and location specificity in their ACORN classification through the use of location-specific algorithms. Also, the ACORN classification, unlike the OACs, is purported to be readily customisable to users' needs and requirements. However, where commercial classifications lead in customisability and data inclusion, they trail in transparency. An obvious explanation for this is the need to maintain a commercial advantage versus their competitors. This behaviour is in direct contrast to the UK OACs which are entirely open.

Further, dependence on infrequently published data is one of the most common criticisms levelled at Census based geodemographic classifications (Gale and Longley, 2013). A prime example of this is the OACs, which are restricted to decennial publication due to their dependence on Census data. Such is the uncertainty, introduced as a consequence of using legacy Census data that many practitioners opt for alternative commercial classifications that tout more timely publication. That said, Gale and Longley (2013) note that the degree of uncertainty in the OACs is not uniform and rather varies in terms of degree, distribution and geodemographic type. Thus, in certain circumstances, the OAC may be valid across its full period of currency.

Furthermore, the majority of geodemographic classifications represent the population solely at their place of residence. This issue is most profound in the case of the OACs, which, as has been noted previously, are based on Census data. The main barrier to addressing this issue is the availability of daytime or workplace data. In the United Kingdom, efforts have been made to address this. In the 2011 Census of Population through the inclusion of a specific question (Q.40 for England and Wales Census 2011) regarding the place of work. The availability of workplace attribution has facilitated the creation of a new Census geography, known as Workplace zones, which were created through the splitting and merging of existing Output Areas. The availability of representative workplace data has facilitated the creation of a new open-source classification named COWZ-EW (Classification of Workplace Zones – England and Wales) (see: Cockings et al., 2015).

A final concern of geodemographic classifications is that of bias in terms of application and interpretation. Vickers and Rees (2006) note that judgement regarding classifications is frequently made based on the name assigned to the groups or subgroups rather than the pen-portraits or summary statistics provided. In a similar vein, it is a common misconception that all those individuals within the group will exhibit the characteristics of the group for which they are assigned; a statistical bias known as the ecological fallacy.

The appropriate categorisation of demographic variables provides a further challenge in the creation and application of geodemographics. Mateos et al. (2009) refer to this issue, in the context of ethnicity, as the Multiple Ethnic Unit Problem. In the case of ethnic origin, individuals completing the UK Census of Population questionnaire are constrained to a limited set of categories or else must record their ethnicity as 'other.' Such specificity does not necessarily align with individuals' self-perception of their identities resulting in them being forced to identify based on the categories imposed (Aspinall, 2012).

2.2.4 New Forms of Data and Population Representations

While new data products, such as Workplace zones, are starting to address the previously discussed limitations, significant work remains necessary. An emerging theme

is the use of so-called 'New Forms of Data'. Examples of such data include those collected via social media, published as Open Data and consumer data collected via businesses (see OECD 'New Data for Understanding the Human Condition' Report) (OECD, 2013). The OECD (2013) report, identifies several challenges in the adoption of new forms of data and identifies a series of potential limitations concerning access, provenance, permanence, comparability, legality, ethics, linkage and structure.

Alongside new forms of data, there is a requirement for novel means to view and interact with demographic data. There have been increased calls for the development of bespoke geodemographic classifications with a focus on mass-participation portals such as the Internet. However, only recently, has such functionality become available through developments in web-mapping technologies and domain specific software libraries such as OpenLayers.js and the Google Maps API (O'Brien and Cheshire, 2015). In addition to incorporating new forms of data, there are further opportunities in terms of enhancing existing technologies. Two specific instances are DataShine, produced at UCL and Smart Steps produced by Telefonica.

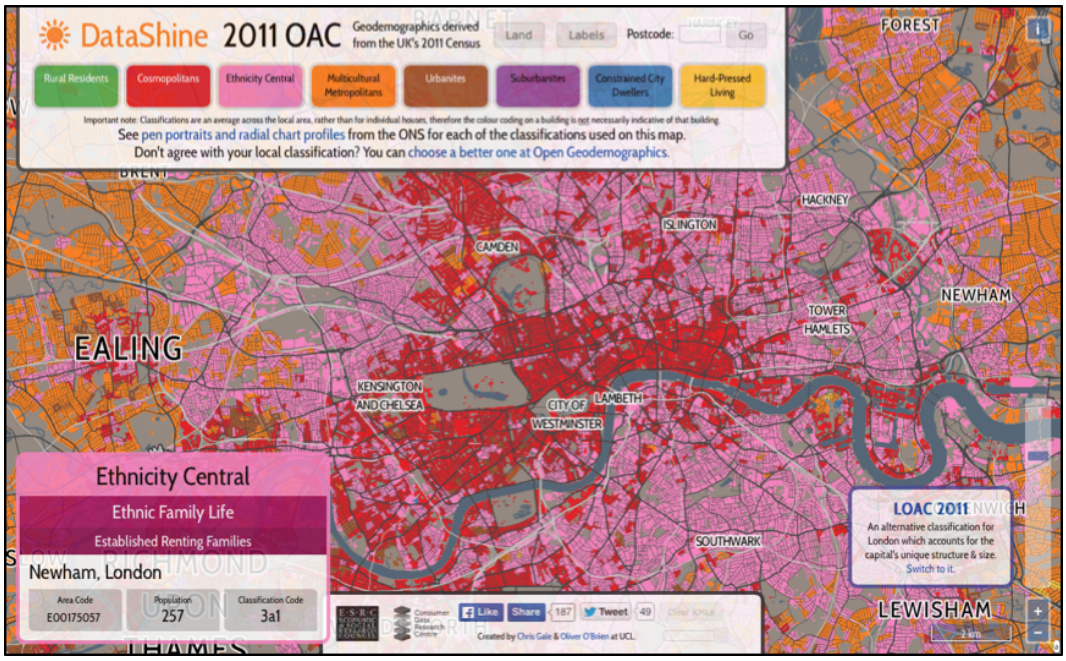


Figure 2.2: Screenshot of the UCL Datashine 2011 OAC web mapping platform (source: http://oac.datashine.org.uk/).

DataShine (see http://www.datashine.org.uk) is an online application for viewing and interacting with various geodemographic and population datasets (O'Brien and Cheshire, 2015). Illustrated in Figure 2.2, DataShine, unlike conventional web mapping platforms, renders geographic datasets in real-time facilitating previously unseen levels of personalisation and interaction. While not addressing all of the critiques of geodemographic, DataShine makes a sizeable step forwards and may in time provide a platform for the creation and analysis of customised geodemographic classifications. A feature of note is 'local area rescaling', which, allows users of the service to recalculate the symbology breaks, such that national trends do not mask local patterns, a common criticism in the visualisation of aggregate population data.

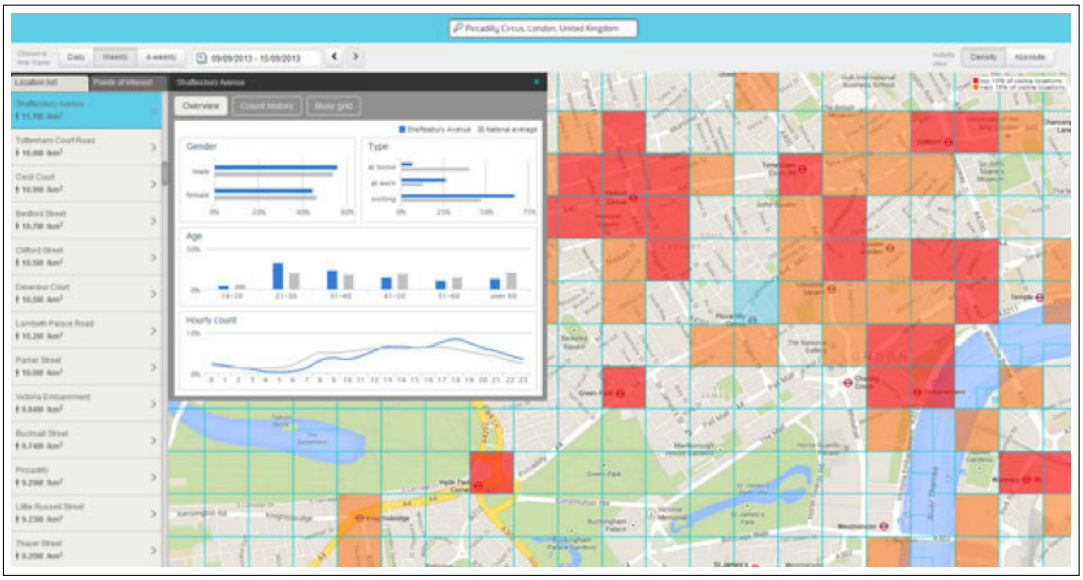


Figure 2.3: Screenshot of the Telefonica's Smart Steps application being employed in the analysis of crime (source: Bogomolov et al., 2014).

Smart Steps (see: http://dynamicinsights.telefonica.com/blog/488/smart-steps-2) which describes itself as a 'Big Data Insight Application' is developed by Telefonica's Dynamic Insight team. Illustrated in Figure 2.3, the application aggregates anonymised data from the O2 mobile phone network using cell-tower locations and identity attribution drawn from customer records. The application facilitates the geo-temporal representation of populations, however, is constrained regarding identity attribution to age, gender and ethnicity. In many respects, Smart Steps may be considered as a new paradigm in geodemographics. Where conventional products

have been rich in attribution but very limited in terms of spatio-temporal resolution, Smart Steps presents a limited pool of variables with rich spatio-temporal resolution. In the case of some applications, the benefits from the richness of the data may outweigh the limitations assumed through a lack of attribution. Limitation aside, Telefonica showcases several successful implementations of the application in the generation of insight. One, in particular, demonstrates the identification of potential customers by the Morrison's supermarket chain. In the Morrison's case study, the objective was to optimise investment in marketing through an enhanced understanding of their current and potential customer base. While not explicitly stated, it is understood that weighted origin and destination data, available through Smart Steps, provide unique insight into store catchments, indicating potential postcodes to be targeted. As part of a trial in the South West, the use of Smart Steps, versus the previously employed mathematical model, resulted in a 150% increase in new and reactivated customers while offering major savings versus a conventional loyalty card scheme or customer relationship management system.

Software solutions such as DataShine and Smart Steps provide new and exciting portals to aggregate and individual level population data, setting new standards in data availability and interaction. The use of such tools is not without challenge, however, often requiring dedicated computer architecture and accessed to privileged data. Further, there are increasing ethical considerations in the use of new and existing data that are being re-purposed beyond their original mandate.

2.2.5 Geodemographics and Population Data

Having discussed the creation and application of geodemographics classifications, it is important to remain mindful of the data employed in their construction. While classifications have increased their diversity in terms of data, there remains an underlying dependency on national statistical datasets such as the UK Census of Population. The data presented in the Census are summaries of individual level population data, aggregated to a pre-specified series of administrative geographies. The approach used in the collection of such data varies between countries, ranging from questionnaires distributed at fixed intervals as in the UK to ongoing Population Reg-

isters in India.

In the UK, the 2011 Census of Population was distributed based on a register of addresses rather than individuals. The register is used first in the distribution of questionnaires and second for the verification of questionnaire completion. Each household is required by law to complete the form for all residents of the household on the night of the census including guests. Once collected, the census forms are processed such that detailed individual-level data are generated. In effect, the result is an inventory of the population that contains specific attribution for the vast majority of the population.

The United Nations define a population register as a mechanism for the continuous recording of statistics for all members of a population (United Nations, 2001). In the United Kingdom, a specialised population register of all those individuals eligible to vote is maintained referred to as the electoral register. Historically, the full register was available for purchase, however, following a change in legislation in 2001 this is no longer the case. Rather, individuals have the opportunity to opt-out of inclusion in an edited version, known as the 'open register', which is made available for general sale. The rate at which individuals opt-out of inclusion in the open electoral register varies significantly across the UK. According to Call Credit, the national opt-out rate in 2013 was 40% down from a high of 46% in 2010. The highest opt-out rate was 80.25% in Blackpool in 2013. Before the introduction of the edited register, third parties could purchase the full register for any purposes such as advertising, address validation and consumer targeting. Following the withdrawal of the full register, various commercial entities, such as CACI Ltd, have sought to create alternatives consumer registers that combine the open registers with alternative sources of data such as from surveys.

A key motivation in the identification of new data is the uncertain long-term future of the Census. As part of cost-cutting measures, the 2021 census will be administered as an online questionnaire for all households with more regular surveys designed to improve annual reporting. Further details on the 2021 Census are available from https://www.ons.gov.uk/census.

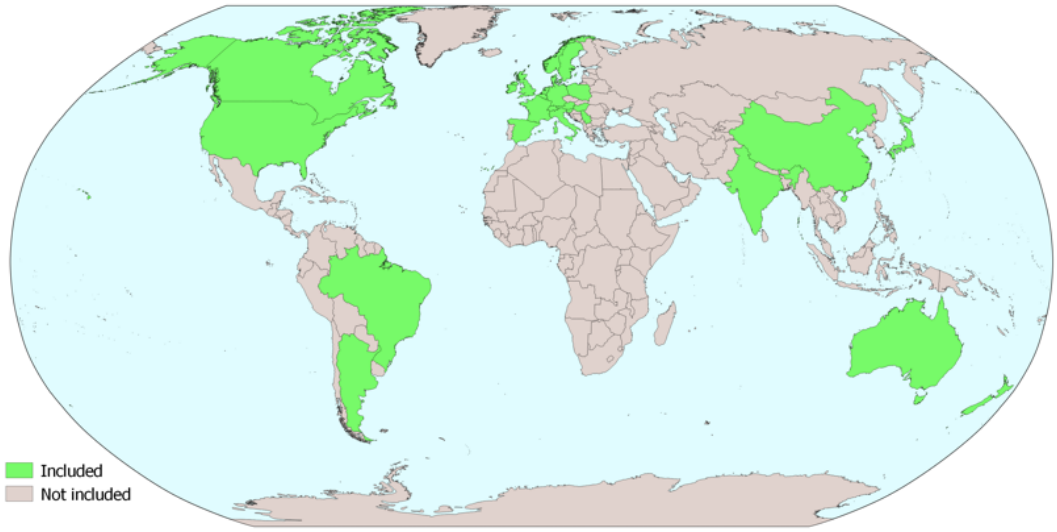


Figure 2.4: World map showing the coverage of the UCL Worldnames Database. Countries shaded in green are included.

Within academia, a growing interest in personal names and their association with demographics has led to various independent efforts to compile large population inventories that span multiple countries. One such example, and central to this thesis, is the UCL Worldnames Database (http://www.worldnames.publicprofiler.org); a composite population inventory for 26 countries, illustrated in Figure 2.4, drawn from the publicly available telephone directory and electoral roll datasets. Based on the summed populations of these countries, the database is representative of approximately 2 billion of the Earth's population. However, as may be evident from Figure 2.4, the coverage of the dataset is limited, failing to account for large regions including the African continent, Central America and large parts Central and Eastern Asia. Such omissions have the potential to cause significant bias in the completion of any global analyses.

2.2.6 Summary

In this section, the concept of geodemographics has been introduced along with a discussion of its strength, weaknesses and opportunities. The relationship between geodemographics, aggregate population data, and in turn individual-level population data has been established. In summary, geodemographics provides a useful means to describe and understand human populations, though is limited by the data used in

their construction. While efforts have been made to enhance the utility of conventional data, such as the introduction of Workplace zones, there remains an increased demand for new data that exhibit a higher degree of timeliness and specificity. That said, developing new forms of data is not without challenge. Often, due to concerns about privacy and disclosure, the level of key demographic attribution is poor. Thus, in seeking to establish a link between conventional and new forms of data, there is a need for a common point of reference. For this, it is proposed that the use of individuals' personal names, which, due to the association between naming conventions and key identity characteristics, may act as a means of linkage.

2.3 Identity

2.3.1 Introduction

Having discussed geodemographics, the focus of this review is shifted to identity. Identity may be considered as the unique combination of attributes of an individual that facilitate their distinction from others. Identity plays a fundamental role in the creation of geodemographic classifications due to its role in social categorisation which in turn forms the basis of individual level population data. Goss (1995) highlights three assumptions of linking identity and geodemographics. First, that individuals personal identities may be aggregated in such a manner as to facilitate the partitioning of the the population into stable homogeneous groups. Second, that individuals' social identities play an important role in their behaviour, and finally, that their location is a key factor in their social identity.

The two main perspectives on identity are Identity Theory from Sociology and Social Identity Theory from Psychology (Hogg et al., 1995). While the concepts have emerged in different disciplines, Hogg et al. (1995) note that the two theories are largely similar with Identity Theory being better suited to "dealing with chronic identities and with interpersonal social interactions" and Social Identity Theory being better suited to "exploring inter-group dimensions and in specifying the socio-cognitive generative details of identity dynamics."

Social identity theory is based on self-perception through their affiliation with

particular groups such as families, teams or ethnic groups (Tajfel and Turner, 1979). Fundamental to this is the concept of in-group and out-group; a reference to individuals who possess the same affiliations and those who do not. Tajfel and Turner (1979) suggest three phases in the definition of group affiliation of which the first two are most pertinent to this thesis. Social Categorisation is simply a stage of grouping in which we identify individuals who exhibit similar behaviours or characteristics to ourselves. Social Identification is the process by which individuals often assimilate the behaviour of the groups with which they are associated. For example, an individual born into a practising christian family may be more likely to adopt the religious beliefs of their family rather than an alternative belief system. Finally, Social Comparison is the process of comparing one's group association that of others and is associated with self-esteem, prejudice and rivalry. Tajfel and Turner (1979) highlights the importance of 'in group' and 'out group' in how we perceive others.

Grotevant (1992) considers identity as the structure out of which individuals interact with the world. This identity is a dynamic representation of an individual that is updated as the individual gains new experience and knowledge. Notably, the situation into which a child is born has a major impact on their social identification. The degree of control an individual has over the specific categories, or identities that they possess varies significantly. Grotevant (1992) proposes the division of these identities into those that are assigned and those that are chosen. Assigned Identities are those that an individual is born with such as age, gender and ethnicity. Chosen Identities are those for which an individual has some degree of control such as personal interests and political view. By its nature, Social Categorisation leads to a hierarchy of nested identities. For instance, while two individuals may both be considered as Christians, they may associate themselves with either being Catholic, Protestant or any of the other religious or cultural denominations. Hence, identity may be considered as a hierarchical structure in which the unique identities enable association with specific groups, while an individual's unique combination of identities makes them unique. This concept is concordant with that of Goss (1995) who noted that geodemographics assumes that individuals' identities may be aggregated to such a

point where they may be considered as stable groups.

An alternative perspective, proposed by Chandra (2012), splits identities into those that are nominal and those that are activated. Nominal Identities are those which an individual holds the appropriate characteristics, however, has not yet self-identified or been identified as possessing. Activated Identities are those which an individual has self-assigned or been assigned to. Chandra (2012) differentiates Activated identities into those that are chosen and those that are assigned. This definition should not be confused with the terms as defined by Grotevant (1992). Rather, in this instance, chosen identities are those that individuals' use for themselves and assigned identities are those used by others. It is noted that a person's chosen and assigned identities will often be independent of one another. Chandra (2012) states that for an individual to hold a particular identity, they must first possess the appropriate characteristics for inclusion. This view is concordant with that of Stone (1990) who considers the formation of each identity in two phases; identity announcement and identity placement. Identity Announcement occurs where the particular identity is created and Identity Placement where the identity is endorsed by others.

2.3.2 Online Identity

Moving from the observable to the virtual world, many of the typical cues associated with identity perception are no longer present. Not only may an individual's virtual identity differ from their true identity, but also they may hold multiple disjoint virtual identities entirely disassociated from their true self. However, while traditional identity cues are lost, Smith and Kollock (1999) identify new cues such as email addresses and signatures that may aid in Social Perception.

The literature surrounding online identity suggests that individuals either self-idealise (Manago et al., 2008) or extend their real identity (Zhao et al., 2008). Back et al. (2010) suggest two theories regarding such identity extension, the extended real life hypothesis and the idealised self hypothesis. In the extended real life hypothesis, it is assumed that individuals' online identities are an accurate representation of their true selves. This similarity may be due to anchors between the identities such as common connections or interests. Conversely, the idealised self hypothesis sug-

gests that individuals present only what is best about themselves. In a similar vein, a phenomenon associated with the online identity is the online disinhibition effect; a behavioural change in which some individuals' personalities shift, such as expressing a greater degree of emotion or sharing a greater volume of personal information (Suler, 2004).

On the whole, there is an apparent shift towards the extension of real identity. It might be argued that this shift is due to the increasingly intertwined nature of peoples' offline and online lives. For instance, users of Facebook and Twitter may act to moderate self-idealisation through the acceptance or rejection of other users' announced identities as noted by Stone (1990) who discussed the concept of identity formation.

2.3.3 Summary

In summary, identity may be considered as the fundamental structure upon which geodemographics analysis and data products are constructed. As a discipline, geodemographics attempts to segment individuals into distinct homogeneous groups based on the presence of one or more unifying social, cultural or physical characteristics. These groups are typically identified based on the analysis of aggregate population data which, as has been elaborated in the above, are simply summaries of the social identities held by a number of individuals. This is particularly pertinent given that the majority of individual-level population data are self-reported.

When considered from the perspective of social media, many of the conventional cues associated with identity perception are no longer present necessitating the inference of key identities based upon the limited data that are available. Consequently, it has been identified that individuals' personal names, which are often available, may be used to provide insight into individuals' identities due to their association with culture, ethnicity and language.

2.4 Personal Names

2.4.1 Introduction

Personal names are a fundamental component of individuals' identities providing a range of distinct functions. The name serves as an indicator of a person's place in socio-cultural space, position within a family, lineage, social status and often, regional or ethnic origin (Shagrir, 2003). The cultural, ethnic and linguistic conventions associated with the ascription of personal names means that the name may often act as an indicator of individuals' social identities (Jenkins, 2014). However, for such association to be exploited, a detailed understanding of the nuances of naming conventions and their relationship to identity must be held. Significant effort has been expended in the investigation of personal names and the inference of specific categories of identity. For instance, in terms of gender (McConnell-Ginet, 2003); ethnicity (Mateos et al., 2011); and the identification of regions (Longley et al., 2011).

In seeking to understand personal naming conventions, there is first a requirement to understand their structure and form. Rather than being fixed, the form of names is often contingent on cultural or legal conventions. Specific aspects of names structure include the unique name parts, the order in which the parts are recorded; the effects of gender; the inheritance of paternal and maternal family names; and the effect of marriage. Within the bounds of this thesis, the primary concern is individuals' given names and family names. However, it is valuable to understand their place within the name as a whole. From now on, family names will be referred to as surnames and given names as forenames.

Forenames act as a means to differentiate between individuals within communities and familial groups. Commonly assigned at birth, forenames are often based on either situation of birth, location, religious teachings or popular culture. For example, the most common forename in the UK, John, was one of Jesus' disciples in Christian teachings. Likewise, the name Mohammad, one of the most popular forenames in the world, originates with the Islamic Prophet. In the majority, forenames are associated with specific genders and thus may be considered a cue in social per-

ception. Beyond gender, forenames are often indicative of age due to the long-term trends in forename popularity. The association between forenames and social class is contentious and less well grounded. Willis et al. (1982) identify various literature supporting the concept that individuals with uncommon forenames are less successful, however, provide limited justification for this behaviour. One hypothesis was that individuals' names, and the connotations that the names hold, may impact upon individuals' self-perception and thus, their behaviours. The principle, in a less academic setting, is demonstrated by The Guardian newspaper which published a Venn diagram of the 30 most common forenames for Guardian staff, FTSE100 directors and prisoners. Illustrated in Figure 2.5, the Venn diagram coincides with some anecdotal stereotypes.

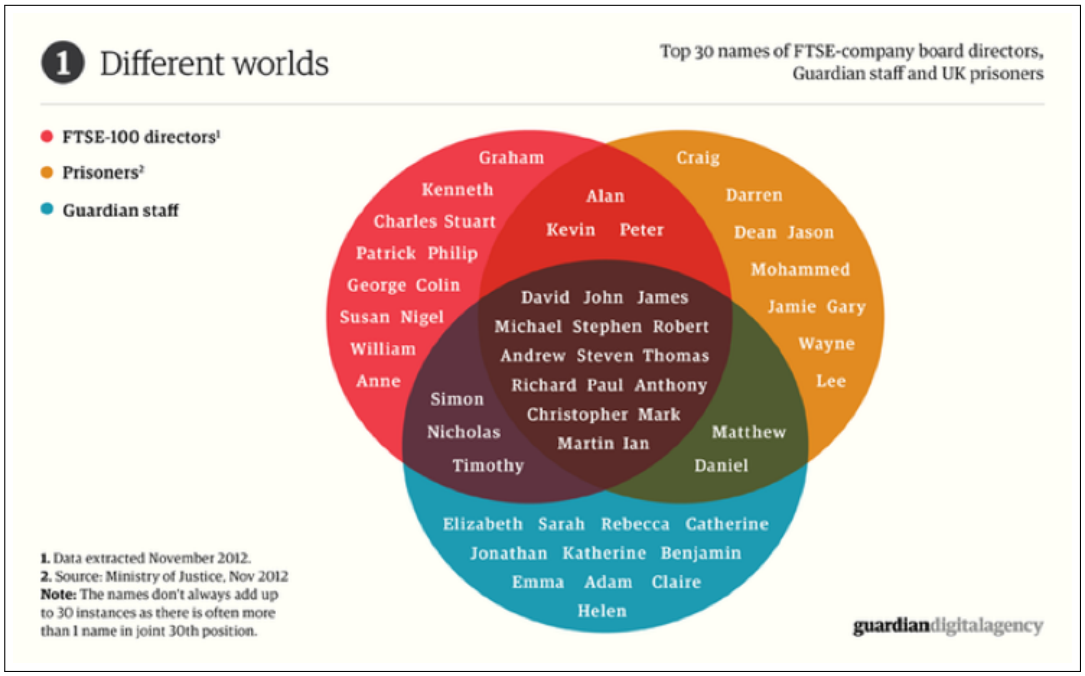


Figure 2.5: Venn diagram showing the 30 most common forenames for FTSE100 directors, The Guardian newspaper staff and prisoners (source: https://www.theguardian.com/news/datablog/gallery/2013/feb/11/whats-in-a-name).

Unlike forenames, surnames are a relatively recent phenomenon. Adopted in Europe during the medieval period, the time of adoption ranges from the 1940s in Turkey to almost 5,000 years ago in parts of China (Jobling, 2001). While surnames have emerged independently, in the majority, they may be categorised using

a straight-forward typology. Surnames typically fall into one of the following categories: toponyms, matronyms, patronyms, nicknames or occupations. Toponyms are names based on some form of location such as the surnames Cheshire and Longley or buildings such as in the case of Church. Nicknames are often based on a characteristic of an individual at the time the name was first coined such as Short or Loud. Occupational names are based on the occupation of the person at the time the name was adopted such as Smith or Fisher. Patronyms and matronyms are names inherited from the father or mother respectively. Patronymic surnames are evident in many cultures, notably the Arabic-speaking countries and Iceland. Arabic surnames are often affixed with bin- or bint translating as 'son of' or 'daughter of'.

In the majority of cultures, individuals' surnames are taken directly from the father. However, in others, such as Slavic and Hispanic, the nuances of inheritance are more complex. In the case of the Slavic countries, surnames are regularly appended with a gender specific identifiers. For instance, the son of Alexander Yordanov would have the same surname while the daughter would have the surname Yordanova. The gender specific affixes are 'ov' and 'ova' respectively. In the case of Hispanic surnames, individuals inherit both their mother's and father's surname with the father's surname preceding the mother's. For example, Cristina Borda Fortuny is the daughter of David Borda Garcia and Laura Fortuny Perez. In some situations, both surnames may be inherited from both parents resulting in a four-part surname. The inclusion of both maternal and paternal surnames provides the potential to better understand the genealogy of an individual as well as providing a means to identify those of mixed ethnicity.

In addition to forenames and surnames, many cultures use additional name parts before, between or after the forename and surname. For example, in Slavic countries, the middle name is patronymic based on the father's forename. Veronika Kamenova Yordanova from Bulgaria is the daughter of Kamen Hristov Yordanov. The order in which the above name parts are structured may be divided into those that conform to the western order where forename precedes the family name or eastern-order in which the surname precedes the forename. In the majority of cases, individuals

are referred to using the western convention though names are often recorded using the eastern convention such that family name may be used as a point of reference. Exceptions to this include China and Hungary in which the surname precedes the forename.

Lastly, it is important to bear in mind the significance of personal titles. For instance, Mr, Mrs and Miss. Where such titles are available they offer an additional level of insight into their bearer identities, beyond gender, titles may provide an indicator of academic or professional accomplishment (Doctor or Professor) or relationship status (Miss, Ms and Mrs).

2.4.2 The Analysis of Personal Names

As should now be clear, personal names offer a unique window into their bearers' social identities, offering the potential for insight into individuals' identities. Accordingly, the analysis of names may offer significant opportunities for investigating the construct and dynamic of social media data in which personal names are often available; a concept successfully demonstrated by (Longley et al., 2015) in the analysis of Twitter demographics in London.

2.4.2.1 Inferring age and gender from personal names

While surnames are commonly inherited, forenames tend to be reflective of popular culture (Suler, 2004). As such, individuals' forenames are often indicative of both age and gender. In the majority, forenames are associated with a single gender. However, an increasing number of names are now gender-neutral. Often, gender-neutral forenames are encountered where individuals have adopted abbreviations of traditional names, such as in the case of Alex or Sam. However, while the association between specific forenames and genders is well documented, the relationship to age is less well understood. In the case of both age and gender, one of the key challenges is the availability of suitably attributed and representative inventories of personal names. Illustrated in Figure 2.6, the FiveThirtyEight Blog provide an analysis of the association between specific forename and age distributions in the United States, based on data from the US Social Services Administration. The authors note that a

more pronounced distribution, such as in the case of Joshua, is far more informative than a flatter distribution such as in the case of Joseph. The FiveThirtyEight example uses inter-quartile range as a means to assign confidence to the age estimates.

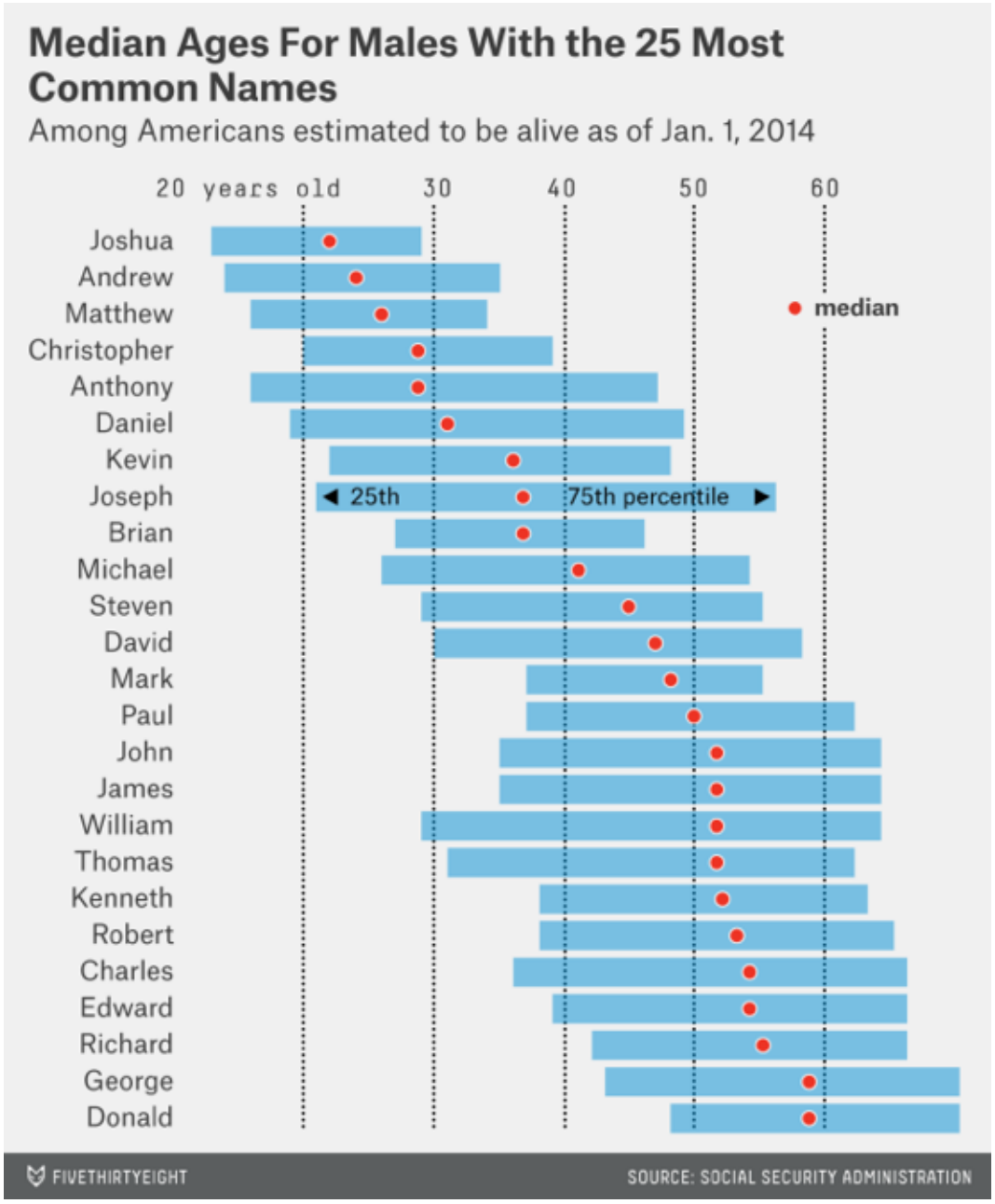


Figure 2.6: Graph from the FiveThirtyEight blog showing the inter-quartile range and median ages for the 25 most common forenames in the United States (Source: Silver and McCann, 2014).

Where such comprehensive data are not obtainable, alternative data may be sourced as a proxy. For instance, Monica, a data product produced by CACI Ltd,

uses frequencies of names by age and gender based on debit card holder records. An enhanced version of the Monica classification by Lansley and Longley (2016a) incorporates birth registration data such that individuals too young to possess debit cards are represented. Finally, the data are standardised such that the data correctly account for the known population. When using such classifications, it is important to remain aware of the inherent limitations assumed through the data choice. First and foremost, the classification is restricted to use within the same sampling frame as the data that were used in its creation. For instance, The FiveThirtyEight classification uses US Social Security Administration data and is thus only applicable to the United States of America. Similarly, the Monica classification is limited to application within the UK. Further, the enhanced Monica classification only provides a single gender option for each forename and thus, must be applied with caution where a name is considered unisex.

Gallagher and Chen (2008) use forenames as a means to improve the quality of facial recognition algorithms when differentiating between two named individuals. The application goes on to provide age and gender estimates that combine both image-based-estimates and the age and gender distribution sourced from the U.S. baby names database. In this analysis, the inclusion of forenames as a prior in the analysis was found to improve the quality of the classification significantly.

2.4.2.2 Surnames and genetics

One of the first proponents of name-based research was George Darwin, son of naturalist Charles Darwin, who employed surnames in investigating the link between marriage between first-cousins and physiological differences in offspring (Darwin, 1875). It was this early work by Darwin that was antecedent to various coefficients of relationship and studies of isonomy. Also, academic interest in surnames has emerged due to the association between specific surnames and distinct genetic markers. This association is due to the hereditary nature of both surnames and the Y-chromosome in the male line of the population (Jobling, 2001). During meiosis, particular portions the Y-chromosome do not recombine and thus are passed from father to son, consequently serving as distinct markers of co-ancestry (Jobling

and Tyler-Smith, 2003). In an investigation by Sykes and Irven (2000) it was found that of 48 individuals bearing the surname 'Skyes', 43.8% bore a specific haplotype suggesting a common ancestor. King and Jobling (2009) suggest some reasons why individuals bearing the same surname may not be genetically related. These include multiple originators of names; non-paternity events; adoption of male children; deliberate name changes; and genetic drift. In the context of identity inference, it should be noted that while two individuals bearing the same surname may not be genetically related, the high rates of endogamy - marriage to individuals within the same cultural group - means that surnames often remain associated with the same cultural groups (Mateos, 2007). However, Model and Fisher (2002) note that endogamy rates are not consistent across ethnic groups.

In the analysis of surnames, the frequency distribution of surnames is such that the majority of surnames may be considered uncommon (Sykes and Irven, 2000) and geographically localised (or regionalised). In practice, the less frequent a surname is, the more likely it is that those bearing the name will be genetically related. For instance, in the case of the surname Smith, is highly likely that there were multiple originators of the name. For instance, the most common UK surname, Smith, is an occupational name adopted initially by metal workers.

Before the work of Jobling and Tyler-Smith (2003), various research was conducted on the use of surnames as indicators of co-ancestry. Building on the original work of Darwin, Lasker proposed a measure of Isonomy – inbreeding – based on the co-occurrence of surname referred to as the 'Lasker kinship coefficient'. A criticism of the Lasker coefficient is its assumption that individuals bearing a surname are descendants of a common ancestor. However, in practice this assumption is flawed as it fails to account for differences in common naming practices. Case in point being the surname 'Smith' which is based on an individual's occupation and therfore had multiple distinct originators. The issue may be observed in the work of Sykes and Irven (2000) who showed just 43.8% of the individuals bearing the surnames Sykes shared a common ancestor. While not discussed in the literature, it may be possible to address this criticism through the filtering of surnames such that

only toponyms – names based on locations – are analysed. Such a process may remove the bias created through common surnames which are known to have multiple originators.

2.4.2.3 Inferring cultural, ethnic and linguistic groups from personal names

A key descriptor of any population is the structure and distribution of cultural, ethnic and linguistic (CEL) groups. This information is valuable in the provision of services, the study of segregation, and general observation of population. Such data are not, however, readily available to analysts at the individual-level due to their personal nature. Rather, the data are published in aggregate form as part of national publications such as the UK Census of Population. Seeking to address this absence, some studies have attempted to distinguish between specific ethnic and cultural groups based on individuals' surnames; for instance, in the identification of individuals of Chinese descent by Quan et al. (2006).

Mateos et al. (2011) sought to extend the specificity and reach of such classification techniques through the development of a Java-based classifier that processed forename-surname pairs against a dictionary of names and associated CEL groups. To create the classification, a bipartite naming network, in which forenames and surnames were represented as nodes and edges was weighted based on the frequency of specific forename-surname pairs was created. The network was then transformed such that the impact of common forenames and surnames in uncommon combinations (e.g. 'John Patel') would not inhibit community detection. Finally, transformed into two single-mode networks, communities were identified using the fast-community algorithm and labelled based on a dictionary of known name-ethnicity associations. Subsequently, the labelled clusters formed the base of the Onomap CEL classification tool. Based on the strongest result, the user is assigned a series of attributes about their individual CEL profile.

When using this method, and name-based inference techniques more generally, it is important to remain conscious of issues relating to acquired identity', the ecological fallacy and the potential for the introduction of other uncertainty. That is to

say that in most cultures, at the point of marriage, the female takes the surname of their partner. In this case, the methodology may assign an incorrect classification. Further, regarding the ecological fallacy, it should be recognised that all individuals of a specific name will not always possess the same CEL characteristics. The issue of uncertainty propagation must also be considered. Given the requirement to model key identities it is quite probable that a proportion of individuals will be incorrectly classified. Subsequently, where inference is made, there is a risk that conclusions are made based on incorrectly attributed data. To minimise the risk it is important that any derived analysis clearly states the limitations of the data and that these incorporated into any interpretation. One approach to reducing the effect of uncertainty is the use of aggregation. While not all individual will be accurately classified, when considered in aggregate form, it is likely that greater accuracy will be achieved. Given that direct validation is not feasible, it is also necessary that a common sense approach is used to help ascertain the validity of any outputs derived from individuals' personal names.

2.4.3 The Geography of Personal Names

By their origin, surnames often exhibit unique spatial patterns. The assumption is that the decedents of those with whom each surname was coined would have resided within a relatively limited geographic region. One of the earliest efforts to record the geographic origin of surnames was the compilation of a dictionary of surnames from the British Isles by Guppy (1890). Guppy recorded details for thousands of common surnames and recorded notes as to their probable origins.

An extension of the isonomy research discussed previously has been on the use of surnames in the identification of natural regions. Devoid of typical geographic constraints, surnames may be indicative of historical regions and migrations. Not only can such a method identify national regions, but also historical regions that may no longer be evident. The approaches broadly involve the creation of a distance matrix based on isonomy and then the subsequent partitioning of this data using clustering techniques. Holloway and Sofaer (1989) calculated the coefficient of isonomy between 12 regions of Scotland producing a regional similarity matrix. Histori-

cally, such studies have been constrained to limited samples of populations drawn from local parish records or telephone directories. However, through increases in the availability of comprehensive population records, and developments in computational power, it is now possible to perform similar analysis at a far greater spatial extent (Adnan et al., 2010). Expanding on existing work into region based surname analysis, Cheshire et al. (2011) were able to create a regional geography through the applications various clustering techniques to the regional distance matrix calculated from the inter-regional Lasker distance matrix. The success of the approach has been demonstrated across a range of countries and regions including Great Britain (Cheshire et al., 2010), Western Europe (Cheshire et al., 2011) and Japan (Cheshire et al., 2014).

2.4.4 Challenges in the Use of Personal Names

When making inference about an individual's identity based on their names, it is important to be aware that a name may have been changed due to marriage or outside circumstance. In the case of marriage, many women take the surname of their spouse (Goldin and Shim, 2004). Beyond marriage, some other circumstances may result in a change of name. For instance, the British Royal Family changed their surnames from Battenberg to Mountbatten due to negative German sentiment during the First World War. Failure to account for such changes has the potential to introduce error when making inference about an individual based on their name.

The conversion of text between scripts has the potential to introduce discontinuity. Such conversions are often in the form of translation, transliteration or transcription. Translation is the conversion of terms in terms of meaning. For instance, the Polish surname Kowal, the Italian surname Ferrari and the British surname Smith are all equivalent referring to a metalworker or blacksmith. Transcription is concerned with the conversion of the sound while transliteration is the literal character by character conversion. The approach used in the conversion of names between formats poses some challenges. For example, the German surname Müller may be represented as Müller, Mueller or Muller dependent on the approach used. The inconsistency in conversion approaches poses significant challenges regarding name

matching when collating names data from multiple sources.

2.5 Conclusions

While traditional forms of data, such as those collected via the Census, retain an important role, there is significant potential for new data in the description of populations. However, rather than replacing traditional forms of data, new forms of data offer the ability to add value through increases in specificity, improved representation of mobility and increased availability and timeliness.

In seeking to develop applications of social media to geodemographics, this chapter has sought to establish a relationship between the two concepts which is illustrated in Figure 1.1. A common theme to geodemographics and social media is the base unit of the individual. In geodemographics, it is the aggregation of individual-level data that are used in the construction of geodemographic classifications, while for social network data, the individual is the natural unit of reference. However, while a common point of reference has been identified, the data from social media often have limited demographic attribution. Thus, it is proposed that such variables may be modelled based on the novel analysis of personal names; the process of which has been outlined previously. Beyond social media, the ability to enrich new forms of data, such as those collected via social media offers the potential to address a range of the current critiques of geodemographics.

Chapter 3

Social Media and Geodemographics Applications

3.1 Introduction

In the previous chapter, the discussion centred on establishing a framework by which social media could be incorporated into the study of geodemographics. The motivation for this was a desire to address limitations in conventional geodemographics assumed through the use of traditional aggregate population data. Subsequently, it was found that New Forms of Data, such as those collected via online social media, could potentially be employed in the form of population data given the application of suitable identity inference techniques. In seeking to exploit said new data, the association between individuals' online identities and conventional individual level population data was established via the novel application of personal names. In turn, it was discussed how this could, through aggregation, be used in the creation of conventional aggregate population data. Thus, building on the previous material, the objective of this chapter is to compile a body of knowledge concerning social media, upon which potential applications to geodemographics and security may be drawn. The chapter will begin with an introduction to social media in its various forms and provide an overview of the key research themes.

3.2 Social Media

In seeking to identify potential applications of social media, it is first necessary to answer the question, 'what is social media?' In the literature, social media is considered a collective term employed to describe a range of web-based applications. These applications typically enable multiple individuals to interact in an online networked environment. When we consider this in the context of social medias' origins, the earliest incarnation of social media were *Bulletin Board Systems*. Such systems enabled users to log in, post statuses and send messages to other users. The services, commonly hosted on private servers, saw rapid decline following the launch of the World Wide Web in 1989 (Kaplan and Haenlein, 2010). More recently, the major innovations in social media have come as a consequence of increased access to high-speed Internet and the introduction of Web 2.0; a paradigm shift concerning how people understood and interacted with the Internet. Key features of Web 2.0 were a change from thick to thin client architectures and a greater emphasis on the user rather than publisher generated content (O'Reilly, 2007). This transition fis reflected in the definition of Kaplan and Haenlein (2010) who define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content."

Since their inception, the number of social media platforms has grown rapidly with a diverse range of specialised and general-purpose platforms now readily available. These platforms range from the inherently social Facebook to the more professionally aligned LinkedIn and video orientated YouTube. Such are the niche offerings of each platform that many individuals choose to adopt a plethora of complimentary, and often interlinked services. The GlobalWebIndex reports that the average Internet user has 5.54 social media accounts of which 2.82 are considered to be active[1] (GlobalWebIndex, 2015). Figure 3.1, provides a breakdown of the GlobalWebIndex data highlighting some distinct features of individuals' network adoption behaviour. First, it is evident that the number of social media accounts is a function

[1] An active user is one who has accessed the service at least once in the preceding month.

of age with older people having few accounts on average. Second, the number of active accounts per user is highest in the 25-34 age bracket. Interestingly, it is this cohort who would have been most technologically active during the time in which the major social networks came into popular use.

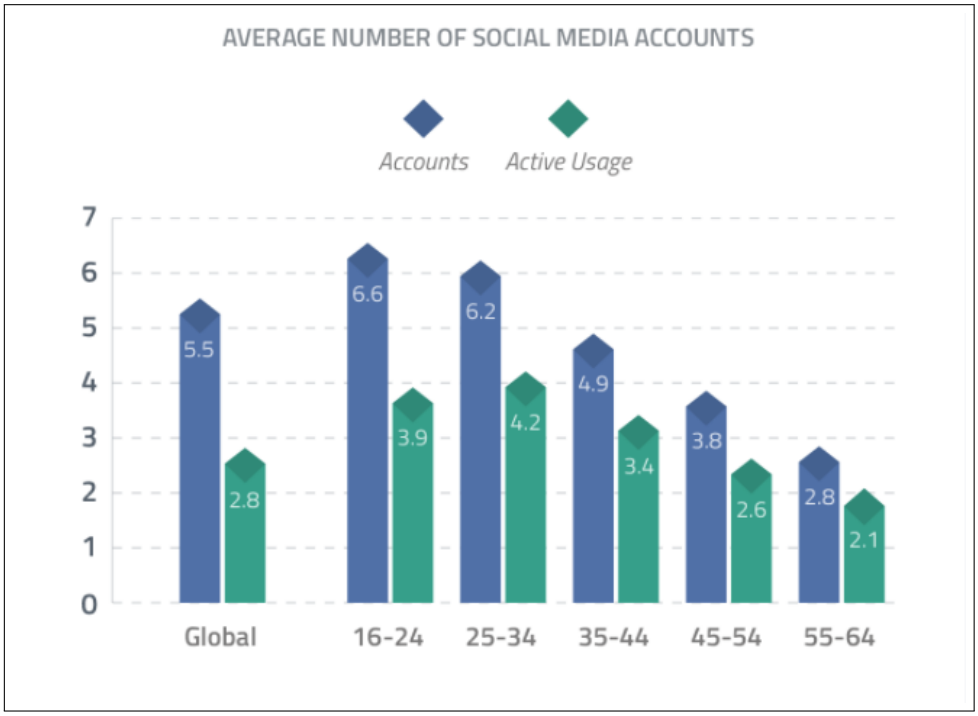


Figure 3.1: Bar graph showing the average number of social media accounts per Internet user broken down by age and active status (GlobalWebIndex, 2015).

In seeking to differentiate between the various social media platforms, Kaplan and Haenlein (2010) proposed a six group classification: Collaborative Projects, Blogs, Social Networking Sites, Content Communities, Virtual Social Worlds and Virtual Gaming Worlds.

- *Collaborative projects*, such a Wikipedia and OpenStreetMap, facilitate the creation and maintenance of various content by communities. In the case of the platforms mentioned above, individuals can create, edit and curate content in line with each project's objectives. Using the example of Wikipedia, as users add new content, changes are logged and reported such that the community may make further additions, retractions or alterations. Key research themes include individuals' motivations to contribute and the verification and

validation of user-generated content.

- *Blogs* are one of the earliest examples of social media and are primarily concerned with the publication of short time-stamped articles that are presented in reverse chronological order. Blogs range in content from those expressing a general opinion on a range of subjects to those that provide more niche material.

- *Social Networking Sites*, such as Facebook, Twitter and Linkedin, are web-based platforms that enable the creation of public profiles, the curation of contacts and communication via messaging.

  The world's largest social network, Facebook, launched in 2004 as a private network and rapidly grew in popularity following its global launch in 2006. As of September 2015, Facebook has an estimated 1.55 billion active users (Facebook, 2015). Arguably a 'Jack of all trades and master of none', Facebook offers a significant range of capabilities ranging from group organisation to event management and photo sharing. A large proportion of Facebook's growth, as is the case with many large social networks, may be attributed to the acquisition of other platforms. For example, Facebook's purchases of Instagram and WhatsApp.

  In contrast to Facebook, LinkedIn targets professional users who are seeking to build professional as opposed to social networks (Papacharissi, 2009). LinkedIn claims 400 million registers users, however, only 24% of these are understood to be active. In much the same way that Facebook identifies potential contacts based on shared interests and mutual connections, LinkedIn suggests connections based on 2^{nd} and 3^{rd} degree relationships and shared skills. Further, the impetus is placed on the individuals' job titles, education and employment as opposed to friends, relationships or photos; in essence creating an alternative professionally aligned virtual identity.

  Twitter, a minimalistic social network, is a micro-blogging platform facilitating the publication of short 140 character messages (Kwak et al., 2010). The Twitter ecosystem revolves around users following those that are of interest

to themselves and vice versa. With over 320 million active users, submitting more than 500 million messages per day, Twitter is one of the largest social networks in the world (Twitter, 2016). A key feature of Twitter is the ability to follow geographically bounded trending topics. Using the location information provided by client devices, Twitter presents the topics, or Trends, which are most frequently discussed. A unique characteristic of Twitter is the non-directional network structure between users. Where the majority of OSNs require bilateral approval for connections to be formed, a Twitter user may follow whomever and whatever they want.

- *Content Communities*, such as YouTube and Flickr, are concerned with sharing individuals' generated content in a community environment. Unlike online social networks, the onus is on the content, though it is often the user that is employed as the point of reference. Increasingly social networking platforms are seeking to absorb such services as is the case with Facebook's video sharing endeavour.

- *Virtual Gaming Worlds* and *Virtual Social Worlds* may be considered as abstractions of reality in which individuals can develop comprehensive online identities. Depending on the platform, the degree of similarity to the observable world varies widely. Of the virtual worlds, Second Life by Linden Labs is arguably the most well known and studied. Second life is a highly complex virtual world in which users create and interact via avatars. Second Life is highly immersive with complex communities, economy and social interactions (Boulos et al., 2007).

From the above categorisation, it is evident that the direction of this thesis is most closely aligned with that of Online Social Networking. Unlike the alternatives, social networks are primarily concerned with the extension of individuals identities into the virtual world and thus, are most likely to be illustrative of individuals' behaviours. In some senses, the data that are generated by online social networks may be considered as the digital exhaust from individuals' routine online activities

(Lupton, 2013). This digital exhaust, when suitably attributed, offers a means to reconstruct the behaviours of individuals and groups in a previously unprecedented manner, not feasible with conventional aggregate population data. It is worth noting that while the collection and analysis of social media data is a relatively recent phenomenon, the analysis of so-called digital exhaust has been a feature throughout the history of computing. A simple example being the collection of system logs to monitor and understand systems' performance. The system log records all events and specific attributes such as date, time and process descriptions. Such logs are not overly dissimilar to social media data where individual activities are recorded alongside date, time, location and user information.

A relatively recent development in online social media has been the incorporation of location. Considered under the umbrella of Location Based Services (LBS), these are "any service that takes into account the geographic location of an entity" (Junglas and Watson, 2008). The degree to which location plays a role in each social media platform varies considerably. For example, Foursquare, one of the earliest platforms to integrate location permits users to 'Check-In' to locations in return for points and badges. Alternatively, Twitter uses location as a means to geographically reference Trending Topics and personalise searches. Likewise, Flickr, a platform for storing and sharing photos, employs location as a tool to add value to its content by allowing users to record the location of their pictures and also to conduct searches based on location. In each case, location is a fundamental component of the products in their present forms.

Unsurprisingly, the growth of LBS has coincided with an increase in smartphone ownership which has allowed a greater proportion of the population to determine their locations accurately and access mobile data services. In 2015, Ofcom, the UK regulator for communications, reported that 66% of all UK adults possessed a smartphone with this figure at 90% in the 16-24 age bracket (Ofcom, 2015). An additional factor influencing the uptake of LBS has been the open development nature of the major mobile phone operations systems: Apple's IOS and Google's Android. Such openness has allowed developers to exploit devices inbuilt capabilities

resulting in a plethora of location aware applications and social media (Li and Chen, 2009).

However, while the growth of LBS and geosocial media is a testament to their success, the ease in which individuals may inadvertently disclose their precise location has led to a growing discourse on the topic of location-based privacy. For example, compromising oneself or others through the inadvertent disclosure of location. Seeking to understand this issue better, Vicente et al. (2011) identify four key aspects of privacy associated with location sharing:

- *Location Privacy* relates to sharing the location of others without their explicit consent. Users may inadvertently share their location at a time that is undesirable. An example of such behaviour is the tagging[2] of individuals at a location or event.

- *Absence Privacy* occurs where sharing an individual's location highlights their absence from a specific location. For example, being tagged at a theme-park when you are supposedly working from home.

- *Co-location Privacy* refers to the ability to establish whether two or more individuals are present at the same place based on shared location information. Again, such an action may be a consequence of being tagged.

- *Identity Privacy* is concerned with the protection and disclosure of identity through information sharing. For example, sharing details such as an address or telephone number.

3.3 Social Network Analysis

Almost ironically, it is the same set of factors considered as privacy issues by Vicente et al. (2011) that make Social Network data so attractive to researchers. Within academia, there has been significant engagement with social media due to the ease in which such insightful data may be gathered. Applications have been pursued in a range of fields ranging from geography (Jurdak et al., 2015) and health (Hawn,

[2]Tagging is the process of linking an individual's virtual identity to a particular event or location.

2009) to security (Briggs and Baker, 2012) and crime (Procter et al., 2013). In the majority of cases, these analyses have been performed on large collections of data harvested via Application Programming Interfaces (API) or 'scraped' from the web. Ellison et al. (2007) attribute this rise in popularity to the "affordance and reach" of the data that are now available. A key advantage of social media data over alternative new forms of data is its ongoing accessibility; a quality often lacking with new data, which prevents replication, extension and the implementation of research outputs. It is these features in particular which push social media data to the forefront in regards to the analysis of new forms of data for demographic research.

In seeking to differentiate between the various aspects of social media analysis Kietzmann et al. (2011) identify seven guiding themes. Illustrated in Figure 3.2, these themes are conversations, groups, identity, presence, relationships, reputation and sharing.

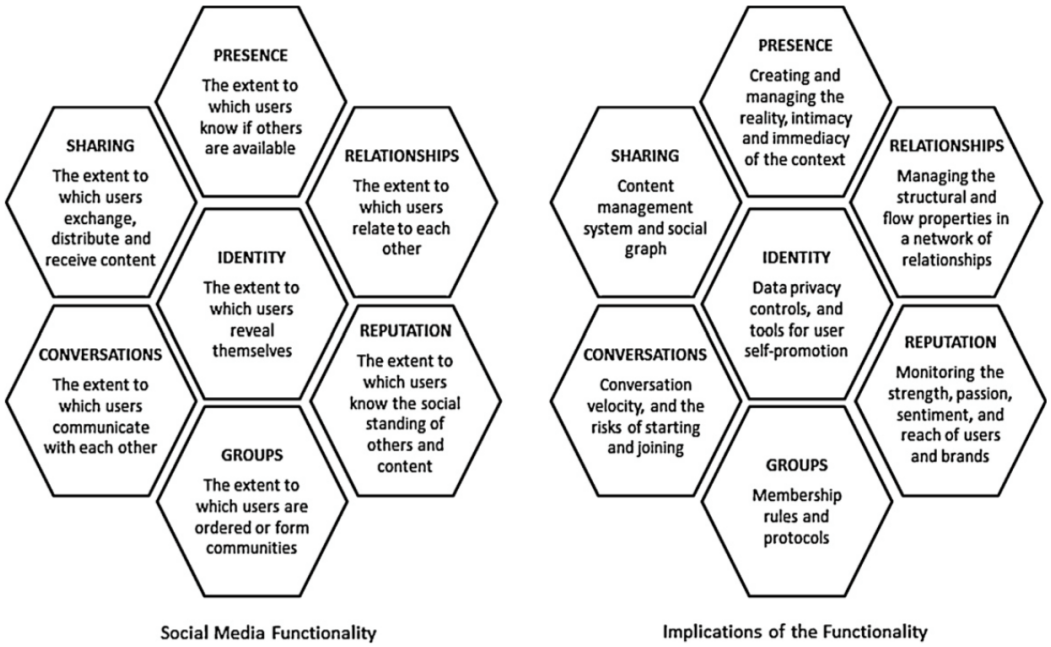


Figure 3.2: Illustration of social media analysis themes highlighting the function and implications of each (source: Kietzmann et al., 2011).

- *Conversations* are focused on the verbal interaction between users.
- *Groups* are focused on the formation and behaviour of groups.

- *Identity* is concerned with the extension of individuals' identities onto the web.
- *Presence* is concerned with how users infer the status of others.
- *Relationships* are concerned with how individuals and groups are connected.
- *Reputation* is concerned with the degree of influence individuals possess.
- *Sharing*: is concerned with the spread of information and misinformation.

3.3.1 Data

Increasingly, online services are capitalising on the commercial value of their digital assets through the provision of their data and tools to third-parties. Such functionality is delivered through the use of APIs which provide a structured means to integrate and utilise the services' data and tools. In the case of Twitter, the API provides access to a range of functions such as user searches, content streaming and message submission. Each function is referred to as an endpoint. Within the bounds of this thesis, we are primarily concerned with the Public Streaming API, which provides access to a portion of the total throughput of Twitter at any given time. The streaming APIs include the 'POST statuses/sample', the 'POST statuses/filter' and the 'POST statuses/firehose' endpoints. The following list provides a brief overview of each.

- The *POST statuses/sample* is a random sample of all public Tweets being submitted. The sample stream is 'Rate limited' to 1% of all Tweets.
  (see: dev.twitter.com/streaming/reference/get/statuses/sample)
- The *POST statuses/filter* allows third-parties to stream a sample of all public Tweets based on some content-based parameters. The filtered stream is also 'Rate Limited' such that the maximum number of Tweets available are equivalent to 1% of all Tweets.
  (see: dev.twitter.com/streaming/reference/post/statuses/filter)
- The *POST statuses/firehose* returns all public Tweets. Twitter notes that few applications will require such a volume of data.
  (see: dev.twitter.com/streaming/firehose)

The Public APIs are provided in a prescribed format in which rate-limiting is imposed to prevent excessive queries. For example, the *GET users/lookup* endpoint is limited to 180 requests per 15 minutes window. Where a user requires enhanced access to Twitter's data, such as in the case of the firehose or historical data, this may be achieved via a third-party distributor. In the case of Twitter, the data are made available via Gnip (see: http://www.gnip.com), a data warehouse that provides a bespoke API. Alongside Twitter, Gnip provides access to Foursquare and number of other major social platforms. It should be noted that Gnip was acquired by Twitter in April 2014.

3.3.2 Applications and Methods

While the applications of social media are broad, many of the commonly applied analytical techniques are applicable across the board. As such, and in line with this thesis' direction, the following discussion will focus predominately on Twitter though drawing parallels with other services where appropriate. Unlike the majority of social networks, Twitter provides quick and comprehensive access to their data. The ease in which large collections of data may be gathered has led many academic studies to opt for Twitter resulting in a large body of literature across a broad range of topics.

Figure 3.3: World map of Facebook friend connections based on a sample of 10 million friend pairs. (source: https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919)

Regarding demographics and security, various social media data have been employed in the observation and modelling of human behaviour. One such example is the Facebook friends graph illustrated in Figure 3.3. Constructed based on a sample of Facebook users, the graph is composed of nodes representing individuals and edges representing connections. Analysis of the graph provides information on social structure, influence and the potential paths by which information may be shared.

3.3.2.1 Migration and mobility

The ubiquitous nature of social media and the ease with which such data may be collected has facilitated various studies investigating human migration and mobility. Using collections of social media data that contain both temporal and spatial attribution, it is possible to infer the activity patterns of specific individuals. Such analysis is often described as trajectory analysis. At the local scale, Adnan et al. (2014) investigated the use of temporal profiles derived from tweeting activity to enquire into the daily and annual rhythm of the major world cities. Interestingly, these analyses were able to identify the impact of the main religious events such as Ramadan. Not only are the data an effective representation of city-wide activity patterns, but they are also unconstrained by conventional geographic boundaries that are restrictive to traditional analysis. An issue which is particularly important in the case of borderless phenomena such as migration.

At a global scale, Hawelka et al. (2014) investigated the use of Tweets in the inferences of global mobility patterns. Using a collection of circa 1 billion Tweets, they identified users' probable countries of origin and subsequently inferred global mobility based on individuals users' tweeting histories. Country of residence was determined based on where each user had tweeted most frequently within the preceding year. A further contribution made in the paper was the identification of functional regions using a network-based community detection algorithms. Hawelka et al. (2014) make reference to differences in how representative Twitter data are through the removal of data for those countries with fewer than 500 resident users. This approach constrains the analysis to just thirty countries.

3.3.2.2 Health

Various studies have explored the use of social media in the analysis of global health patterns. In 2008, Google sought to monitor the spread of influenza in real-time based on the analysis of specific search term frequencies (Ginsberg et al., 2009). Specific queries, which corresponded with historical influenza data were identified and subsequently employed in monitoring activities. Numerous criticisms of the method have emerged since publication. However, this has not deterred others from attempting to implement the approach using social media data such as from Twitter. Achrekar et al. (2011) demonstrated a high degree of correlation between Flu-related Tweets and reported Flu-data. Achrekar et al. (2011) subsequently used the volume of Tweets to improve upon existing models rather than use Twitter data in isolation.

Of note, and relevant to the above, Paul and Dredze (2011) note that while Twitter may be suitable for monitoring some aspects of public health, it will not always provide a valid conclusion. In their analysis, it was found that there were insufficient Tweets for analysis to be significant, and in the case of more localised conditions, conversations by non-affected individuals had the potential to introduce significant bias.

3.3.2.3 Crime and security

In June 2014, it was reported that over half of calls passed to frontline police in the UK were related to online social networks (BBC, 2014a). However, due to the categorisation employed in the recording of crime data, accurate statistics were not available as social media crimes are recorded under the same categorisation as traditional crimes. Ellison et al. (2007) identify some specific risks to social media users related to the sharing of personal information including identity theft, stalking and bullying.

In real terms, social media has formed the centre of several high-profile crimes. Notably, the case of the Robin Hood Airport 'Twitter joke trial' in 2010 and the McAlpine Libel case in 2012. In January 2010, Paul Chambers, a passenger due to fly from Robin Hood International Airport, sent a 'joke' tweet threatening to blow up the airport if his flight was cancelled. Chambers was subsequently arrested and

charged with sending a menacing communication under the Communications Act 2003. While the conviction was overturned following its second High Court appeal, the trial provided a tangible demonstration of the link between actions by the real and virtual self (Kelsey and Bennett, 2014). The McAlpine Twitter libel case followed allegations by the BBC that a senior Conservative Party Member of Parliament had been involved in historical sex abuse. At the time, a large number of social media users falsely implicated Lord McAlpine with notable individuals including Sally Bercow, wife of the then House of Commons speaker. While the majority of minor users were let off with a request to donate £25 to the Children in Need charity, the prosecution successfully pursued the libel case against Bercow receiving an undisclosed settlement within the courts. In both cases, it was highlighted that the general public had a limited understanding as to what constituted a crime on social media; a problem which is ongoing.

In contrast to the use of social media as a facilitator of crime, police in the UK and other nations are increasingly integrating social media as a means for research, communication and outreach (Crump, 2011). While the use of social media by UK Police initiated in 2008, the use of such services only came to prominence during the 2011 London riots. Following the events, significant research was conducted using the social media data generated in an attempt to reconstruct the events. Meanwhile, social media was highlighted as a critical weapon in the rioters' arsenal, facilitating coordination and orchestration between disparate rioting groups. The role of social media in such high-profile events has raised questions in government regarding the suppression of certain online media during times of civic unrest (Casilli and Tubaro, 2011).

Social media is also playing an increasingly prominent role in security. In this context, security may be considered as "the state of being free from danger or threat" (Oxford University Press, 2010). SOCMINT or Social Media Intelligence was first mentioned following the 2011 London Riots (Omand et al., 2012). Following the shooting of Mark Duggan in August 2011, social media activity rapidly increased and retrospective analysis has suggested aspects of the riots could have been better

tackled using such data. However, at the time, the Police acknowledged that they did not have the physical, or personal capacity to gather intelligence through social media (Omand et al., 2012).

At a more extreme level, several nations have temporarily blocked access to online social media as a means to suppress activism. One of the most recent example being Turkey (Genç, 2014). Online social media is permanently blocked or heavily restricted in some nations, e.g., North Korea. Ironically, bans on social media often lead to users finding alternative means to circumvent the blocking technologies.

During the 2010 Egyptian uprising, the following Tweet by the 'rebels' was widely circulated.

> "We use Facebook to schedule the protests, Twitter to coordinate and YouTube to tell the world" (Khondker, 2011).

While the author remains anonymous, the statement echoes the behaviour of many of the countries involved in the Arab Spring uprisings. In this instance, social media enabled the 'rebels' to broadcast their messages to the world; in effect, increasing global awareness. Significant discourse in the media surrounds the use of OSNs by terrorist organisations such as the Taliban and Islamic State. Previously, many of these organisations have used Facebook, Twitter and YouTube to share videos and news regarding their activities. In some case, this has resulted in the individual sites blacklisting users sharing such content. However, this approach has led to displacement rather than reduction; a phenomenon also witnessed in crime prevention (Cornish and Clarke, 1987). Case in point is the shift of Islamic State to social networks such as Diaspora (see: https://www.diasporafoundation.org); an online social network with a decentralised control structure. This feature allowed Islamic State to shift their online presence to the platform without the censorship applied by the mainstream social networks (Diaspora, 2016). In a separate incident, a Taliban commander believed to be in Afghanistan sent a geotagged Tweet from within Pakistan (BBC, 2014b). While the addition of location was dismissed as an "enemy plot", the fact that the event occurred highlighted the potential for social network analysis in monitoring security concerns.

3.3.3 Challenges and Limitations

While infrequently discussed, social network analysis is susceptible to various challenges and limitations. First and foremost, that the individuals who use social networks are not necessarily representative of the population. Rather, they are a self-selecting group which have opted to register and utilise the platform being studied. Consequently, any conclusions from the data are limited in applicability to the population from which the samples were drawn. The issue is compounded by the fact that limited information is publicly available regarding the demographic breakdown of each social platform. For instance, the geographic data that are available via the Twitter API are from a self-selecting subset of an already self-selected group of the population. In other words, the data are that are available are from those individuals who have chosen to use Twitter and, in turn, decided to share their location. Consequently, the results may not be generalised to the population as a whole. Seeking to understand this issue better, (Longley et al., 2015) attempted to quantify how representative the Twitter users were versus the underlying population. The analysis was conducted across the Greater London area and employed personal names as a means to infer Twitter users' ages, genders and ethnicities. In turn, theses data were examined against data from the 2011 Census of Population.

Further challenges encountered in the analysis of social media are the impact of fake accounts and the tourist/constrained study area effect. Fake users, or bots, are present in most forms of social media. On Twitter, bots are often employed to deliver marketing and will regularly follow users and send Tweets. That said, such accounts are not necessarily malicious. For example, the @DearAssistant Twitter account uses Wolfram Alpha to answer questions such as "How many days until Christmas?" Nonetheless, when analysing social media, it is advisable that such accounts are omitted to limit the potential for bias. The omission may be accomplished through the use of blacklisting or filtering based on some form of a textual or spatial parameter.

The second issue, referred to here as the tourist effect, is associated with how social media data are assembled. When streaming Twitter data, it is normal practice

to constrain the sampling frame to a limited geographic extent. The motivation being to minimise the volume of data collected and increase ease of data handling and manipulation. While this is advantageous in some senses, it can result in analysis failing to capture the true behaviour of the study participants. For example, when analysing data recorded in London, a proportion of the users identified will not be resident in the study area, let alone the country.

A third issue is that of multiple locations associated with each user. Lacking a single point of references impacts how easily the data may be referenced against conventional aggregate and individual level population data in which each observation is linked to a single location. For example, comparison against traditional population data is challenging as users are liable to have multiple locations associated with their accounts. Of these data points, there are limited means to identify which, if any, of the locations, coincide with the user's actual place of residence.

3.4 Ethics

While the potential of social media data is clearly evident, limited consideration of these factors appears obvious. In a review of 380 studies citing Twitter, Zimmer and Proferes (2014) found that only 4% considered the ethical implications of the data being analysed and how it was collected. Of the sixteen papers to acknowledge ethical issues, six mentioned ethical approval; five acknowledged the ethical questions, and five stated that no ethical issues were present. The justification typically used to is that the data are in the public domain and also that users have consented to their data being used as part of the registration process. For example, the Twitter Terms of Service (TOS) state that the 'data may be used for many purposes beyond users' direct experiences'. In many cases, this results in users being unaware that their data are being used in research or marketing.

In the case of the major social media platforms, data remains the property of the submitter. In the case of Twitter, the TOS state that 'You retain your rights to any content you submit, post or display on or through the Service'. However, by agreeing to the TOS you are giving the right to Twitter to 'use, copy, reproduce,

process, adapt, modify, publish, transmit, display and distribute such content in any and all media or distribution methods (now known or later developed).' In essence, while the user retains ownership of their data, Twitter has full rights to its use and distribution. In effect, the collections of data held by academics are being used under licence and may no way be considered property. To administrate this, Twitter required developers to possess authentication tokens for access to the API. Available to developers, the tokens provides a means to manage rate-limits and monitor requests. In certain scenarios, such limits have proven restrictive leading to individuals circumventing the conventional techniques and scraping the desired content. Not only does this approach breach most platforms TOS, but it may also amount to theft as the data, as has been previously established, remains the users' property.

One of the main challenges in the analysis of Twitter is the restrictions placed on the collection of data. As has been noted, the free API limits the user to real-time collection making the compilation of historical data potentially expensive and time-consuming. Further, Twitter prohibits developers from sharing data which has been collected via the API. It is likely that such a restriction is designed to prevent competition and maintain the value of the companies data assets. Twitter's 2014 Q4 financial statement recorded data licensing income as $70 million, roughly 10% of its $710 million quarterly revenue.

A further issue in the analysis and collection of social media data is that of consent. The term consent is used in reference to the participants of a study giving their permission for their data to be used in the analysis. The issue of consent received significant public scrutiny following the revelation that Facebook, in partnership with Cornell University, conducted a massive psychological experiment without gaining informed consent. The study altered the volume of positive or negative posts for 689,003 individuals to determine if emotional contagion occurred online (Kramer et al., 2014).

Consent is considered as either being informed or assumed. Informed consent regards the process of gaining full permissions from study participants having provided appropriate and accessible information about how and for what their data will

be used (Crow et al., 2006). Assumed consent makes the assumption that individuals have consented to the use of their data based on having agreed to the original, TOS. Assumed consent is generally used in the analysis of social media data where it would be challenging or impossible to gain the consent of all study participants. One source of guidance on the ethical use of OSN data is published by ESOMAR, the 'World association for market, social and opinion research.' ESOMAR provide guidelines that state where meaningful consent is not available, or may cause inconvenience to the user, that the research may progress, however, should report only depersonalised information (ESOMAR, 2011). Depersonalised information being data in which it is impossible to identify the individual from the data that are published. This approach is further encouraged where data are either not depersonalised at the point of collection or where minors may inadvertently be included in the collected data; a feature which is relevant to Twitter which has an unenforced lower age limit of thirteen.

A further concern in the analysis of social network data is that of individuals' privacy. In particular, the disclosure of Personally Identifiable Information (PII). PII are attributes that pertain to individuals' identities such as age, name, gender and ethnicity. While the ESOMAR guidelines make some recommendations for the removal of PII they do not address the spatial data content. With the accuracy of GPS-enabled devices regularly +/- 10 metres, it is entirely possible to determine the specific locations and activity patterns of individual users.

One area in which the issue of privacy has been explored is the reporting and publication of crime data by the UK Police. The UK Police Data Service (see: http://www.data.police.uk), which provides monthly records of crime, use a series of steps to anonymise the individual, event type, and location. In particular, the service aggregates all location information to an area, which includes six or more postcode units. The final data are assigned a point location over either a road or public building. While thorough, aspects of the approach have been criticised for introducing the potential for interpretation bias with a large number of crime events appearing to emanate from a single point (Chainey and Tompson, 2012). Alternative

methods of spatial anonymisation may be seen in the publication of the UK Census. In this case, individual data are aggregated to output area level: the smallest census unit. The average output area population being 309 individuals as of the 2011 Census. The use of statistical geographies enables the linkage of aggregate data to high-quality national statistics for assessment and comparison.

3.5 Conclusions

The aim in writing this chapter was to establish a body of knowledge regarding social media from which potential applications to geodemographics could be drawn. On investigation, it was found that the form of social media most relevant to geodemographics was Online Social Networks. As was discussed, there exist a broad range of online social network, each which possess unique and potentially valuable characteristics for the observation of human populations. Facebook was recognised for the wealth of semi-structured personal data, however, due to limitations in the ability to harvest large volume of data was deemed unfit. Twitter, was chosen on the basis that it offered the best balance between accessibility and expression of identity. Further, various investigations have sought to model human behaviour using Twitter data with varying degrees of success. However, beyond token reference, have failed to acknowledge the limitations of such analysis. In particular, the self-selecting nature of the Twitter sample versus the underlying population.

While the derivatives of Social Media analysis will never share the provenance of conventional studies, the opportunities which the data present are undeniable. Being able to establish individuals' identities has significant implications for the exploitation of social network data as, identified by (Kietzmann et al., 2011), identity is central to the analysis of social networks. Further, we are no longer constrained to single countries or regions as is the case with traditional aggregate population data. Thus, given proper consideration, the analysis of Twitter data offers a means by which we may empirically examine stocks and flows of population. Not only concerning static representations but also at a dynamic scale in both time and space.

Being a global platform, Twitter has the potential to enable analysis of a range

of spatial scales from the local to the global. Further, as the users of Twitter often use their personal names, there is the potential to model, and by effect differentiate by demographic types. Thus, based on the above, there is clear evidence to support the incorporation of Social Media into the study of geodemographics and security.

Chapter 4

Database Creation, Linkage and Validation

4.1 Introduction

Having identified Twitter as a potential resource for the production of individual level population inventories there remains the issue of transforming the data from their raw unprocessed form into functional invidual level population inventories. Regarding the structure of this analysis the framework of Mitchell (2005), recommended by De Smith et al. (2011), is employed. The framework is composed of seven stages: framing the question; understanding the data; choosing a method; calculating the statistics; interpreting the statistics; testing the significance of the statistic; and examining the results. Thus, seeking to frame the question, the first requirement is to understand the form and function of a population register entirely. Concerning answering this question, the United Nations provide a useful definition which is as follows:

> "...a mechanism for the continuous recording of selected information pertaining to each member of the resident population of a country or area, making it possible to determine up-to-date information about the size and characteristics of the population at selected points in time. Because of the nature of a population register, its organisation, as well as its operation, should have a legal basis. Population registers start

> with a base consisting of an inventory of the inhabitants of an area and their characteristics, such as date of birth, sex, marital status, place of birth, place of residence, citizenship and language. To assist in locating a record for a particular person, household or family in a population register, an identification number is provided for each entity". (United Nations, 2001)

Based on the United Nations' definition we may draw several key criteria in regards to the desirable form of the proposed inventories. First and foremost, a baseline of all inhabitants should be established. These data should, in turn, be supplemented with essential identity characteristics including as age, gender and place of residence. In addition, these data should be assigned some form of identification number such that it may be possible to differentiate between individuals.

Considering the above criteria in the context of creating the Twitter-derived population registers several potential challenges are evident. First, the UN definition considers a population register as being inclusive of the entire resident population of a country or area. Such coverage will not be achievable with the Twitter data. Thus, to address this issue, and mitigate potential confusion, the Twitter registers will henceforth be referred to as inventories. Second, while some identifying characteristics may be inferred based on individual user's accounts, such identifying data are rarely explicitly stated by the user or the organisation responsible for collecting them. Therefore, it is important that such data, when inferred about individual users is acknowledged and correctly interpreted.

Bearing the above in mind, the objective of this chapter is to create a series of population inventories that take raw Twitter data and through a series of heuristic and data-mining techniques, create individual level population inventories that take a form that is broadly consistent with the UN definition. The chapter will initiate with as a section concerned with the data that are to be employed in the construction and validation of the Twitter inventories. The subsequent sections will be concerned with a) the methods used in the creation of the Twitter inventories and b) with the validation and verification of the new inventories. The chapter will conclude with a

discussion of the results and series of initial conclusions.

4.2 Data

Having framed the problem, the next phase in the analysis is concerned with developing an understanding of the data. Seeking to assemble the population inventories several distinct datasets are required. First is the corpus of geotagged Tweets. Second is an administrative boundary dataset suitable to serve as the geographic reference for the new inventories. Finally, is a suitable form of reference data against which the Twitter-derived inventories may be assessed. In the case of each data source, efforts are made to describe and where necessary validate the quality of the data.

4.2.1 Twitter Data

The Twitter dataset used within this chapter, and the thesis as a whole, is a corpus of 1.4 billion Tweets submitted by 24.4 million users. Obtained via the Twitter Streaming API, the data represent a full year spanning the period between December 2012 and the January 2014 inclusive. The full temporal coverage of the data is indicated in Figure 4.2 and the spatial coverage in Figure 4.1. The data were obtained via the Twitter API and stored in a PostgreSQL relational database (see: http://www.postgresql.org). The 'POST statuses/filter' API was used to harvest the data from Twitter in real-time. The 'POST statuses/filter' endpoint is a specific version of the sample stream API in which a series of parameters are specified such that only specific data are returned (Twitter, 2015). These parameters include, but are not limited to, keywords, user ids and locations.

An important point of clarification is on what data are accessible through the Twitter Streaming API. A common misconception in the use of the sample and associated filtered stream is that only 1% of the total throughput of Twitter are ever accessible. Rather, the data that are available are described as up to the equivalent of 1%. As such, rather than receiving just the 1% geotagged Tweets from the already reduced 1% sample stream, the API will return the majority of geotagged Tweets from the Twitter feed; assuming that they never exceed the 1% of the total threshold. Further, with most estimates of the proportion of Tweets to contain geographic at-

tribution centring on 1% it may be presumed that the majority of geotagged Tweets are available for collection. Morstatter et al. (2013) recorded the filtered streaming API returning 90.10% of all geotagged Tweets versus the full stream.

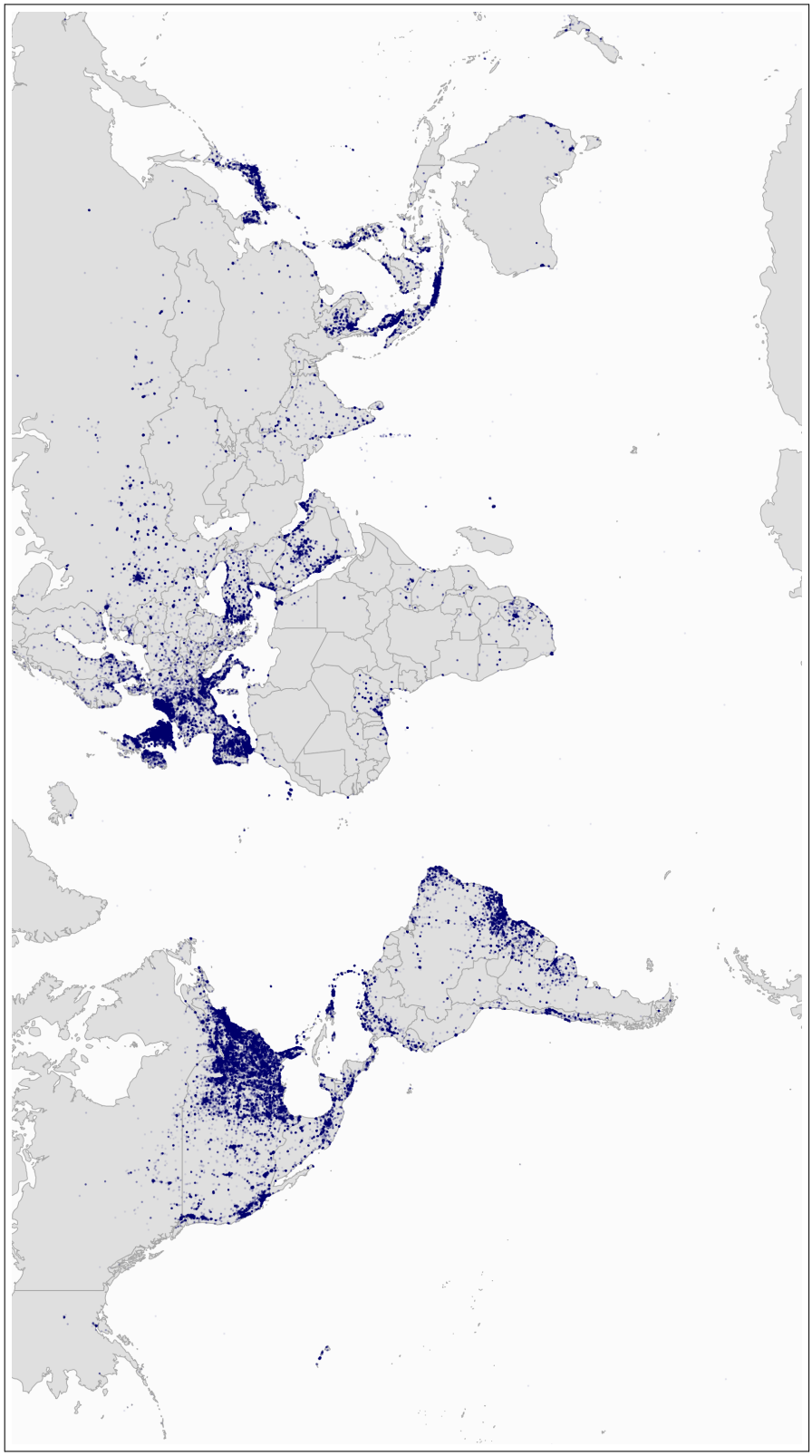


Figure 4.1: A random sample of 1 million Tweets drawn from the full Twitter dataset. Each Tweet is represented by a blue dot.

Unfortunately, as is apparent in Figure 4.2, the data collection application failed on several occasions during the recording period. The application for harvesting Tweets was implemented on a Windows Server using the Java Programming Language. Due to unforeseen circumstances the server was reset at various periods resulting in the application failing. In hindsight, it would have been beneficial to implement redundant systems such that the risk of data loss was reduced. The missing data are evident where a cell is not filled or records a low frequency of Tweets. While it may have been possible to recover the data that were missed, the costs of such an exercise would have proven prohibitive. Though Twitter does provide an API for the collection of historical data the use of rate limiting would have been a significant barrier; in practice, a query for each unique user would have been necessary.

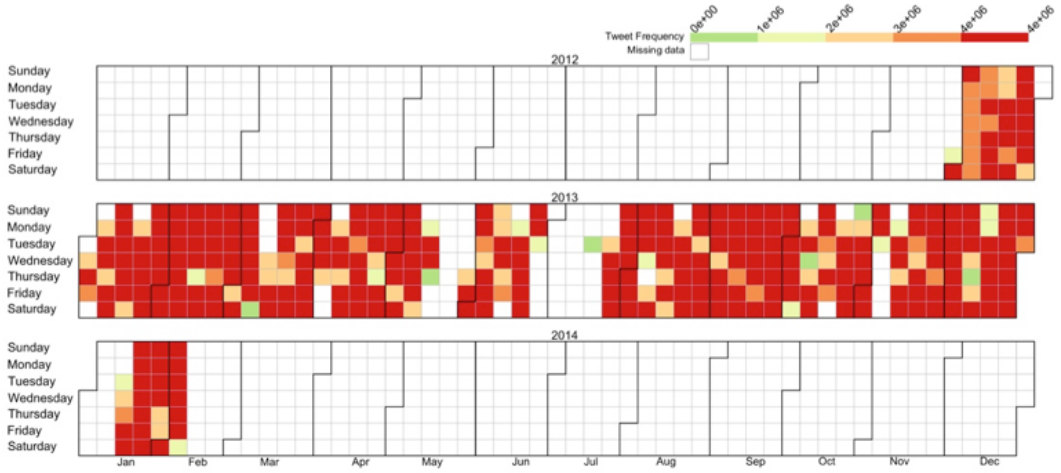


Figure 4.2: Heat-map calendar showing the temporal coverage and completeness of the Twitter dataset. Empty cells indicate that no data were collected on these days.

Also, it is worth noting that the volume of data harvested from Twitter was such that commonly used desktop approaches to data storage and manipulation were not possible. Beyond simply storing the data, use of the PostgreSQL database offered multiple advantages. First, the database facilitated the creation of optimised data structures, through indexes, which significantly reduced data access times. Second, through the addition of the PostGIS extension, it was possible to perform a significant proportion of the required GIS functionality introducing significant time-saving in

the spatial analysis workflows. The attributes recorded in the Twitter dataset are detailed in Table 4.1.

Table 4.1: Descriptions of variables collected through the Twitter API.

Variable	Description
Userid	A unique identifier assigned to each user
Language	The language which the user has set for the service
Location	The value the user has entered into the optional location field
Name	The user's username e.g. 'Alistair_Leak'
Screen name	The public facing name the user has chosen eg 'Alistair Leak'
Time zone	Time zone in which the user tweeted
Latitude	The latitude at which the Tweet was submitted
Longitude	The longitude at which the Tweet was submitted
Timestamp	The date and time at which the Tweet was submitted
Tweet text	The 140 character Tweet text
Status id	A unique identifier assigned to each Tweet

4.2.2 Administrative Boundary Data

Selection of a suitable administrative geography for use in the construction of the proxy population inventories appears quite trivial on the surface. However, in practice, such an exercise is potentially complex and time-consuming. In seeking to identify suitable administrative boundary data for all countries, foreseeable challenges include limited access to data, incompatible data formats and projections and also inconsistent nomenclature. With this in mind, it was deemed most appropriate to find a pre-existing database of global administrative areas.

Thus, the administrative boundary dataset employed within this analysis was the GADM 2.0 global administrative boundary dataset version 2.0 (available from http://www.gadm.org). The GADM dataset, illustrated in Figure 4.3, is a seamless global database of administrative areas. A key feature of the GADM dataset is the use of a standardised naming structure that lends itself to automated processing. At its coarsest, GADM level 0 represents the outline of countries and as the level is increased so is the granularity. The degree of granularity varies between countries and appears to be dependent on the availability of data. However, taking the example of the UK, the highest granularity, level 2, is consistent with 'districts and boroughs'. That being said, in the context of the analysis to be performed, the benefits of the consistent data structure far outweigh the drawbacks.

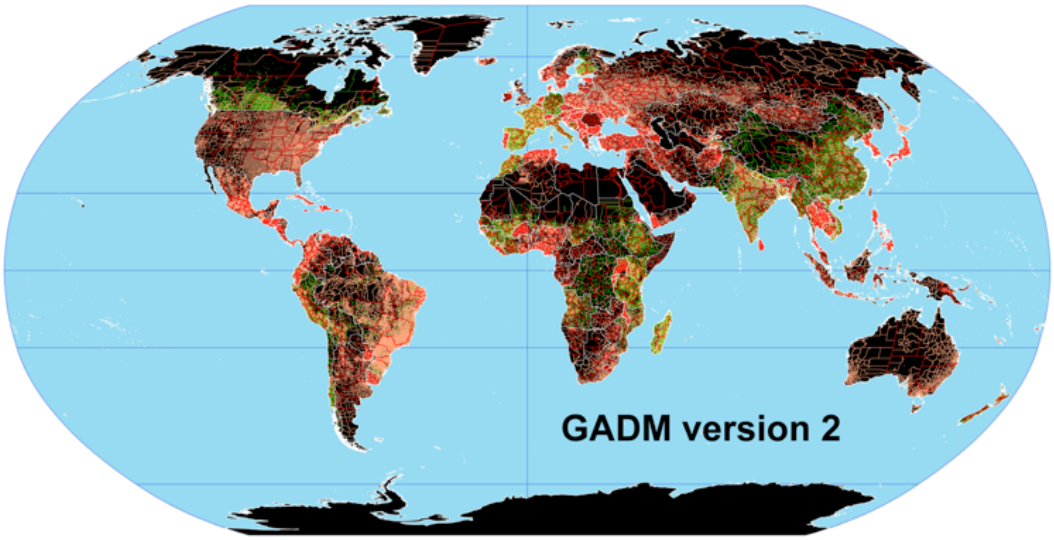


Figure 4.3: Map of the GADM 2.0 administrative boundary dataset. The GADM data provide a standardised global geography based on known administrative regions (source: GADM, 2012).

4.2.3 The Worldnames Database

The UCL Worldnames Database was chosen as the most suitable reference for the validation of the inventory creation framework due in part to the ease with which it could be obtained. Besides, the Worldnames Database was compiled by the UCL Department of Geography and is arguably the most complete inventory of individual names ever constructed. The Worldnames Database is a compilation of publicly available electoral roll and telephone directory datasets representative of approximately two billion of the Earth's population. However, being that the data used in the database have been compiled over a period of years and drawn from a multitude of sources invariably resulting in variations regarding completeness, coverage and quality both within and between countries. With this in mind, an audit of the Worldnames Database was performed, the aim of which was to assess the quality of the component datasets systematically. Specific objectives included the clarification of the date of publication; the format of the data; and the representativeness of the data regarding the underlying population.

An audit of the data was performed on a country-by-country basis. In each case, the most common surnames as recorded in the database were compared against alternative sources of names data. Early in the completion of the audit, it was found that

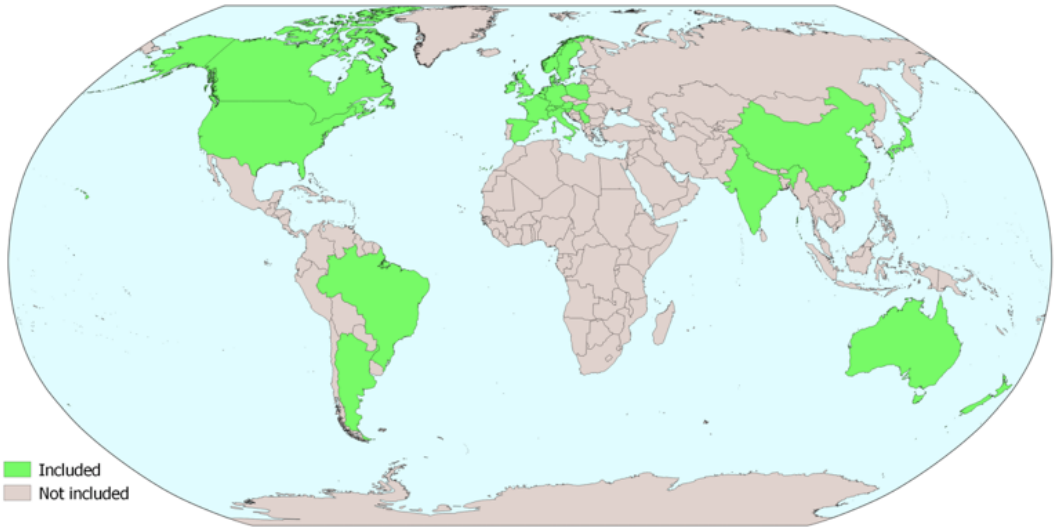


Figure 4.4: Coverage of the UCL Worldnames Database. The countries highlighted in green are those which are included.

only limited official records of surnames within countries were publicly available. In the majority of cases only the top 10, 20, 50 or 100 names could be obtained; limiting the potential options for statistical comparison. Such was the constraint regarding reference data that a method, which worked with limited data, was necessary. Consequently, three validation metrics were employed: 1) the proportion of overlap between the reference and Worldnames registers, 2) the Pearson's correlation coefficient between the top sets of names and 3) the Spearmans's rank correlation between top sets of names.

In the case of both the proportion of overlap and rank correlation, the comparisons were performed on the top 10, 20, 50 and 100 most common surnames. The proportion of overlap was considered as the number of names shared between the two sets of data irrespective of rank while the rank correlation between the two datasets was calculated using the Spearman's Rank correlation coefficient. The rank correlation was performed on pairs of observations within the top n values. The full data audit is reported in Appendix B and a summary of results of the data audit are summarised in Table 4.2.

In the case of both comparison techniques, the approaches are based on the frequency distribution of surnames. Figure 4.5 is a plot of the frequency of unique

Table 4.2: Summary of the Worldnames Database data audit
p-value < 0.05 *, < 0.005 **, < 0.0005***.

ISO-3	Overlap (%)				Correlation rho				Quality
	Top10	Top20	Top50	Top100	Top10	Top20	Top 50	Top100	
ARG	100	80	84	96	0.64*	0.90***	0.92***	0.90***	Good
AUS	90	100	98	98	0.98***	0.97***	1.00***	1.00***	V. Good
AUT 1	100	100	-	-	0.90**	0.98***	-	-	Good
AUT 2	90	95	88	88	0.67	0.90***	0.96***	0.88***	V. Good
BEL	90	90	92	90	0.97***	1.00***	0.97***	0.98***	V. Good
BGR 1	70	50	58	50	0.57	0.82*	0.92***	0.80***	Poor
BGR 2	50	55	52	55	0.9	0.96***	0.85***	0.58***	Poor
BRA	60	60	68	72	0.89*	0.55	0.56**	0.55***	Poor
CAN 1	10	20	-	-	-	-0.22	-	-	Poor
CAN 2	20	50	46	55	1	0.08	0.202	0.31*	Poor
CHE	60	50	60	58	0.71	0.6	0.41*	0.41**	Poor
DEU	100	95	98	97	1.00***	1.00***	1.00***	1.00***	V. Good
DNK	100	95	98	95	1.00***	1.00***	0.99***	0.99***	V. Good
ESP	90	95	86	80	0.93***	0.99***	0.99***	0.98***	V. Good
FRA	90	95	94	93	0.53	0.91***	0.92***	0.94***	Good
GBR	100	100	90	94	0.98***	0.98***	0.96***	0.96***	V. Good
HUN	90	70	78	71	0.43	0.78**	0.74***	0.81***	Good
IND 1	10	35	26	29	-	0.14	0.57**	0.56**	Poor
IND 2	10	15	20	23	-	1	0.35	0.4	Poor
IRL	100	95	98	97	1.00***	1.00***	1.00***	1.00***	V. Good
ITA	80	100	88	93	0.93**	0.93***	0.97***	0.96***	V. Good
JPN	90	95	92	95	0.90**	0.81***	0.93***	0.86***	V. Good
LUX	70	75	72	68	1.00***	1.00***	0.84***	0.90***	Good
MLT	90	100	98	93	0.98***	0.98***	0.97***	0.99***	V. Good
NLD	80	90	82	86	0.76*	0.94***	0.95***	0.94***	V. Good
NOR	100	95	94	97	0.98***	0.99***	1.00***	0.98***	V. Good
NZL	70	65	-	-	0.96**	0.99***	-	-	Poor
POL	50	60	68	62	0.2	0.38**	0.54**	0.62***	Poor
SER1	90	75	84	89	0.77*	0.88***	0.75***	0.89***	Good
SVN	90	80	80	85	0.97***	0.95***	0.96***	0.87***	V. Good
SER	80	75	74	82					Good
SWE	100	85	86	74	0.99***	1.00***	0.95***	0.96***	V. Good
USA	70	75	80	82	0.96**	??????	0.51**	0.86***	Good
MNE	70	75	70	74					Good
UNK	-	-	-	-	-	-	-	-	-

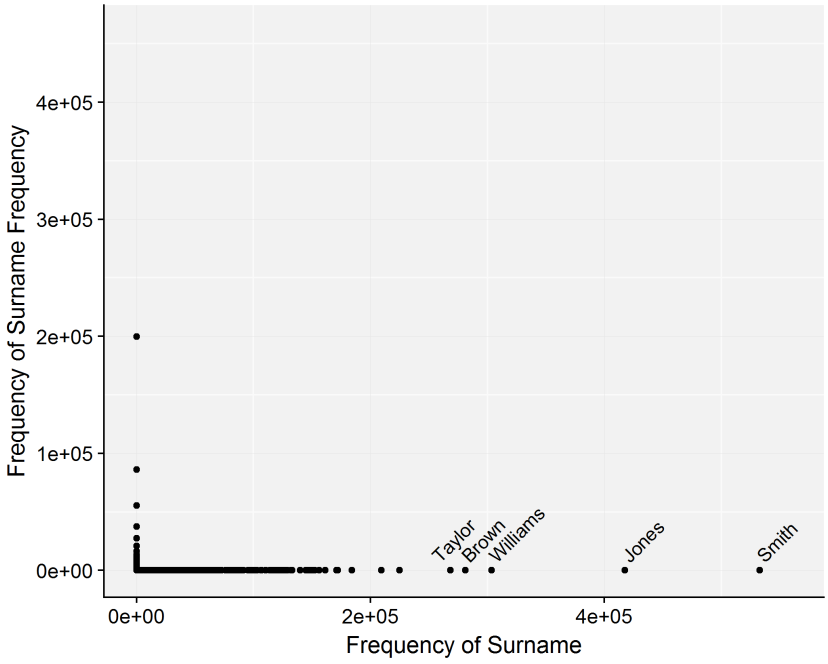


Figure 4.5: Plot of the frequency of surname frequencies in the UK.

surname counts based on data from the 2013 Consumer Register. The distribution of surnames is known to conform to a power law in which there are very few common surnames and many uncommon surnames. A feature of this distribution is that the most common names are usually very distinct in terms of their order; a feature that is evident in Figure 4.5. Thus, it may be thought that agreement of the most common names is a good early indicator as to the representativeness registers as a whole. The audit of the Worldnames Database registers covered 28 countries and revealed a number of issues in the data.

Countries for which the population registers were deemed good included: Argentina, Australia, Austria, Belgium, Denmark, France, Germany, Hungary, Ireland, Italy, Japan, Luxembourg, Malta, Netherlands, Norway, Montenegro, Serbia, Spain, Slovenia, Sweden, the United Kingdom, and the United States of America.

Countries for which the population registers were deemed less than good included: Bulgaria, Brazil, Canada, India, New Zealand, Poland and Switzerland.

In the case of the data for Brazil and India, it was found that only a limited number of cities were represented introducing a regional bias. In the case of the New Zealand data, it was found that the data were compiled over a series of years between 1897 and 1992. In the case of Switzerland and Canada there appeared to be a cultural bias in the data that may have been a consequence of irregular regional recording. The Canada data appeared to omit several common ethnic names, and the Swiss data appeared to be biased towards the German-speaking portion of the population. No specific explanations for the Polish and Bulgarian data could be found.

Completion of the Worldnames Database audit proved to be highly informative in terms of understanding the data quality and completeness. In several cases, it was found that the data were only representative of specific regions potentially leading to an ecological fallacy where those individuals recorded were considered representative of the whole population.

4.3 Inventory Creation and Validation

Having introduced the problem and available data, the next phase of the analysis was the development of an appropriate framework for creation and validation of the Twitter-derived inventories. In Section 4.3.1 the framework for the creation of the inventories is detailed and the methods used in the subsequent validation form section 4.3.2.

4.3.1 Inventory Creation

4.3.1.1 Identification of users' places of residence

Unlike the Worldnames Database, in which each person is recorded at a single location, users of Twitter are often associated with multiple locations. Thus, there is a requirement to be able to identify which, if any, of a user's location information, truly relates to their place of residence. Potential sources of user location include their declared locations, information embedded within their Tweets and locations that are recorded within the meta-data of individual Tweets.

If we first consider users' declared locations, these attributes are provided by the user at the time of registration and are not constrained to actual locations or naming conventions. Consequently, the locations are often ambiguous, imprecise or aspatial (Graham et al., 2014). Examples of users' declared locations include 'England', 'Global' and 'Liverpool'. As may be evident, these locations are valid yet ambiguous and imprecise limiting their usefulness in terms of compiling population inventories. An alternative to users' declared locations are the location details that are embedded within individual Tweet's meta-data. Where a user chooses to 'Share their location', Twitter records the location as it is provided by the device being used to access the service; in the case of smart phones, this may be via GPS, cell location, or WiFi. Where location data are available, the attribution is embedded within the Tweet's meta-data. Unlike users' declared locations, these data are less open to ambiguity or manipulation.

In seeking to assign individuals to single unique locations based on the spatial information embedded within their Tweets it must be recognised that a typical user's

recorded activities will not be constrained to a single location. Rather, an individual will travel within and between multiple localities as part of their routine activities. Thus, there is a requirement to be able to identify which, if any, locations associated with a specific user are at their normal place of residence.

It may be argued that knowledge of each user's precise place of residence is not necessary. In many cases, knowledge of users' locality, region, or country may be sufficient. For example, in the case of the Onomap CEL classifier, the clustering is entirely aspatial and is solely concerned with individuals' forename-surname pairs. Similarly, much of the existing data in the Worldnames Database is recorded at a very aggregate level. A further and critical consideration is the data that are available for the analysis. Due to the constraints of data availability, and the lack of precision regarding users declared locations, it was decided to pursue the use of the spatial attribution embedded within individual user's Tweets.

The intuition in the following analyses is that a person will tweet most frequently within the area in which they are resident. Seeking to implement this intuition the first consideration is what is meant by the term 'area' and how can the data be formatted to suit this. Geotagged Tweets report a high degree of precision (often within a matter of metres) that does not lend itself to simple aggregation. One approach to addressing the issue of aggregation may be the use of geographic binning in which Tweets are spatially joined to some form of administrative geography. i.e. rather than using the longitude and latitude data provided in their raw form, the Tweets are spatially joined to an administrative geography such as Output Areas in the UK. Through a reduction in the precision of the location information for each Tweet, the degree of accuracy and ease of aggregation may be increased.

The methodology employed in the identification is illustrated in Figure 4.6 and implemented as follows:

1. All users within the Twitter dataset are identified and a subset created based with a minimum threshold of five Tweets within the dataset.

2. All Tweets by the previously subset users are spatially joined to the relevant administrative geography.

3. For each user:

   (a) Once spatially referenced, Tweets for each user are aggregated based on the region in which they emanated.

   (b) A user is assigned a location assuming that they have five or more and greater than 50% of their total Tweets in the same administrative unit.

4. The process is repeated at a series of spatial resolutions such that inclusion in the final register may be maximised.

Concerning supplementing the Worldnames Database, the administrative geography chosen for the initial analysis was the GADM administrative dataset introduced in Section 4.2.2. The GADM dataset comprises a hierarchical geography in which the finer geographies are nested into regions and countries. Such a data structure aids in the creation of new inventories that allow for the maximisation of inclusion of users at each spatial resolution. For instance, a user may not meet the '5 or more Tweets and 50% of total Tweets' at Postcode level, however, may do so at a less granular resolution such as at the regional or national level.

The accuracy of the location assignment algorithm was assessed in the UK using a stratified sample of users across the 192 GADM administrative regions. The stratification was conducted in such a manner that a minimum of 1,000 users were selected with a lower threshold of one user from each administrative region. For each user, the declared location was manually geo-coded and compared to the user's estimated location as determined by the location assignment algorithm. The results of the assessment were recorded using a contingency table such that the overall classification accuracy could be determined. The use of a contingency table offered several advantages beyond just measuring overall accuracy. More specifically, the contingency table facilitated the calculation of accuracy adjusted for chance and also the identification of specific areas of miss-classification (De Smith et al., 2011).

Table 4.3 provides a summary of the confusion matrix, indicates that the overall classification accuracy was 84% though, it should be noted that this score was calculated excluding those users who had aspatial (39%) or ambiguous (12%) declared

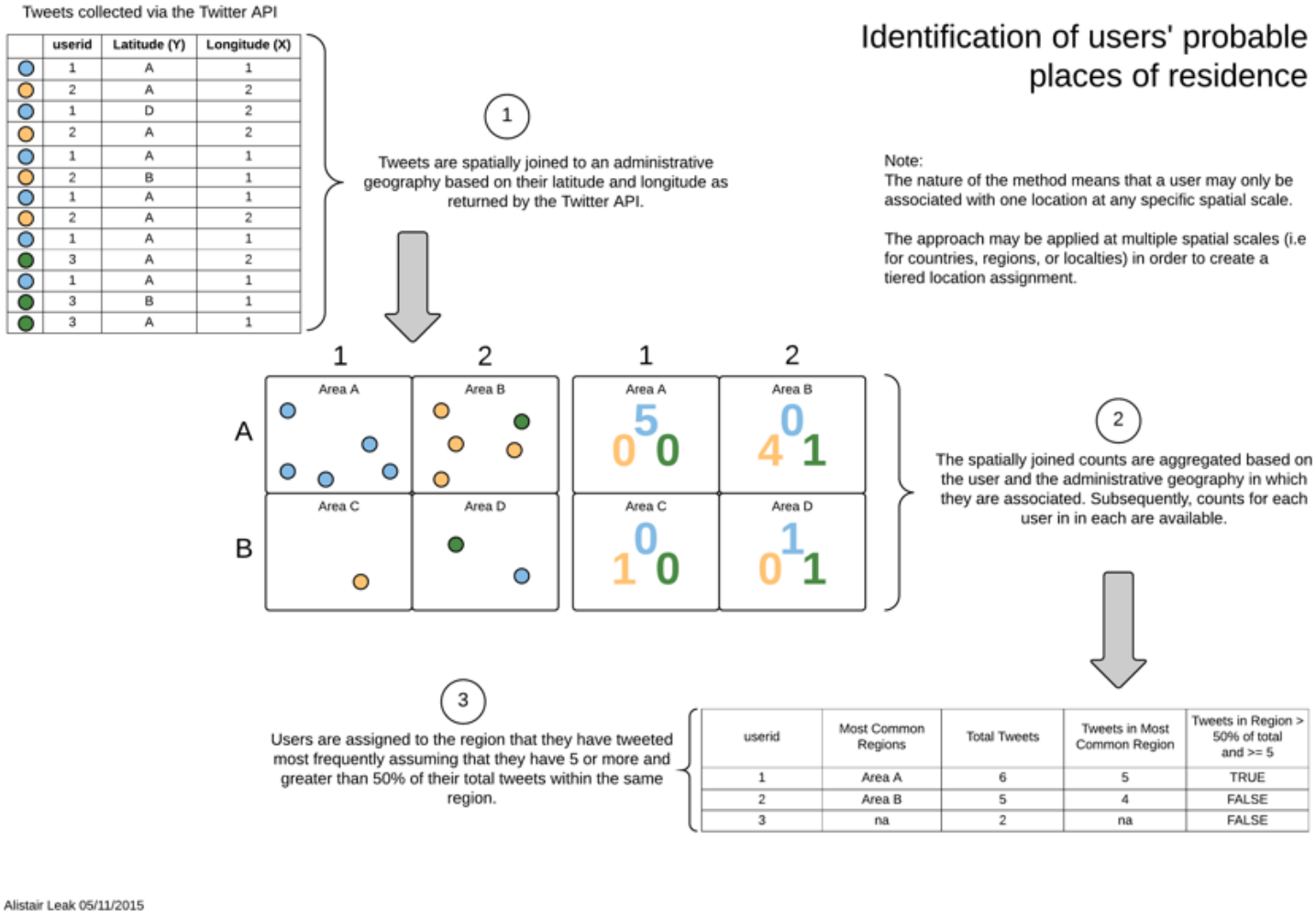


Figure 4.6: A graphical illustration of the identification of users' places of residence.

locations. The Kappa Index value was only marginally lower than the overall classification accuracy, suggesting that very few correct classifications occurred through chance and vice versa. Ambiguous locations were seen most commonly in large metropolitan areas which were subdivided into multiple districts or boroughs; most notably in the case of London. While it may have been desirable to automate this process, a significant number of users' declared locations required some form of human intervention that would not have been feasible using a solely machine based approach.

While the results are shown in Table 4.3 are calculated as a global measure it is also possible to calculate the equivalent statistics on a case-wise basis using a variation of the Kappa Index. However, achieving this would have required a significantly larger stratified sample to be taken, and the expected usefulness was deemed unlikely to merit the effort required.

Table 4.3: Summary results of the confusion matrix used in the assessment of the location assignment algorithm.

Metric	Value
Sample Size	1,073
Ambiguous locations	133
Aspatial locations	416
Usable locations	524
Overall Accuracy	0.84
Kappa	0.83

4.3.1.2 Extraction of users' personal names

Having identified the likely residential areas of users' the next phase in the creation of the individual level population inventories was the extraction of users' probable names. As indicated previously, there are two possible identifiers in the Twitter data that may be used; the users' usernames and their screen names. The users' usernames are a distinct identifier used to differentiate between specific accounts on the service while the screen name is any string chosen by the user. Unlike the username, there is uniqueness constraint imposed on users' screen names, however. Consequently,

in the knowledge that many individuals share the same name, users' screen names were deemed the most appropriate for extraction of users' names.

The extraction of Twitter users' real names from their screen names poses several challenges. First, users' screen names are provided as a single string variable devoid of logical partitioning regarding conventional name parts. Second, the screen name variable may contain multiple non-name components such as personal titles, number and symbols. To compound this issue further, there is no guarantee that the screen name is a real name or a real person.

Considering first the issue of the single string screen name, a typical name might be 'Mobiles UK', 'Dr John Smith', 'Rhiannon de la Merr' or 'Jane Doe 1990'. In these fictitious examples, various of the previously mentioned challenges are evident; notably, the inclusion of personal titles, dates and non-real names.

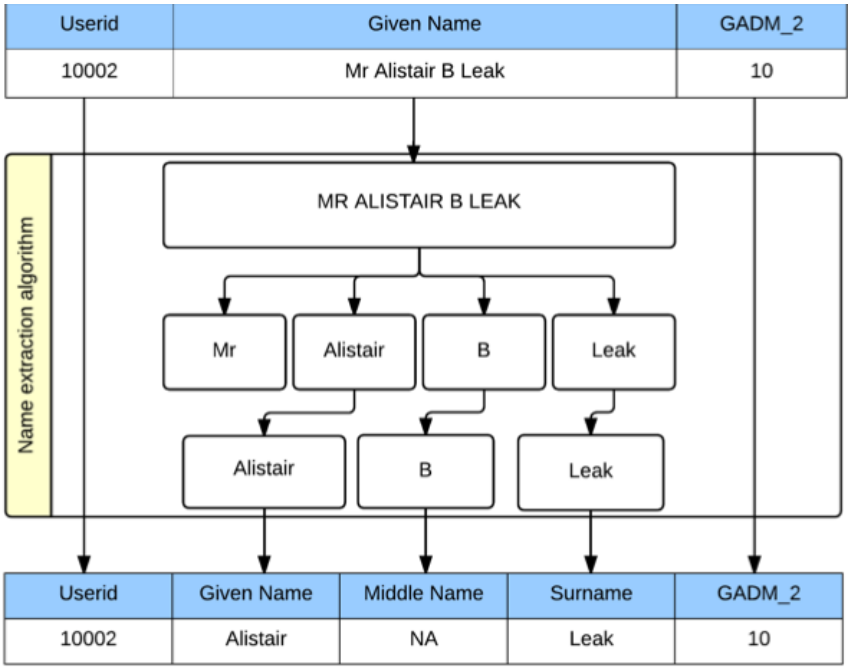


Figure 4.7: A graphical representation of the name extraction process.

Looking to address the first two concerns, personal titles and non-alpha characters, a heuristic technique, based on a method by Longley et al. (2015), was used in the extraction of the name parts. The method, illustrated in Figure 4.7, assumes a western naming order where forename precedes surname. The method is applied as follows:

1. All non-alpha characters are removed from the user's screen name.
2. Screen names are split into multiple tokens based on white spaces.
3. Tokens are compared against a list of titles and surname prefixes.
4. Surname prefixes are affixed to the family name segment.
5. Cleaned surnames are recorded against the user's id.

Table 4.4: Examples of forenames and surnames extracted from Twitter users' screen names using the names extraction process.

User id	Screen name	Forename	Surname
1	Dr James Cheshire	James	Cheshire
2	Rhiannon de la Merr	Rhiannon	De la Merr
3	Bieber Believer 99	Bieber	Believer

Table 4.4 provides a sample of users' screen names and the forenames and surnames that were extracted using the previously discussed heuristic. It should be noted that this method is not able to differentiate between real and non-real users as in the case of 'Bieber Believer'.

4.3.1.3 Treatment of gendered names

As noted in the review of personal names, in certain cultures surnames can be affixed with gender specific identifiers. For example, the Polish names Kowal(ski) and Kowal(ska) are masculine and feminine forms, respectively, of the same name. They are not, however, associated with the similar sounding Polish surname Kowal. Concerning computerised analysis, the difference in spelling has the undesirable effect of inferring that the names originate from two different familial groups. Thus, the decision was made to convert all gendered names, where possible, to their masculine form. This process entailed replacing the feminine affix (i.e –eva, -ova, -ska etc) with the equivalent masculine affix (i.e –ev, -ov, -ski). The gender-standardised names were recorded as a new variable such that the gender information inherent within the names was not lost.

4.3.2 Inventory Validation

Having identified the probable residential location of the Twitter users in Section 4.3.1.1 and extracted their probable personal names in Section 4.3.1.2 the final set of inventories successfully address the requirements drawn from the United Nations' register definition in Section 4.1. However, while the new individual level population inventories address the requirements of a proxy regarding data structure, little is known as to their representativeness regarding the underlying population. Thus, the following section provides details and justifications for the methods employed in the initial validation of the Twitter inventories. In seeking to validate the Twitter inventories a number of key population metrics are explored. These metrics include the overlap in the most commonly occurring surnames, geographic distribution and surname composition.

The first validation exercise was applied to two countries, the UK and Spain, such that the two largest naming groups (Western and Hispanic) could be included. For reference, the equivalent individual level population inventories were drawn from the Worldnames database. In each case, the analysis was performed at a series of geographic resolutions based on the GADM administrative boundary dataset.

4.3.2.1 Common names

As discussed previously the most commonly occurring surnames in countries tend to be clearly distinct regarding frequency versus the majority of the population. Such is the distinction between these names that it is thought that a comparison of the most common names may be a good initial indicator of inventory performance. As this is a preliminary test, only the top 10 names were analysed.

4.3.2.2 Geographic distribution

Understanding the geographic distribution of geotagged Twitter users versus the reference population is a useful means of identifying regions of notable under and over representation. Geographic distribution is quantified as the ratio of the observed Twitter population versus the expected Twitter population.

$$p_i^T = \frac{p_i^R}{P^R} P^T \tag{4.1}$$

The expected Twitter population is calculated using Equation 4.1 where p_i^T indicates the estimate Twitter population at i, p_i^R is the reference population at i, P^R is the total reference population and P^T is the total Twitter population. The Location Quotient is subsequently calculated as the number of Twitter users identified divided by the expected Twitter population as defined in Equation 4.1.

4.3.2.3 Compositional similarity

In the context of this thesis, compositional similarity is considered as the amount of overlap of names between the Twitter-derived and reference inventories. It is hypothesised that a Twitter inventory that is compositionally similar to the reference is more likely to be representative of the true population than one that is not.

A number of measures are available for quantifying the degree of similarity between two sets of population inventories. The majority of these methods are found within the field of ecology. However, a number of similar methods exist within the names literature. For example, Cheshire et al. (2010) made use of the Lasker kinship coefficient as a means to measure the similarity between groups of names. In the case of the work by Cheshire et al. (2010), the objective was the creation of an inter-regional distance matrix used in the subsequent partitioning of geographic space.

$$R_i = \sum \frac{N_{s1} N_{s2}}{2 N_1 N_2} \tag{4.2}$$

Equation 4.2, the Lasker Kinship coefficient is calculated as the sum of the count of surnames in sample one multiplied by the count of surnames in sample two divided by two times N1 and N2 which are equivalent to the size of each sample respectively (Lasker, 1977). The Lasker Kinship coefficient is widely used within the names analysis literature.

An alternative method drawn from ecology is the Morisita-Horn Index of overlap (Horn, 1966). Like the Lasker Kinship coefficient, the Morisita-Horn Index is calculated independently for each region. A key quality of the Morisita-Horn index

of overlap is its ability to deal with populations of significantly differing sizes and diversities. This quality is confirmed by Wolda (1981) who conducted an in-depth review of commonly used similarity statistics.

$$C_H = \frac{2\sum_{i=1}^{S} x_i y_i}{(\frac{\sum_{i=1}^{S} x_i^2}{X^2} + \frac{\sum_{i=1}^{S} y_i^2}{Y^2})XY} \tag{4.3}$$

Equation 4.3, the Morisita-Horn index returns a linear value between 0 and 1 where 0 indicates no overlap between the two populations and 1 indicates the identical composition of names.

- S is the number of unique surnames shared between the two populations.
- x_i and y_i are the number of individuals sharing a specific surname in region X and Y.
- X and Y are the number of unique surnames in regions X and Y respectively.

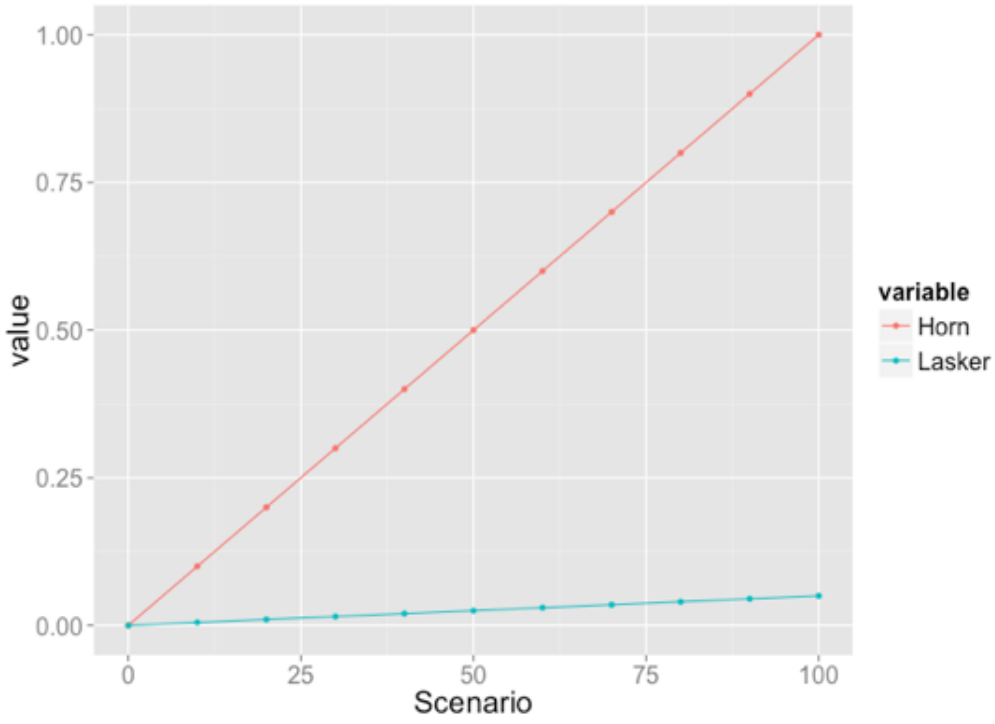


Figure 4.8: A line graph plotting the Lasker Kinship coefficient and Morisita-Horn index of overlap for a series of simulated populations.

Seeking to understand further the difference between the Lasker Kinship coefficient and Morisita Horn index of overlap a set of testing data was simulated. The testing set was composed of fourteen scenarios, each of which contained two sets of surnames with a known degree of overlap. The first twelve scenarios tested the coefficient's response to overlap and the final two scenarios tested relative population size. In the overlap scenarios, overlap ranged from 0 to 100 percent in increments of 10. Figure 4.8 shows that in practice both the Lasker Kinship coefficient and Morisita-Horn index of overlap exhibit a linear response, however, the Morisita Horn index is more easily interpreted due to its limits being fixed between 0 and 1.

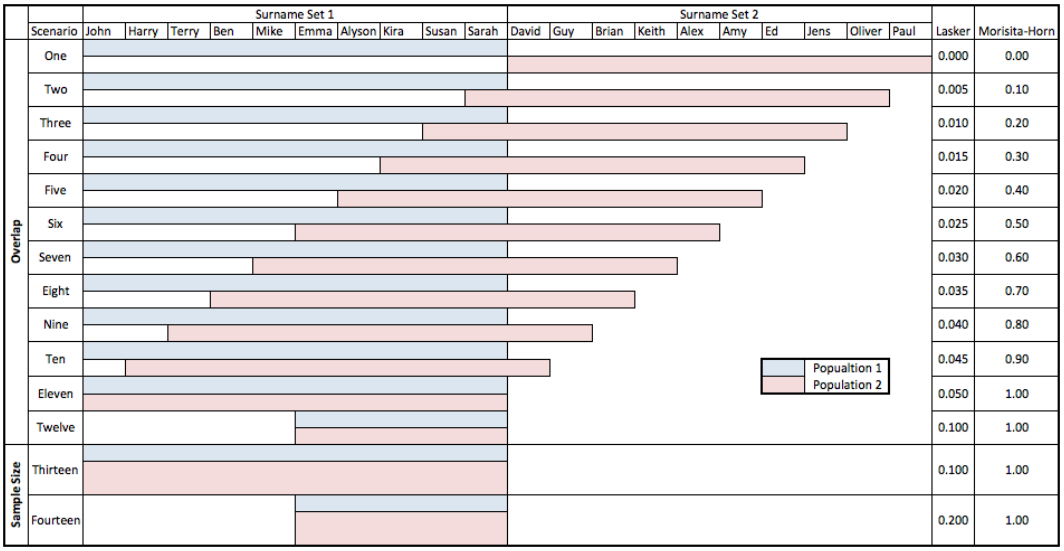


Figure 4.9: A graphical representation of the Lasker Kinship Coefficient verses the Morisita-Horn Index of Overlap.

Further to the above, Figure 4.9 demonstrates that the Lasker Kinship coefficient is significantly impacted by the relative size of the populations. If we compare test eleven for overlap and test thirteen for sample size, it is evident that both surname sets are composed of the same names in the same proportions. However, the calculated Lasker Kinship coefficient is 0.05 where the populations are the same size and 0.10 where the populations are different sizes. Concerning interpreting these data, the Lasker Kinship coefficient could, therefore, be challenging, as, without knowledge of the population sizes, the values are meaningless. In contrast, the Morisita-Horn Index returns a value of 1 irrespective of the number of surnames

where the two populations exhibit identical structure.

Thus, having proven the usefulness of the Morisita Horn Index of Overlap in the measurement of similarity between regions, the measure is be applied to the two countries at GADM levels 0, 1, 2 and 3 (for Spain) such that an understanding of the individual countries and effects of scale may be attained.

4.4 Results

4.4.1 Common Names

Table 4.5: Comparison of surname ranks between the Worldnames Database reference data and Twitter-derived individual level population inventories for the UK (top) and Spain (bottom).

Surname	Reference rank	Twitter rank	Difference
Smith	1	1	0
Jones	2	2	0
Williams	3	3	0
Brown	4	5	-1
Taylor	5	4	1
Davies	6	6	0
Wilson	7	7	0
Evans	8	8	0
Thomas	9	9	0
Johnson	10	11	-1

Surname	Reference rank	Twitter rank	Difference
Garcia	1	1	0
Fernandez	2	6	-4
Gonzalez	3	5	-2
Rodrigues	4	4	0
Lopez	5	2	3
Martinez	6	7	-1
Perez	7	8	-1
Martin	8	10	-2
Gomez	9	9	0
Ruiz	10	11	-1

4.4.2 Geographic Distribution

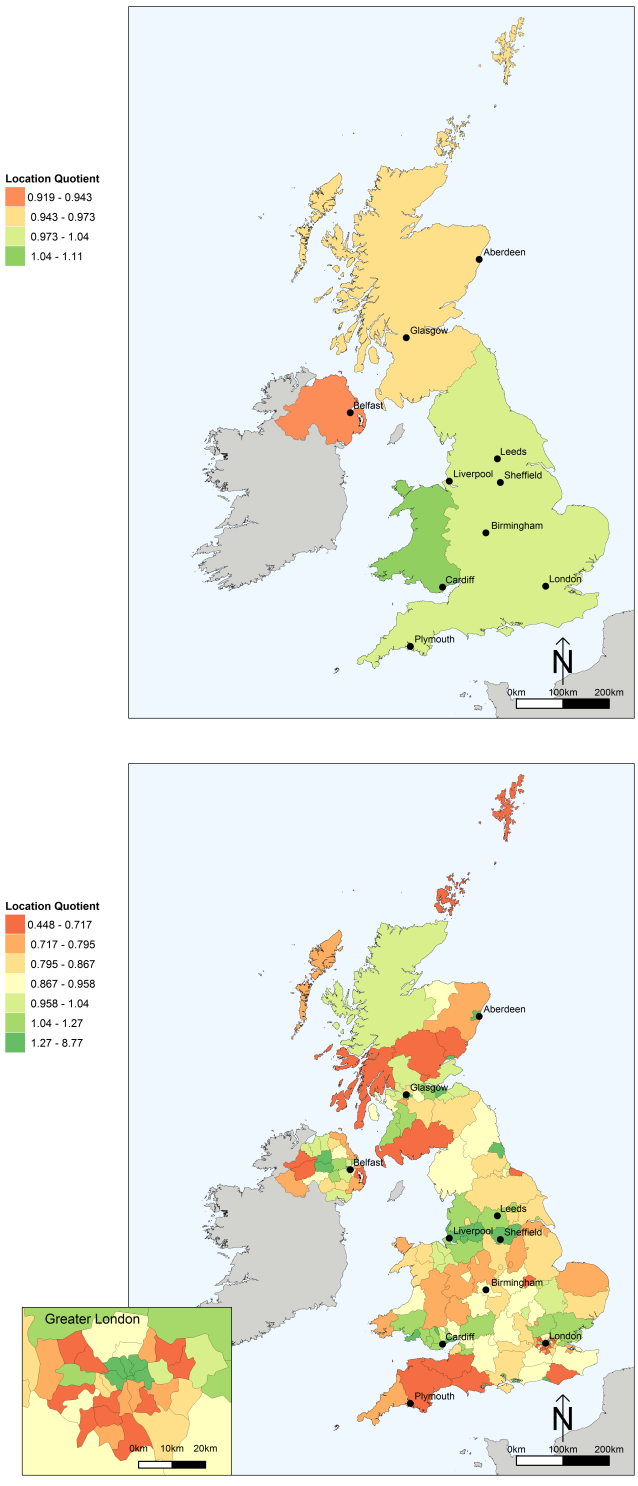


Figure 4.10: Maps of Location Quotient for the UK at GADM level 1 (top) and 2 (bottom).

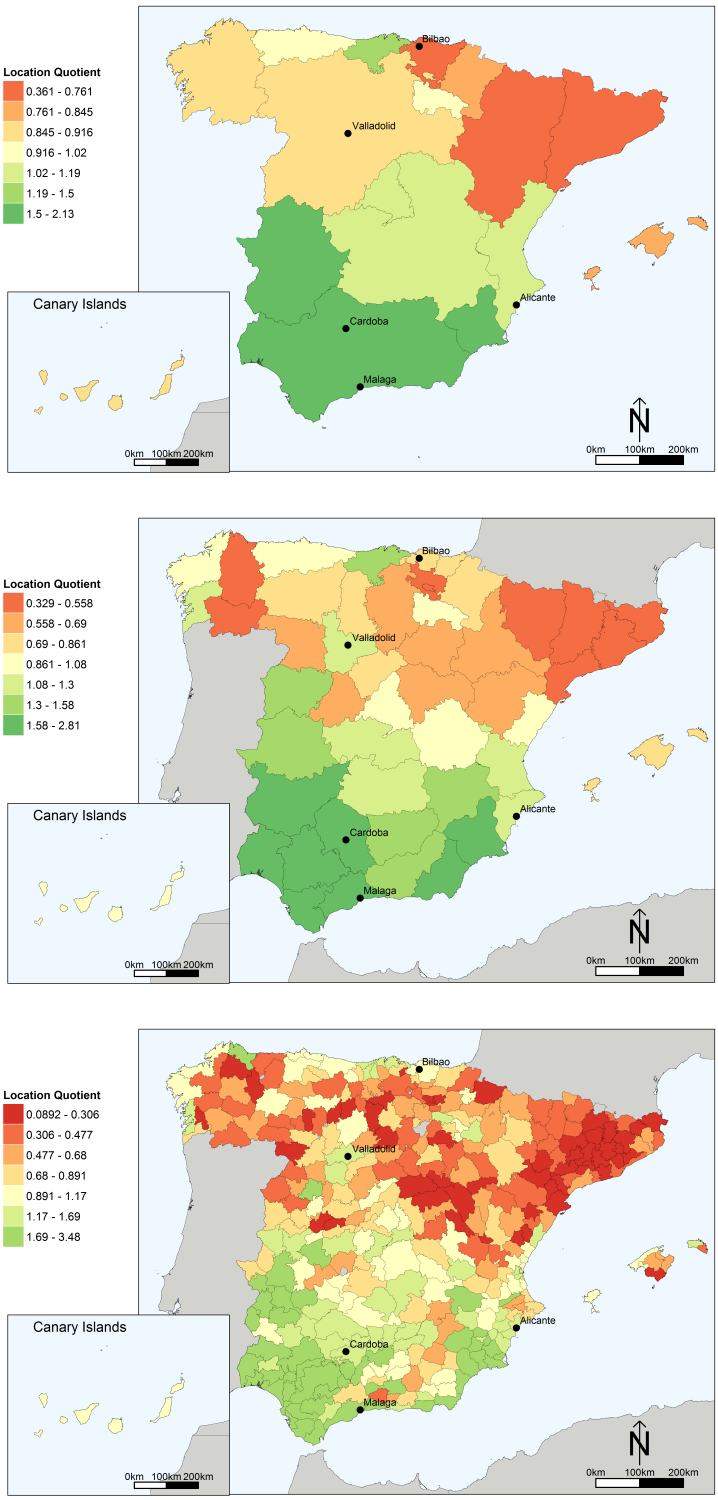


Figure 4.11: Maps of Location Quotient for Spain at GADM level 1 (top), 2 (middle) and 3 (bottom).

4.4.3 Compositional Similarity

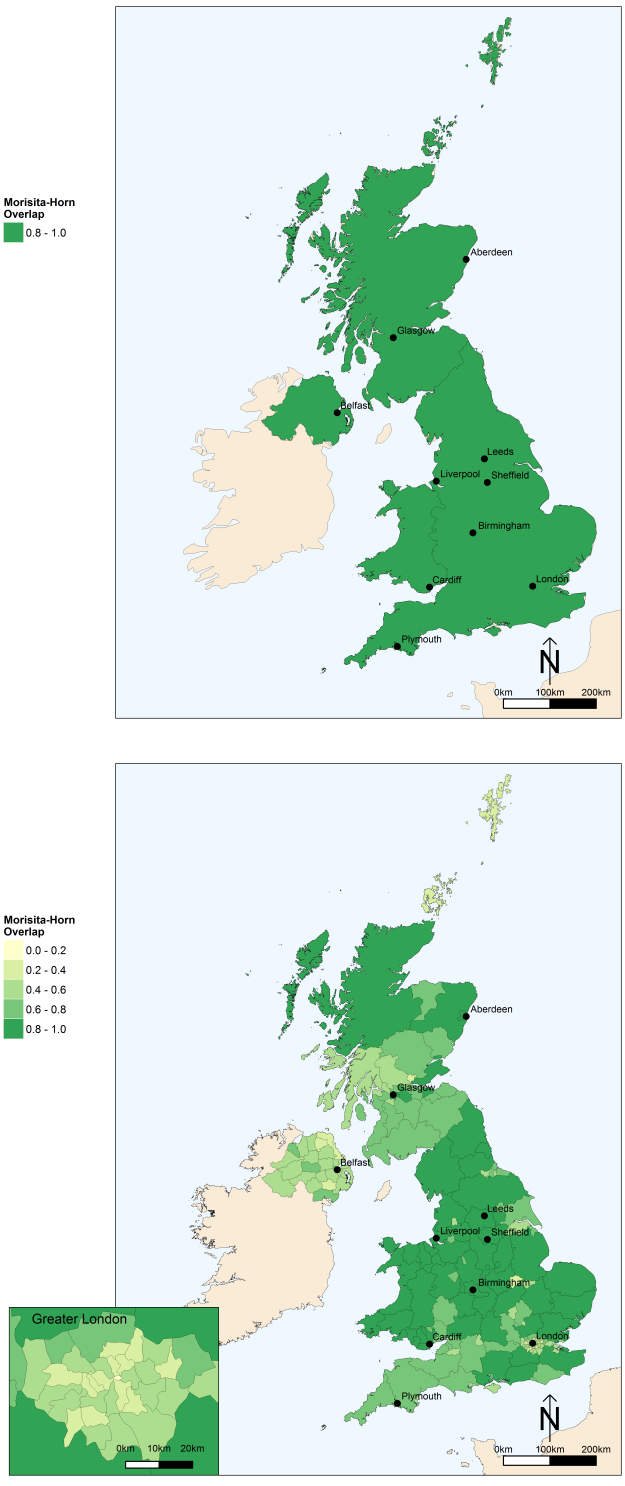


Figure 4.12: Maps of Morisita-Horn similarity analysis for the UK at GADM level 1 and 2.

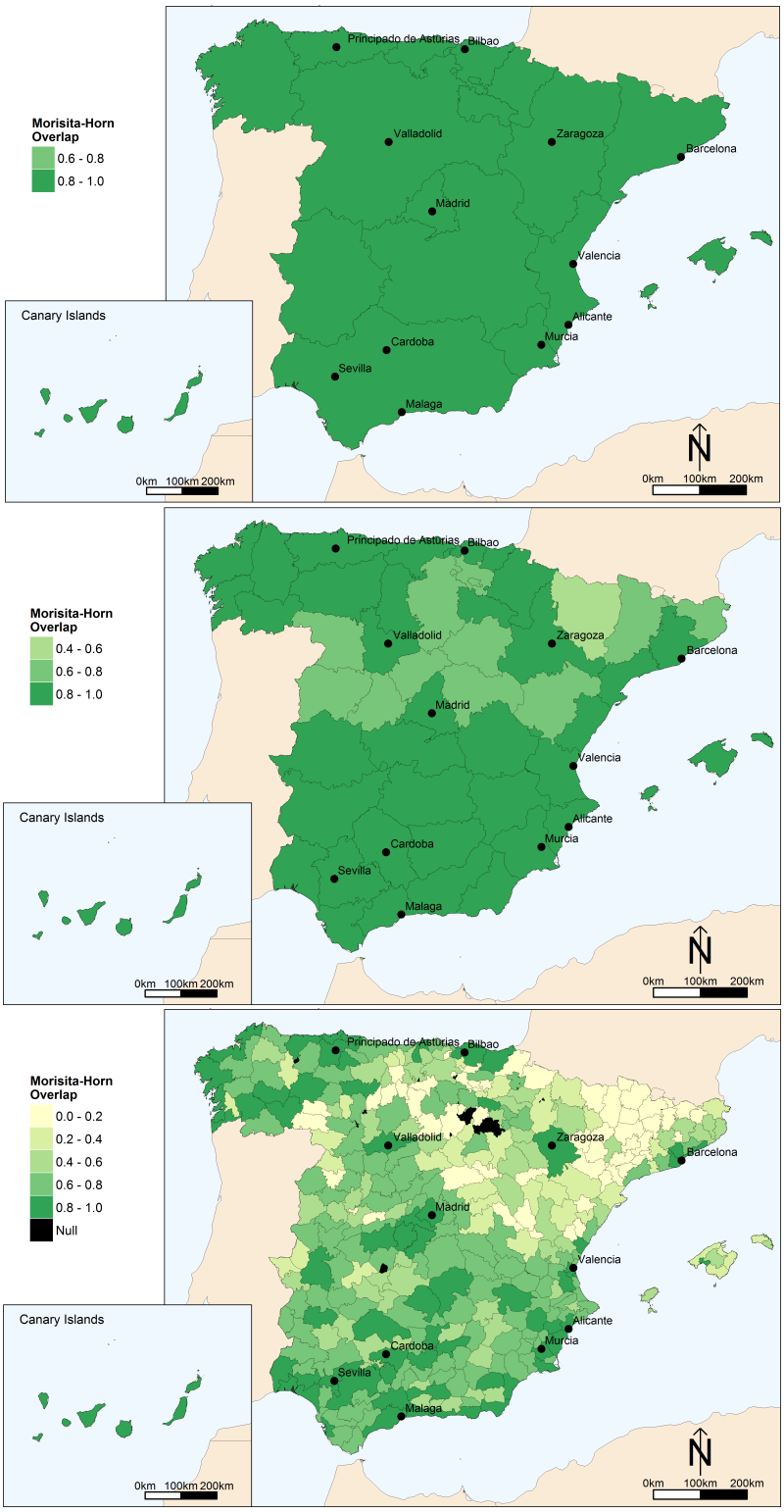


Figure 4.13: Maps of Morisita-Horn similarity analysis for Spain at GADM level 1, 2 and 3.

Table 4.6: Table of results of Morisita-Horn analysis for the UK and Spain at GADM levels 1, 2 (and 3 for Spain). n indicated the number of distinct regions at the specified scale.

UK					
GADM Level	Valid Locations	Min	Mean	Max	Null
0 (n=1)	416,819		0.983		0
1 (n=4)	413,461	0.911	0.966	0.992	0
2 (n=192)	346,900	0.096	0.665	0.968	0

Spain					
GADM Level	Valid Locations	Min	Mean	Max	Null
0 (n=1)	261,131		0.914		0
1 (n=18)	248,190	0.677	0.859	0.938	0
2 (n=51)	241,706	0.568	0.824	0.935	0
3 (n=368)	227,546	0	0.555	0.923	19

4.5 Discussions

Returning to the framework outlined by Mitchell (2005), the final steps in our analysis are concerned with the interpretation of results and their subsequent interrogation. At the outset, it was decided that the new individual level population inventories should take the form described by the United Nations population register definition. The definition specified a base list of 'each member of the resident population of a country or area' and subsequently the attribution of location and other identifying characteristics. In practice, however, the nature of the Twitter data required a slightly more nuanced approach. More specifically, there is not initially an explicit list of individuals who may be considered as the base population. Rather, what may be drawn from the data is a distinct list of user identifiers and screen names associated with these identifiers. As Twitter allows its users to change their screen names, there was an issue of identifying which, if any of the screen names were correct. Thus, in meeting the UN definition, a two-pronged approach was required. The first prong being the identification of those individuals believed to be the 'resident population' and the second being the extraction of the users' actual names from the data that were available. Once complete, the two outputs were combined into a functional individual level population inventory in which each user was attributed with a unique

identifier, forename, surname and location.

The identification of users' residential locations was performed first as it allowed for the feasibility of the register creation framework to be assessed. Also, by omitting those individuals who would not have been assigned a location, the number of users whose names had to be extracted was reduced. As is reported in Table 4.3, the allocation of users to an administrate geography using the proposed methodology appeared to be successful with an overall classification accuracy of 83.9% at GADM level 2 when users were assigned to the GADM geography in the UK.

However, the location estimation method was not without its flaws. A possible limitation of the approach used was the incorporation of all Tweets irrespective of the time-of-day or day of the week that they were submitted. It may be argued that inclusion of Tweets submitted during standard working hours (8 am – 6 pm, Monday - Friday) could have introduced a bias towards individuals' work locations and away from their areas of residence. This issue is evident in Figure 4.11 that shows the LQ of Twitter users' versus the reference population in the UK. In The City of London, the LQ is 8.77 indicating that there are 8.77 times the Twitter users than would be expected given the reference population. This observation is undesirable yet unsurprising. According to 2014 data published by the Greater London Authority, the daytime population of the City of London (excluding weekdays in term-time excluding tourists) was 431,384 while the residential population was just 7,947. When we compare this with the results of the LQ analysis, where the LQ was 8.77 when comparing the Twitter inventory to the reference, the equivalent LQ (workday versus residential population) would have been 54.3. Therefore, the degree of over-representation was far less than might have been expected.

In many respects, the issues identified here may be attributed to the effects of scale and the division of space into areal units. A case in point is Greater London which is disaggregated into its 33 administrative Boroughs. With this in mind, it is important that we remain aware of the purpose for which the inventories are being composed. Previously there was discussion regarding the creation of the Onomap classification of individuals' cultural, ethnic and linguistic groups, and also of the

UCL Worldnames Database. In the case of Onomap, the classification does not accommodate geographic context; in the case of the Worldnames Database, the highest level of geography is NUTS3 that divides the UK into 139 regions; less than the 192 in the GADM dataset.

Use of a less granular geography has some advantages. First, as the granularity is reduced so are the numbers of journeys individuals are likely to make between regions. Thus, assuming users tweet at their location of residence, a greater proportion of their total Tweets will originate from the same area. Table 4.6 provides evidence that the decision to assign individuals to locations at multiple spatial resolutions was good. The alternative was to simply aggregate those users who were located at the highest spatial resolution. Use of the tiered assignment resulted in an increase in inventory inclusion of 16.8% and 7.4% for the UK and Spain respectively when going from GADM 2 to 0.

While the focus of the discussion has focused predominantly on the UK, many of the features discussed in the UK are also evident in the Spanish case study. However, unlike in the UK, Figure 4.11 suggests a significant spatial trend for both GADM levels 1, 2 and 3 in Spain. Moving from the Southwest to the Northeast the LQ shifts from notable over-representation to notable-under-representation. One possible explanation for this, as the mapped data would suggest, is that use of Twitter is lower in the Northeast of Spain and higher in the Southwest. An alternative possibility is that there may be regional variations in the reference data coverage. Unlike the consumer register used in the UK analysis, a national telephone directory is used in Spain. The size of these two population inventories is 54.29 and 10.4 million equivalent to 85% and 22% of the populations respectively for the UK and Spain. It might be argued that while the compositional similarity analysis is performed between the existing Worldnames data and the Twitter inventories that the LQ analysis is conducted against Census data or national population estimates.

Figure 4.13 that maps the Morisita-Horn similarity scores in Spain shows a similar spatial pattern to that observed in Figure 4.11. Once again, it would appear that the Twitter population are fairly representative of the population in the Southeast

and less representative in the Northwest. As in the discussion of LQ, there is a possibility that this pattern is a consequence of the uneven representation of the population in the telephone directory.

4.6 Conclusion

The objective of this chapter was to develop a framework for the creation of proxy population inventories derived from the geotagged Twitter data, the format of which was based on the definition of population registers by the United Nations. To this end, the objective has been achieved with the creation of feasible population inventories for both the Spain and the UK. However, while successful in principle, it was evident that the register creation framework was heavily dependent on the quality and volume of underlying data.

Consequently, in the knowledge that Twitter usage in the UK and Spain is relatively high, it is fair to presume that the inventory creation framework is unlikely to be so successful across all countries for which new inventories are required. Regarding one of this thesis' aims to supplement a seamless Worldnames Database, this outcome is disappointing. However, while the goal of complete global coverage may not be attainable, it would appear that there are still opportunities for inclusion of Twitter-based population inventories to address missing data within the database. Thus, moving forwards, the next step in for this thesis is the creation of equivalent individual level Twitter population inventories for all countries for the purpose of increasing the Worldnames Database coverage. Initially, the existing Worldnames Database inventories will be used for the validation of the proxy inventories' validation. Once the Worldnames data is exhausted as a reference, new methods, which will form the bulk of the following chapter, will seek to infer the performance of the Twitter inventories in their absence. The goal in this analysis will be to provide new population inventories where country data are not presently available.

Chapter 5

Towards a Seamless Worldnames Database

5.1 Introduction

Following the successful creation and validation of the individual level Twitter-derived population inventories, the focus of this chapter is the application of the method on the global scale. The aim is twofold. First, to supplement the existing Worldnames Database and second, to better understand the global geography of Twitter providing a point of reference for the future analysis of Twitter data. As motivation, an supplementation of the Worldnames Database offers several distinct opportunities. First, the inclusion of additional data has the potential to improve accuracy and precision within the database. In particular, regarding the geographic bias introduced as a consequence of limited data coverage. In its present form, the Worldnames Database tells us which of the Worldnames' countries an individual is most likely to originate from rather than the actual country from which they have originated. Second, the availability of individual-level names data, as opposed to aggregate surname counts, may prove advantageous in the development of the Onomap CEL classification tool. Finally, the creation of a global population inventories, inclusive of key identity and mobility attribution may provide a new platform to investigate the stocks and flows of human populations at a range of spatiotemporal scales.

Regarding the goals of the chapter, the approach should be considered in two phases. In the first the goal is to use the population inventories already present within the Worldnames Database to validate the Twitter population inventory creation framework. The assumption here is that the existing data are the most complete and therefore the best choice of reference data. This assumption is thoroughly tested in section 4.2.3. The second goal of the chapter is concerned with supplementing the Worldnames Database using the individual level population inventories derived from Twitter. It should be noted that when we refer to supplementing the Worldnames Database, this is in reference to adding data for countries where no existing data are currently available.

In seeking to realise these opportunities, it was recognised, based on work performed in the previous chapter, that the individual level Twitter-derived population inventories were not uniformly representative, with notable variation within and between countries. In the case of the two validation countries, Spain and the United Kingdom, it was possible to assess said variation through comparison against pre-existing national population inventories. However, in moving forwards, the dependence upon pre-existing data is unrealistic. Rather, the availability of fully inclusive individual level national population inventories is significantly limited, necessitating an alternative means of inventory validation. Consequently, it was proposed that a model is designed in such a manner as to inform the decision as to whether the Twitter-derived inventories were suitable proxies for conventional individual level population inventories, or not. In doing this, it was decided that the Morisita-Horn Index of Overlap, previously employed in quantifying the degree of overlap between sets of names, should be incorporated as the dependent variable in the analysis. In doing this, it is worth reiterating that the Morisita-Horn Index of Overlap is acknowledged for its ability to deal with samples of significantly differing size and diversity and also exhibits a linear progression (Wolda, 1981).

5.2 Methods and Materials

5.2.1 Regression Analysis

In producing the model, it was proposed that a multiple regression analysis be performed with the aim of developing a parsimonious model capable of inferring the probable representative ability of each national scale individual level population inventory. The modelling process would be divided into a series of distinct phases: variable identification; variable preparation; model selection; model diagnostics; and model application. In each phase, background and justification is provided for the methods and data employed.

5.2.2 Variable Identification

As intimated, the first phase of the modelling process was the identification of a collection of potential independent variables. In seeking to identify said variables, efforts were made to source literature that identified national level environmental and cultural factors associated with the adoption of online social networks and where possible, Twitter. Given the volume of literature published regarding the analysis of online social networks, only a limited number had considered or sought to identify such factors. It is presumed that this is due to the tendency for analyses to be focused on limited geographic extents such as major world cities or singular countries. Nonetheless, one study, completed by Jan et al. (2015), based on Malaysian Muslims' adoption of online social networks, proposed a conceptual framework of national level factors which may be associated with the adoption of online social networks. The study considered both practical factors such as access to the Internet and cultural factors such as the use of online social networks by brands for customer engagement. The main branches of the framework were Social, Technological, Educational and Brand/Product communication. These factors were drawn from five general themes: Demographics and Socioeconomic; Social Information Sharing; Technological Advancement; Knowledge Allocation; and Product and Brand communication. In seeking to incorporate the above factors, each was explored and, where appropriate, potential variables identified.

In the identification of the variables, it should be highlighted that the major barrier to incorporation was the availability of data published in a consistent format between countries. Such a constraint placed a significant limitation on possible data providers. Given this constraint, it was decided that the World Bank DataStore (see http://www.data.worldbank.org) provided the best possible data resource. The decision was based on the large number of variables available, the global coverage of the data, and the availability of detailed source, methodology and attribution data.

Social: Demographics and socioeconomics

In outlining their conceptual framework, Jan et al. (2015) highlighted demographic and socioeconomic factors as the fundamental drivers of individuals' adoption of online social networks. They discussed how both age and gender play a major role in how and for what purpose individuals use online social networks. For example, in the context of users' privacy preferences, Li and Chen (2010) note that females tend to express a greater degree of concern regarding their online privacy and consequently, are less inclined to share their personal information online. The regard for privacy is also evident regarding age with older users being less willing to disclose their personal information.

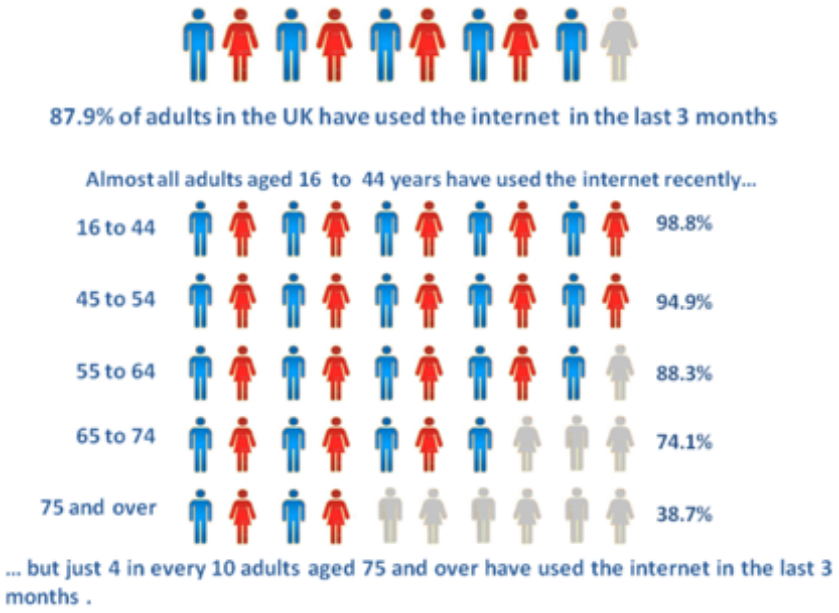


Figure 5.1: Graphical depiction of Internet usage by age in the United Kingdom (ONS, 2016)

Regarding individuals' access to the Internet, age also plays a role. As may be evident in Figure 5.1, a graph of Internet use by age in the UK, Internet use decreases notably with age. While Internet usage in the 16 to 44 and 45 to 54 age bands is $\geqslant$ 94.9%. this figure rapidly declines with only 38.7% of over 75s making regular use of the Internet. Thus, assuming that such a trend extends beyond the UK, it may be deduced that a country with a younger population will possess a greater proportion of the population with access to the Internet. That said, this assertion does not account for differences between economies and the more general states of development. Consequently, it was believed that such an indicator, while interesting, should it be included, may introduce unnecessary uncertainty which is not appropriate given the limited number of observations available.

Considering next the quantification of socio-economic development, a commonly employed metric is the Gross Domestic Product (GDP). GDP, in its simplest form, is a measure of the total economic output of a country for a given period. In the case of the World Bank data, GDP is considered a measure of the sum of value generated by the population plus product taxes and minus the subsidies not associated with the aforementioned products. In seeking to address the issue of differences in population size, GDP is commonly reported on a per capita basis. Though standardised, this metric is not necessarily appropriate as it fails to account for the relative purchasing power per monetary unit between countries. Consequently, a more appropriate statistic is GDP per capita standardised by purchasing power parity (PPP). With the data normalised to account for purchasing power, this is the most appropriate technique to account for individuals' abilities to purchase technology or access to utilities such as the Internet. GDP per capita with PPP at current international dollars data are available for the World Bank DataStore. It should be noted that while the data are for 2013, the "current" international dollars refers to 2011.The data may be accessed at the following URL: data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD and are credited to the World Bank, International Comparison Program Database.

Technological

The technological theme is concerned with specific developments that have enabled

individuals to access the Internet and by proxy, online social networks. Examples of such factors include Internet access and cell phone ownership. Access to the Internet is an obvious facilitator of social media uptake. Unsurprisingly, without some form of access to the Internet, such as via desktop terminals or mobile phones, individuals are unable to utilise said services. Internet access data are available via the World Bank which reports the proportion of the population per 100 with access to the Internet via a range of Internet-capable devices which includes cell phones, desktop terminals and games consoles. The data may be obtained at the following URL: data.worldbank.org/indicator/IT.NET.USER.P2 and are credited to the World Bank, International Comparison Program Database (International Telecommunication Union, 2013a).

Alongside Internet Access, mobile phone usage is a potentially valuable indicator. As stated previously, a large proportion of individuals access online social networks via their mobile devices. In 2014 Twitter reported that 80% of UK users accessed the service via their mobile devices and 70% used mobile devices as their primary means of engagement. The equivalent percentages for Spain were 80% and 69% respectively and 68% and 60% for France (Twitter, 2014). More recently, in 2016, Twitter reported that 83% of its traffic was via mobile devices (Twitter, 2016). While smartphone ownership data are not explicitly available via the World Bank, data are available on the number of cell phone contracts per 100 of the population. The data cover both prepaid and postpaid subscriptions on devices that can transmit voice. These data are unfortunately not constrained to those devices with access to the Internet as might be desirable. The data may be accessed at the following URL: data.worldbank.org/indicator/IT.CEL.SETS.P2 and are credited to the World Bank, International Comparison Program Database (International Telecommunication Union, 2013b).

Educational

Regarding education, Jan et al. (2015) discuss the shift of individuals from conventional news content providers towards online social networks for the purpose of consuming and publishing news media and educational content. For example, many

individuals use the Twitter social network to harvest news from a range of individuals who can meet their niche requirements. Hermida et al. (2012) note that as individuals become increasingly familiar with online social media, so their confidence in crowd-based information increases. In a similar sense, this shift in the consumption of online information may be observed in growing acceptance of Wikipedia, an online crowd-sourced encyclopaedia. Unfortunately, data regarding the use of social media for the consumption of news and education are not available via the World Bank, and thus a variable is not proposed. That said, the influence of media consumption is recognised and may be employed for the purpose of interpreting the model outcomes.

Brand/Product communication

Regarding brand and product communication, Jan et al. (2015) discuss the shift of major brands to online social networks. The premise is that as brands work to increase their social profiles through marketing, they will promote the uptake of online social networks. Such is the power of social media that brands are investing increasingly large sums of money on targeted advertising seeking to exploit the availability of demographic data as a means to explicitly market their products and brands. Regarding the identification of potential model variables, brand and product communication has limited potential due to the lack of a consistent national indicator of brand expenditure or investment. Theoretically, a measure of social capital, the value of the social media services, may be quantified but this in itself could form the basis of an entire thesis.

At a coarse scale, data from eMarketer, an online market research company, shown in Table 5.1, provides a useful indicator of the geographic variation in social media investment. While the data are not standardised by population, the difference in investment by region is stark. Case in point is the difference in spending per capita between North America and the 'Middle East and Africa'. With populations of 352 and 409 million in 2013 respectively, investment per capita equates to $14.03 versus $0.17. While this calculation does not account for PPP, a clear difference is evident. Should this data have been available at the national or sub-national level, this may

have proven to be a valuable addition to the pool of potential variables.

Table 5.1: Spending on social media advertising by region (source: eMarketer.com, 2015)

	Social network ad spending ($ billions)		
	2013	2014	2015
North America	4.94	7.71	10.10
Asia-Pacific	3.25	5.18	7.40
Western Europe	2.34	3.68	4.74
Latin America	0.35	0.54	0.68
Central & Eastern Europe	0.41	0.52	0.61
Middle East and Africa	0.07	0.11	0.16
Worldwide	11.36	17.74	23.68

Other factors

While the conceptual framework of Jan et al. (2015) provided some useful direction in regards to the identification of parameters, it remains the case that they are not unique to the Twitter social network and further, do not account for the effects of population naming structure.

If we consider first, the lack of specificity towards the Twitter social network, the conceptual framework of Jan et al. (2015) does not account for the unique geographic distribution of adoption of each social network. There is, therefore, a requirement for a variable or variables able to capture individuals' specific propensity towards the Twitter network. Of the publicly available data that may address this, arguably the most relevant are social media uptake or social media penetration statistics. In 2013, PeerReach, an online market research organisation, published Twitter penetration statistics for the 23 countries in which there were greater than 800,000 active users (PeerReach, 2014). The penetration was reported as the percentage of Twitter users versus the size of each countries online population. These statistics, while interesting, are of limited use due to their limited coverage. Nonetheless, by virtue of the data compiled in the creation of the Twitter inventories, it was possible that a proxy statistic be produced able to provide a similar indicator of social network penetration.

While actual counts of active users are not available, counts of users identified within each country as part of this thesis are. Consequently, a modified penetration

measure was proposed in which the number of users identified within each country was dived by the total population with access to the Internet. In this case, both population and Internet usage statistics were sourced via the World Bank. A second version of the above is also proposed in which the data are not standardised to account for Internet usage. The purpose of this simplified measure is to avoid the introduction of bias in those countries where a small proportion of the population have access to the Internet. For example, in Eritrea where fewer than 1% of the population have access to the Internet, the effect of only a few Twitter users would be significantly magnified. In effect, the assumption is that the new statistic would provide an indication of the proportion of the population who are users of Twitter and that the greater the percentage of the population to use Twitter, the better their names will represent the population.

Beyond those factors associated with social network adoption, it is also recognised that certain features of the population structure might impact on how representative the individual level Twitter-inventories are of the underlying population. For instance, the effect of national-scale surname diversity on the number of names required to depict the population accurately. Thus, efforts were made to identify population factors which may impact upon inventory performance. Considering the effect of surname diversity, the assumption is made that the lower the overall diversity of names at the national level, the fewer the total number of individuals that will be required to represent the population accurately. Case in point is China where just 100 surnames account for approximately 85% of the total population (Liu et al., 2012). With so few names and the consequent higher proportion of the population to bear these names, the surnames have a higher likelihood of being correctly stratified. To put this in perspective, based on data from the 2013 Consumer Register, 18,088 unique surnames are required to account for 85% of the population. Consequently, it may be assumed that a greater number of individuals would be needed to depict the population name structure accurately.

In seeking to understand this relationship better, an investigation was performed to determine the association between the number of names, surname diversity and

between-group surname similarity. For each country included in the analysis, ten samples of size 10, 100, 1,000 and 10,000 were drawn, and the similarity between each of the samples and the respective reference population calculated. In some senses, the results generated may be considered as a theoretical maximum similarity score based on sample size given that the samples are being drawn from the primary population. Subsequently, through plotting the results, it was possible to examine the relationship between the three features.

Figure 5.2, presents the outcome of the investigation as a faceted box plot split by country and sample size. The countries are ordered by surname diversity as calculated using the top 1,000 names from top to bottom with colour providing an indication of the surname diversity value. The plot provides a valuable illustration of the effect of both sample size and surname diversity and in turn, the effect on between-group similarity.

If we consider first the effect of sample size, it is evident that as the size of the sample is increased, the mean similarity score increases with the biggest difference occurring in those countries with the lowest surname diversity. This observation supports the hypothesis that in countries where surname diversity is low that fewer individuals are required to depict the population accurately. The second observation is that as the size of the sample is increased, the variation in the calculated similarity values decreases. In effect, as the sample is increased, so the impact of chance is reduced. Regarding the analysis, the results have a number of implications. Most significantly, it is clear that both sample size and surname diversity have an impact on the between-group surname similarity. For this reason, it is proposed that both surname diversity and number of valid Twitter users be included in the set of candidate parameters.

Regarding the inclusion of surname diversity, it should be recognised that the values are calculated based on the existing Worldnames Database inventories. Consequently, should surname diversity prove to be a significant factor, an alternative source of data will be required. This, as has been discussed previously remains a major constraint to the parameter's inclusion. At the time of writing, the

Figure 5.2: Faceted box-plots showing distribution of Morisita-Horn values when samples of given size are taken from the reference dataset. Each box is coloured based upon the proportion of the population represented by the top 1,000 most common surnames.

single globally inclusive database of common surnames and their frequencies is http://www.forebears.io. However, while the data are published in such a manner

as to promote confidence, the provenance of the data across countries remains questionable. Nonetheless, given the potential significance of the parameter, the diversity measure, based on the proportion of the population covered by the top 100 names will be included in the model selection process. The reason to choose 100 names is that Forebears.io, the potential source of global names data, only publishes the top 100 or 200 most common names for each country.

Final variables

Having investigated the factors which may be relevant to predicting the Twitter inventories utility, the following set of variables were identified and, where necessary, derived. The variables are:

- X_1 Proportion of users versus the national population.
- X_2 Proportion of users versus the national population with access to the Internet.
- X_3 GDP per capita adjusted for purchasing power parity (Int. Dollars).
- X_4 Number of cellular subscriptions per 100 of the population.
- X_5 Surname diversity within the top 100 names. The higher the value, the lower the surname diversity.
- X_6 Number of Twitter users identified.

The dependent variable in the analysis, Y, is the calculated Morisita-Horn Index of Overlap employed in the previous chapter to quantify the relationship between the individual level Twitter and reference population inventories. The observations for Brazil, Bulgaria, Canada, India, New Zealand, Poland and Switzerland, marked with asterisks, are omitted from the modelling due to their questionable provenance (discussed in Section 4.2.3).

Considerations in the use of the Worldnames Database Data

A key consideration within the subsequent analysis is the reliance on existing Worldnames Database population inventories for the purpose of validating the Twitter-based inventories. Of particular relevance is the calculation of the national-level

Morisita-Horn Index values employed in this chapter. When calculating the national-level Morisita-Horn similarity index, there is an assumption that the Worldnames Data are an accurate representation of the population naming structure. Should this structure be incorrect, this will introduce uncertainty into the similarity comparison and potentially impact upon the reported similarity scores. A number of explanations exists as to why the reliability of the Worldnames population inventories may be of questionable provenance. These issues include that the data have been collected at different periods in time, the data are often compiled from different sources, the data may be collected in an inconsistent manner, the data may possess some degree of demographic bias, and finally that the statistics employed may be affected by differences in population sample size and diversity. While it is not possible to make an explicit validation of these data, various steps have been taken to ensure the provenance of the data.

In the first instance a audit was performed to assess each Worldnames Database Population inventory against a series of key criteria. The objective of the exercise was to determine the source of the data, the time of collection, the number of individuals reported within the data and also, the correlation between the most common names and alternative sources of names data. Knowledge of such information is key to making informed observations based on the data being analysed. The results of this benchmarking exercise are reported in appendix B.

In terms of of sample size and diversity, specific consideration was given to the choice of a suitable similarity measure. Specifically, the decision was made to use the Morisita-Horn index of overlap. Of the various similarity measures available, the Morisita-Horn index was recognised in the literature for being independent of sample size and diversity (Wolda, 1981). Given that both the Twitter-derived and Worldnames individual-level population inventories vary dramatically in terms of size and diversity this decision was critical to the effective calculation of similarity measures. The specific features of the Morisita-Horn Overlap Index were examined in Chapter 4, Section 4.3.2.3. Regarding the use of the most frequently occurring surnames as an indicator of an inventories likely provenance, it should be remem-

bered that surnames typically conform to a power law distribution. That is to say that the most commonly occurring names are typically well stratified and consistent in order. This feature of surname distribution is highly valuable given that most publicly available data on surname frequencies is constrained to a limited set of the most frequent names. Typically the top 10, 100 or 500.

While various limitations have been raised, it should be remembered that the data within the Worldnames Database are arguably the most complete and comprehensive the are currently within the public domain. Further, given our prior knowledge of the data, and the results of the benchmarking exercise, we can employ the data with a degree of confidence.

5.2.3 Variable Preparation

Having identified a suitable pool of variables, the next phase in the analysis was variable preparation. The objective in this phase being to develop an understanding of the data and, if necessary, make any alterations or transformations to aid in their application and interpretation. The raw consolidated input data for the model are shown in Table 5.2.

Table 5.2: Table showing the raw data to be employed in the model selection process. Those data marked with an asterisk are to be omitted from the model creation framework having been of questionable provenance.

ISO3	morisitaHorn	usersByPop	usersByOnlinePop	gdppcintdol	cellular2013	usercount	surnamDiv100
ARG	0.845	0.001587	0.002649	22,404.26	162.53	67,489	0.302
AUS	0.838	0.001502	0.001809	45,476.98	106.84	34,719	0.225
AUT	0.324	0.000324	0.000402	47,416.29	156.23	2,750	0.129
BEL	0.712	0.001197	0.001457	43,057.20	110.90	13,384	0.073
BGR*	0.521	0.000149	0.000280	16,573.47	145.19	1,081	0.462
BRA*	0.175	0.001184	0.002321	15,816.80	135.31	241,944	0.757
CAN*	0.788	0.002173	0.002533	44,281.27	80.61	76,409	0.114
CHE*	0.663	0.000542	0.000628	59,351.42	136.78	4,387	0.136
DEU	0.675	0.000219	0.000260	44,184.82	120.92	7,976	0.128
DNK	0.722	0.000795	0.000840	45,681.11	127.12	4,465	0.533
ESP	0.914	0.004626	0.006457	32,842.43	106.89	215,644	0.396
FRA	0.548	0.000636	0.000777	39,157.67	98.50	41,984	0.079
GBR	0.983	0.005514	0.006138	39,111.23	124.61	353,629	0.203
HUN	0.509	0.000252	0.000347	24,037.17	116.43	2,491	0.289
IRL	0.946	0.004911	0.006277	47,599.68	105.48	22,584	0.364
ITA	0.866	0.001040	0.001779	35,707.83	158.82	62,632	0.065
JPN	0.790	0.000144	0.000161	36,618.31	116.32	18,395	0.344
LUX	0.050	0.000431	0.000459	95,928.60	148.64	234	0.084
MLT	0.594	0.000716	0.001039	29,525.71	129.76	303	0.780
MNE	0.396	0.000472	0.000782	14,623.75	159.95	293	0.329
NLD	0.853	0.002813	0.002994	47,954.50	116.16	47,273	0.125
NOR	0.816	0.001605	0.001689	66,817.17	116.27	8,154	0.246
NZL*	0.799	0.001184	0.001431	37,096.10	105.78	5,261	0.182
POL*	0.363	0.000086	0.000137	24,493.76	149.08	3,267	0.097
SRB	0.393	0.000540	0.001048	13,668.12	119.39	3,866	0.238
SVN	0.195	0.000298	0.000409	29,097.58	110.21	613	0.117
SWE	0.878	0.002288	0.002414	45,067.44	125.53	21,967	0.359
USA	0.886	0.003869	0.004595	52,660.30	97.08	1,224,333	0.160

In seeking to understand the data and variable relationships better, a scatter plot

matrix was created. Shown in Figure 5.3, the plot provides a range of useful insight including variable density distributions; paired scatter plots with the OLS line of best fit and the Loess line; and also, a naive R-squared value and indicator of significance. It should be noted that in the case of the R-squared and significance indicators, that the regression assumptions have not been formally verified and consequently, the values must be considered only as guides.

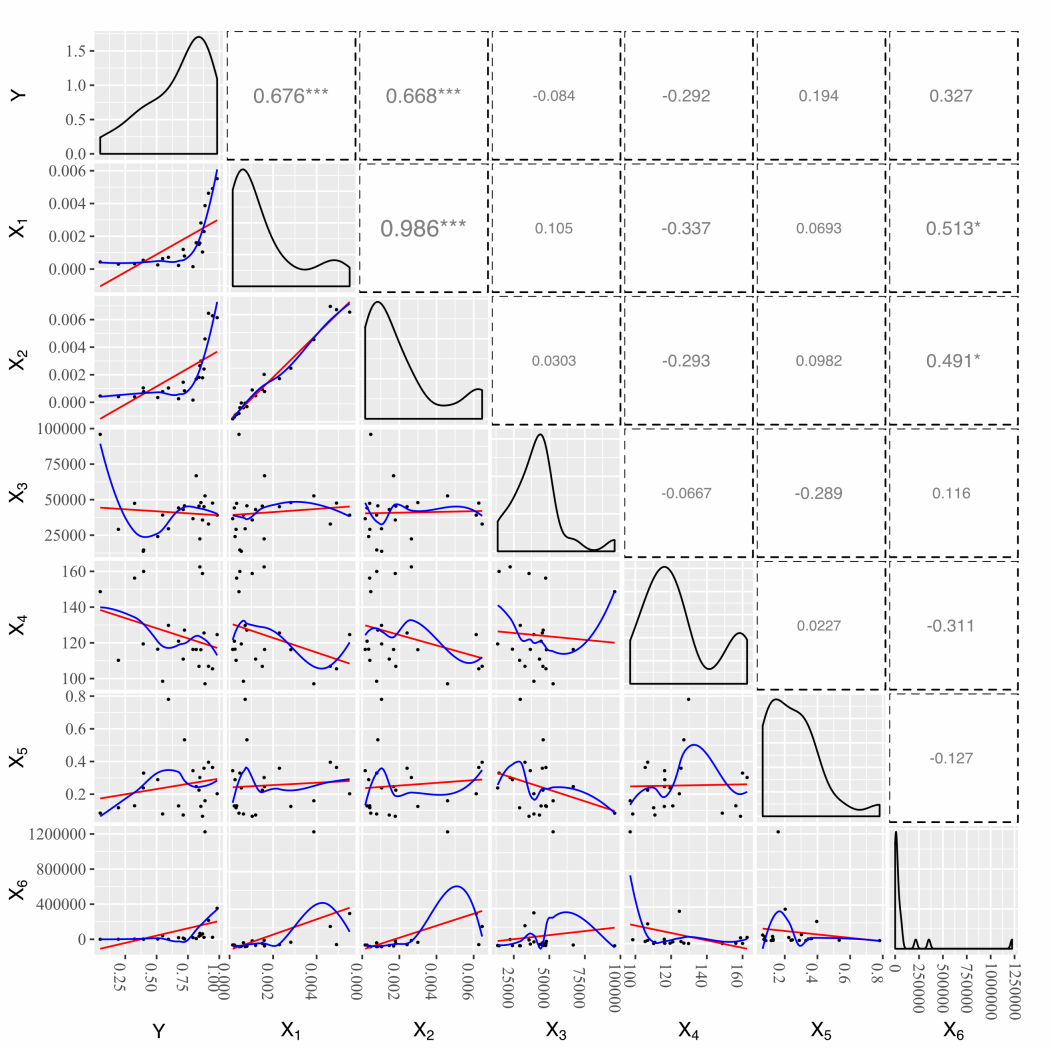


Figure 5.3: Scatter plot matrix of the dependent and independent variables identified as being related to social media uptake and surname structure. The matrix includes the linear and Loess fits, the kernel density of each variable and a naive R-squared value.

If we first examine the diagonal axis, it is evident that each of the variables exhibit some degree of positive or negative skew with notable negative skew in the

case of Y and positive skew in the case of parameters X_1 through X_6. Further, in the case of parameters X_1 through X_6 there appears to be some degree of bi-modality. Considering next, the area below the diagonal, the scatter-plots provide a useful indicator of the between-parameter relationships. In the case of Y and both X_1 and X_2, there is clear evidence of a curvi-linear relationship suggesting a potential association. The relationship between Y and X_6 is also apparent. Lastly, in the area above the diagonal, we can see observe an apparently strong correlation between Y and both X_1 and X_2. Also, we can see there is a very strong positive relationship between X_1 and X_2, not surprising as they are both derived from the same data. Furthermore, there is evidence of a weak positive relationship between X_6 and both X_1 and X_2. Again, it should be noted that X_1 and X_2 are standardised versions of X_6 and thus it is not appropriate that both parameters be incorporated into the final model.

Having examined the input parameters, attempts were made to normalise the data such that the linearity between the dependent and independent variables might be improved. In support of the visual inspection of the data, the Shapiro-Wilk's test of normality was employed as a formal means of assessing whether the data deviated significantly from the assumptions of normality. In this analysis, the null hypothesis was that the samples were drawn from normally distributed populations. Given an α of 0.05, where the p-value < 0.05 then the null hypothesis may be rejected. Where the p-value is $\geqslant 0.5$, it is not possible to reject the null hypothesis that the data are drawn from a normally distributed population.

Table 5.3: Table showing the summary statistics from the Shapiro-Wilk's test of normality for the potential model variables.

Variable	W	p-value	Null Hypothesis
Y Morisita-Horn	0.89787	0.02695	Reject
X_1 Users By Pop.	0.80093	0.0005175	Reject
X_2 Users By Online Pop.	0.79497	0.0004178	Reject
X_3 GDP pc ppp	0.89143	0.02011	Reject
X_4 Mobiles per 100	0.88358	0.01415	Reject
X_5 Surname Diversity	0.87229	0.008639	Reject
X_5b Surname Diversity	0.85936	0.004992	Reject
X_6 Raw User Count	0.40186	1.777e-08	Reject

Table 5.3 presents the output of the Shapiro-Wilks tests including the W and associated p-values. As may be evident, all of the variables tested have p-values $<$ 0.05 and thus must be considered to be drawn from a non-normally distributed population. Consequently, seeking to normalise the data, an exponential transformation was applied to Y and natural log transformation applied to parameters X_1 through X_6. The Shapiro-Wilk's test was subsequently repeated and the outcome investigated.

Table 5.4: Table showing the summary statistics from the Shapiro-Wilk's test of normality for the post-transformed potential model variables.

Variable	W	p-value	Null Hypothesis
$\exp Y$ Morisita-Horn	0.933	0.1417	Accept
$\log X_1$ Users By Pop	0.96366	0.5667	Accept
$\log X_2$ Users By Online Pop	0.96561	0.6102	Accept
$\log X_3$ GDP pc ppp	0.93455	0.1526	Accept
$\log X_4$ Mobiles per 100	0.91405	0.0573	Accept
$\log X_5$ Surname Diversity	0.96418	0.5782	Accept
$\log X_5 b$ Surname Diversity	0.96744	0.6519	Accept
$\log X_6$ Raw user count	0.9668	0.6371	Accept

Shown in Table 5.4, it is evident that the transformations have effectively normalised the data. For each of the model parameters the p-value > 0.05 meaning that the hypothesis that the data were drawn from a normally distributed population may not be rejected. Consequently, the transformed data were re-plotted and the relationships between the variables re-examined.

Figure, 5.4, the scatter-plot matrix of transformed variables clearly illustrates the effect of transformation. Notably, there was an apparent improvement in the relationship between Y and both X_1 and X_2 and a dramatic improvement in the relationship between Y and X_6. A further observation, in the case of Y and both X_1 and X_2, was the emergence of several outliers deviating from the main distribution. These outliers were Japan, Luxembourg and Germany.

5.2.4 Model Selection

In possession of a suitable collection of explanatory variables, the next phase of the analysis was the model selection. In this phase, the objective was to identify a suitably parsimonious model in which model complexity and explanatory power

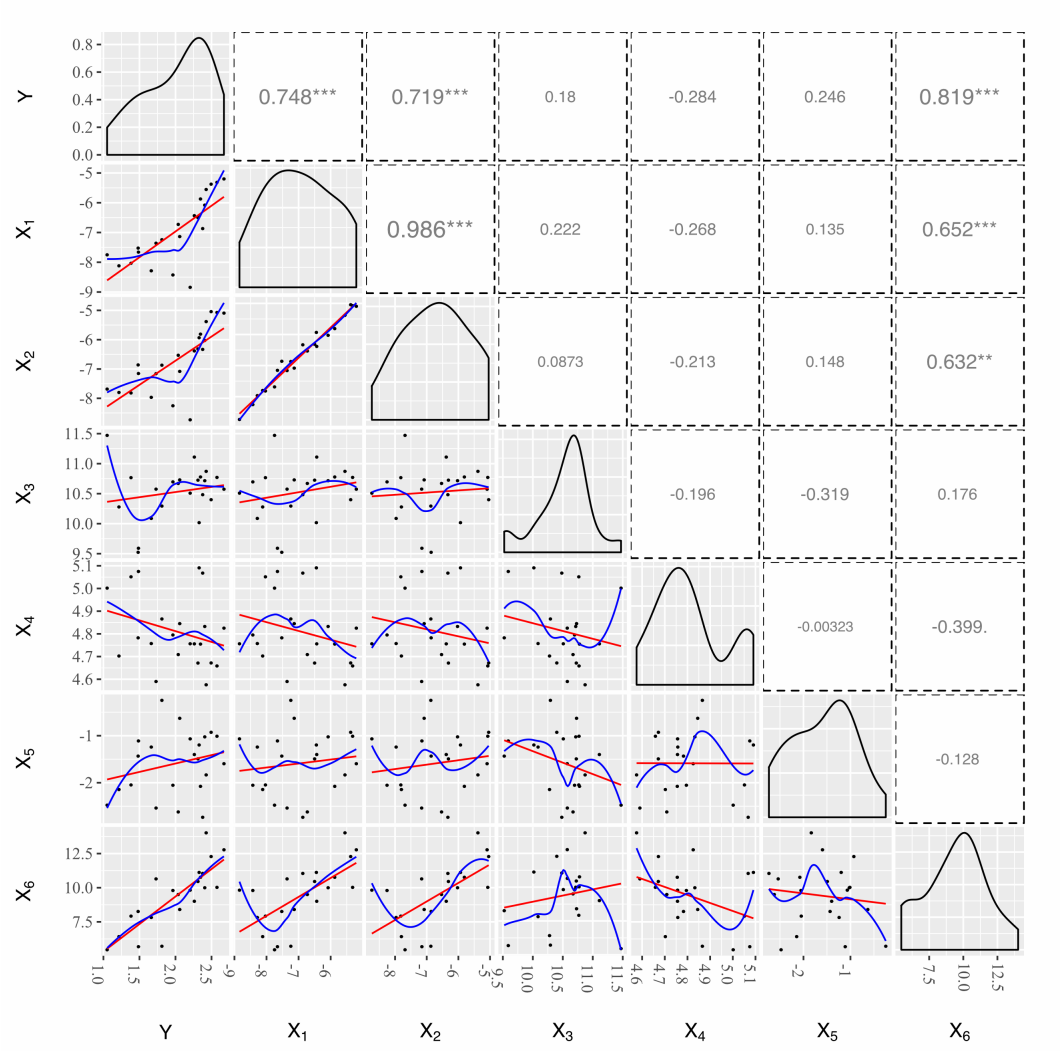


Figure 5.4: Scatter plot matrix of the transformed variables identified as being related to social media uptake. The matrix includes the linear and smoothed fits, the kernel density of each variable and the naive R-squared value.

were balanced. In seeking to obtain the optimum model, a number of model selection techniques were considered. These methods included manual selection, step-wise regression and all-subsets regression. In each case, the objective was to identify the optimum model given the data available.

Manual selection, as the name implies, involves the analyst making the decision regarding the construction of models based on their own experience. While suitable in certain circumstances, for example in the case of few potential parameters, the manual selection technique is inefficient and lacks robustness. Consequently, it is increasingly common to employ autonomous or semi-autonomous model se-

lection techniques such as Step-wise and all-subsets regression. Step-wise regression, commonly employed in ecology, is based on either the process of the addition of less-correlated variables (forward selection), the removal of less-correlated variables (backwards elimination) or a combination of both. The approach is very efficient when working with large numbers of potential parameters. However, it is noted by Whittingham et al. (2006) that there are some significant limitations to such approaches. These include bias regarding parameter estimation, inconsistent implementations between statistical packages and an inappropriate focus and reliance on the optimum model. In preference, Whittingham et al. (2006) suggests the use of the all-subsets regression family for model selection.

All subsets regression involves performing a regression analysis on all possible combinations of independent variables as a means to identify the best possible model. The best model is identified using a range of model success criterion. Whittingham et al. (2006) notes that while all-subsets regression does consider all possible models, the technique should not be seen as a 'shot-gun' approach. Rather the approach should be considered as a systematic means to identify the best model based on a set of the variables that have been selected due to their known association with the phenomenon being studied. The all-subsets method has some advantages over the more conventional step-wise regression techniques. Notably, that the outcome is not impacted on by the order in which the model parameters are arranged. Nonetheless, while more thorough, the nature of the approach is far more computationally intensive than its step-wise equivalent, thus providing the motivation for some to choose alternative model selection techniques. Using all-subsets regression there are $(2^p) - 1$ possible models where p is the number of parameters (excl. the intercept), the number of possible models is exponential necessitating a simplified means of identifying the optimum model. For the six independent variables identified, the number of possible combinations is 63. Such is the exponential growth of parameter combinations that all-subsets regression is not necessarily suited to modelling where a high number of parameters are present.

In the case of both approaches, various model fit parameters are available which

enable the analyst to differentiate between models. These statistics include the R-squared, Adjusted R-squared, Mallow's Cp and the Bayesian Information Criterion (BIC). The adjusted R-squared value is a modification of the traditional R-squared and will only increase where a parameter improves a model more than may be expected by chance alone. Consequently, the adjusted R-squared is far better suited to comparing models with differing numbers of the parameters than the standard R-squared value. The Bayesian Information Criterion (BIC), like the adjusted R-squared, employs a penalty term such that models with more parameters are penalised. The model with the lowest BIC is preferred. Mallow's Cp is a model selection technique proposed by Mallows (1973) designed to assess model fit where models have differing numbers of explanatory variables. In the case of Mallow's Cp, the smaller the value, the more precision is exhibited. A rule-of-thumb in the interpretation of Mallow's Cp is that Cp will be equal to the number of model parameters.

Having considered both the potential strengths and weaknesses of the model selection techniques, the all-subsets approach was chosen. Given the number of observations and explanatory variables in the analysis, it was felt that this approach offered the most comprehensive solution. The all-subsets regression was implemented in R using the Leaps package (Lumley, 2009).

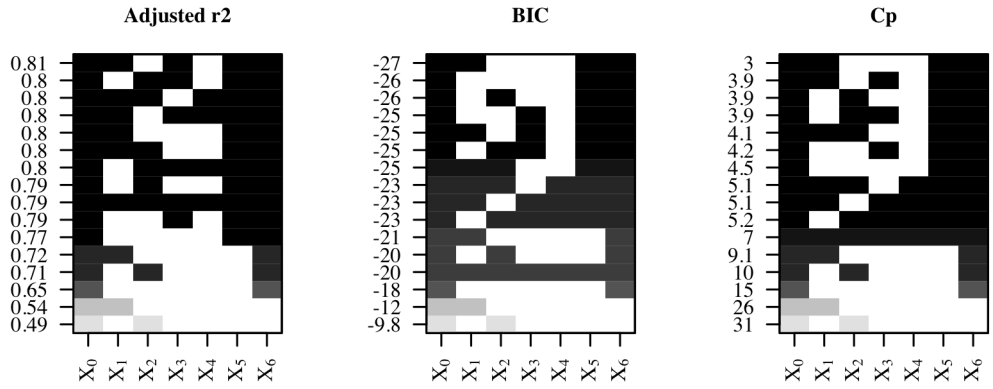


Figure 5.5: Results from the all-subsets regression analysis. The best three models for each subset size are shown based on each of the three model success criterion: Adjusted R-squared, BIC and Mallow's Cp.

Figure 5.5 provides a graphical illustration of the all subsets regression output

for each of the aforementioned model selection criterion. For each criterion, the three best models for each number of model parameters is shown. Regarding interpreting the plots it should be recognised that the regression assumptions are not explicitly verified, and thus it is not possible to determine the optimum model from the output alone. Consequently, it is necessary that each model is independently examined before acceptance. Furthermore, bearing in mind the limited number of observations (n=22), it is desirable that the number of parameters in the final model is kept to a minimum. While many rules-of-thumb exist regarding the minimum ratio of observations to independent variables, few are explicit. For example, Harrell (2001) suggests a minimum of ten observations per variable, while Austin and Steyerberg (2015) suggests as few as two observations per variable may be sufficient. In light of this, it is proposed that the final model should preferably have two or fewer and no more than three independent variables. Consequently, considering the output of the all-subsets regression in the context of the above, the following potential models are identified:

Based on the Adjusted R-squared plot, the optimum models appear to be $(Y=X_1+X_5+X_6)$, $(Y=X_3+X_5+X_6)$ or $(Y=X_5+X_6))$. Based on the BIC, the optimum models appear to be $(Y=X_1+X_5+X_6)$, $(Y=X_5+X_6)$ and $(Y=X_2+X_5+X_6)$. Finally, based on Mallow's Cp, the optimum models appear to be $(Y=X_1+X_5+X_6)$, $(Y=X_1+X_5+X_6)$, $(Y=X_3+X_5+X_6)$ and $(Y=X_5+X_6)$. Across the three criterion outcome plots, the top 3-parameter model is $(Y=X_1+X_5+X_6)$ and the top two-parameter model $(Y=X_5+X_6)$. Consequently, these two models are investigated in terms of their assumptions and utility.

Table 5.5 presents the regression summary for the three-parameter model $(Y=X_1+X_5+X_6)$. From the summary, it is evident that the while the model has an Adjusted R-squared 0.803 and is, in itself, statistically significant, that the usersByPop variable is not statistically significant. Further, the Link Function assumption is not acceptable. Thus, omitting the usersByPop variable, we are left with the second proposed model $(Y=X_5+X_6)$.

Table 5.6 presents the regression summary for the two-parameter model $(Y=X_5+X_6)$. The summary indicates an Adjusted R-squared value of 0.78, that all of the

Table 5.5: Regression model summary based on the optimum three-parameter model.

	Dependent variable:
	Y Morisita Horn
X_1 Users Versus Population	$0.111^{\bullet}$
	(0.058)
X_6 User Count	0.140^{***}
	(0.027)
X_5 Surname Diversity (n=100)	0.206^{**}
	(0.070)
X_0 Constant	1.790^{**}
	(0.581)
Observations	22
R^2	0.831
Adjusted R^2	0.803
Residual Std. Error	0.208 (df = 18)
F Statistic	29.503^{***} (df = 3; 18)

	Value	*p-value*	*Assumption*
Global Stat	9.31199	0.053757	Acceptable.
Skewness	0.08441	0.771409	Acceptable.
Kurtosis	0.64780	0.420899	Acceptable.
Link Function	7.51410	0.006122	Unacceptable.
Heteroscedasticity	1.06568	0.301922	Acceptable.

Note: $^{\bullet}p<0.1$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

parameters are statistically significant and that the assumptions are all met. Furthermore, the model RMSE was calculated as 0.105005 once the dependent transferable was transformed back to its original state. In effect, the model states that the usefulness of each Twitter-derived population inventory in describing the 'observable' population is greatest where the number of Twitter users present within a country is high and the surname diversity within the country is low.

In practice, however, having successfully demonstrated the inclusion of the surname diversity parameter in the model, a key challenge is raised that the data employed are derivatives of the existing Worldnames Database population inventories.

Table 5.6: Regression model summary based on the optimum two-parameter model.

	Dependent variable:		
	Y Morisita Horn		
X_6 User Count	0.176^{***}		
	(0.021)		
X_5 Surname Diversity (n=100)	0.245^{**}		
	(0.072)		
X_0 Constant	0.747^{**}		
	(0.222)		
Observations	22		
R^2	0.796		
Adjusted R^2	0.775		
Residual Std. Error	0.222 (df = 19)		
F Statistic	37.122^{***} (df = 2; 19)		
	Value	*p-value*	*Assumption*
Global Stat	4.51871	0.34033	Acceptable.
Skewness	0.00498	0.94372	Acceptable.
Kurtosis	0.82896	0.36257	Acceptable.
Link Function	2.76265	0.09649	Acceptable.
Heteroscedasticity	0.92213	0.33692	Acceptable.
Note:	$^{\bullet}$p<0.1; *p<0.05; **p<0.01; ***p<0.001		

Thus, as was discussed previously, it is necessary that an alternative source of data be sourced.

In this case, it was found that the limited data provided by the Forebears.io online surname database had the potential to fulfil this requirement. Unlike the World-names Database, the Forebears.io dataset contains limited sets of names data for all countries. This data provides a means by which country level surname diversity may be estimated. However, given that limited information exists regarding the provenance of the Forbears.io data, it was necessary that an independent validation be performed on the data. Consequently, a web-scraping algorithm was implemented in R using the rvest package (Wickham, 2016). For each of the 227 countries/regions published by Forebears.io, the table of the top 200 names was extracted and the pro-

portion of the population represented by the top 100 names calculated based on the summation of the provided frequency data.

Recognising that the data provided by Forebears.io are in themselves derived from a range of sources, a one-tailed Pearson's product-moment correlation was performed to assess the relationship between the Forebears and Worldnames derived surname diversity measures. For the test, the assumptions are that the data are either interval or ratio; a linear relationship exists between the variables; there are no extreme outliers; and the data are approximately normally distributed. Performed on the log-transformed variables, the assumptions were found to have been met and the outcome to be therefore valid. The test, the results of which are shown in Table 5.7, returned a Pearson's correlation of 0.96 with a p-value < 0.001 indicating a very strong positive correlation that was statistically significant. Consequently, it was considered that the Forebears-derived surname diversity values were a valid proxy for their Worldnames-derived counterparts. It was thus possible that surname diversity values may be incorporated for the majority of countries worldwide and therefore, that the diversity variable could be included in the final model.

Table 5.7: One tailed Pearson's product-moment correlation summary indicating a very strong positive correlation between the two surname diversity variable sets.

Pearson's product-moment correlation	
Alternative Hypothesis (1-tailed)	True
t	14.4***
Observations	22
DF	20
Correlation	0.9550
95 Percent Confidence Interval:	0.9067 - 1.0000
Note:	$^{\bullet}p<0.1$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Having established that the surname diversity statistics calculated from the Forebears exhibited a strong positive correlation, the model was recalculated with the new surname diversity data. Based on the model summary, shown in Table 5.8 it was found that the surname diversity parameters remained significant based on an α of 0.05 though the Adjusted R-squared value decreased from 0.78 to 0.73. Fur-

thermore, all of the model assumptions remained acceptable. The RMSE for the updated model was 0.1151 which is only marginally higher than the original model RMSE of 0.1050. The substitution of the surname diversity variable was therefore considered a success.

Table 5.8: Regression model summary based on the substitution of the Forebears-derived surname diversity values.

	Dependent variable:
	morisitaHorn
X_6 usercount	0.181^{***}
	(0.024)
X_5 SurDivForebears	0.236^{*}
	(0.090)
X_0 Constant	0.706^{***}
	(0.243)
Observations	22
R^2	0.759
Adjusted R^2	0.733
Residual Std. Error	0.242 (df = 19)
F Statistic	29.898^{***} (df = 2; 19)

	Value	*p-value*	*Assumption*
Global Stat	6.4053	0.17086	Acceptable.
Skewness	0.6314	0.42682	Acceptable.
Kurtosis	0.5215	0.47019	Acceptable.
Link Function	3.1699	0.07501	Acceptable.
Heteroscedasticity	2.0824	0.14900	Acceptable.

Note: $^{\bullet}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Thus, the final model was:

$$\exp(Y) = \beta_0 + \log(X_5)\beta_5 + \log(X_6)\beta_6 \tag{5.1}$$

Where Y is the estimated similarity as represented by the Morisita-Horn Index of Overlap, β_0 is the intercept, X_5 is the surname diversity and X_6 is the number of valid Twitter-users.

5.2.5 Model Diagnostics

Having identified a suitably parsimonious model, the next concern was that of diagnostics. The diagnostic process was designed to critically examine the model, its assumptions and the underlying data such that predictions and inference could be made in the best of confidence. The process included an assessment of the model assumptions; an examination of the model outputs; and the completion of a cross-validation exercise to assess the model's generalisability and mitigate the risk of over-fitting.

Having formally assessed the model assumptions by way of the Global Validation of Linear Model Assumptions test, influence plots where employed as an efficient means to analyse the effect of observations on the model outcome. Influence plots provide a simple yet effective means to assess the impact of individual observations on fitted regression models. The plot provides an indication of those observations which exhibit excessive leverage via the Hat Value, outliers via the studentized residuals and influence on the model parameters is indicated by the size of the circles are based on the Cook's distance. Used in partnership with the model summary it is possible to make a number of key observations. For example, observations with a studentized residual $\pm$ 2 are considered as outliers and observations where the hat-values exceed 0.2 are considered to exhibit high leverage.

Figure 5.6 a, the plot based on the original surname diversity value, suggests that none of the observations are outliers though observation 15 has significant leverage. Figure 5.6 b, the plot for the final model, suggests that beyond observation 15, observation 19 (Serbia), with a studentized residual of -2.3, is an outlier. However, given that there is no justification for the omission of the variable it would be inappropriate for the observation be omitted from the analysis.

In addition to the influence plot, it is possible to examine the difference between the calculated and fitted values of the dependent variable. Figure 5.7 provides an illustration of these differences through plotting the original similarity score, the fitted similarity score and the difference emphasising whether the fitted value is greater or less than expected. From the Figure, there appears to be some degree of over-

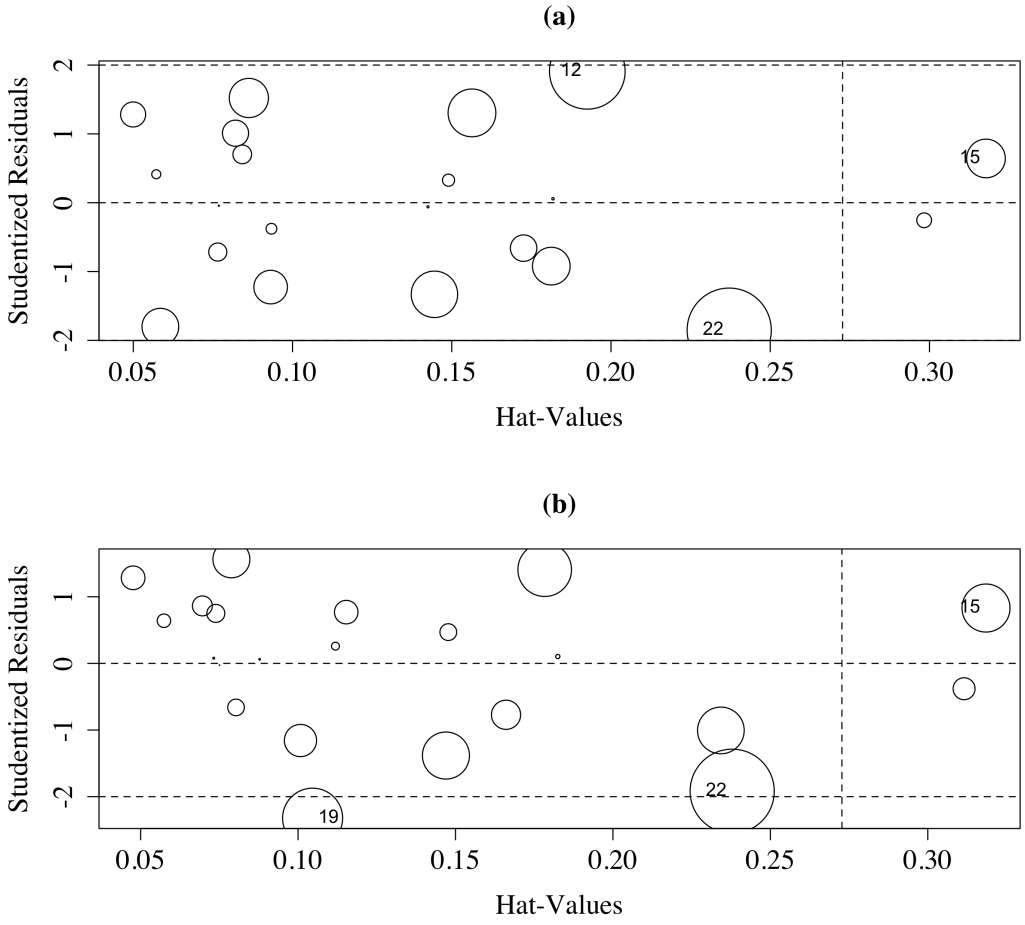


Figure 5.6: Influence plots showing the studentized residuals, Hat Values and Cook's Distance from the regression analysis based on the use of the original surname diversity (a) and Forebears-derived (b) surname diversity measure.

estimation in the lower range and under-fitting in the upper range. The transition point between the over and under estimation is approximately 0.7.

Cross-validation

Shrinkage is a term used to refer to the reduction in the coefficient of variation when a model is applied to new data. In seeking to quantify the effect of shrinkage, it is normal practice to perform some form of cross-validation. In its basic form, cross-validation may be considered as an assessment of a model's accuracy across a range of samples drawn from the original pool of observations. In practice, a sampling strategy is employed, either randomly or systematically, such that the initial observations are split into a training set used to build the model and a testing set

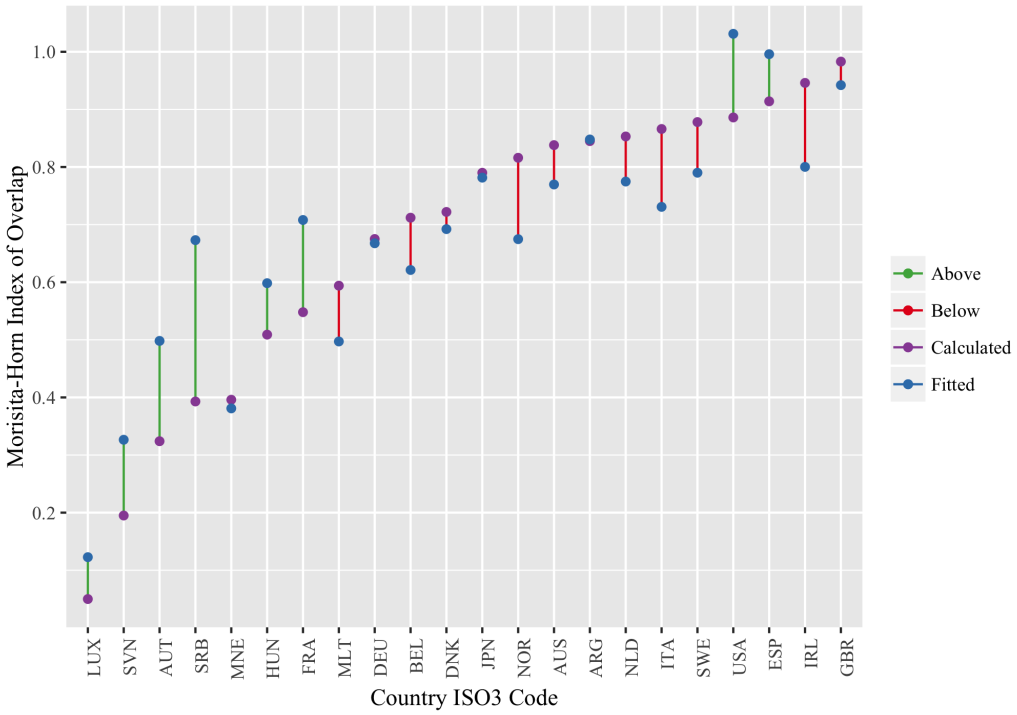


Figure 5.7: Plots showing the difference between the calculated and estimated Morisita-Horn similarity values.

designed to assess the model's performance. The process is repeated in such a way as to assess the model across a range of splits. Cross-validation is commonly employed where the number of observations is limited, and the alternative hold-out approach may be detrimental to the outcome. A further limitation of the hold-out approach is that the validation is susceptible to the original split employed in the data.

Cross-validation techniques may be divided into those that are exhaustive and those that and non-exhaustive. Exhaustive techniques include Leave one out cross validation (LOOCV) and Leave p-out cross validation (LpOCV). Non-Exhaustive methods include k-fold cross validation. The exhaustive techniques assess every possible model based on the value of p which is specified. In K-fold cross-validation, the data are randomly assigned into K sets. Subsequently, for K-iterations, the model is trained on K-1 sets and tested on the final set. The output of the analysis is the mean across each of the K tests. It should be noted that as the partitions are generated randomly, the output will vary between calculations and the test should be repeated multiple times to gain an understanding of the outcome. The number of repetitions

is generally dependent on the stability of the model, and the amount of variance experienced between tests.

LOOCV, as the name suggests, involves testing a model for all possible combinations of training and testing data. In some senses, LOOCV may be considered as a K-fold validation where K = number of observations. LOOCV is commonly employed where the number of model observations is low, and omission of multiple observations would have a significantly detrimental impact on the model. This technique is equivalent to K-fold cross validation where k is equal to the number of observations. Similarly, leave p out cross validation tests every possible combination where p-observations are used for testing, and $n - p$ observations are employed for the model testing. As may be imagined, increasing p makes the model increasingly pessimistic.

Given that the number of observations is limited, it was proposed that the LOOCV and K-fold cross validation be performed.

Table 5.9: Table showing the results of the LOOCV exercise for the final regression analysis.

LOOCV	
Original R-squared	0.759
Original Adjusted R-squared	0.733
LOOCV R-squared	0.677
RMSE	0.1331

LOOCV and k-fold cross validation

Table 5.9 provides a summary of the LOOCV exercise. It can be seen that the LOOCV R-squared is 0.677 indicating a difference of 0.082. Given that some consider LOOCV overly optimistic, a second assessment was performed using k-fold cross validation. Given that k-fold cross-validation assigns the observations to groups on a random basis, the outcome will vary each time the test is repeated. Consequently, the test was repeated 100 times and the mean cross-validated R-squared reported.

Table 5.10, provides an indication of the model shrinkage with the original R-squared dropping from 0.759 to 0.673. With the difference between the origi-

Table 5.10: Table showing the summary of the k-fold cross-validation.

k-fold cross-validated R-square	
Original R-square	0.759
Original Adjusted R-squared	0.733
k	11
n (repetitions)	100
Mean Cross-Validated R-square	0.673
Change	0.086

nal and mean R-squared being 0.086, the small decrease in the model R-squared suggests the model is fairly generalisable. In effect, the cross-validated R-squared implies that 67.3% of the variance may be explained by the model. Importantly, the cross-validated R-squared remained similar in the case of both validation scenarios suggesting the model performs well in terms of generalisability.

Relative weights

Finally, the relative importance of each model parameter is calculated. In much conventional analysis, it is necessary that the dependent variables are standardised such that the coefficients provide an indication of the relative importance of each parameter. While such an approach remains popular, the calculation of relative importance is increasingly accepted. Relative importance was calculated based on the methodology of Johnson (2000). The calculation takes into account both the effect of each predictor in isolation and its effect in the combination of any other predictors. The results indicate that of the %75.9 of variance explained by the model, 93.6% is explained by the number of users, and 6.4% is explained by the surname diversity. In practice, this difference in relative importance suggests that the number of Twitter users is most significant in terms of how well the 'true' population naming structure is represented by the Twitter-derived population inventories.

5.3 Model Application: Results and Discussion

At the outset, the objective was to develop a method by which the probable representativness of the Twitter-derived population inventories could be estimated such

that new inventories could be used to fill gaps in the Worldnames database at the national scale and further enhance the Onomap CEL classification tool. Note, it was not the objective to replace or supplement the existing national-level population inventories, rather, it was to provide a suitable source of reference data where no alternative was available. In seeking to achieve this, a model was proposed for the purpose of identifying factors which may indicate the efficacy of Twitter in representing the population name structure within specific countries. Consequently, factors were identified based on two themes: Those which impact on the adoption of online social networks and those which impact on how many names are required to represent populations. In identifying such factors, the major challenge was the lack of data published consistently on the global scale. Consequently, a set of candidate parameters were determined using data from the World Bank DataStore and also from the raw Twitter-population inventories.

In seeking to identify the optimum model, an all-subsets regression was performed such that all possible models could be assessed. The final model was identified as $(Y=X_5+X_6)$ where X_5 was the number of valid users and X_6 was the surname diversity. Proving to be statistically significant, the required data were compiled and, in the case of the surname diversity values re-sourced. Subsequently, the model was applied and both the fitted values and prediction intervals calculated.

5.3.1 Geography of Twitter

As a first step in investigating the outcome of the regression analysis, a series of maps were constructed designed to illustrate the lower, fitted and upper C_H values. On initial inspection of the results, it was evident from the fitted C_H values that the potential utility of the Twitter-derived population inventories varied significantly across the globe. In seeking to understand this distribution, a series of maps were produced showing the fitted model values and also the upper and lower prediction intervals. The maps were overlaid with the second tier UN GeoScheme boundary data for the purpose of providing a concise reference structure.

The maps, shown in Figures 5.8 through 5.10, provide a valuable illustration of the geographic distribution of the fitted values clearly indicating a geographic trend

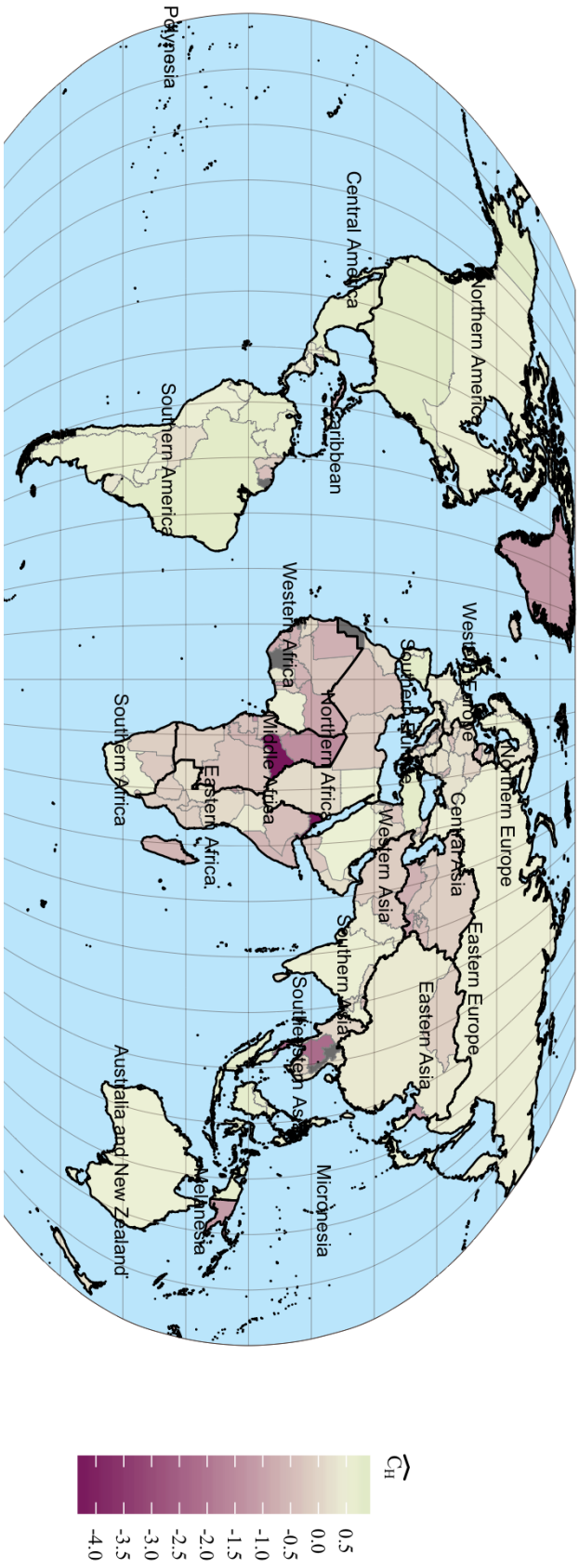


Figure 5.8: World map showing the model lower values.

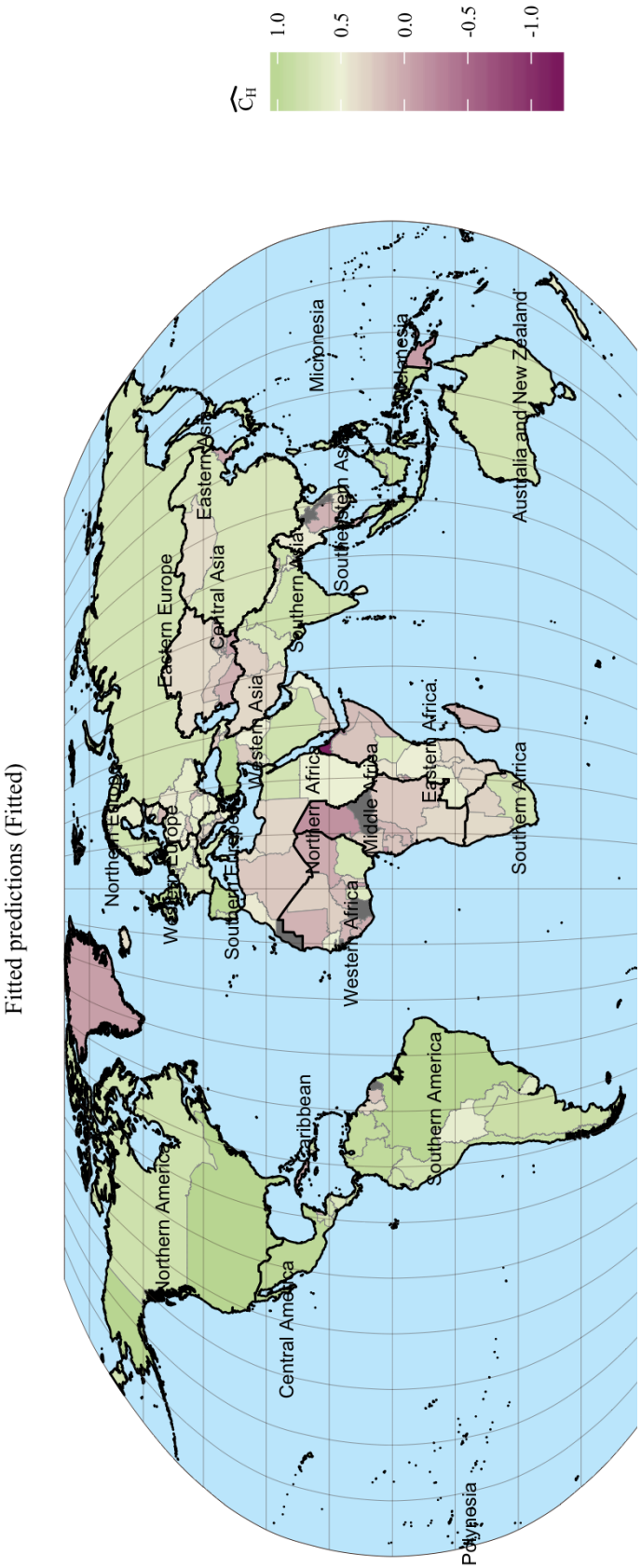


Figure 5.9: World map showing the model fitted values.

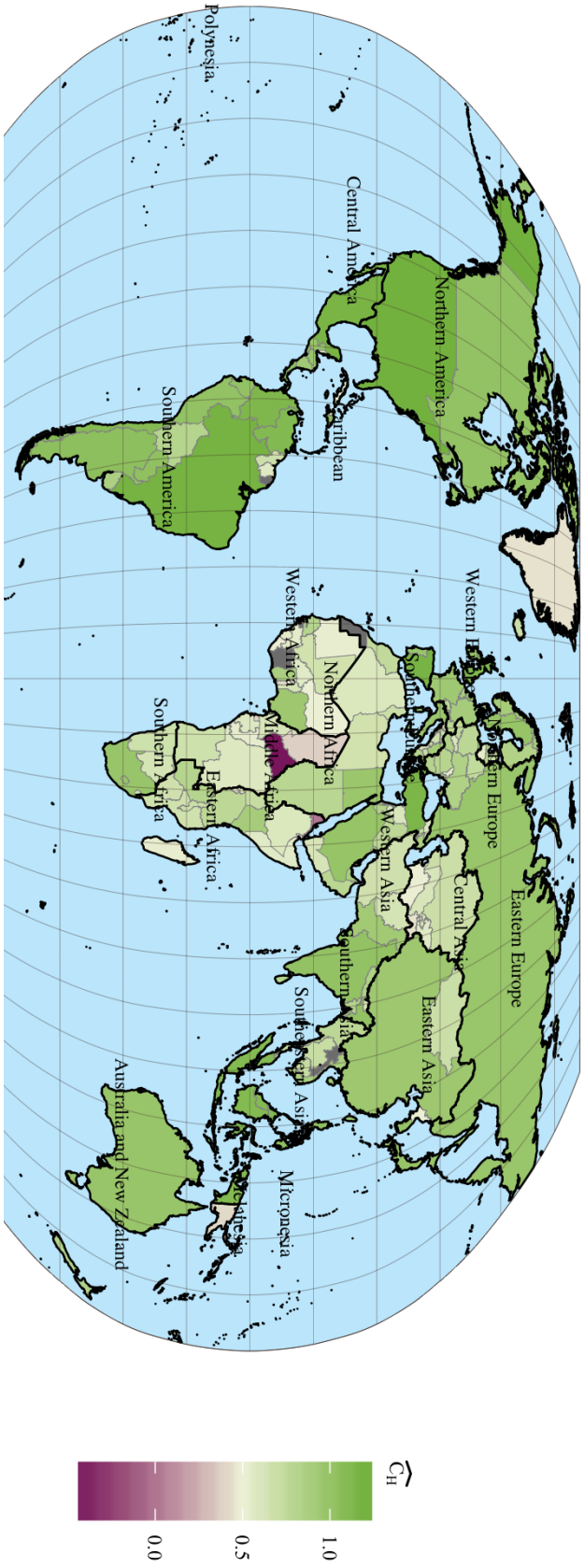


Figure 5.10: World map showing the model upper values.

in the utility of the Twitter-derived inventories.

Furthermore, the observable patterns of the fitted C_H values appeared concordant with the regions defined by the UN GeoScheme. Given this apparent agreement, and for want of a suitable structure for the explanation, the top tier UN regionalisation was subsequently employed. The top tier consists of seven regions: Antartica, Africa, the Americas, Asia, Europe, Oceania and Other.

As an initial step in understanding the results, a box and whisker plot was created for the fitted model values faceted based on the UN GeoScheme region classification.

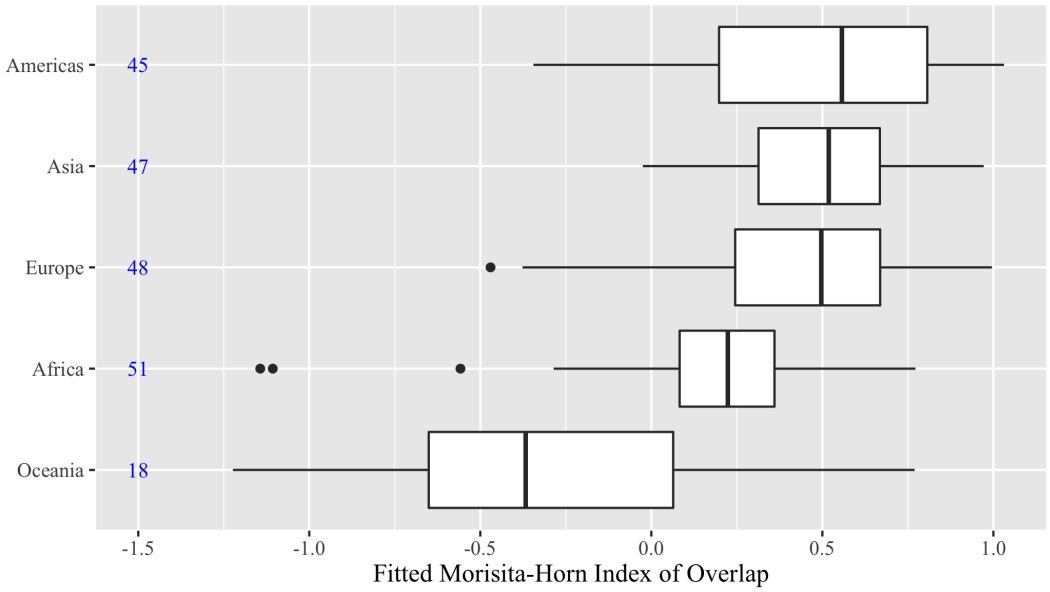


Figure 5.11: Faceted box and whisker plots showing the varying distribution of similarity estimated between each of the five main global regions. The number of countries in each group shown in blue.

Figure 5.11, the faceted box plot of fitted C_H values, provides a useful illustration of the distribution of values within and between regions highlighting an apparent skew toward the Americas, Asia and Europe. The degree of variation within each region provides a useful indicator as to overall regional performance. It is likely that the amount of variance observed, particularly in the case of the Oceania region is due to the diversity within the associated countries in terms of technological provision, economies and more generally, regional trends in the adoption of online social media.

5.3.2 Common Names

A further opportunity to assess the national Twitter inventories is through comparison against existing sources of names data. A process identical to that employed in the audit of the UCL Worldnames Database. In completing such a validation exercise, it is necessary that the reader remains mindful of the limitations inherent in all forms of names data. In the case of forbears.io, the majority source of the 'official' top 10 names in the subsequent tables, the surname counts are clearly scaled samples such that they appear to represent the entire population; arguably delivering a false sense of precision and accuracy. That said, Forebears.io is the only service to publish names data for almost every country in the world, and thus, the data are employed. For the sake of efficiency, only the top three inventories, based on the fitted Morisita-Horn Index of Overlap are reported.

5.3.2.1 Africa

Countries:

AZE, ARM, BHR, BGD, MMR, BRN, KHM, LKA, CHN, AFG, BTN, CYP, GEO, IND, IRN, ISR, IRQ, JPN, JOR, KGZ, PRK, KOR, KWT, KAZ, LAO, LBN, MNG, OMN, MDV, MYS, HKG, MAC, PSE, NPL, PAK, QAT, PHL, SAU, SGP, SYR, THA, TJK, TUR, TKM, UZB, VNM,YEM, IDN, ARE and TLS.

Table 5.11: Table showing the top 10 names in the three best performing countries in Africa.

South Africa C_H = 0.7720999			
Twitter		Forebears	
Surname	Count	Surname	Count
KHUMALO	115	NAIDOO	300,130
NDLOVU	107	BOTHA	274,396
NKOSI	95	SMITH	233,996
DLAMINI	93	GOVENDER	212,829
SMITH	87	PILLAY	198,144
BOTHA	80	VAN DER MERWE	192,209
NAIDOO	69	NEL	188,428
WILLIAMS	68	PRETORIUS	184,903
MOKOENA	66	JACOBS	182,456
MKHIZE	62	COETZEE	178,931

Egypt C_H = 0.7530496			
Twitter		Forebears	
Surname	Count	Surname	Count
MOHAMED	255	KHAN	416,957
AHMED	247	AHMADI	325,530
ALI	122	SAFI	238,848
HASSAN	120	NOORI	205,315
ADEL	113	AHMAD	182,854
KHALED	108	AZIZI	181,588
GAMAL	101	RAHIMI	176,843
ASHRAF	95	AHMADZAI	172,414
IBRAHIM	92	SADAT	141,095
TAREK	91	AMIRI	125,910

Nigeria C_H = 0.7324573			
Twitter		Forebears	
Surname	Count	Surname	Count
IBRAHIM	95	LAWAL	843,038
BOY	89	AJAYI	821,489
EMMANUEL	89	ADEBAYO	765,398
SAMUEL	87	BELLO	743,696
MICHAEL	79	OJO	683,326
BELLO	67	ADEYEMI	620,816
BOI	66	BALOGUN	557,390
DANIEL	66	IBRAHIM	499,618
JOHN	62	OKAFOR	485,557
YUSUF	60	ABDULLAHI	476,998

5.3.2.2 Americas

Countries:

ATG, ARG, BRB, BMU, BHS, BLZ, BOL, BRA, CAN, CHL, CYM, COL, CRI, CUB, DMA, DOM, ECU, SLV, GUF, FLK, GRD, GRL, GTM, GUY, HTI, HND, JAM, MTQ, MSR, MEX, ABW, AIA, SUR, NIC, PRY, PER, PAN, PRI, KNA, LCA, TTO, USA, URY, VCT, VEN, VGB, VIR, GLP, ANT, SPM, TCA, MAF and BLM.

Table 5.12: Table showing the top 10 names in the three best performing countries in the Americas.

United States C_H = 1.0312202			
Twitter		Forebears	
Surname	Count	Surname	Count
SMITH	9,179	SMITH	2,552,459
JOHNSON	6,990	JOHNSON	1,967,023
JONES	5,542	WILLIAMS	1,609,082
WILLIAMS	5,309	BROWN	1,482,001
BROWN	5,199	JONES	1,455,165
GARCIA	4,642	MILLER	1,203,150
MILLER	4,540	DAVIS	1,172,346
RODRIGUEZ	4,448	ANDERSON	872,825
MARIE	4,135	WILSON	860,309
MARTINEZ	4,088	TAYLOR	787,071
Brazil C_H = 1.0098336			
Twitter		Forebears	
Surname	Count	Surname	Count
OLIVEIRA	4,515	SILVA	5,073,774
SANTOS	3,321	SANTOS	3,981,191
SILVA	3,228	OLIVEIRA	3,738,469
LIMA	3,203	SOUZA	2,630,114
RODRIGUES	2,977	RODRIGUES	2,399,459
SOUZA	2,644	FERREIRA	2,365,562
ALVES	2,417	ALVES	2,264,282
FERREIRA	2,211	PEREIRA	2,251,864
MARTINS	2,110	LIMA	2,020,288
COSTA	2,056	GOMES	1,697,130
Mexico C_H = 0.9878378			
Twitter		Forebears	
Surname	Count	Surname	Count
GARCIA	3,164	HERNANDEZ	2,534,379
GONZALEZ	2,521	GARCIA	2,416,128
HERNANDEZ	2,463	LOPEZ	2,151,072
RODRIGUEZ	2,190	MARTINEZ	2,151,046
MARTINEZ	2,171	GONZALEZ	2,093,124
LOPEZ	2,039	RODRIGUEZ	1,741,540
SANCHEZ	1,835	PEREZ	1,614,913
RAMIREZ	1,565	SANCHEZ	1,537,179
FLORES	1,332	RAMIREZ	1,294,894
PEREZ	1,239	FLORES	1,110,320

5.3.2.3 Asia

Countries:

DZA, AGO, BEN, COG, COD, BDI, CMR, TCD, COM, CAF, CPV, DJI, EGY, GNQ, ERI, ETH, GMB, GAB, GHA, GIN, CIV, KEN, LBR, LBY, MDG, MLI, MAR, MUS, MRT, MOZ, MWI, NER, MYT, NGA, GNB, REU, RWA, SYC, ZAF, LSO, BWA, SEN, SLE, SOM, SDN, TGO, STP, TUN, TZA, UGA, BFA, NAM, SWZ, ZMB, ZWE, SHN, ESH.

Table 5.13: Table showing the top 10 names in the three best performing countries in the Asia Region.

Turkey C_H = 0.9717326			
Twitter		Forebears	
Surname	Count	Surname	Count
YLMAZ	3,881	YILMAZ	756,629
OZTURK	2,802	KAYA	665,898
KAYA	2,773	DEMIR	583,657
DEMIR	2,180	CAN	429,235
CELIK	1,970	AYDIN	407,994
AYDN	1,919	ARSLAN	393,914
AHIN	1,899	SAHIN	388,918
OZDEMIR	1,870	YILDIZ	379,765
YLDRM	1,852	YILDIRIM	375,956
YLDZ	1,810	ÖZTÜRK	370,471
Indonesia C_H = 0.9158336			
Twitter		Forebears	
Surname	Count	Surname	Count
PUTRI	6,800	SARI	902,933
PUTRA	4,267	SETIAWAN	630,683
SARI	4,177	LESTARI	623,178
PRATAMA	3,317	HIDAYAT	506,648
SAPUTRA	2,780	SAPUTRA	506,430
DEWI	2,680	WATI	494,134
LESTARI	2,390	RAHAYU	493,953
KURNIAWAN	2,297	SANTOSO	406,886
PRATIWI	2,278	WAHYUNI	402,339
SETIAWAN	2,176	KURNIAWAN	384,105
Malaysia C_H = 0.9141692			
Twitter		Forebears	
Surname	Count	Surname	Count
ISMAIL	578	TAN	713,765
AHMAD	516	LIM	609,577
LEE	474	LEE	545,429
AZIZ	445	MOHAMED	487,454
TAN	423	WONG	449,385
LIM	377	NG	373,400
RAHMAN	343	CHONG	274,498
ABDULLAH	299	AHMAD	244,524
AZMI	293	ABDUL	229,195
AZMAN	275	CHAN	221,657

5.3.2.4 Europe

Countries:

ALB, BIH, BGR, DNK, IRL, EST, AUT, CZE, FIN, FRA, DEU, GRC, HRV, HUN, ISL, ITA, LVA, BLR, LTU, SVK, LIE, MKD, MLT, BEL, FRO, AND, GIB, IMN, LUX, MCO, MNE, ALA, NLD, NOR, POL, PRT, ROU, MDA, RUS, SVN, ESP, SWE, CHE, GBR, UKR, SMR, SRB, VAT, SJM, GGY, JEY.

Table 5.14: Table showing the top 10 names in the three best performing countries in Europe.

Spain C_H = 0.9958624			
Twitter		Forebears	
Surname	Count	Surname	Count
GARCIA	6,261	GARCIA	1,489,445
LOPEZ	4,036	GONZALEZ	932,929
SANCHEZ	4,018	RODRIGUEZ	930,332
RODRIGUEZ	3,753	FERNANDEZ	926,719
GONZALEZ	3,684	LOPEZ	893,278
FERNANDEZ	3,593	MARTINEZ	844,540
MARTINEZ	3,535	SANCHEZ	824,073
PEREZ	3,260	PEREZ	800,021
GOMEZ	2,335	GOMEZ	504,099
MARTIN	2,246	MARTIN	496,403

United Kingdom C_H = 0.9421356			
Twitter		2013 Consumer Register	
Surname	Count	Surname	Count
SMITH	3,822	SMITH	532,928
JONES	3,262	JONES	417,703
WILLIAMS	2,275	WILLIAMS	303,581
TAYLOR	1,993	BROWN	281,152
BROWN	1,909	TAYLOR	268,352
DAVIES	1,751	DAVIES	224,886
WILSON	1,452	WILSON	209,168
EVANS	1,339	EVANS	184,143
THOMAS	1,247	THOMAS	172,178
JOHNSON	1,129	JOHNSON	170,887

Ireland C_H = 0.8001966			
Twitter		Forebears	
Surname	Count	Surname	Count
MURPHY	361	MURPHY	56,815
KELLY	292	KELLY	42,550
BYRNE	258	BYRNE	39,400
WALSH	237	RYAN	35,939
RYAN	223	WALSH	35,453
OBRIEN	145	DOYLE	26,363
DOYLE	137	O'BRIEN	23,896
MURRAY	127	O'CONNOR	21,816
OCONNOR	121	LYNCH	20,800
NOLAN	117	O'SULLIVAN	20,145

5.3.2.5 Oceania

Countries:

ASM, AUS, SLB, COK, FJI, FSM, PYF, GUM, KIR, NCL, NIU, MNP, NFK, VUT, NRU, NZL, PNG, TKL, TON, TUV, WLF, WSM, PCN, PLW and MHL.

Table 5.15: Table showing the top 10 names in the three best performing countries in Oceania.

Australia C_H = 0.7696712			
Twitter		Forebears	
Surname	Count	Surname	Count
SMITH	253	SMITH	184,513
WILLIAMS	152	JONES	95,808
LEE	134	BROWN	89,518
JONES	128	WILLIAMS	88,727
WILSON	111	WILSON	77,239
NGUYEN	107	TAYLOR	75,521
TAYLOR	100	LEE	58,147
BROWN	97	JOHNSON	54,396
THOMAS	80	ANDERSON	54,227
RYAN	77	WHITE	51,916

New Zealand C_H = 0.5700322			
Twitter		Forebears	
Surname	Count	Surname	Count
SMITH	49	SMITH	16,920
BROWN	30	WILLIAMS	10,002
LEE	23	JONES	9,912
WILSON	23	WILSON	9,874
TAYLOR	22	BROWN	9,533
HARRIS	15	TAYLOR	9,297
WILLIAMS	15	ANDERSON	7,704
CLARKE	14	SINGH	7,446
JONES	14	SCOTT	7,389
THOMAS	14	THOMPSON	7,337

Guam C_H = 0.3903865			
Twitter		Forebears	
Surname	Count	Surname	Count
CRUZ	10	CRUZ	3,575
MARIE	5	SANTOS	1,776
SANTOS	5	PEREZ	1,429
GARCIA	4	SABLAN	1,287
GUERRERO	4	DUEÑAS	1,279
AGUON	3	CAMACHO	1,271
CASTRO	3	BLAS	1,223
FLORES	3	LEON	1,200
GIRL	3	AGUON	1,176
RIVERA	3	FLORES	1,152

5.3.2.6 Other

Countries:

CCK, ATA, BVT, ATF, HMD, IOT, CXR, UMI, SGS and TWN.

Table 5.16: Table showing the top 10 names in the best performing country in Other.

Taiwan C_H = 0.7156517			
Twitter		Forebears	
Surname	Count	Surname	Count
CHEN	157	LIN	1,549,426
LIN	125	CHANG	1,096,361
HUANG	87	HUANG	1,095,086
LEE	79	WANG	941,361
WANG	74	WU	842,855
CHANG	66	LIU	668,386
LIU	47	HSU	645,266
YANG	40	YANG	576,834
HSU	39	TSAI	547,514
TSAI	36	CHENG	415,981

Having examined the comparison tables of most common surnames, several trends are evident. Notably, that with the exception of Africa and Asia, there was a good degree of overlap between the Twitter-derived names and the Forebears counterparts. Thus, considering the usefulness of the fitted C_H values is in terms of indicating probable inventory success, it would appear that the degree of correlation among the common names decreased in an ordinal manner. This observation acts to support the belief the model is a good indicator of national-level inventory performance.

That said, on examination of Table 5.11, the names tables from Africa, the agreement between the most common names is low, while the fitted C_H values indicate strong performance. Given the number of valid users identified in within each country, it would suggest that the issue was manifest in either the way in which individuals report their own names or alternatively, in how individuals' names were extracted. In regards to the the processing of individuals' names, the extraction framework is based on western naming conventions in which individuals' have a forename and surname. In many African countries, notably those with large Muslim and Arabic populations it is common to use the father's forename in place of a

surname. I.e., Mohamed bin Ahmed. In this case, the final string, reported as the surname would be Ahmed. Such a feature of the algorithm, and more broadly the cultural and regions tendencies in the ascription of names highlights the importance of human intuition and judgement in the analysis of names data.

Considering the Americas and Europe as a singular entity, both the fitted values and the name comparison tables provide a reason for optimism. In both cases, the three top countries are either English-speaking or Hispanic. In the case of all six comparison tables, there is a high degree of overlap and also agreement in name order. It should be noted that with the exception of Ireland, the fitted C_H values are in the mid-to-high nineties.

In a similar scenario to Africa, of the three countries examined, only the comparison for Turkey proved effective. In interpreting this, it is important to bear in mind the regional differences in naming conventions. In the case of Indonesia, it is common for individuals to have no surnames and in the case of Malaysia surnames are generally patronymic. The exception being Turkey which introduced surnames in 1934. Previously, It was common practice, as in many Muslim countries to adopt the father's forename.

In the case of Oceania, it is evident from Figure 5.11 that the region exhibits the highest degree of variation in terms of the Twitter inventories explanatory power. That said, the data for both Australia and New Zealand do show some degree of overlap. Interestingly, in the case of all three countries, the number of valid Twitter users was relatively small with just 1,500 and 1,164 users per million of the population for Australia and New Zealand respectively versus 5,514 per million in the UK. Data for Guam have a limited level agreement. However, the small number of Twitter users means that the inventory lacks the definition/structure that would be desired.

Of the countries under the description 'Other', Taiwan is the only country to be included in the analysis. It should be noted that Taiwan would, in ordinary circumstances, be considered as part of Asia. Regarding success, the overlap in names for Taiwan was high. The case, as was so in Turkey was the common use of western style surnames. A further advantage, given the limited number of users identified

(n=2,466), is the low national surname diversity meaning that fewer names are required to depict the populations naming structure accurately.

5.3.3 Discussion

From the preceding analysis, it would appear evident that the Twitter-derived population inventories do offer an alternative source of national-scale individual level names data. However, regarding interpretation, it is important that we remain conscious of the limitations inherent in the data and analysis. In the first instance, the analyses are based on a broad range of data drawn from a multitude of official and unofficial sources. Second, both of the independent variables are derivatives, in the case of the raw counts of Twitter users this is based on the analysis performed within this thesis and the surname diversity values are derived from published lists of names data. Third, The reference data, whilst considered the most complete and representative available, are in themselves liable to uncertainty. Finally, the set of observations used in the construction of the model were not a stratified sample with the notable omission of any data from Africa; a feature evident in Figure 4.4. In all cases, efforts have been made to verify the quality of the data and the effects which may occur as a consequence. However, in practice, there are barriers to such an exercise. The key point here is that we must recognise the limitations of such data for the study of populations and ensure that these limitations are clearly communicated such the any subsequent analysis is founded on the correct beliefs and assumptions.

In seeking to develop the predictive capability of the model in regards to the expected utility of the Twitter inventories a number of potential extensions may be possible. These include the addition of further variable and the use of alternative modelling approaches better able to capture the variation observed within the data. In terms of the identification of variables the primary challenge was the availability of suitable indicators at the global scale; many of the data, such as social media penetration were restricted to a limited pool of countries. This issue was exacerbated by variables which were globally available, however, collected in an inconsistent manner. Choice of variables was consequently limited to major data portals such as the WorldBank Data Store. In some cases it may have been possible to include

additional variables at a finer spatial resolution such that regional variation in effectiveness could have been better understood. Given the ecological fallacy, if only a small segment of the population are well represented, this may be disguised for the country as a whole. Such an adjustment may have proven useful in countries which exhibit a high degree of segregation or variation in wealth and provision of services. It may be prudent to perform such analysis on an as and when required basis for specific target countries.

Arguably, the most effective means of improving the Twitter inventories would be to include data collected over a greater time period. Extension of the time window would lead almost certainly increase the number of Twitter users within a country increasing the pool of individuals for the inventory creation. That said, in those countries with few Twitter users, arguable the most appropriate response is to seek alternative forms of population data. Likewise, the overall modelling process may be improved in the future through the addition of further reference population inventories if and when these become available.

Having discussed the model outcome in the context of the six UN GeoScheme regions, it is evident that the inventory creation framework is effective. However, as has been seen in the case of Africa and Asia, both the number of Twitter users and the ability to handle more complex naming structures has posed a barrier. More specifically, it is clear that the methodology performs best in countries with large numbers of users and which typically assign personal names in line with the western order in which the personal name precedes the surname. While this aspect of the analysis places a constraint on the potential utility of the Twitter inventories, a number of significant opportunities remain.

Regarding the inventory creation framework, the primary opportunity is to develop the name extraction algorithm such that it can better account for cultural naming conventions. For example, in the handling of double-barrelled names and names which are patronymic. In a similar vein, there is the potential to create sub-national/regionalised population inventories in those countries where the national similarity is high, and also there is sufficient volume of viable Twitter users. Ex-

amples of such countries include Brazil, Mexico and Turkey. Having created the population inventories, it is possible to export each inventory at a range of increasingly fine spatial scales. Good opportunity to do this for Brazil as this addresses the issues with the current Worldnames dataset which is limited to just three cities. Furthermore, should the three telephone directories be aligned with the relevant GADM regions then it will be possible to calculate the similarity between these datasets and get an idea of how well the inventories are likely to perform within the country. If it is good, this is a useful addition to the Worldnames database and will provide a useful illustration of how the data may be employed.

The next major opportunity is regarding applications. In possession of an enhanced collection of names data and an improved understanding of the global distribution and representativeness of the Twitter population, there is potential for analysis to be conducted at a range of spatial and temporal scales previously unimaginable. The Twitter-derived population inventories, where they are deemed to be representative of the population, provide an opportunity not only to examine stocks of the population but also to observe the population in a dynamic manner where population stocks may be observed in both space and time. Furthermore, the analysis offers an improved understanding of the global patterns of Twitter adoption. Such information is significant in that it provides justification for social network data to be exploited in a greater number of countries and situations. This, in itself, has laid the path for further research investigating demographics and security across a plethora of new countries and regions.

A third opportunity, beyond indicating the probable representative ability of national level Twitter-derived inventories, is the opportunity to perform data standardisation. For instance, for the purpose of performing international mobility analysis between regions where the popularity of Twitter varies. Such an approach may address limitations in the inference of international mobility from social media as experienced by Hawelka et al. (2014) in their analysis of global migration using Twitter data.

Double-barrelled names

Double-barrelled surnames present a unique challenge in terms of working with surname data. These challenges are manifest in both the extraction of names and also their interpretation for the purposes of analysis. The impact of double-barrelled surnames varies significantly among countries with the percentage of double-barrelled surnames ranging significantly.

Considering first the extraction of individuals' names, specifically in the case of Twitter, the '-' symbol is often used for the purpose of decoration as opposed to being an indicator of a double-barrelled surname. Consequently, the algorithm interprets the first of the two names as being the penultimate string and the second as being the surname. In practice, however, it may be argued that the surname is highly likely to be indicative of either the mother or father's ethnicity; thus minimising the potential for bias.

Regarding analytical interpretation, double-barrelled surnames raise further challenges. First, the majority of analytical techniques consider the surname as a single string and match said strings on a like for like basis. In effect, the analysis is not able to account for the separate names employed. This is a typical example of a situation in which a human interpretation differs from that of a machine. Three possibilities exist to address this issue. First, the approach utilised by Cheshire and Longley (2012), is to omit uncommon names and double barrel names. Second, various fuzzy matching techniques exist, though, in practice, such approaches may result in unrelated or nuanced surnames being incorrectly associated. For example, *Sharples* and *Sharpless*. Alternatively, would be to omit the double-barrelled surnames, or alternatively, treat each surname part in isolation. In some senses, such an approach is analogous to the Onomap methodology in which both forename and surname are processed individually, and then the assignment is made on the portion with the strongest association with any CEL group.

5.4 Conclusions

In this chapter, the objective was to develop a global database of names through the creation of national level population inventories derived from geographically references Twitter data. The motivation being to develop a supplement the existing Worldnames Database. The approach was to apply the methods developed in the previous chapter for Spain and the United Kingdom to the remainder of the world. Proving successful, the new inventories would hold the potential to fill gaps in the current database and further to add to the pool of forename-surname data available for the the Onomap classification.

In practice, however, having previously observed that the Twitter-derived inventories ranged dramatically in terms of their performance, it was proposed that a model is created with the purpose of predicting probable register success based on a series of associated cultural and environmental factors. The final model, with an Adjusted R-squared value of 0.733, was based on the number of Twitter users identified within a country and also the surname diversity based on the 100 most commonly occurring surnames. In effect, the model states that the more Twitter users are resident within the country, the better those individuals' names will depict the population as a whole. The exact number will be higher where surname diversity is high and lower where surname diversity is low. For example, fewer individuals are required to represent the population of China than are necessary to represent the population of the United Kingdom. This concept relates back to the idea of 'How many names are enough?' that was discussed in Section 5.2.2.

Based on the analysis that has been performed in this, and the preceding chapters, it is evident that utility of Twitter-derived population inventories varies significantly within and between countries. In many cases, the Twitter-derived inventories are either poor proxies for the observable population or, in some cases, alternative sources of names data may be sourced. That said, such a perspective is based on the principles of conventional geodemographics and more specifically the static population representation. What this chapter has demonstrated is that in certain circumstances, the Twitter-derived inventories are a powerful descriptor of national

populations and may, therefore, provide a new platform for the analysis of population dynamics where such analysis has previously been unfeasible. Examples of such locations include Brazil, Mexico and Turkey. Furthermore, the data generated are, in many regions borderless facilitating the analysis of populations without the constraints imposed by conventional forms of population data.

Chapter 6

Twitter in the UK: A Basis for Analysis

6.1 Introduction

At the outset, the objective of this thesis was to assess the potential applications of new forms of data to the study of geodemographics and security. In providing justification for such an objective, a critical evaluation of current geodemographic data and practises was completed. The key limitation identified was the ongoing dependence of conventional demographic products on static population representations such as the UK Census of Population. In light of this, it was proposed that new forms of data, in this case, drawn from online social networks, should be investigated for their potential utility in the description of the stocks and flows of population.

Seeking to exploit the potential manifested within the Twitter data, a framework was developed for the construction of static individual level population inventories based on the analysis of a global corpus of geotagged Tweets. The proposed framework was based on a series of criteria guided by the United Nations definition for population registers (United Nations, 2001). Specifically, to create an 'individualised data system' able to provide 'a mechanism of continuous recording' for the population. In this instance, the objective was that for each Twitter user, the following attributes would be recorded: a unique identifier, forename, surname, age, gender, ethnicity and probable geographic location. The product of this framework

being a richly attributed inventory of Twitter users for each world country. Recognising that the popularity and, by effect, the usefulness of Twitter varies within and between countries, an assessment of the candidate population inventories was performed. Based on the measurement of similarity between populations measured as surname composition, the analysis provided insight into the global geography of Twitter and delivered some initial insight as to the potential for Twitter data to be employed in investigating specific countries. However, whilst this approach to validation functioned well in the context of a global assessment, it was unsuitable for the more nuanced aspects of population structure.

In light of the above, the objective of this chapter is to develop a framework by which the representativeness of the Twitter inventories may be assessed against key demographic markers. For the purpose of developing the framework, an assessment was performed using data for the United Kingdom. The justification for this choice included the popularity of Twitter within the UK, the author's prior knowledge of the UK, and also, the abundance of detailed population data already available which may be used for reference. A further motivation for this thesis is the potential value of this information within industry and academia. Understanding the demographic composition of Twitter users within a country or region may help inform the validity of future studies, or provide a means of standardisation such that any Twitter-derived insight may be suitably evaluated. However, while this assessment is centred on the UK, the methods, given adequate consideration, may be applied to any country in which sufficient volume of Twitter data is available.

6.2 UK-Wide Validation and Benchmarking

In seeking to establish the Twitter-derived inventories as a viable alternative to conventional aggregate population data, there is a requirement to quantify how representative the data are of the observable population in regards to a series of important demographic markers: age, gender, ethnicity and geographic distribution. A previous study by Longley et al. (2015) sought to achieve a similar goal, but lacked the potential to be more widely applicable having only used data from the Greater Lon-

don extent. Thus, in delivering this chapter, it should be recognised that this study is not simply a reproduction of the work by Longley et al. (2015). Rather, it should be considered as an extension, building upon both the coverage of the analysis and also the efficacy in which various methods are applied. The focus on Greater London by Longley et al. (2015), while valid, limited the applicability of the research. Given the heterogeneity of the UK population, there is a risk that certain biases, unique to London, are propagated to the rest of the UK. By constraining the study and not accounting for the Tweets by users submitted beyond this area, there is a high likelihood that individuals who are not residents, such as tourists and commuters, within the city are included in the analysis. Consequently, when comparing the data to the UK Census of Population, there is a risk of not comparing like for like. Conversely, the approach employed here seeks to study only those individuals who are believed to be resident in the study space. The second major difference is the unit of analysis employed in this study. Longley et al. (2015) consider each Tweet independently, whereas the method demonstrated here is based on the user. Placing the focus explicitly on the Tweets, without taking into consideration who they are sent by, presents a potential source of bias whereby highly active users may overwhelm the behaviours of less active users. Such a consideration is particularly important when seeking to assess the representativeness of such data. An individual who tweets prolifically and who is a member of a minority group may have a significant impact on how the group as a whole is depicted.

It should be noted that the issue of data scope raised in the above is not limited to this study and rather is endemic within social media based analyses. Possible explanations for this behaviour include that of individual researcher's capacity to collect and manipulate the large volume of data and also the method by which Twitter data are harvested. When using the Twitter Streaming API, it is common practice to specify a geographic extent from which the data are to be collected. Doing so prevents the researcher from being inundated with data and, often, can achieve their analysis goals using standard desktop infrastructure. Researchers typically specify this extent based on the study area assuming that this is all the data that they will

need. The challenge then occurs when a researcher wishes to supplement or enrich their data based on a specific pool of users or any other requirements. The nature of the Twitter API is such that making historical queries is heavily constrained via the implementation of rate limits. Such limits prevent Twitter serving excessive volumes of data and assist in adding value to the data; historical Twitter data may be purchased from Gnip (see: www.gnip.com); a data reseller. Such limits may be circumvented. However, this is questionable both legally and ethically. Consequently, without significant financial outlay, it is near impossible to create a truly representative historical dataset. Given these challenges, in the majority of studies, these issues are either not recognised, overlooked or ignored.

Having identified a range of limitations common to social media analysis, this chapter, seeks to set a new precedent in how Twitter data are understood and consequently employed. At the very minimum, key questions will be answered such as: is the sample representative of the population I wish to study? Are the sample residents of the study area? Or, if not, where is it that they have originated? In essence, the following analysis presents a more nuanced approach to the identification and attribution of Twitter users building upon those methods set out by Longley et al. (2015) and others. Furthermore, by taking into account the limitations discussed in the above, it is believed that this analysis has a multitude of benefits. First, through addressing the above, a more accurate understanding of the Twitter demographic may be obtained. Second, by better understanding the Twitter demographic, an improved understanding may be obtained in regards to whom it is that the Twitter users are representative. In combination, it is believed that this will aid in establishing Twitter's status as a source of population insight.

6.3 Population Benchmarking

Having discussed the limitations previously, the following approach was taken to the construction and validation of the Twitter inventories. In the first instance, all users to be observed within the bounding box of the UK were identified. In turn, these data enabled the issue of non-residents and tourists being included in the analysis

to be addressed. In doing this, the objective is to isolate those individuals who are residents of the UK and, equally importantly, determine those users who are not. To understand this, the first phase of analysis was the identification of those users who are UK nationals. To achieve this, the methodology developed in Chapter 4 was implemented such that a dataset containing all of the Tweets submitted by UK users could be created. The dataset comprised 98,049,142 Tweets submitted by 495,159 users. The Tweets were subsequently joined to the UK Output Area geography and also to the GADM boundary dataset. In combination, such geographic reference provided a means to identify both individuals' countries of residence and also the locality within which they are likely to reside. Regarding the above there are two key considerations. First, the analysis of individual user's Tweets will be indicative of their country of residence and not necessarily their country of origin. Second, it should be recognised that the hierarchical nature of the location allocation method functions in such a manner as to maximise the total number of users successfully located at each geographic scale.

In the first instance, individuals were processed to determine their probable countries of residence. Of the 495,159 users, 373,456 (75.4%) were successfully allocated to a country. Those users with fewer than five Tweets within a particular country, or, less than 50% of their total Tweets account for the 24.6% decrease. It should be noted that where an individual met the criteria for association, but where the Tweets occurred outside any countries borders, the location was recorded as NA.

The bar graph of users' probable countries of origin, shown in Figure 6.1, provides a useful illustration of the top 10 countries in which the identified Twitter users are believed to be resident. It should be noted in interpreting this that the analysis does not account for differences in popularity of Twitter between countries (discussed in Chapter 5), or account for the likelihood of individuals having access to data services while travelling. Consequently, it should be assumed that some variation in the relative proportions and ranks is likely to occur. That said, several observations can be made based on the data. First, the top five countries in order of magnitude are Ireland, France, the USA, Spain and Italy. This behaviour is unsur-

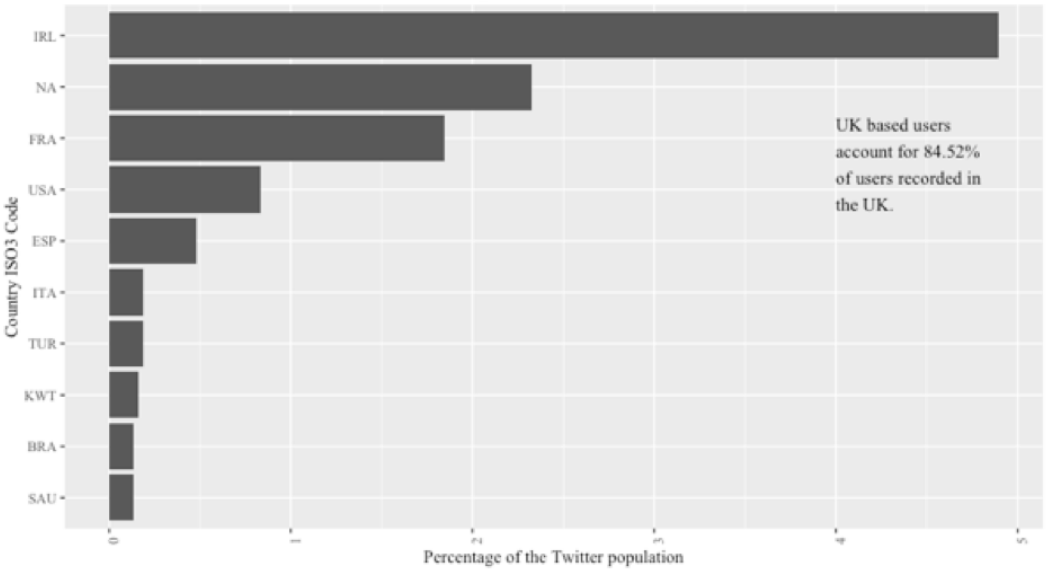


Figure 6.1: Percentage of Twitter population by country excluding those believed to be resident within the UK. Note, 'NA' indicates the proportion of individuals not assigned to any country.

prising given the geographic proximity between these countries. More surprising, however, is the presence of Kuwait, Brazil and Saudi Arabia in the top tier ranks. The driver of such popularity is likely the degree of popularity of Twitter within these countries. Thus, the relative order and magnitude of countries observed are likely to differ with the key factor being the popularity of Twitter rather than the raw number of travellers. Seeking to assess the validity of the above, data on the top 10 countries to visit the UK was sourced from the ONS Travel Trends report (ONS, 2013). This data, illustrated in Figure 6.2, depicts the number of visitors to the UK in 2013 who have resided in the country for at least one night. An assessment of the data indicates some overlap between the two datasets.

If we consider the top seven countries from Figures 6.1 and 6.2, it may be observed that France, the USA, Ireland, Spain and Italy are consistent to both datasets. Furthermore, except Ireland, the order remains consistent. It is likely, however, that the high proportion of Irish individuals observed is a consequence of the UK bounding box overlapping Ireland's eastern coast. The most notable omission from the top countries seen is Germany. This feature may be explained by Twitter's lack of popularity in Germany, however. Examined in the previous chapter, the rate of users

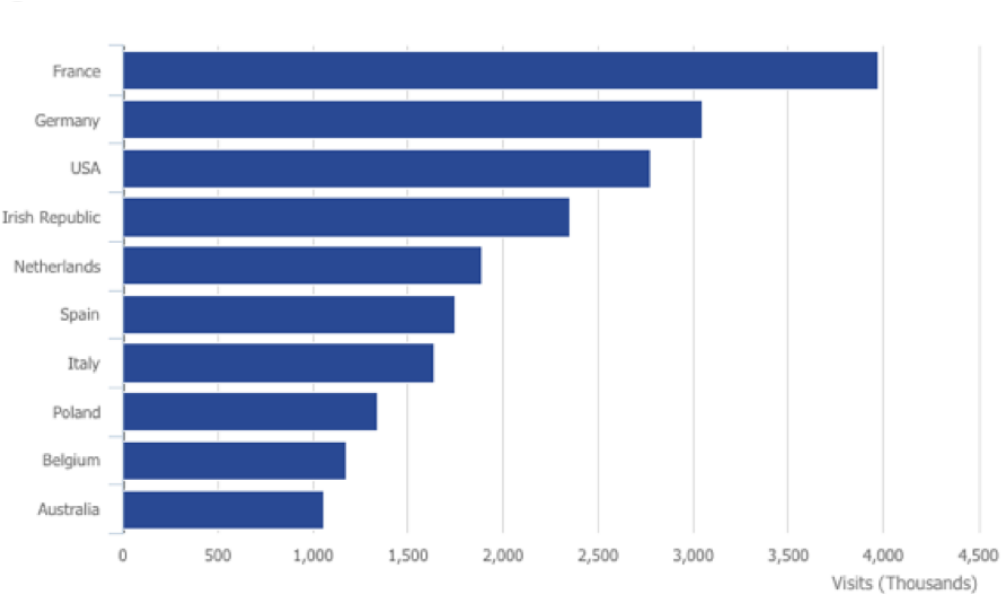


Figure 6.2: Visitors to the UK by country in 2013 as recorded by the ONS (2013)

per 1,000 of the population was just 0.22 in Germany versus 5.51 in the UK. Similarly, Poland has just 0.09 users per 1,000 of the population. It might be argued that standardisation using the country-level popularity of Twitter calculated in the preceding chapter is possible. However, in doing this, a certain degree of risk exists. With some 76 countries having fewer than 100 associated Twitter users and 20 having fewer than five, a small change in the number of users could have a significant effect on the standardisation procedure. In practice, it would seem more prudent to consider the relative popularity of Twitter as part of the interpretation rather than as a step in data processing.

At the conclusion of this section, we are in possession of the probable nationality of 75.4% of those individuals identified within the geographic extent of the UK. Of these, 373,456 users (84.8%) are believed to be resident in the UK. For this, each user will have a minimum of five Tweets and greater than 50% of their total Tweets submitted within the extent of the UK. Having identified those users who are believed resident within the UK, a dataset of Tweets associated with those users was created. Thus, we are in the best possible position to assess the representativeness of the data against the observable UK population as recorded in the UK Consumer Register.

6.4 Name Extraction

Having identified those users believed to be resident in the UK, the next consideration is the extraction of their personal names. Their names being the linchpin upon which the following identity attributes are inferred. Individual users' names were extracted using the approach outlined in Chapter 4. Initially, this methodology was designed in such a manner as to be globally applicable and lacked the finesse which could be achieved given a more bespoke analysis. In the following, a discussion is provided concerning the extraction of individual users' personal names. In particular, challenges will be discussed in the context of the UK-specific inventory. These challenges include individuals having multiple names associated with their identities and the identification and removal of non-personal accounts.

Considering first the issue of individuals bearing multiple names. While the name extraction algorithm processes the screen name(s) associated with each user, limited information is available concerning which, if any of the names correspond to the user's true forename and surname. The challenge arises when a specific user has utilised multiple screen names over the period of data collection. The scale of the problem may be examined through tabulation of the number of unique screen names associated with each user id. Table 6.1 presents a frequency table on the number of different screen names held by the UK-based Twitter population.

Table 6.1: Frequency table reporting the number of different screen names held by UK-based Twitter users.

Number of Screen Names	n	%
1	321,807	76.3
2	54377	12.9
3	19,274	4.6
4 - 5	13,693	3.2
6 – 10	7,902	1.9
11 – 109	2,942	0.7

While it is evident that the majority of users have retained a single screen name, it is clear that having two or more different names is not uncommon. Various scenarios exist in which a user may alter their names including alternating between full

and partial names, a change of relationship status resulting in a name change, or the user using their screen name to display alternative information such as to make a statement. Such behaviour raises several distinct challenges. Primarily, how does one determine which, if any, of each user's screen names, are their legitimate name? It was found that by first processing the screen names, often the extracted forename and surname were identical and that the previous distinction was a consequence of additional grammar within the name. A hypothetical example being 'James Smith 2012' and 'James Smith'. Once processed, both screen names would have been processed as 'James' and 'Smith' for forename and surname attributes respectively.

Table 6.2: Frequency table reporting the total number of segments within the screen-names of UK-based users.

Segments	Frequency
0	1,135
1	218,335
2	359,467
3	59,339
4	13,596
5	2,936
6	931
7	552
8	295
9	225
10	152

Beyond consideration of the number of screen names held, another major consideration is the names' form. Unlike many other social networks, Twitter has no requirement to report distinct forename or surname data. Rather, users are allowed to use any string of up to 20 unique characters. In Table 6.2, the frequency table of name tokens, provides a useful illustration of the distribution of the number of words within those screen names of UK-based Twitter users. The Table indicates that while the modal number of segments is 2 (359,467), a large number of users (218,335) also have single word names. Thus, approximately 65.8% of screen names may be processed. The major omissions are those names with a single segment mentioned previously. The challenge in the use of these names is to determine

whether they are names or not. A further consideration is what techniques can be applied to these assuming they are forenames or surnames. A more complex issue, not directly addressed in this analysis, is the issue of individuals who switch between the use of nicknames and full names. For Example, alternating between 'Bob Jones' and 'Robert Jones.' One approach to addressing this concern would be through the use of a name lookup table such as is published by the Web Science and Digital Libraries Research Group at Old Dominion University[1]. Compiled using genealogical data, the dataset provides a reference between some 1,600 names and common alternatives. Several commercial alternatives also exist which purport to possess significantly more names. Two companies which provide such services are http://www.BasicTech.com and http://www.PeacockData.com. An alternative approach would be to take the longest of the names identified. In the case of UK names, it is common practice that nicknames are abbreviations or contractions of full names. However, while this might work in the case of abbreviations (i.e., Rob and Robert), it would not be suitable regarding rhyming names (i.e. Bob and Rob)

Seeking to reduce the occurrences of non-human users, a blacklist of words was created for forename and surnames respectively. To optimise this process, the top 1,000 forenames and surnames were isolated respectively and then manually processed to assess their validity. Using the top 1,000 names for each name token collectively accounted for 79.4% of forenames and 46.5% surnames. The list of blacklisted words identified is shown in Table 6.3.

6.5 Demographic Assessment

In the previous section, a dataset was compiled composed of Tweets submitted by individuals believed to be UK residents. This dataset contained 35,835,966 Tweets submitted by 273,411 unique users. This section is delivered in two parts. First, the inventory enrichment framework is implemented such that each user, where possible, is attributed with key demographic markers. Second, the augmented inventory is compared against the 2011 Census of Population. The comparison will focus on

[1] https://github.com/carltonnorthern/nickname-and-diminutive-names-lookup/blob/master/names.csv

Table 6.3: Blacklisted words found in personal names.

Forenames	Surnames
THE, NA, DJ, BIG, PRINCESS, IM, ST, IG, LOVE, LONDON, MY, BLACK, THAT, YOUR, GAME, CAPTAIN, HI, IN, OLD, UK, QUEEN, WEATHER and AAP	NA, LTD, XX, FC, UK , CLUB, LONDON, PARK, READ, GUY, BOY, LOVE , ME , GIRL , MAN, XXX , WEATHER, BITCH, CO, JOBS, DESIGN, XO, EVENTS, INN, KITCHEN, CAFE, IRELAND, RICHMOND, SCHOOL, ENGLAND, HOTEL, SHOP, CURRY, JR, PT, GROUP and LIFE

the UK as a whole and the Greater London Area. The latter providing a means to compare the data against the work of Longley et al. (2015). The comparison will focus on three key metrics: age, gender, ethnicity and geographic distribution. For the purpose of analysis, census data for each England & Wales, Scotland and Northern Ireland were sourced.

In the subsequent analysis, each user is assigned an age, gender, ethnicity and geographic location based on their personal names and historical tweeting activity. Here it is important to remain mindful that the assignment is based on the general characteristics of those individuals as expressed in the source data and thus biases in the original data and heuristics will affect the assignment of demographic characteristics to individuals. Consequently, it is important that the heuristics and source data are disclosed and due diligence be completed. In the subsequent sections, an assessment is made of each classification tool such that any manifest biases can be identified.

6.5.1 Age and Gender

With regard to the inference of users' genders, various approaches have been employed across the literature. Such approaches range from the analysis of language expressed within individual Tweets to the use of specific gender identifiers such as personal names. The use of personal names, while arguably the most accurate, is constrained by the availability of suitable reference data. The majority of studies which have utilised such an approach are based in the United States where name and

gender data are made publicly available by the Social Security Administration[2]. In the case of Longley et al. (2015), a bespoke classification system for age and gender, based on CACI's Monica dataset, was employed. The exact method is outlined by Lansley and Longley (2016a). Constructed using data from credit card applications and augmented with birth certificate data, the enhanced Monica classification provides a valuable indication of age and gender structure. The classification is based on individuals' forenames and is founded on the premise that particular forenames are commonly associated with a single gender and further, that they tend to exhibit patterns of use over time (Gallagher and Chen, 2008). Further support to name-based approaches to age estimation is that they provide an additional means by which non-personal individuals may be omitted. However, it should be recognised that such an approach may ignore users whose names are not present in the reference tables. Such a limitation may lead to the omission of human users with less common names.

Prior to the application of the Monica classification, a validation of the method was performed. The purpose being to identify any pre-existing bias within the classification tool. To do this, the classification was applied to the 2013 Consumer Register provided by CACI Ltd. The objective being to ensure that the outcome accurately depicted the true form of the population before conclusions are drawn on the age structure of the Twitter users. In the case of both the validation against the Consumer Register and the Twitter individual level population inventories, the Monica classification was adjusted to model the correct age range. In the case of the Consumer Register, the classification is limited to those aged 15 and over, and for the Twitter classifier, the lower age limit was set as 10. The results of the classification test using data from the UK Consumer Register are illustrated in the form of a population pyramid below.

Figure 6.3 provides a demonstration as to the effectiveness of the enhanced Monica classification in the modelling of populations based on individuals' names. While some discrepancies do exist, for example in the 15-19-year-old female group,

[2]The Social Security Administration name data are available from 1881 and include counts of names by gender. The highest geographic resolution is State level. The data are available from https://www.ssa.gov/oact/babynames/limits.html

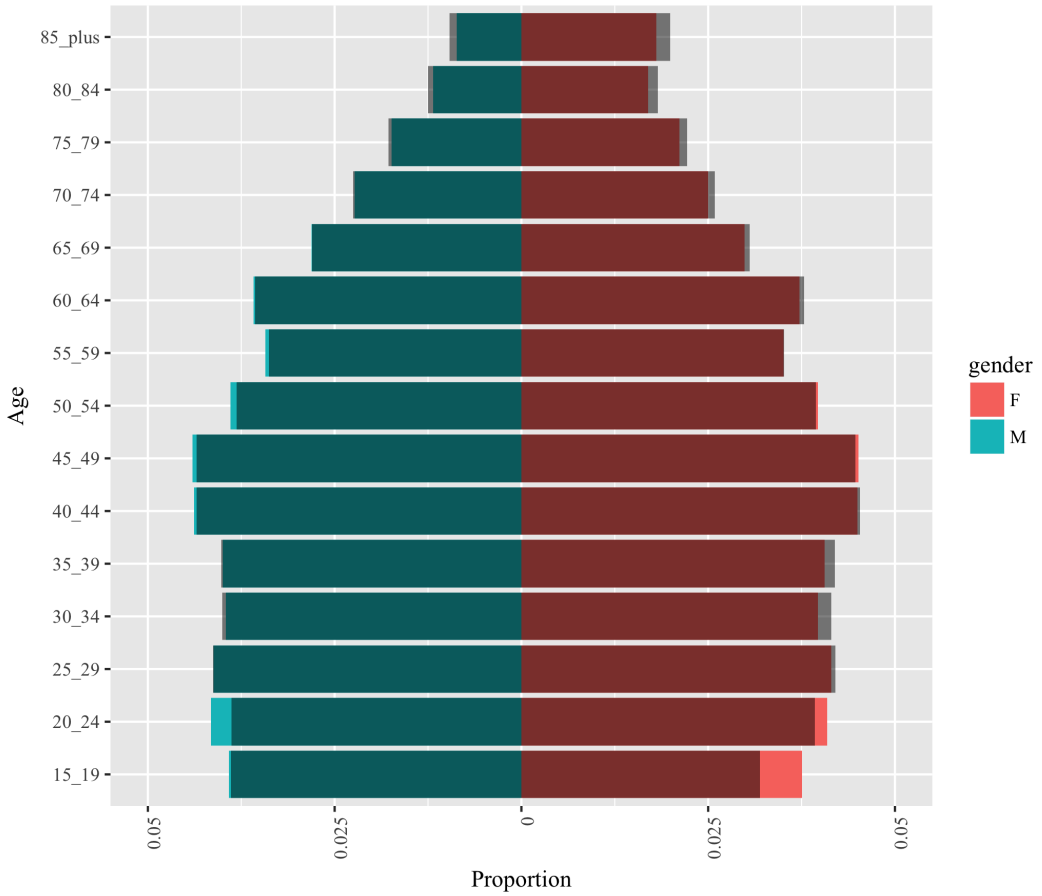


Figure 6.3: Population pyramid based on the Consumer Register versus the equivalent data sourced from the 2011 Census of Population. The Census data are depicted in grey and the equivalent Consumer Register data in red and blue.

the overall agreement between the two profiles is significant. Consequently, when applying the Monica classification, we can have confidence that the age structure observed is an accurate representation rather than simply a manifestation of bias inherent within the classification itself. In the knowledge that the Monica classification is an effective means of inferring age and gender, the first assessment was to determine if a gender bias existed within the Twitter inventory. The comparison is performed for the UK as a whole and London.

Tables 6.4 and 6.5 and provide a useful summary of the gender bias observed in the UK and London more specifically. In particular, it may be observed that their exists a consistent gender bias towards males who are typically over-represented by approximately 15%. Across the full population this appears to be fairly consistent.

Table 6.4: Table showing the proportion of UK users by gender versus the population data for 2013.

	Twitter		UK		Quotient
Female	111,235	0.40 (0.45)*	32,572,781	(0.508)	0.886
Male	133,485	0.49 (0.55)*	31,532,873	(0.492)	1.118
Unknown	3,141	0.01		n/a	
NA	25,550	0.1			
Total	273,411	1			

Table 6.5: Table showing the proportion of London users by gender versus the population data for London 2013.

	Twitter		London		Quotient
Female	13534	0.37 (0.44)	4,251,200	(0.505)	0.87
Male	17136	0.46 (0.56)	4,165,335	(0.495)	1.13
Unknown	633	0.017		n/a	
NA	5591	0.15			
Total	36894	1			

Though, as may be observed in Figures 6.3 and 6.5, the behaviour exhibits a clear association with age.

Having established that the gender bias at the UK scale is towards male users, the next assessment is concerned with age structure. It is well recognised that social network usage, specifically in regards to Twitter, is most concentrated within the younger age cohorts. This pattern is clearly evident in Figure 6.4 in which one can observe significantly better representation in the younger age bands. Interestingly, the transition between over and under representation occurs at different ages based on gender with males transitioning in the 35-44 age bracket while females transition in the 30 – 39 age bands. Such a gender divide may be indicative of differences in activity between genders. Beyond these transitional points, the decrease in usage is more rapid for females with the male bias becomes increasingly dominant over the age of 40.

Figure 6.5, the comparative gender plot, provides a clear depiction of the relationship between age and gender bias in the UK. In the youngest band, 10-14, both genders are represented relatively equally, however, above this, there is consistent

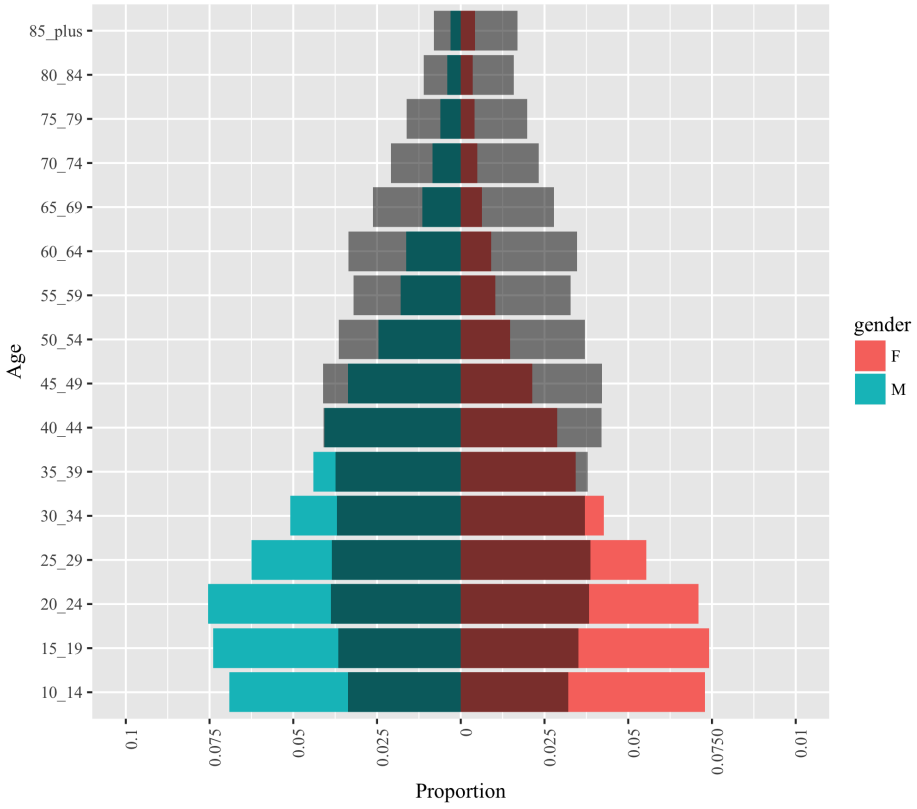


Figure 6.4: Population pyramid of Twitter users in the UK versus the equivalent ONS data for 2011. The ONS data are depicted in grey.

over-representation of males. The exception is in the 85 plus group where females become more dominant. This is unsurprising given the relatively greater life expectancy of women.

Having observed the general UK trend, the same comparative analysis was performed for those individuals users believed resident within Greater London. The purpose of this comparison being two-fold. First, as a means to validate this analysis against that of the Longley et al. (2015) study and second, as an initial assessment as to whether there is any evidence of a geography to how well Twitter depicts the population. Considering first age-based patterns, the transition between over and under-representation is younger in London than the UK: 25-29 for males and 20-29 for females. Regarding general population structure, it would appear that while age and gender structure vary around the country, the demographic of Twitter users remains largely consistent. Typically, one would observe a largely young population with a general decline beginning in the late 20s. Also, around the early 30s, the

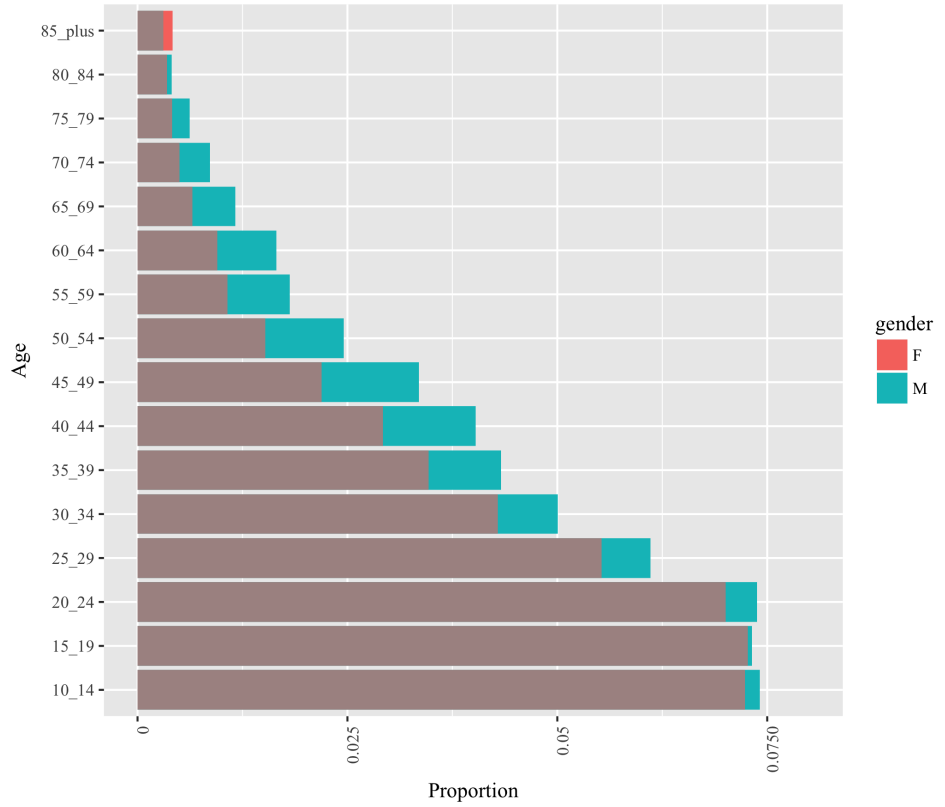


Figure 6.5: Gender comparison population pyramid for the UK Twitter Population.

Twitter population becomes increasingly dominated by males.

Comparison of the London scale outcome against the results from the Longley et al. (2015) study highlights various similarities and differences. While the observed gender bias was practically identical, the age distribution differs. Notably, whilst Figure 6.6 indicates the 15-19 group as being similar to the 10-14 and 20-24 age groups, the study indicates this group to be approximately 100% larger. This is surprising given that both studies relied upon the same corpus of Tweets and also employed similar approaches to the estimation of users' ages. One possibility is that average ages were used rather than the age distributions, or the analysis was based on Tweets rather than users. This could have a significant impact due to common names which exhibit relatively flat distributions being interpreted based on just the most common age. Examples of such names include James and David, both which have remained popular.

Extending the analysis, it would be interesting to understand differences in be-

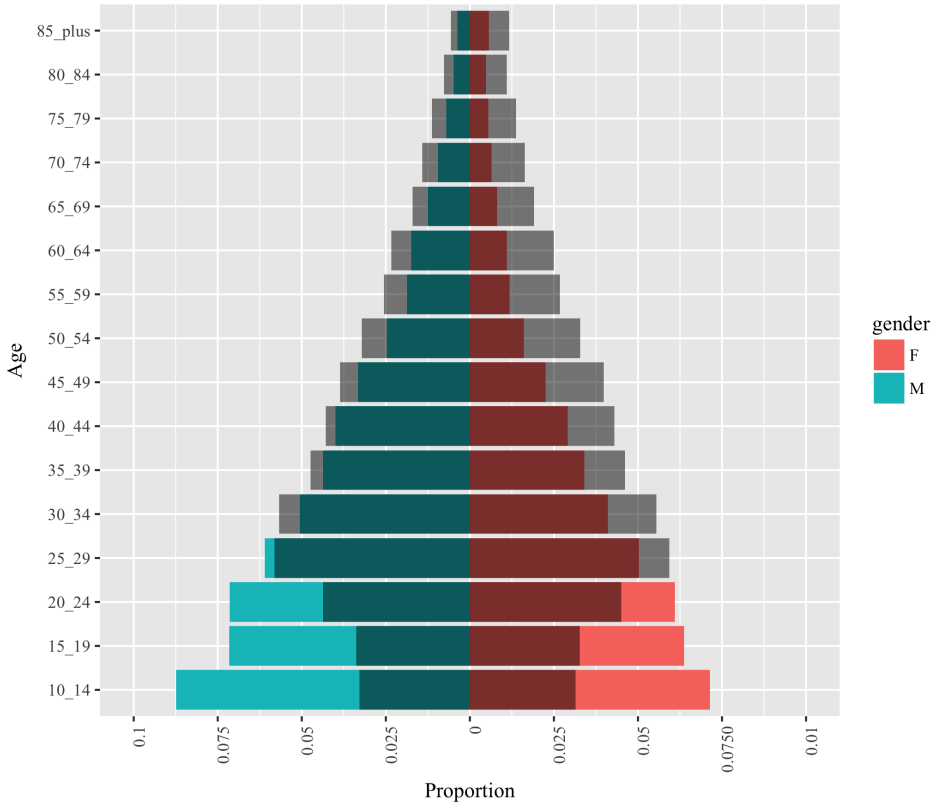


Figure 6.6: Population pyramid of Twitter users in London versus ONS data for 2011. The ONS data are depicted in grey.

haviour between genders and the number of Tweets submitted. For example, does gender influence the frequency of tweeting and do different age groups tweet at different rates. This in itself may provide a useful indicator of how Twitter may be applied in studying population stocks and flows. For example, while considered as a whole may appear to be representative across a broad spectrum of the population, if one specific group, for example under 25s, tweeted at a much greater rate then this could potentially lead to a biased sample being utilised within the study.

Importantly, the comparative analysis demonstrated that there is a geography to the distribution of age and gender within the UK regarding the true population. This is, however, less pronounced in the Twitter data which appears to have a consistently younger user base. Further, there is an apparent geography to the uptake of social media by each gender. London, for example, appears to have a greater bias towards males than is observed in the UK. This observation highlights the importance of understanding local social media demographics when making any decisions based

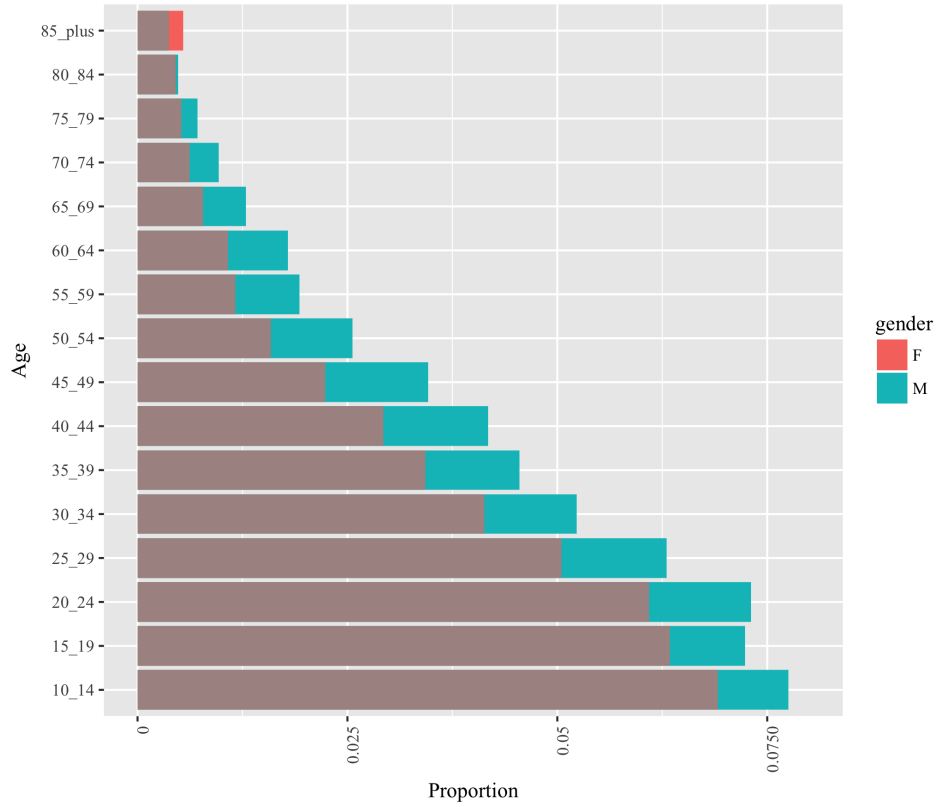


Figure 6.7: Gender comparison population pyramid for the London Twitter population.

on the analysis of the data.

6.5.2 Ethnicity

Having established the degree to which the Twitter-derived population inventory is representative regarding age and gender, the next concern is that of ethnicity. The availability of data about ethnicity enables analysts to better explore patterns of behaviour.

A major concern in the application of the Onomap classification is the degree to which the observed results are an accurate depiction of the population, or rather, an illustration of preexisting bias within the Onomap classification itself. Should this bias be the case, there is a significant risk of misinterpreting, and by effect, misrepresenting the actual ethnic composition of the Twitter population. In seeking to determine the extent of this bias, it was proposed that the Onomap classification be first applied to the 2013 Consumer Register. The justification for such a move, as was the case with assessing Monica, was that the Consumer Register is arguably the

most inclusive record of the population in the UK. Though, in utilising the Consumer Register, it is recognised that the Edited Electoral Register, the mainstay of the Consumer Register, is known to under-represent various minorities.

For the purpose of comparison, a lookup table was created between the Onomap CEL Groups and an aggregate classification of UK-wide Census ethnicity groups. An introduction to the Onomap classification is provided in Section 2.4.2.3. In essence, the Onomap classification scores individuals forenames and surnames using a heuristic algorithm which assigned individuals to a specific CEL group based on the strength of association. The aggregate classification of ethnicity was created for the purpose of linking the three different UK Censuses. While each of the three administrations collects data on ethnicity, the classification employed is not consistent. The relative percentages of each population by ethnic group are shown in Table 6.6. To aid in comparison, the quotient for each group is reported. A quotient value of greater than 1 indicates a higher proportion that expected and less than 1 indicated less.

Table 6.6: Ethnicity breakdown comparison between the 2013 Consumer Register and 2011 Census of Population.

Aggregate Census Group	CR2013 Onomap	Census 2011	Quotient
White - All - Gypsy - Traveller - Irish Traveller	92.42	87.2	1.06
Asian - Asian British - Indian	1.97	2.30	0.86
Asian - Asian British - Pakistani	1.79	1.90	0.94
Black - African - Caribbean - Black British	0.85	3.00	0.28
Asian - Asian British - Other Asian	0.72	1.40	0.51
Asian - Asian British - Bangladeshi	0.40	0.70	0.57
Asian - Asian British - Chinese	0.32	0.70	0.46
Mixed - Multiple Ethnic Groups	0.00091	2.00	0.00
Other Ethnic Group	1.52	0.90	1.69

Inspection of Table 6.6 provides a range of insight regarding the effectiveness of the Onomap tool in the classification of ethnicity. In general, with the exception of the 'Mixed Multiple Ethnic Group' category, the Onomap classification has performed well. The explanation for the 'Mixed Multiple Ethnic Group' is that Onomap tends towards discrete classification or other, rather than suggesting 'Mixed'. This is evident from the 'Other Ethnic Group' category. The main group, 'White', is slightly over-represented in the Consumer Register which is unsurprising given that this group is the best represented in the electoral roll. It is likely that the

over-representation in this group is responsible for the slight under-representation in other groups, specifically, the 'Indian' and 'Pakistani' groups. The other three Asian groups are also under-represented in the Consumer Register with a typical quotient of around 0.5. The consistency between the Asian groups suggests a systematic under-representation. Lastly, the 'Black' group is the most significantly under-reported with a quotient of 0.28. Given that the Consumer Register is so comprehensive, it would appear that the differences in classification between the classified Consumer Register are a result of the Onomap classification tool rather than an issue with the register itself.

Table 6.7: Census response rates for England and Wales by Ethnic Group 2011 (ONS).

Ethnic Group	Persons (%)	Males (%)	Females (%)
White: English/Welsh/Scottish/Northern Irish/British	95.1	94.4	95.9
White: Irish	94.0	92.4	95.5
White: Gypsy or Irish Traveller	90.1	88.9	91.2
White: Other White	90.3	88.0	92.4
Mixed/multiple ethnic groups: White and Black Caribbean	83.4	81.7	85.0
Mixed/multiple ethnic groups: White and Black African	82.8	80.6	85.0
Mixed/multiple ethnic groups: White and Asian	85.4	83.3	87.5
Mixed/multiple ethnic groups: Other Mixed	82.5	79.9	85.0
Asian/Asian British: Indian	94.3	92.9	95.7
Asian/Asian British: Pakistani	93.5	92.5	94.6
Asian/Asian British: Bangladeshi	92.5	91.1	94.1
Asian/Asian British: Chinese	84.6	81.2	87.6
Asian/Asian British: Other Asian	85.1	81.9	88.1
Black/African/Caribbean/Black British: African	88.2	85.8	90.5
Black/African/Caribbean/Black British: Carri bean	91.9	90.3	93.4
Black/African/Caribbean/Black British: Other Black	64.0	60.1	68.1
Other ethnic group: Arab	72.4	68.7	77.5
Other ethnic group: Any other ethnic group	74.0	69.5	79.4

Table 6.8: Estimated electoral registration rates of Census respondents by ethnic group 2011 (ONS).

Ethnicity	Cases	Estimated Registration Rate (%)	95% Confidence Interval (%)
White	35,158	88.8	87.8 - 89.7
Mixed	735	79.3	75.2 - 83.4
Indian	1,763	85.4	82.7 - 88.2
Pakistani	1,203	80.5	75.9 - 85.0
Bangladeshi	609	79.7	74.4 - 84.9
Other Asian	881	80.4	76.1 - 84.6
African	963	75.4	71.2 - 79.6
Caribbean	912	84.1	80.9 - 87.3
Other Black	190	75.5	68.1 - 82.9
Other	345	78.7	72.6 - 84.9
Unknown	478	73.2	67.3 - 79.0

That said, the observations in the above are, in part, supported by known under-reporting in the Census of Population and the Electoral Register. Tables 6.7 and 6.8

provide a valuable illustration of differences in census and electoral roll completion/registration dates from the Census collection for England and Wales. Noticeably, that in both cases, the 'Black' groups are typically the most under-represented. Furthermore, comparison of the two tables suggests that under-reporting and under-representation are a consistent issue in the collection of national datasets.

In 2010, the Electoral Commission published a report on the completeness and accuracy of electoral registers in Great Britain. Whilst it is noted in the report that the work is based on an unrepresentative sample of electoral regions, the results do provide some insight into the patterns of representation. Notably, under-representation of the 'Black and Minority Ethnic' residents. Approximately 31% of those individuals in the 'Black and Minority Ethnic' groups are unregistered versus 14% in the 'White British' group. Based on this observation, it might be argued that the under-representation identified in the 2013 Consumer Register is not a consequence of classification, and rather, that the group is notably under-represented in the register as a whole.

If we make the assumption that the bias in the Onomap classification is systematic, it may be more appropriate to compare the Twitter ethnic composition against the comparable data from the Consumer Register. In effect, both datasets have been processed under the same bias assumption. An added advantage of such an approach is that the assessment of representativeness can be conducted on an annual basis rather than on a decennial basis in line with the publication of the Census.

Table 6.9: Ethnicity breakdown comparison between the UK Twitter Population and 2011 Census of Population.

	Twitter Onomap (%)	Census 2011 (%)	Quotient
White - All - Gypsy - Traveller - Irish Traveller	93.36	87.2	1.07
Asian - Asian British - Indian	1.36	2.30	0.59
Asian - Asian British - Pakistani	1.11	1.90	0.58
Black - African - Caribbean - Black British	0.75	3.00	0.25
Asian - Asian British - Other Asian	0.77	1.40	0.55
Asian - Asian British - Bangladeshi	0.23	0.70	0.33
Asian - Asian British - Chinese	0.54	0.70	0.77
Mixed - Multiple Ethnic Groups	0.0006	2.00	0.00
Other Ethnic Group	1.88	0.90	2.09

Table 6.9 presents the comparison between the Onomapped Twitter inventory and the aggregate 2011 Census data. Here, a similar pattern of over and under-

representation is observed with the collective 'white' group over-represented and the Black and Bangladeshi Groups being most under-represented. This view again supports the hypothesis that comparison against the classified Consumer Register is a more appropriate means of validation.

Table 6.10: Ethnicity breakdown comparison between the UK Twitter Population and the 2013 Consumer Register.

	Twitter Ono. (%)	CR 2013 Ono. (%)	Quotient
White - All - Gypsy - Traveller - Irish Traveller	93.36	92.42	1.01
Asian - Asian British - Indian	1.36	1.97	0.69
Asian - Asian British - Pakistani	1.11	1.79	0.62
Black - African - Caribbean - Black British	0.75	0.85	0.88
Asian - Asian British - Other Asian	0.77	0.72	1.07
Asian - Asian British - Bangladeshi	0.23	0.4	0.58
Asian - Asian British - Chinese	0.54	0.32	1.69
Mixed - Multiple Ethnic Groups	0.0006	0.0009	0.62
Other Ethnic Group	1.88	1.52	1.24

Table 6.10 provides the comparison between the individual level Twitter population and 2013 Consumer register. Here it is evident that some of the bias previously observed has decreased in magnitude; most noticeably in the case of the collective 'White' group. Consequently, the quotients suggest an improvement in how well the Twitter population depicts the 'observable' population. In fact, the 'Black' Ethnic Group, previously the most under-represented, appears to be closer to what would be expected. That said, with the exception of the 'Chinese' and Other Asian groups, there remains a consistent level of under-representation.

In Figure 6.8, the bar plot shows the breakdown of ethnic groups in the UK based on the three different population representations: Twitter, the 2013 Consumer Register and the aggregate Census data. The Consumer Register and Twitter both having been processed with Onomap. Considering a comparison between the Census and Consumer Register data, shown in Table 6.6, which there is consistent under-representation in all groups except the 'White' group and the 'Other' category. That said, the order of magnitude by which said discrepancies varies notably. To indicate the degree of difference, the quotient is reported as the Consumer Register percentage divided by the Census percentage. Consequently, a positive value indicates that the group is over-represented in the Consumer Register and vice versa.

In summary, the analysis of how representative the Twitter population are re-

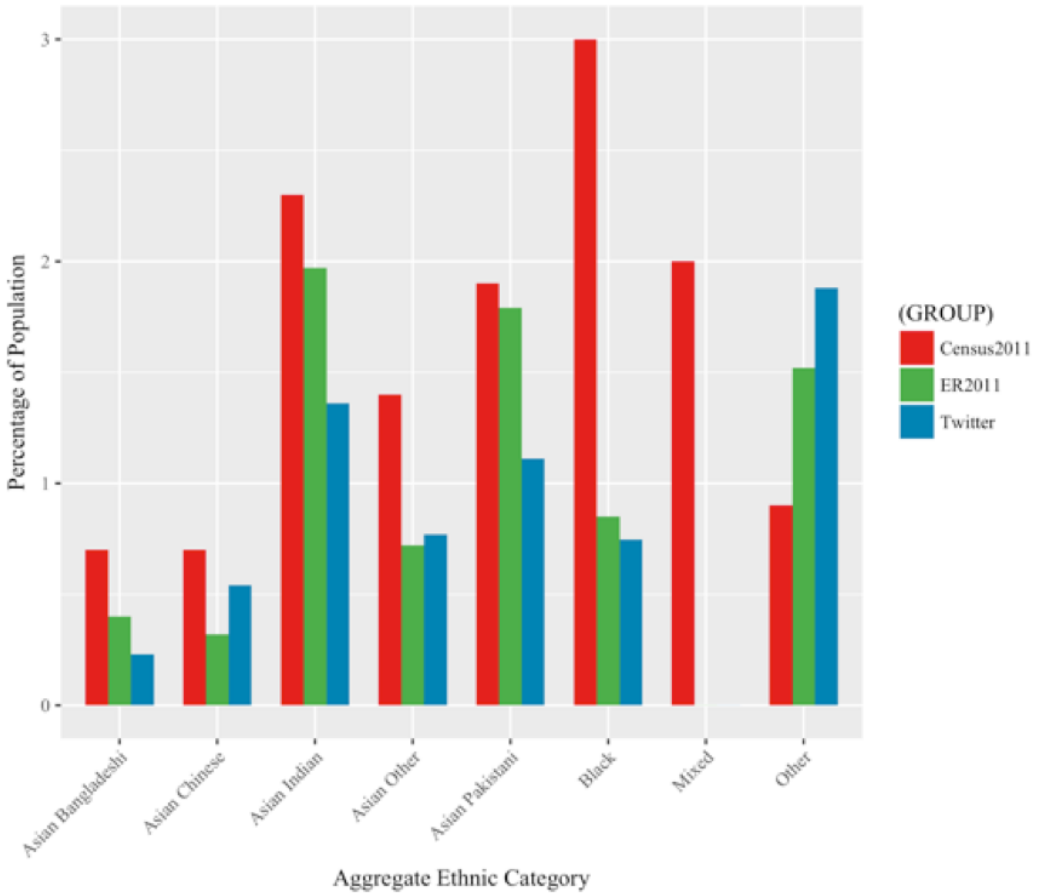


Figure 6.8: Plot showing the ethnic breakdown between the UK Census of Population in 2011, the 2013 Consumer Register and the Twitter population. Note that the data for the White Ethnic Group are omitted. The figures for the White group were 87.2% (UK Census), 92.42% (CR2013) and 93.36% (Twitter).

garding ethnicity in the UK has been quite inconclusive. In seeking to validate the Onomap tool, it was observed that the outcome was significantly different from what would have been expected given the equivalent data sourced from the UK Census of Population. Consequently, caution should be exercised when making any conclusions regarding how well Twitter represents the population regarding ethnicity. This area of ethnicity classification, in particular, would be a valuable future research direction.

6.5.3 Geographic Distribution

Having assessed the Twitter population regarding age, gender and ethnicity, the final consideration is that of geographic distribution. The purpose here is twofold. One,

to gain an understanding as to the geography of Twitter users within the UK and two, to ascertain and inform others on what is a suitable scale for analysis. In seeking to examine this, the Twitter population inventory was processed such that it depicted the population at a range of spatial scales. As was noted previously, the methodology is hierarchical in nature meaning that as the resolution is decreased, the valid population will increase in size. This analysis was performed at the following scales: Output Area, Lower Super Output Area, Ward, Middle Super Output Areas and Local Authority and Unitary Authority. It is important to note that as the number of areas is reduced, not only are more users successfully assigned to a specific region, the number of individuals within regions will rise meaning a potentially larger and more representative population.

In many respects, the decision over what scale of analysis should be employed is guided by the phenomenon that is being studied. However, as has been indicated in the above, the greater the level of granularity, the lower the quality of the Twitter-inventories. Thus, there is likely to be a requirement for compromise in any analysis that is performed. Often, we are simply interested in determining regional views or sentiment. In such cases, the finer geographies such as OA and LSOA are irrelevant and may lead to omission of useful data or false precision. Geographies such as Unitary Authorities and Districts and Travel to Work areas then become more appealing.

Table 6.11: Count of valid Twitter users in the UK at a range of spatial scales.

Geography	Users	Areas	Users per Area
OA	128,974	232,033	0.6
LSOA	170,484	42,622	4.0
Ward	184,961	9,199	20.1
MSOA	190,190	8,484	22.4
LAUA	235,278	394	597.2

For reference, a study by Mislove et al. (2011), analysing age, gender and ethnicity in the United States based on Twitter used the geographical unit of counties. Given that Mislove et al. (2011) identified 3,279,425 unique Twitter users, and there are 3144 US counties, this equates to approximately 1,043 Twitter users per county.

This number was considered sufficient to draw out national insight. Table 6.11 provides an illustration of potential samples sizes given varying scales of analysis in the UK. From the table, it is evident that while the Twitter population is relatively large, once split across the various administrative spaces, the data rapidly become sparse. In this case, Local Authority and Unitary Authority districts would appear to be the highest practical resolution for the study of population stocks. The following two maps present the LQ across the UK for the UK as a whole and an excerpt for London.

Figure 6.9, the LQ map provides a useful illustration as to the geography of Twitter use in the UK. Areas marked in green indicate a greater than expected number of Twitter users while red indicates fewer. Examining the UK as a whole, it is clear that there exist a geographic behaviour with a roughly N to South progression from over to under-representation.

An investigation conducted by British Telecom in 2012 interviewed approximately 2000 individuals regarding their use of and preference for Online Social Media. The analysis suggested that Scotland has the highest proportion of Social Media users in the UK with 48% making regular use of the Internet. These figures were 43% in Wales, 20% in Wales, 33% in the North East, 39% in the South West and 45.7% in London. While these figures cannot be substantiated, they do provide some further context to the patterns observed.

Looking more closely at London, illustrated in Figure 6.10, it is evident that a degree of banding exists with general under-representation. The exception to this being Westminster and the City of London which are for this analysis considered as a single region. It is likely that differentiating between Westminster and the City of London would highlight more significant over-representation within the City of London. The bias found in this area is undoubtedly associated with the high daytime population associated with the large influx of workers.

The five areas with the highest over-representation are Cardiff, Gloucester, Liverpool, Swansea and Lincoln. A feature across the five regions is that the areas have younger than average populations. Given that Twitter is biased towards a younger

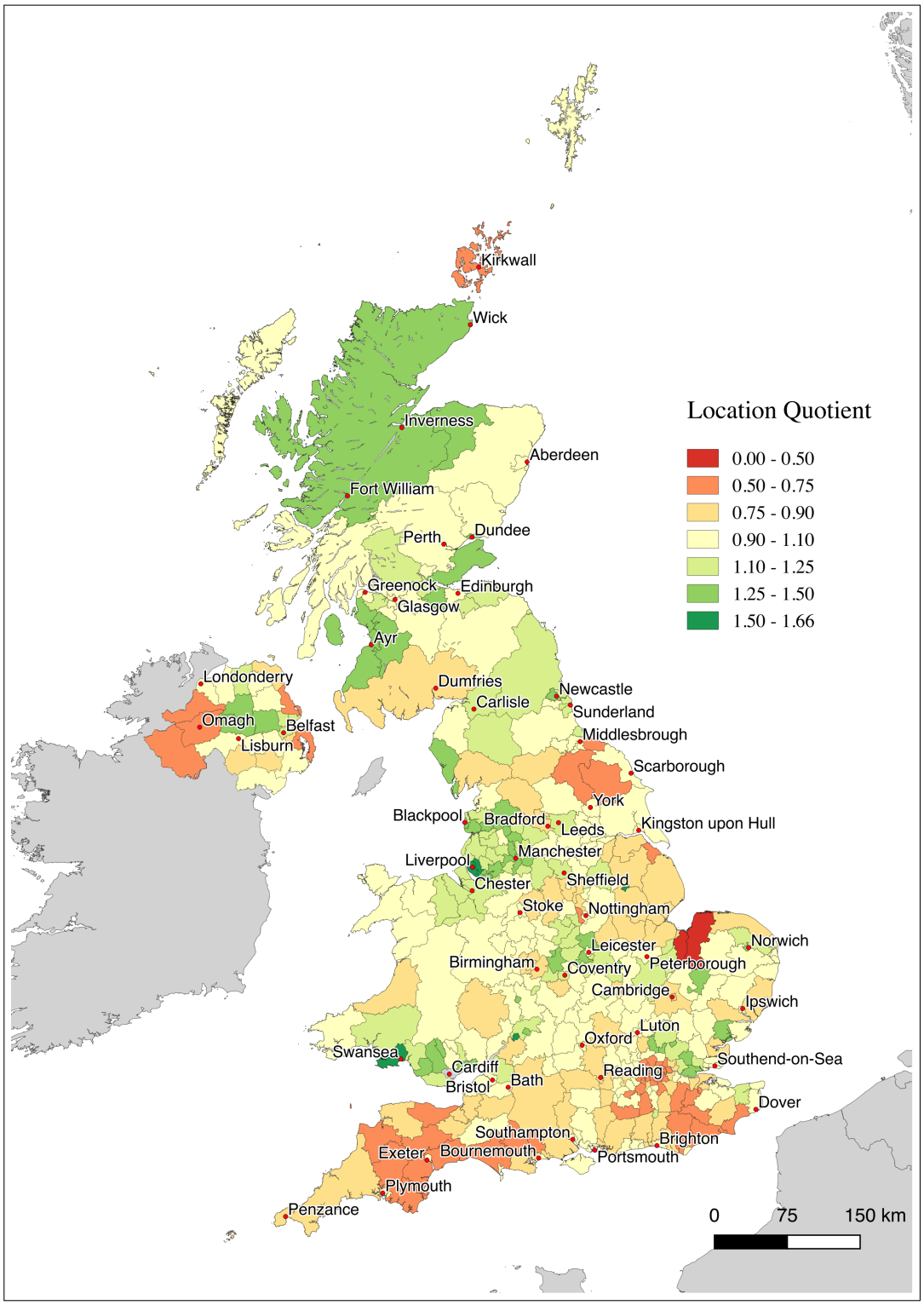


Figure 6.9: Map showing the LQ of Twitter Users versus all usual residents as recorded in the 2011 Census of Population.

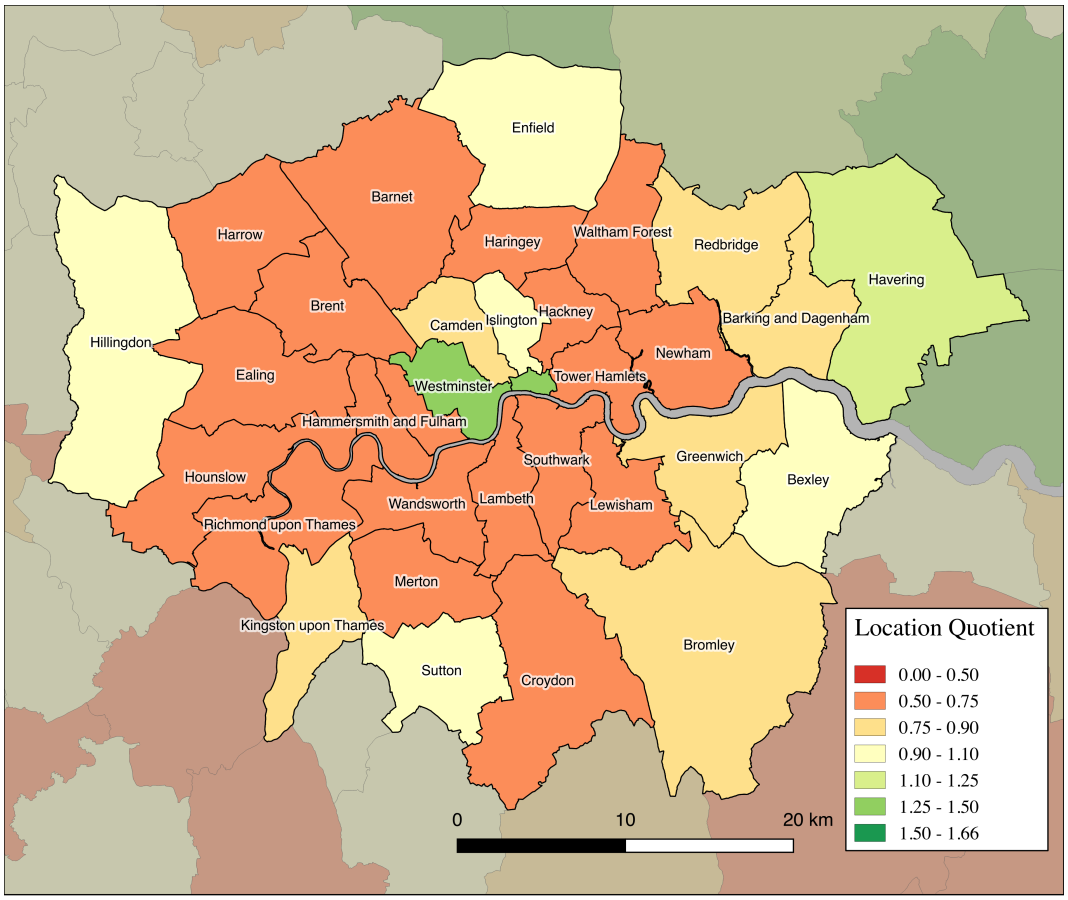


Figure 6.10: Inset of UK-wide map (Figure 6.9) showing the LQ of Twitter Users in London versus all usual residents as recorded in the 2011 Census of Population.

demographic the effect is magnified. In each case, the regions have significant student populations. The effect being to artificially inflate the size of the population. However, the students are in effect temporary migrants leaving the place of education on completion of their studies. The effect is to introduce a greater number of individuals who are likely to use social media services.

The five areas with the greatest under-representation are King's Lynn and West Norfolk, Wandsworth, Brent, Harrow and Kensington and Chelsea. Considering the four London regions, the key factor influencing the comparison is the degree to which the resident population travel beyond their region of residence. In each of the four cases, a significant portion of the population travels to the City of London for work. Consequently, a proportion of those individuals resident within each of the areas are likely to have been miss allocated reducing the size of the observed Twitter

populations. Conversely, this would have led to the over-representation within the City of London. The authority of King's Lynn and West Norfolk was the most significantly under-represented by Twitter. Unlike areas with similar populations, such as Poole with 399 Twitter users, just 2 users were assigned to the region.

As has been discussed at various points over the course of this thesis, the spatial units employed presents a distinct challenge in the application of geotagged social media. While it is desirable that all data be considered in the identification of individual's probable places of residence, where individuals move between regions for work, as in the case of the City of London, it is not uncommon for individuals to be misallocated. This issue is arguably most evident in large metropolitan areas such as Greater London.

The result of this analysis reinforces the idea of employing Travel to Work Areas for the analysis of Twitter data. As noted previously, Travel to Work areas have several key advantages. Constructed using Census commuting data, the Travel to Work Areas split the UK into 228 regions. In each case, 75% or more of individuals live and work within the same area. Further, London, an area previously discussed as being the most biased, is represented as a single region. Lastly, the classification is constructed through the aggregation of existing census geographies which provides a valuable means to link data.

6.6 Discussion

A common criticism of social media based analysis is that the data are not representative of the population as a whole. Significant analysis has been and continues to be performed with little or no regard to these limitations. Too often, data collected via online social networks are labelled under the heading of Big Data and are consequently considered rich. This idea appears to resonate with the early statement by Wired Magazine proclaiming that Big Data marked the end of theory and that in turn, the scientific method was now obsolete (Anderson, 2008). In practice, once the data are processed and filtered for the study, they are often sparse and provide a much poorer descriptor than may at first have been hoped. Bollier and Firestone

(2010) note that Big Data is driven to a greater degree by storage capability and hype more than by the superior means by which the data may be analysed. Case in point being the data employed in the completion of this thesis. From a pool of 1.4 Billion Tweets, submitted by 24.4 million users, just 206,825 of those users are incorporated into the final analysis of the UK. However, had the original dataset not been collected, a significant amount of valuable information would have been missed.

Furthermore, the observation is made that the data are a self-selecting sample of an already self-selected group. In the case of this thesis, these data are from those individuals who have chosen to use the Twitter social network and further, have made the active decision to share their location information. Such behaviour leads to bias within the sampling frame which, unlike a random or stratified sample is not easily quantified by the researcher. In the case of Twitter, the situation is further exacerbated by the lack of any explicit demographic tagging beyond names. This thesis, along with select other studies, has sought to address this criticism through the development and implantation of methods to infer key identity-related information. Subsequently, these analyses provide a means by which others may make more informed decisions regarding the analysis which they wish to perform. To illustrate the significance of the above, two cases are provided. The first looking at the outcome of the 2016 UK vote on EU membership, the second being the 2015 UK General Election. Both examples in which Twitter-based analysis incorrectly predicted the vote outcome.

First, given a theoretical goal to predict the outcome of the 2016 UK Vote on EU membership, a researcher may wish to analyse sentiment on social media concerning each outcome and use this as the basis of their prediction. In this situation, it is quite probable that the researcher would make the incorrect conclusion that the Remain Vote would be a clean win. Why is this? Survey data collected by Statista (2016) clearly indicate an age divide in voter preference with those below 50 being largely pro Remain and over 50's largely pro Leave. The problem then arises that Twitter is over-representative (based on age) of the Remain campaigners and under-representative of the Leave campaigners.

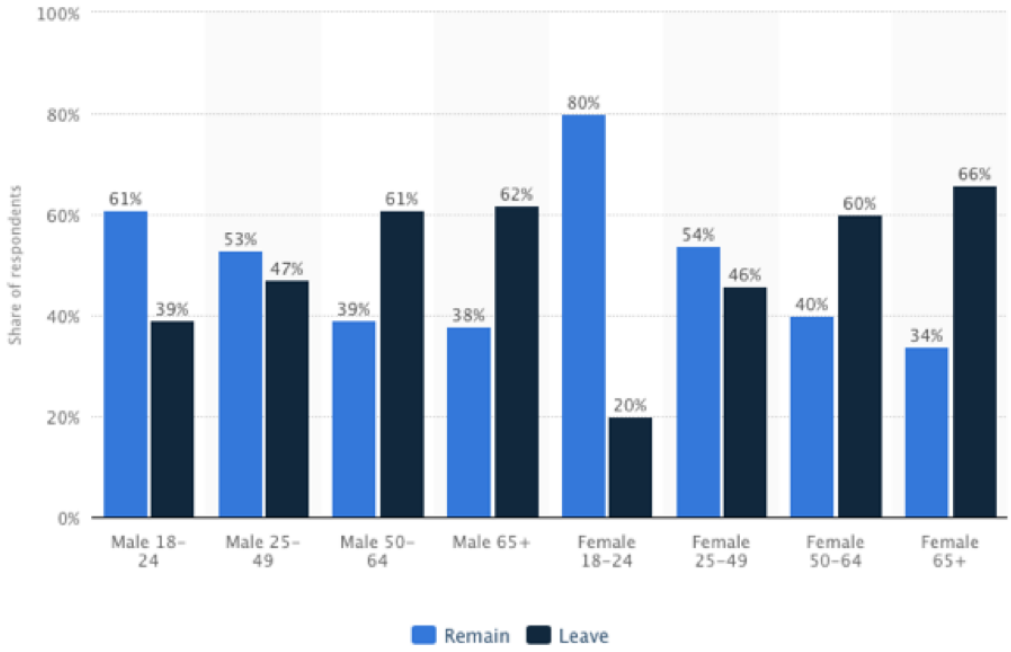


Figure 6.11: Voting preference by age and gender in the UK EU Referendum 2016 (source: Statista, 2016).

Had the researchers had access to this information, they would then have been able to weight their results based on what is known about the age bias, or, select an alternative form of data collection for their study. Such age and gender transitions regarding voting preference are by no means unique. A similar age transition is observed in voting behaviour in the UK. Typically, younger voters express a greater preference for the left which steadily transitions towards the right as voters get older. That is not to say that voting preference changes, rather, the behaviour may be associated with the shift of specific generations.

Table 6.12: Voting preference by age in the 2015 UK General Election (Ipsos Mori, 2015).

	Conservative (%)	Labour (%)	Lib. Dem. (%)
18-24	27	43	5
25-34	33	36	7
35-44	35	35	10
45-54	36	33	8
55-64	37	31	9
65+	47	23	8

Data from Ipsos Mori (2015), shown in Table 6.12, illustrates the percentage of the vote for three of the main UK political parties based on age. It can be seen that while the Liberal Democrats remain relatively consistent across years, a major transition occurs between the Conservatives and Labour. Similar to what was discussed regarding the EU vote, the data available from Twitter over-represents the 18 – 34 age bands and is consequently more likely to represent left-leaning views.

A perfect illustration is provided in a study by Burnap et al. (2016) seeking to predict the outcome of the 2015 UK General Election. Conducted before the election, the study used semantic analysis of political Tweets as a means to predict the eventual victor. Their analysis suggested a labour working majority 21 seats. In reality, the Conservative won the election with a working majority of 12. Whilst this observation is made with the benefit of hindsight, had the study taken account of the demographic structure of Twitter users in the UK and the association between age and voting preferences, it is entirely possible that they would have accurately predicted the referendum outcome. This would have been achieved by applying a weighting in favour of the conservatives who would appear to have been under-represented in the Twitter population. This could have been further developed through applying weightings in a gender-specific manner.

The second consideration for much social media analysis raised and addressed in this study is that of inclusivity. It was noted how, in much analysis, the data employed in the study are constrained solely to the study area. Such an approach raises various issues. First, an assumption is made that all individuals identified within the study area are, in fact, resident. This is clearly not the case. In the case of this analysis, it was shown that just 84.52% of those Twitter users recorded within the bounding box were, in fact, UK residents. It is quite probable that this figure is higher in London. This is not to say that the inclusion of individuals who reside outside the UK is a negative feature of the data. Rather, it provides a useful means to subset the population such that it is comparable to conventional population data. It is important to note that such insight would not have been possible had the original data collected been constrained to the extent of the UK.

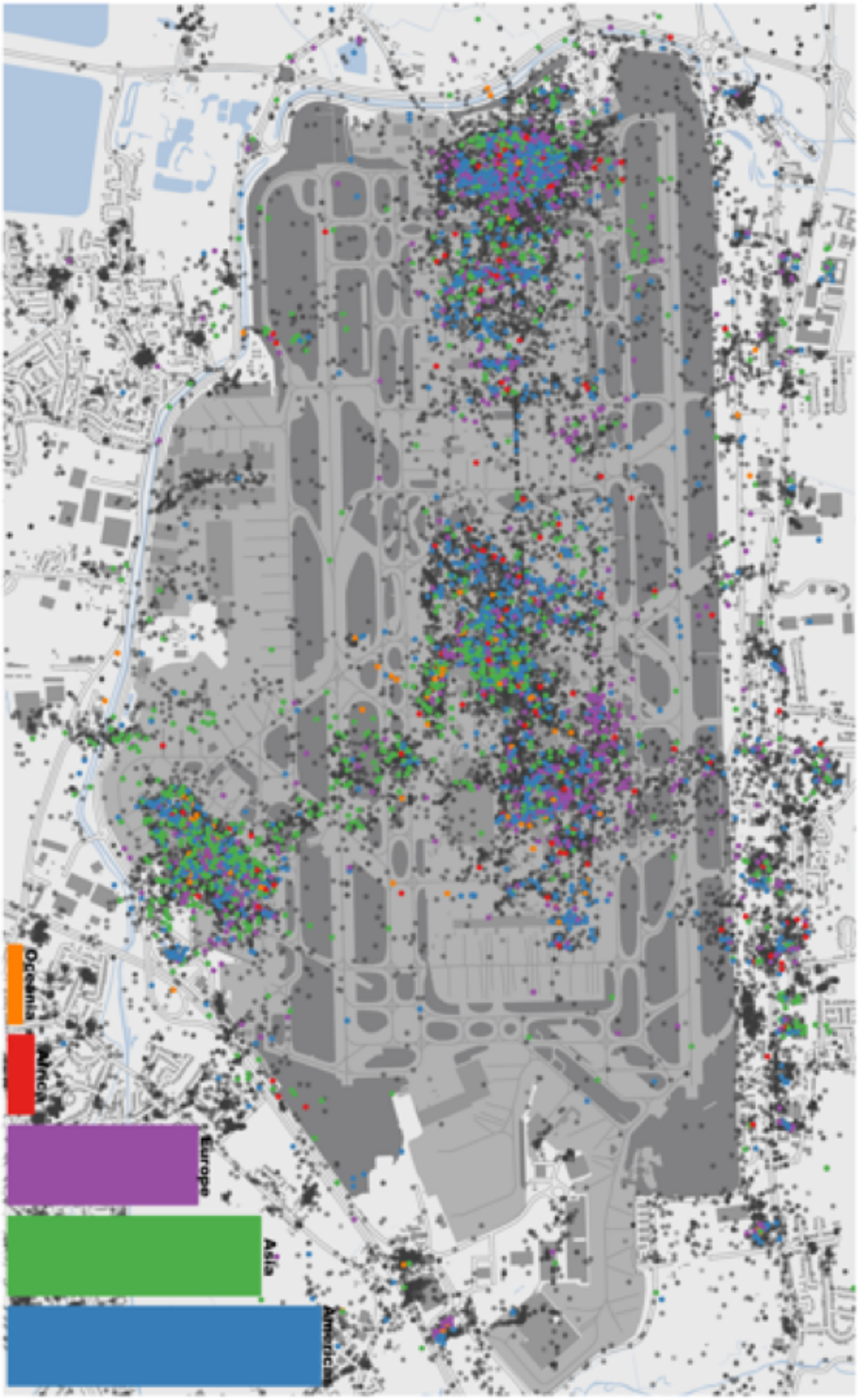


Figure 6.12: Map of Heathrow Airport in the UK. Individual Tweets, shown as points, are coloured by the global regions from which each user is believed resident.

Figure 6.12 provides a novel illustration of the additional insight which may be achieved given a fully inclusive dataset. The map shows the immediate area surrounding Heathrow airport superimposed with Tweets coloured based on the users' countries of residence aggregated to global regions. While this is just an illustration, such an approach could be feasibly employed in a range of applications to monitor international population stocks and flows and also to gain a better sense of place and space. For example, investigating the local catchment of retail centres. Such analysis highlights the importance of taking a holistic view when analysing social media data. As discussed previously, the nature of the Twitter API and associated rate-limits means that in the majority of cases, supplementing the data is not feasible. Consequently, when one is deciding to study a particular phenomenon, it is best that the largest collection window possible is employed.

The final consideration discussed is related to the unit of analysis to be employed. The two units available are either the Tweet or the user. Both approaches are valid. However, use of the wrong measure has the potential to introduce bias. Using the Tweet enables the analyst to track behaviour and may be useful in tracking the general behaviour of the population. Conversely, use of the user has the advantage that you may gain a more holistic view of the user, drawing on a potentially larger pool of data when making conclusions about the user and their behaviour. It could be argued that a hybrid approach is most valid in which the data on each user is employed in the assignment of general characteristics, and subsequently, that these characteristics can be used to add value to the individual Tweets. For example, Figure 6.12 relies on each user's Tweets to determine their probable countries of residence but then uses this information to augment the Tweet level analysis.

A key limitation in the analysis of Twitter data are the constraints regarding the sharing and publication of data. While the Twitter API does provide a means by which data may be collected, sharing of the resultant data are prohibited. Further, the sole means to collect large volumes of data is the streaming API which is limited to providing real-time data. Consequently, without significant financial outlay, legacy data are not easily accessible in bulk. This issue is a major concern for academics

who often constrain their data collection to a specific study window.

6.7 Conclusions

In this chapter the objective was two-fold. First, to assess the overall representative capability of the Twitter-derived population in the UK and second, to make recommendations as to how social media, specifically Twitter, should be employed in academia and more generally in industry. In regards to its representative ability, evidence was found in support of the anecdotal beliefs that Twitter is utilised by a younger generation with a bias towards male users. This behaviour was found to be consistent across the whole population. The analysis of ethnic representativeness was inconclusive with some concern raised regarding the approach used for classification. While the breakdown of groups was in the right order of magnitude, the validation of the method using the Consumer Register did not support a belief that the classification was sufficiently nuanced for the purpose of this study. In regards to geographic distribution, it is clear that Twitter it not equally representative across the UK. Rather, there are clear regions of over and under representation. Key factors affecting this include both regional differences in the demographic structure such as age, gender and ethnicity along with outside factors such as short-term education migration.

Given that this information is now available, should researchers choose to recognise it, some of the limitations which have held social media back as a demographic data source may now be addressed. The key point is that for the successful and effective analysis of social media a systematic approach must be taken. The first consideration of which should be, to what degree is Twitter representative of the population to which I wish to generalise my results.

Chapter 7

Social Media Demographics

7.1 Introduction

Until now, this thesis has sought to investigate the utility of social media derived population inventories in the depiction of stocks and flows of ill-defined and self-selecting segments of the population. Throughout this, a somewhat negative impression of social media has been depicted regarding its ability to capture the true breadth and diversity of the usual observable population. This view has resonated with much anecdotal evidence regarding the representativeness of social media data (Mislove et al., 2011; Longley et al., 2015). That said, the research has shown that in certain circumstances, the data may still provide the capacity to generate useful insight. In particular, there is potential to investigate general patterns of human mobility within the population. Given sufficient recognition of the limitations and biases, the data may provide a novel means by which human mobility patterns may be observed and investigated. Various literature exists regarding the application of social media data to the study of movement in space and time such as Hawelka et al. (2014) who explored the use of geotagged Tweets in modelling international travel patterns, Jurdak et al. (2015) who explore the use of geotagged Tweets in the modelling of travel behaviour within and between cities, and Lloyd and Cheshire (2017) who incorporated Tweets into the identification of retail centre locations and the subsequent derivation of their catchments. At the finest scale, Lansley and Longley (2016b) explored the potential for Twitter data for the detection and measurement of footfall. A key fea-

ture of the Lansley and Longley (2016b) analysis was the suggestion that textual data mining could be used as a means to better understand individuals' behaviour.

In this chapter, the objective is to demonstrate how Twitter and the methods developed during this thesis may be implemented in developing dynamic demographic insight in a novel context. This potential is demonstrated through a case study based on London's four largest airports. The study demonstrates the construction of demographic profiles for each airport, delivers an overview of the passengers' collective spatio-temporal mobility patterns, and finally, outlines the strengths, weaknesses and opportunities of such an approach. The chapter is arranged into three parts. First, the focus is on replicating the types of insight which may be obtained using traditional forms of data. In the second, the focus will shift to identifying insight not possible through the use of such data. In the final section, the discussion will determine the strength, weaknesses and opportunities for Twitter in the acquisition of actionable population insight. It should be noted that the methods implemented here are easily transferable and could potentially be incorporated into a data dashboard or analytics toolkit.

7.2 Case Study: London Airports

For the purpose of this study, the analysis is conducted on London's four largest airports: Heathrow, Gatwick, Stansted and Luton. In 2013, these four airports handled over 135 million passengers flying to destinations across the globe. Situated near London, UK, the four airports cater to a range of different transport requirements with Heathrow processing the bulk of extra-European travel while the three remaining airports are targeted more generally at destinations within Europe. The location of each airport and their relative passenger numbers is shown in Figure 7.1. Note, that due to the temporal coverage of the Twitter data employed in the study, the analysis, where possible draws on passenger data and other statistics from 2013.

Several motivations exist for the use of airports within the case study. First and foremost, airports are a dynamic yet relatively constrained environment for the observation of stocks and flows of populations. Given the nature of airports and the

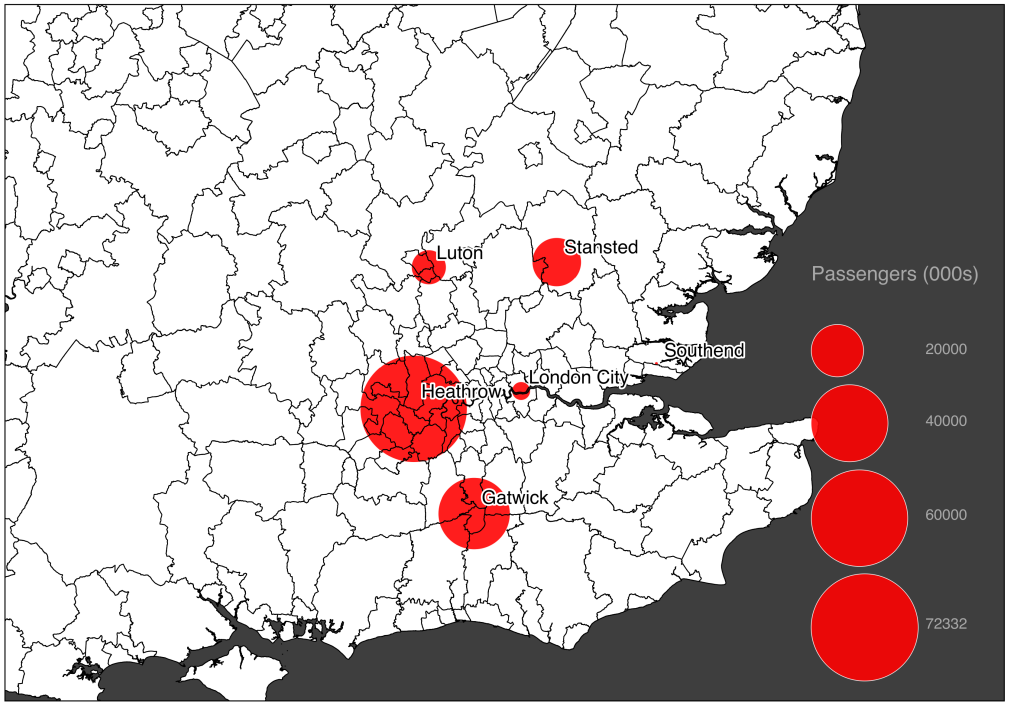


Figure 7.1: Map showing the locations of the six major London airports and the number of passengers in 2013.

requirement to effectively move individuals from arrival at the airport to their departure gates and the equivalent handling of arriving passengers their is significant interest in being able to observe and model individuals' behaviour. Within the airport ecosystem various stakeholders have a vested interest in this information. From a security perspective, it is important to understand general population dynamic, the ease at which individuals can move through the airport and where they may have originated from or be travelling to. From a retail perspective, there is an interest in understanding the demographic composition of airport travellers over time such that sales and income may be maximised. Finally, from an administration and reporting perspective it is important for each airport to understand the geography and demographic of its customers. Such information can be incorporated into service provision and infrastructure planning to optimise the airports function.

Second, from the perspective of future work, many parallels may be drawn between the airport environment and other locations in which there is a need to understand the stocks, flows and attitudes of the population. Proving useful in the airport con-

text, such methods may be readily applied to locations such as retail centres, sporting venues or alternative transport hubs. The key advantage of the airport over the alternative candidates is the abundance of data in the public domain concerning airport activities. The availability of such data provides a foundation for which the results of the subsequent analysis may be verified. Those data that are available include aggregate passenger statistics published by the UK Civil Aviation Authority, and a selection of marketing statistics published by the various airport groups such as JCDecaux[1]. These data are typically very aggregate in both space and time providing only a generalisation of airport activity. More precise and informative data no doubt exist internally within airports, however, are not made available within the public domain; a further justification for social media based insight generation.

7.2.1 Data

The data employed in this case study are drawn from the same global corpus of Tweets first introduced in Chapter 4. In the case of each of the four airports, the bounding box was calculated, and in turn, this was used to identify those users who have tweeted within the airport perimeter. Subsequently, all Tweets by those users identified were obtained from the database. As has been noted at various points throughout this thesis, use of a data-rich approach enables the identification of individuals' nationalities, areas of residence and to some extent, their general behaviours. In turn, heuristics may be applied to the data in the knowledge that they are being implemented in an appropriate manner. A key example being the Monica age and gender classification which has been constructed using UK data and may therefore only be applied to residents of the UK.

7.2.1.1 Tweet Maps

As a primary indicator of the extent of activity associated with each airport, a 10% sample of the geotagged Tweets for each airport is taken and plotted on a world map. These maps provide both an indication of density and coverage.

[1] JCDecaux is an international marketing organisation heavily involved in transport marketing. An example of the passenger profiles may be viewed at http://passengerprofiles.jcdecauxairport.co.uk/project-view/heathrow/

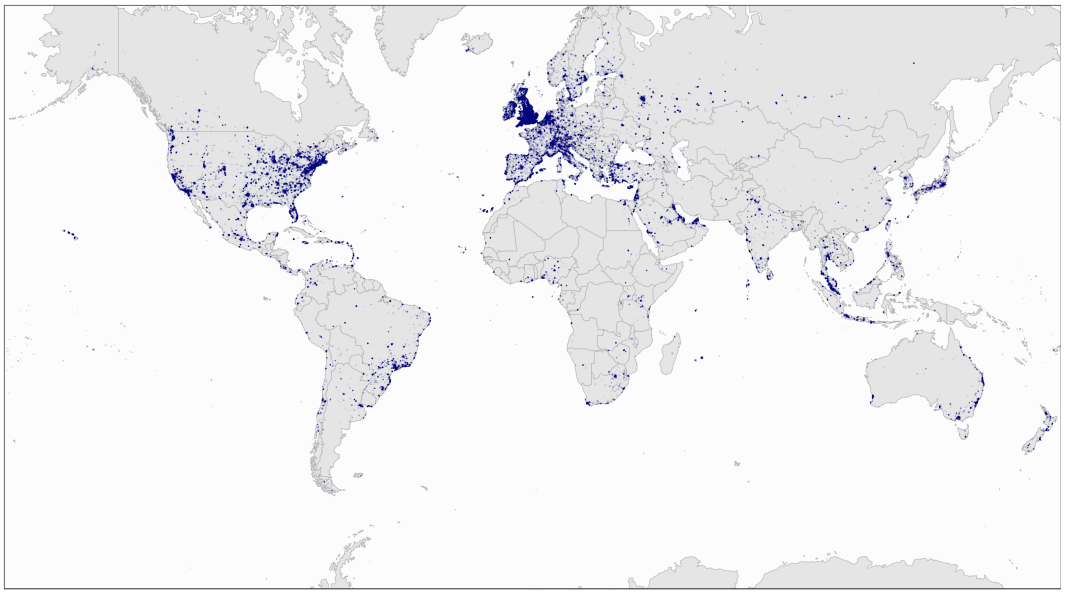


Figure 7.2: Map showing 10% sample of Tweets submitted by those Twitter users identified within the Heathrow Airport extent.

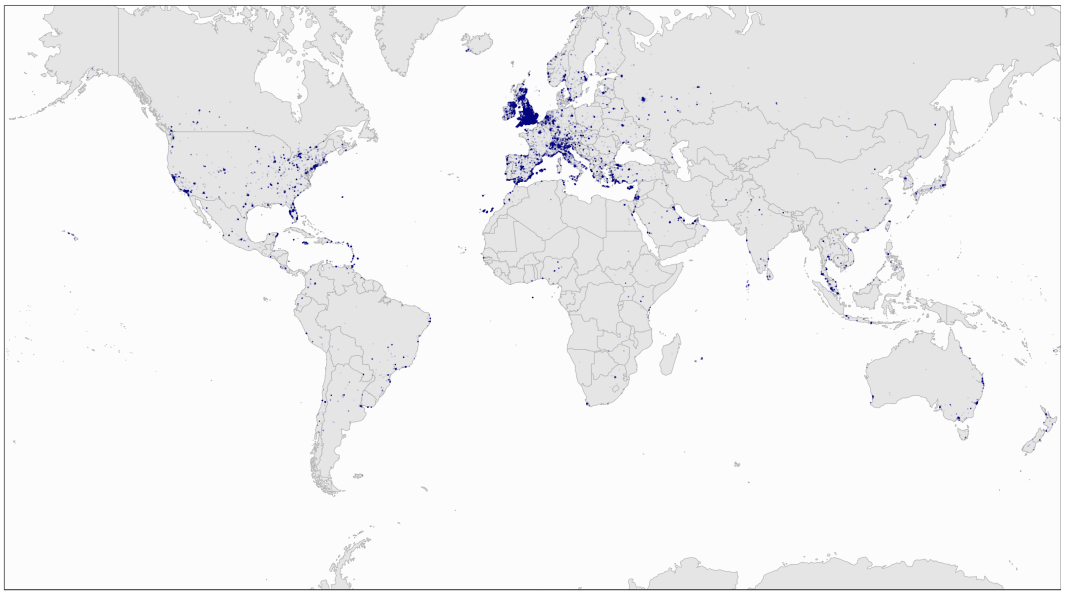


Figure 7.3: Map showing 10% sample of Tweets submitted by those Twitter users identified within the Gatwick Airport extent.

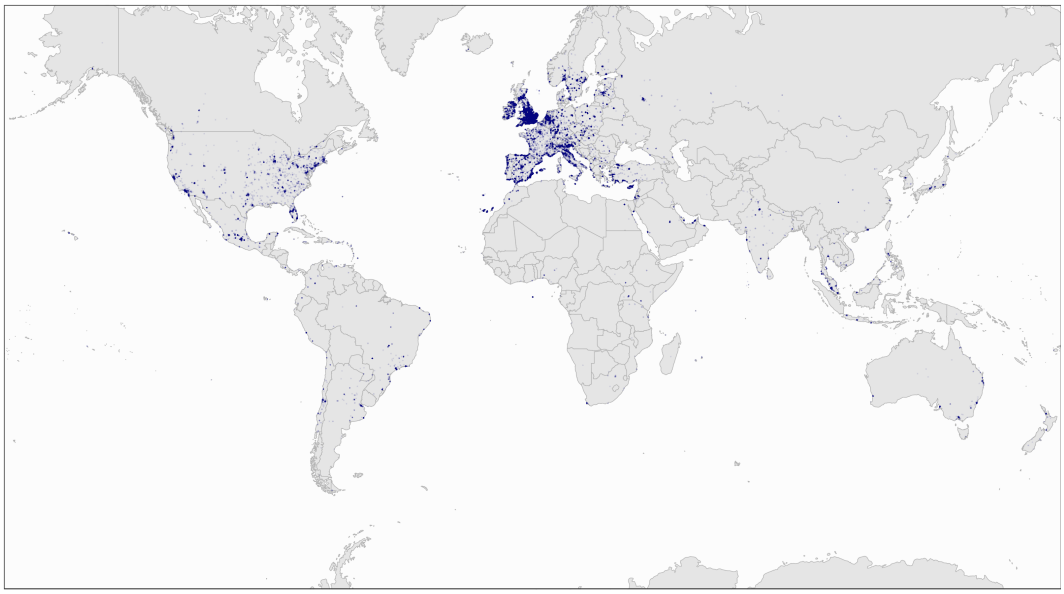


Figure 7.4: Map showing 10% sample of Tweets submitted by those Twitter users identified within the Stansted Airport extent.

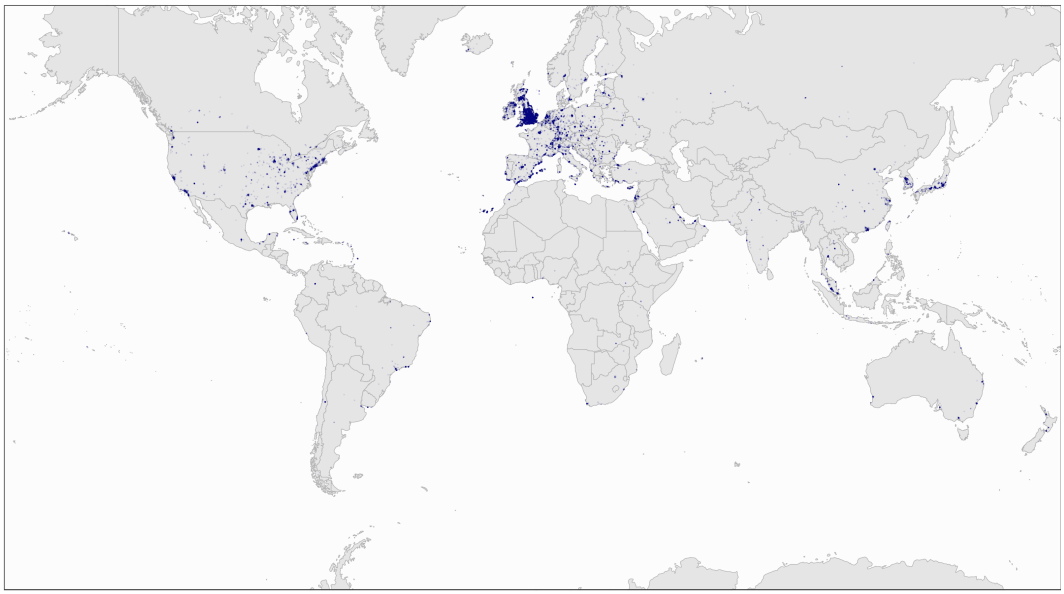


Figure 7.5: Map showing 10% sample of Tweets submitted by those Twitter users identified within the Luton Airport extent.

The four maps, depicted in Figures 7.2 through 7.5, provide an initial indication as to the geography of the tweeting activity associated with each of the four airports. It should be recognised that this viewpoint does not explicitly depict the origins and destinations of passengers, rather, it represents locations visited by those individuals within each airport perimeter. These locations may be indicative of individuals' usual places of residence, the areas which said individuals inhabit during their routine activities or random locations which they have visited. Further, analysis of the data may provide insight into users' other international travel behaviour providing a means by which different airports/travel methods may be connected. Furthermore, it should be recognised that the data may include false or misleading data where Tweets have been submitted with fabricated or adjusted locations.

Consistent across the four datasets is a large number of Tweets within the bounds of the UK, a feature which is unsurprising but a useful reality check. Considering first Figure 7.2, Heathrow, it is clearly evident that this is the dominant airport with dense coverage across the populated regions of the Americas, Europe, Asia and Australasia. Figure 7.3 and 7.5, Gatwick and Stansted, appear to be relatively global with Stansted having the greater European focus of the two. Stansted is the home of the largest low-cost carrier in Europe, Ryanair. Luton, depicted in Figure 7.5, the smallest of the four airports, has some global coverage, however, appears most strongly concentrated in the central European band of countries. Given that the Twitter data span a whole year, it is unsurprising that Tweets are observed beyond each airport's typical regions of operation.

While it is possible to make some inference regarding individuals' collective behaviours based on the raw data, they are of limited value. However, their value may be realised through the application of heuristic techniques able to infer individuals' geographic attributes and personal identities. In this case study, the factors considered are nationality, age, gender, ethnicity and place of residence.

7.2.2 Nationality

Inferring the nationality of social media users is not in itself novel. However, it does play an important role in regards to the subsequent inference of individuals' personal

identities. Lacking this information, it is not appropriate to infer individuals' ages or the purpose of their travel. Further to improving how the various heuristics are applied, knowledge of nationality facilitates differentiation between those who are residents and those who are not. This is particularly relevant given the high proportion of people travelling through the four airports who are not residents of the UK. Heathrow passengers, for example, are composed of approximately 60% non-UK residents. As in previous Chapters, individuals' nationalities are inferred based on their historic tweeting activity. The condition that five or more and greater than 50% of an individual's total Tweets is applied. The nationality analysis was performed for each of the four airports and is presented below.

Figures 7.6 and 7.7 illustrates the top 20 nationalities observed within each airport excluding those users believed resident within the UK. The four graphs provide further insight as to the breakdown of passengers across the four airports. Notably, based on the Twitter data, Heathrow possesses the both the greatest frequency and proportion of travellers of non-EU nationality. This, given the airport's previous recognition as an international hub, is unsurprising. In the case of the three other airports, it is evident the top nationalities observed are more strongly concentrated in Europe. In the interpretation of the above, it must be remembered that some factors will influence the relative position and magnitude of each nationality. The key factor in this being the relative popularity of the Twitter social network. Where Twitter lacks popularity within a given country, for example, Germany, the number of passengers is likely to be less well represented whereas in Kuwait, a country in which Twitter is very popular, passengers are likely to be better represented. Theoretically, given that we have some information regarding the popularity of Twitter, the counts could be standardised to represent the exact composition of the population better. Such an approach, however, is liable to introduce a significant degree of error due to the occurrence of small numbers in the case of both the global assessment of Twitter popularity and also, in the counts of nationalities observed at each airport.

Table 7.1 shows the breakdown of travellers at each of the four airports by UK residency as determined from Twitter versus the same metrics recorded by the UK

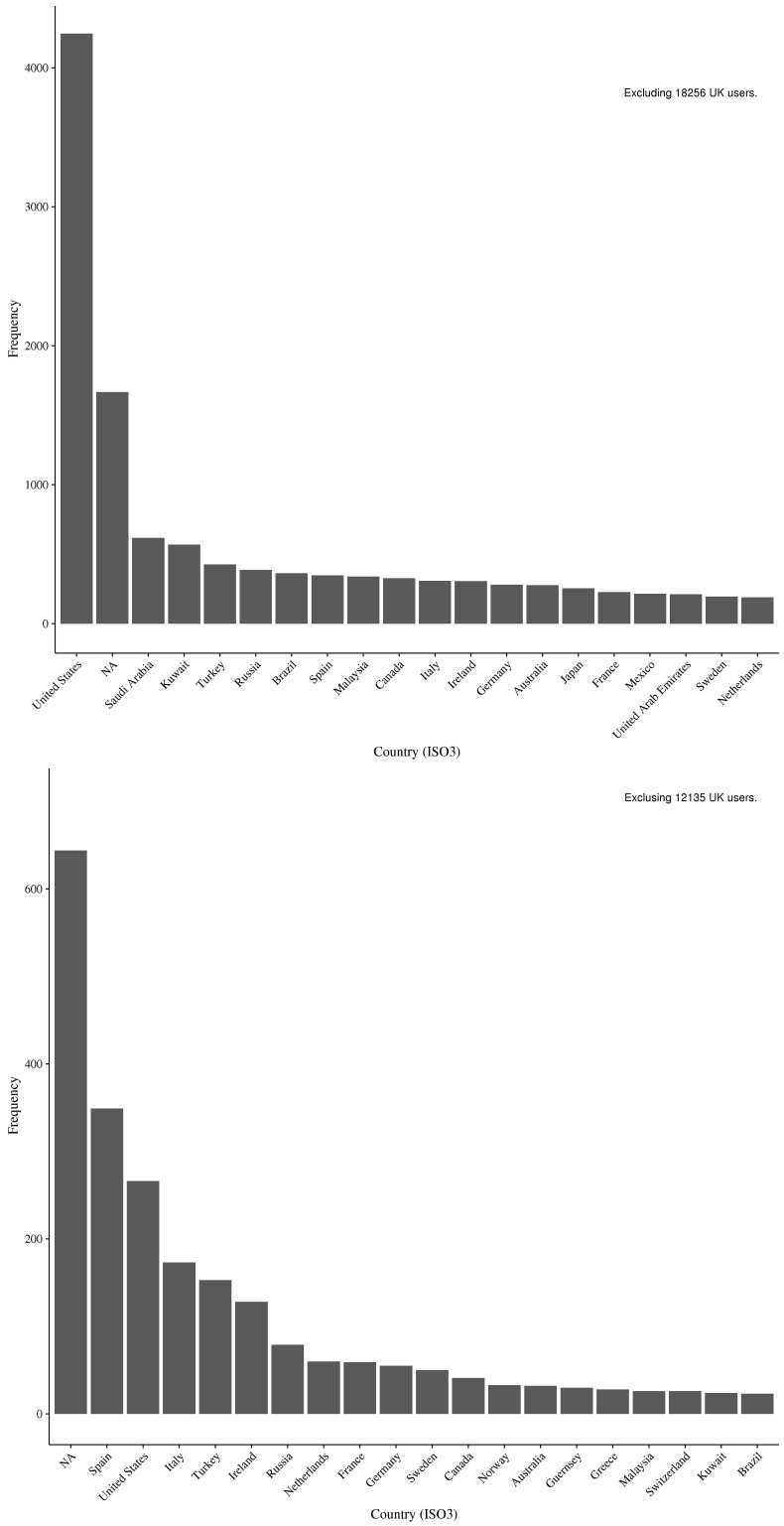


Figure 7.6: Bar plot showing the inferred nationality of individuals identified at Heathrow (top) and Gatwick (bottom).

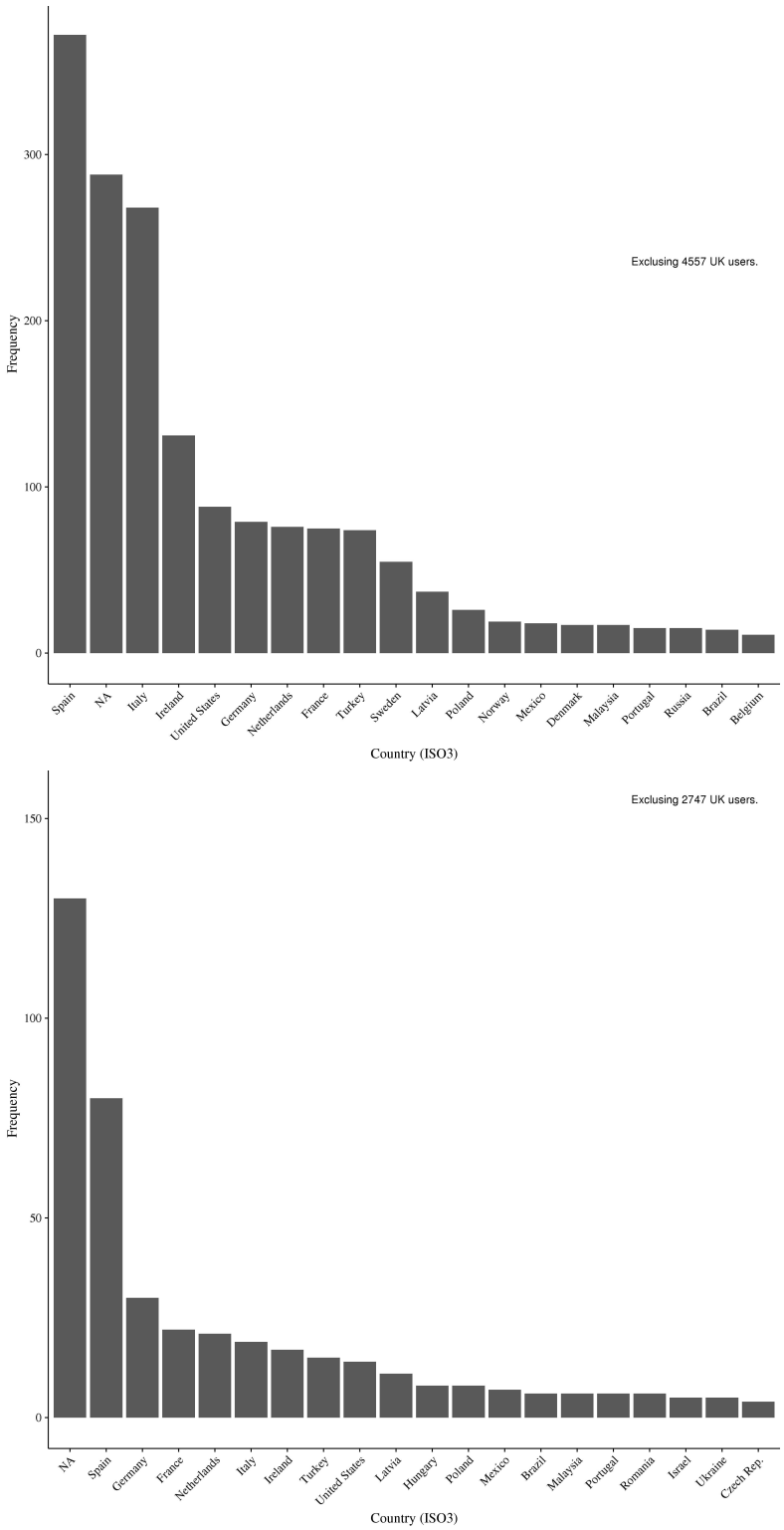


Figure 7.7: Bar plot showing the inferred nationality of individuals identified at Stansted (top) and Luton (bottom).

Table 7.1: Comparison of UK-resident vs. non-UK-residents at each airport based on CAA and Twitter Data.

	Twitter				CAA					
	UK		Foreign		UK		Foreign			
	n	%	n	%	n (000)	%	n (000)	%	UK Qu.	Foreign Qu.
Heathrow	12,261	59.75	8,258	40.25	29,523	40.35	43,642	59.65	1.48	0.67
Gatwick	11,755	84.92	2,088	15.08	27,342	72.17	10,544	27.83	1.18	0.54
Stansted	4,466	73.14	1,640	26.86	11,663	58.61	8,236	41.39	1.25	0.65
Luton	2,804	89.36	334	10.64	7,931	76.26	2,469	23.74	1.17	0.45

Civil Aviation Authority. Based on the results shown, it is evident that Twitter systematically under-represents foreign travellers. The UK Quotient indicate by what factor Twitter over-represents UK residents, and the Foreign Quotient indicates by what factor Twitter under-represents non-UK residents. There are several possibilities as to why this behaviour may occur. First, the Twitter data are inclusive of those individuals who are employed at each airport. Second, non-UK residents are less likely to have access to mobile data services. A further factor influencing the above may be individuals' countries of origins. Given that the popularity of Twitter varies between countries, and that the relative proportion nationalities within each airport differ, the composition of travellers is likely to have some impact on how well represented foreign tourists are. It is evident from the above that there is a striking over-representation of UK residents and under-representation of non-UK residents. The difference between these two groups appears to be systematic.

The reason both Heathrow and Stansted may better represent tourists is likely due to the popularity of Twitter within the main origin/destination countries: Spain, Italy, Ireland and, in the case of Heathrow, the United States.

7.2.2.1 Age and Gender

Having established those who are UK residents and those who are not, it is possible to apply the age and gender heuristics. As noted previously, such analysis is limited based on the nationality of the individuals being studied. While name genders remain relatively consistent between countries, age profiles are distinct. Consequently, using the Monica classification, it is possible to ascertain gender for all users but age just for those individuals believed resident within the UK (Lansley and Longley, 2016a).

Table 7.2: Gender balance at London airports based on Twitter users genders.

	All users				UK Residents			
	Male		Female		Male		Female	
	N	%	N	%	N	%	N	%
Heathrow	8,786	63.61	5,026	36.39	5,224	62.71	3,106	37.29
Gatwick	6,564	57.1	4,931	42.9	5,397	56.32	4,186	43.68
Stansted	2,746	56.64	2,102	43.36	2,018	56.91	1,528	43.01
Luton	1,484	59.81	997	41.29	1,271	58.92	886	41.08

Table 7.2 provides an overview of the gender divide observed in each of the four airports. From the data, it is evident that a significant gender skew exists with between 55% and 60% of passengers being identified as male. Limited public data exist to verify the above.

Table 7.3: Table showing male and female divide at each of the four London airports as recorded by the UK CAA as part of their annual passenger survey.

	All Usual Passengers			
	Male		Female	
	N	%	N	%
Heathrow	26,382	55	21,578	45
Gatwick	19,071	51.43	18,014	48.57
Stansted	11,043	51.53	10,387	48.47
Luton	6,138	51.08	5,878	48.92

The most reliable data regarding passenger data sourced from the UK Civil Aviation Authority 2015 Passenger Survey Report (CAA, 2015). The data, based on a survey of 118,491 individuals was collected using a stratified sample designed to account for carrier, route and quarter such that it accurately captured the full passenger demographic inclusive of seasonal factors. Comparison between Tables 7.2 and 7.3 highlights a notable male bias across each of the four airports. This bias is likely due to the existing male bias associated with the use of Twitter. The amount of bias present is reported in Table 7.4. A value of 1.15 indicated 15% greater than expected Males.

During the UK Wide benchmarking exercise, it was found that a gender bias existed within the Twitter data. The male bias at the UK scale being 13%. Notably,

Table 7.4: Gender bias observed in each of the four London airports.

Airport	Male Bias
Heathrow	1.15
Gatwick	1.11
Stansted	1.09
Luton	1.17

the degree of bias reported in Table 7.4 for each airport does not differ significantly from this. To fully understand the gender bias present within each airport, it would be necessary to assess whether or not a bias existed for each of the nationalities represented within the Twitter dataset.

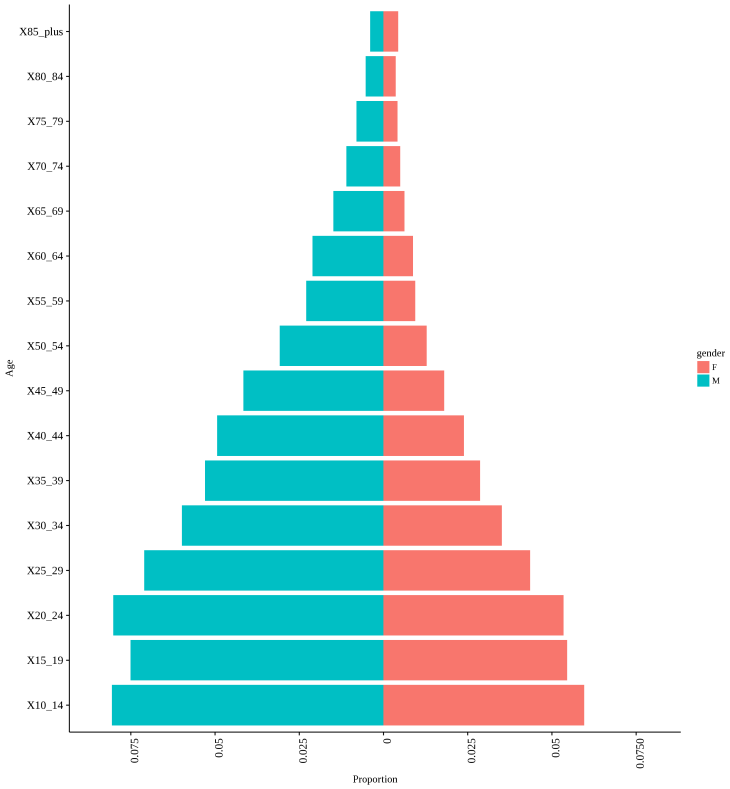


Figure 7.8: Population pyramid for UK-based Twitter users identified at Heathrow.

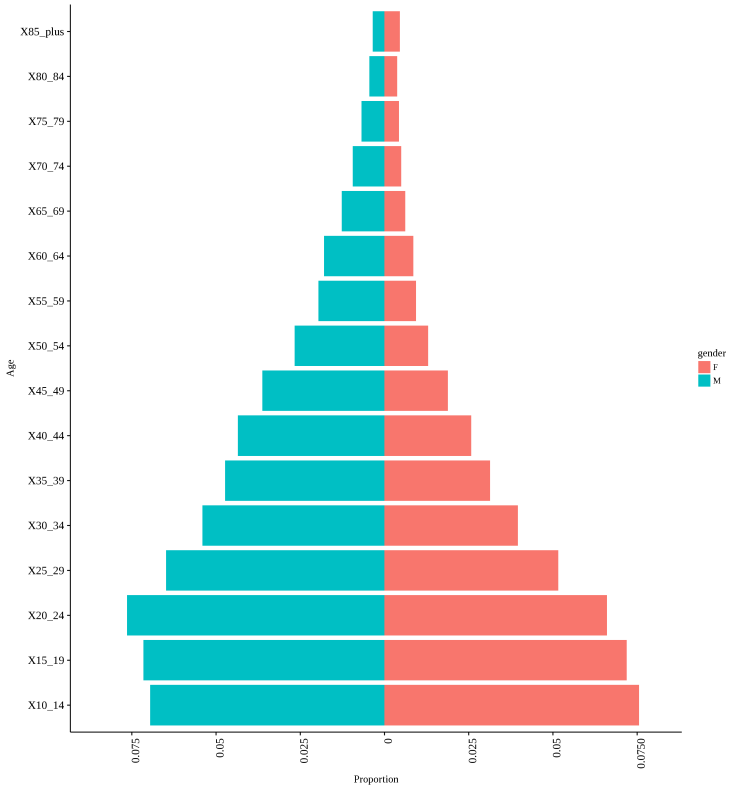


Figure 7.9: Population pyramid for UK-based Twitter users identified at Gatwick.

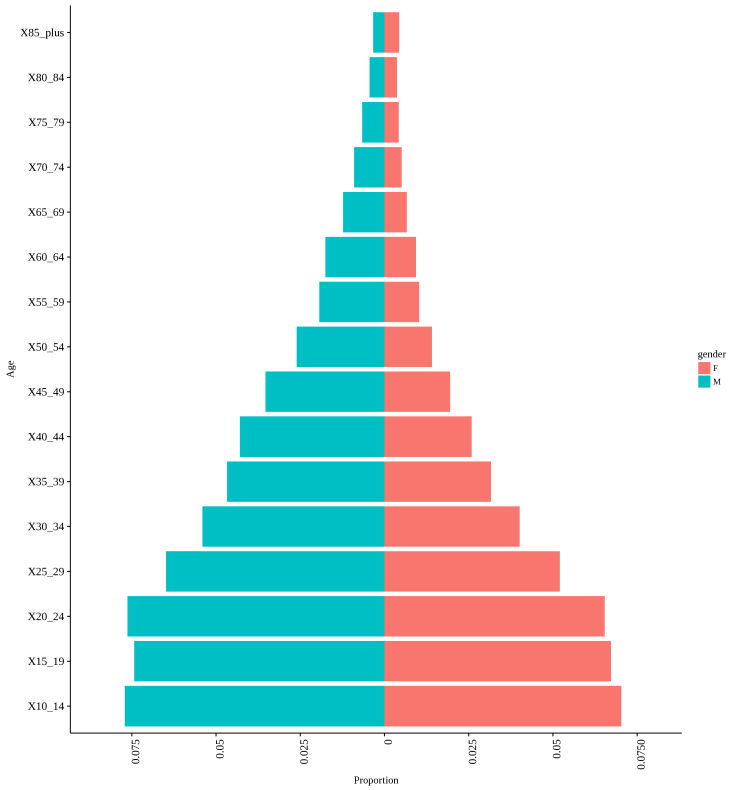


Figure 7.10: Population pyramid for UK-based Twitter users identified at Stansted.

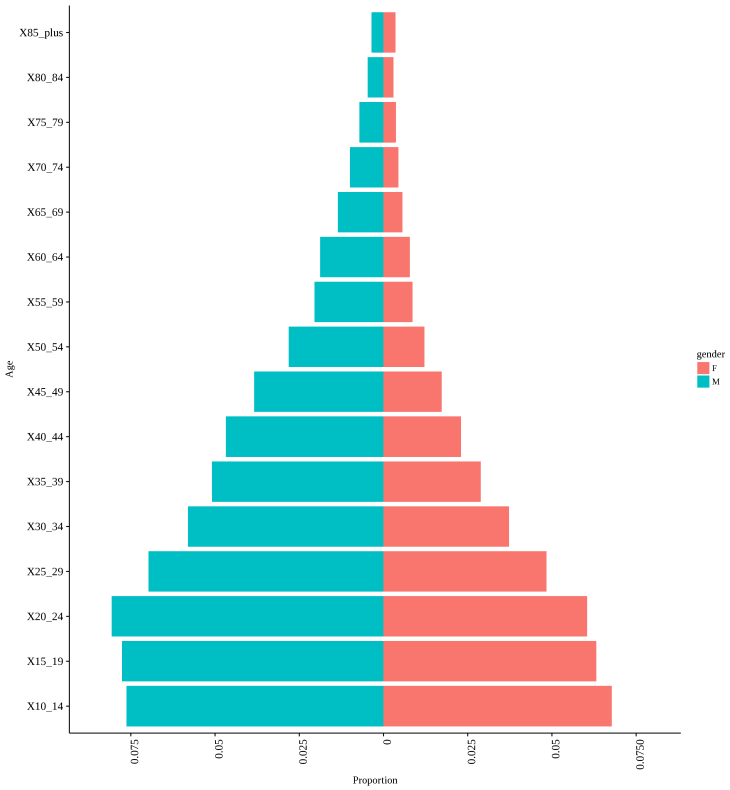


Figure 7.11: Population pyramid for UK-based Twitter users identified at Luton.

Having established gender composition of those individuals within the airport, Figures 7.8 through 7.11 depict the age distribution of Twitter users at each of the four London airports. The population pyramids observed are fairly consistent between airports with just a slight variation. Considering the previous benchmark of Twitter in the UK, it would appear that the profiles are fairly consistent with what was observed for London. However, as was discussed in the previous chapter, the Twitter-based population pyramid for London was only marginally different to that for the UK as a whole. Considering that the age/gender structure of London is quite distinct from the UK as a whole, it would suggest that age standardisation may not be possible.

A key limitation raised in this section is the constraint imposed through the use of UK-centric identity classification tools. This is particularly pertinent in terms of age and gender. As discussed previously, the Monica Classification is built based on data collected within the UK and is thus limited in applicability to this region. Within the literature various alternative approaches to age and gender inference have been

demonstrated based on the analysis of individuals' language. This is a potentially valuable extension to the analysis would may enable greater insight to be generated in terms of those individuals passing through each of the four airports.

7.2.2.2 Ethnicity

Having established the age and gender of Twitter users recorded in each of the four airports, the next consideration is ethnicity. Understanding the ethnic breakdown of individuals' who occupy each airport offers several unique opportunities. Knowledge of individuals' ethnic groups may be used to improve the provision of key services, or as part of general activity profiling. The use of ethnicity/nationality-based profiling is not uncommon as part of security. Frederickson and LaPorte (2002) suggests that knowledge of an individual's personal identities is an important aspect of airport security and a valuable complement to existing security processes which are increasingly automated. It should be noted that the process by which ethnicity is inferred, based on names, means that that the method is not suitable for individual level profiling and rather may only be used to form a general impression of a group of individuals. Seeking to understand ethnic composition in each of the four airports, the users identified within each airport were processed using the Onomap CEL classifier. The identified users split based on being the UK and non-UK residents. The quotient is calculated in the case of each group such that distinction between the two passenger groups may be made.

Table 7.5 provides a breakdown of travellers at each UK airport by Onomap group. The data are split by those who are believed to be residents of the UK and those that are not. In the case of each airport, a quotient is calculated between the UK and non-UK users. Where a quotient value of 1 is observed, the proportion of users of the Onomap group is equal. A quotient of 0.5 indicates approximately half while 2.0 indicates double.

The first observation is that across each of the four airports, the proportion of those individual's codes as either Celtic or English, the two main UK groups are consistently lower in the non-UK users. This is a valuable reality check confirming a difference exists between the two groups. Second, across each of the four airports,

Table 7.5: Breakdown of airport passengers by Onomap CEL group differentiating by those who are believed to be UK residents and those that are not.

	Heathrow					Gatwick					Stansted					Luton				
	UK		Non. UK			UK		Non. UK			UK		Non. UK			UK		Non. UK		
Onomap Group	n	%	n	%	∅	n	%	n	%	∅	n	%	n	%	∅	n	%	n	%	∅
African	156	1.25	111	1.32	1.06	96	0.8	20	0.94	1.18	34	0.75	16	0.96	1.28	31	1.11	4	1.2	1.08
Celtic	2147	17.24	883	10.48	0.61	2195	18.35	244	11.41	0.62	767	16.89	132	7.89	0.47	484	17.26	29	8.68	0.5
E. Asian & Pacific	411	3.3	382	4.54	1.38	348	2.91	79	3.69	1.27	135	2.97	53	3.17	1.07	99	3.53	9	2.69	0.76
English	6138	49.28	2493	29.6	0.6	6650	55.59	563	26.32	0.47	2315	50.97	423	25.27	0.5	1440	51.36	83	24.85	0.48
European	676	5.43	899	10.67	1.97	565	4.72	296	13.84	2.93	333	7.33	344	20.55	2.8	174	6.21	62	18.56	2.99
Greek	82	0.66	65	0.77	1.17	51	0.43	27	1.26	2.93	26	0.57	6	0.36	0.63	16	0.57	2	0.6	1.05
Hispanic	315	2.53	646	7.67	3.03	265	2.22	221	10.33	4.65	170	3.74	218	13.02	3.48	68	2.43	37	11.08	4.56
International	84	0.67	75	0.89	1.33	74	0.62	13	0.61	0.98	34	0.75	19	1.14	1.52	19	0.68	7	2.1	3.09
Jewish & Armenian	18	0.14	42	0.5	3.57	7	0.06	4	0.19	3.17	3	0.07	2	0.12	1.71	18	0.64	1	0.3	0.47
Muslim	86	0.69	114	1.35	1.96	63	0.53	19	0.89	1.68	23	0.51	12	0.72	1.41	95	3.39	43	12.87	3.8
Nordic	766	6.15	1130	13.42	2.18	346	2.89	193	9.02	3.12	174	3.83	110	6.57	1.72	5	0.18	4	1.2	6.67
Sikh	58	0.47	154	1.83	3.89	46	0.38	55	2.57	6.76	28	0.62	34	2.03	3.27	17	0.61	1	0.3	0.49
South Asian	108	0.87	32	0.38	0.44	46	0.38	18	0.84	2.21	20	0.44	11	0.66	1.5	43	1.53	2	0.6	0.39
Unclassified	266	2.14	188	2.23	1.04	143	1.2	25	1.17	0.98	41	0.9	18	1.08	1.2	252	8.99	40	11.98	1.33
Void	950	7.63	1044	12.4	1.63	860	7.19	311	14.54	2.02	363	7.99	242	14.46	1.81	43	1.53	10	2.99	1.95

there are consistently more individuals of the European and Hispanic Onomap CEL groups. Third, in several circumstances, there is an apparent significant high proportion of particular Onomap CEL groups. These quotients are typical of the smallest groups where a small number of users are recorded. Lastly, in the case of several Onomap CEL groups, Luton bucks the general trend. In the case of E. Asian and Pacific, Jewish and Armenian and Sikh. Important to remember that of the four airports, Luton has the smallest proportion of foreign travellers. CAA data records that just 23.74% of passengers are not resident in the UK. This small proportion of passengers leads to a notable issue of small numbers.

In the interpretation of the above, it should be remembered that various biases are manifest within the Onomap classification. Such biases, however, are only realised where the comparison is made against administrative data such as the UK Census of Population. Thus, while the relative magnitude of comparison between the groups may vary, where both groups are classified using the Onomap classification tool, the comparison is possible.

7.2.3 Mobility

Having established the demographic profile of those individuals identified within the four airports, the next focus is on establishing their general mobility behaviours. As an initial step in the investigation of mobility patterns, the processed Twitter data are employed in the identification of airport catchment areas.

7.2.3.1 Airport Catchments

Beyond simply understanding the demographic of those individuals identified within each airport, the processed Twitter data provides various means by which broader behavioural trends may be observed. Such analysis is only possible because individual users have multiple locations associated with their identities over time. Most relevant to determining catchments, individuals' probable places of residences may be inferred. Given this information, it is possible to make a prediction as to the catchment of each airport. An airport's catchment area is the area in which the majority of passengers are resident and is often directly related to an airport's market

share (Lieshout, 2012). Typically, catchment area analysis is performed using origin-destination data and relies on the availability of passenger origin data. However, as Lieshout (2012) notes, such data is largely unavailable in the public domain, limiting the potential for such analysis to be performed. While Lieshout (2012) suggests a model based alternative, here, we propose that the processed Twitter data may be a suitable proxy for regular passenger travel data. Lloyd and Cheshire (2017) investigated the use of geotagged Tweets in the identification of retail centres and their respective catchments employing a home-range estimation technique to differentiate between primary, secondary and tertiary catchment areas.

For the purpose of this demonstration, the LQ is employed as an efficient way of identifying whether or not there exists a geography to UK-based passengers observed within each of the four airports. The Location Quotient is calculated as:

$$LQ_i = \frac{p_i/p}{P_i/P} \tag{7.1}$$

Where:

LQ_i = Location quotient for region i

p_i = Twitter population (n users) in region i

p = Total Twitter population (n users)

P_i = Census population in region (n) i

P = Total Census population (n)

The analysis is performed at Local Authority level which divides the UK into 404 distinct regions. Data on all usual residents was sourced from inFuse, a data portal published by the UK Data Service (see:). An LQ value of 1 indicates the expected proportion of passengers assuming the number of passengers is homogeneous. LQ <1 indicates fewer passengers that would be expected and LQ > 1 indicates a greater number of travellers than might be expected.

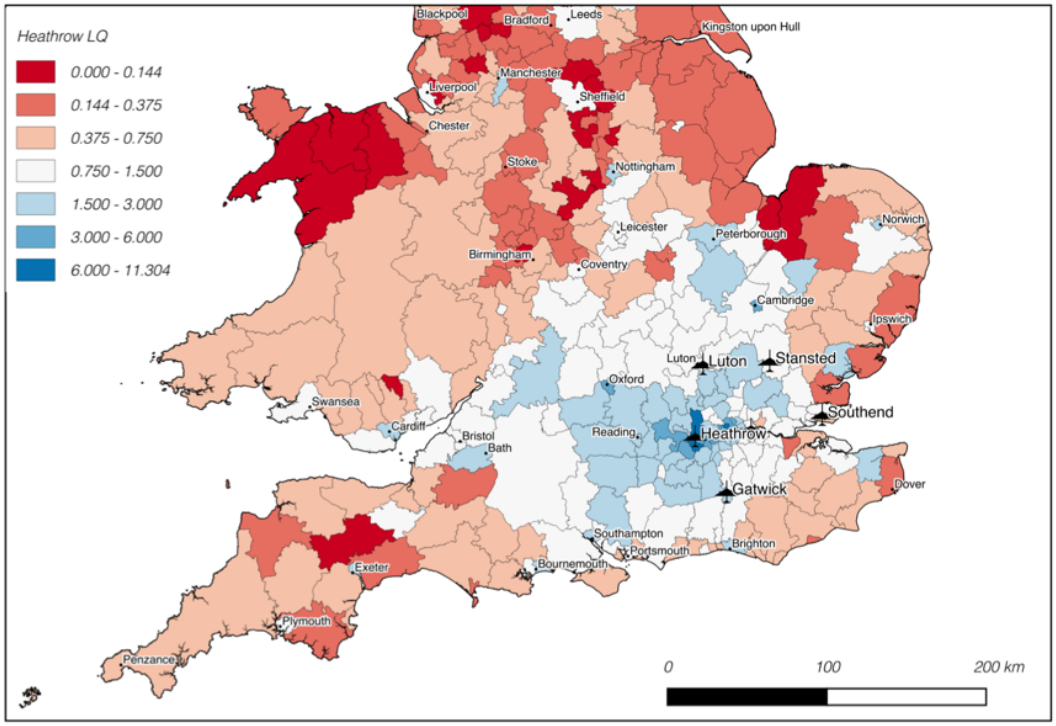


Figure 7.12: LQ map of the areas of residence for those UK-based individuals identified within Heathrow.

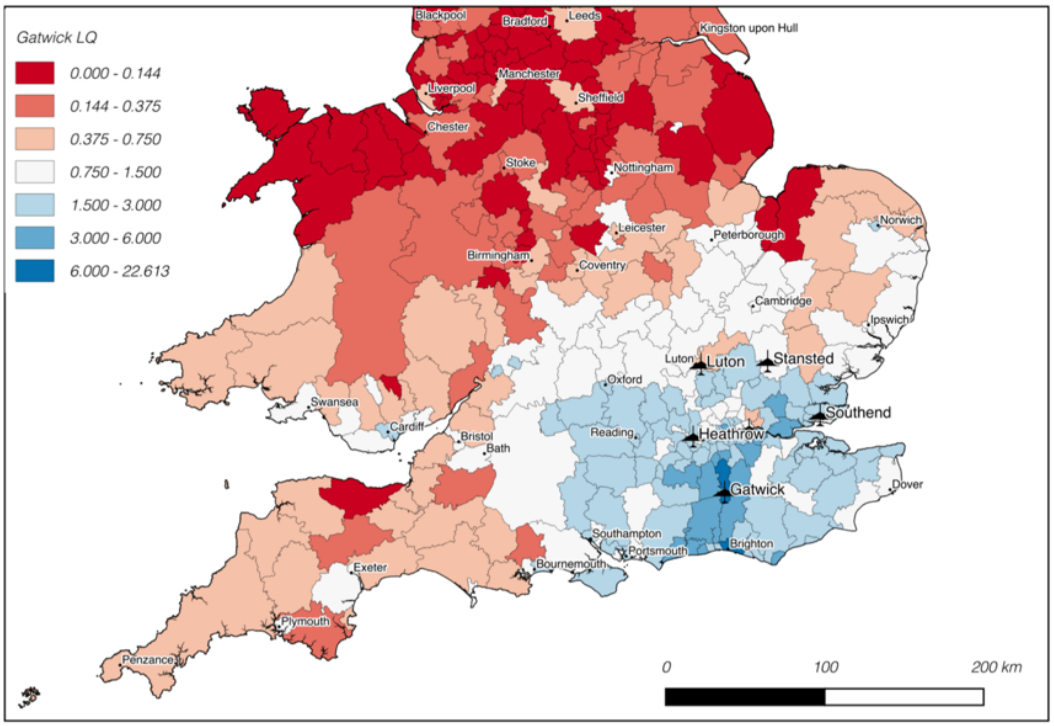


Figure 7.13: LQ map of the areas of residence for those UK-based individuals identified within Gatwick.

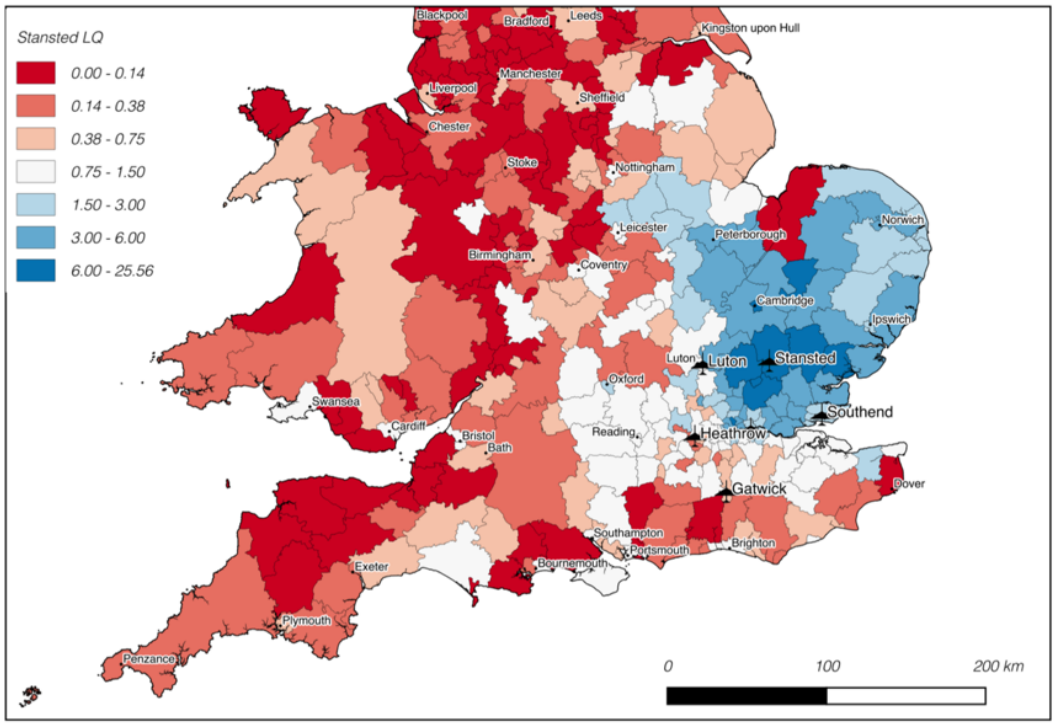


Figure 7.14: LQ map of the areas of residence for those UK-based individuals identified within Stansted.

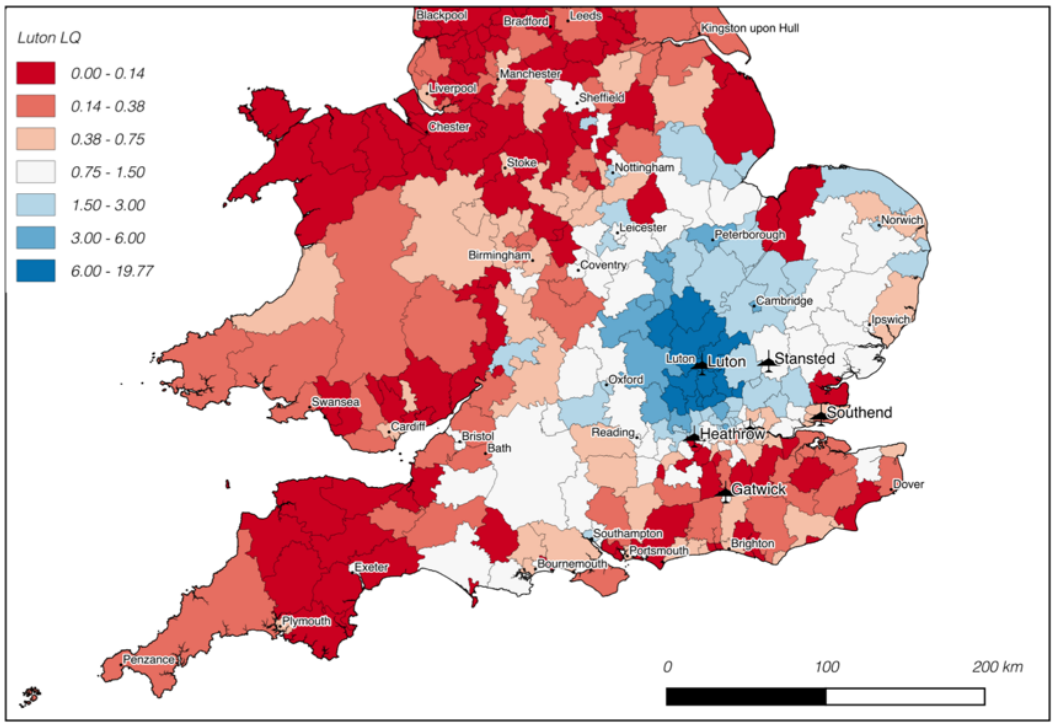


Figure 7.15: LQ map of the areas of residence for those UK-based individuals identified within Luton.

For the purpose of comparison, airport catchments produced by the UK Civil Aviation Authority are shown in Figure 7.16 (CAA, 2011). Based on data from the 2010 CAA Passenger Survey, the catchment maps are indicative of the areas in which 80% of UK-based passengers originate for each airport. The analysis identifies distinct clustering around the four airports emanating out from London in a direction relative to airports position from the city centre. Also evident is the interplay between the four airports in which the various catchments overlap. Such behaviour is typical in multi-airport regions.

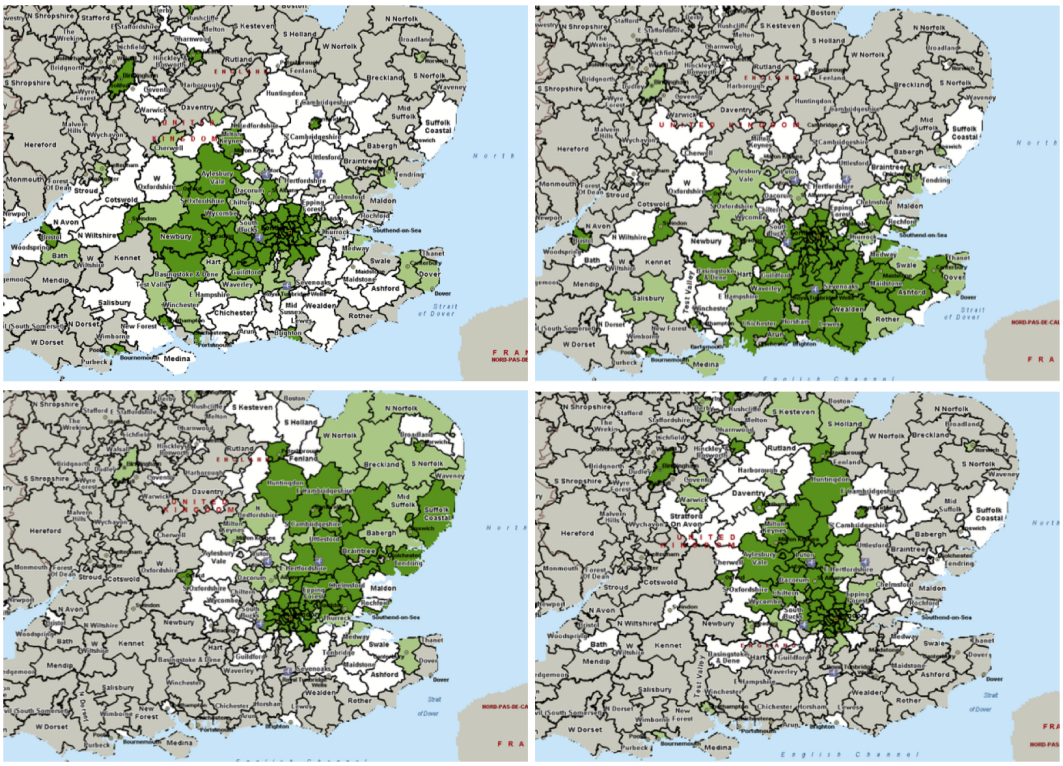


Figure 7.16: CAA maps of overall historical catchment areas for Heathrow (top left), Gatwick (top right), Stansted (bottom left) and Luton (bottom right) (CAA, 2011). For each airport, 70% of passengers are indicated by the dark green areas, 80% in light green and 90% in white.

Comparison between the four LQ maps and the catchment maps produced by the UK CAA indicate a high level of agreement suggesting that such an approach may be valid. However, lacking the original data, an empirical comparison is not feasible. That said, the significance of this analysis is that a similar outcome has been generated without access to proprietary or commercial data typically associated with

such an enquiry.

While the analysis does show promise, a key limitation is an inability to differentiate between those individuals whose employment is associated with each airport and those who are transient. Using the example of Gatwick, approximately 24,000 individuals are employed within the airport campus while an average of 97,000 passengers pass through the airport each day[2]. These individuals, who are likely to be resident within the immediate vicinity of each airport, who are also likely to use Twitter, are likely to inflate how many passengers are originating within the immediate vicinity of each airport.

Theoretically, one may extend the above analysis by instead investigating the spatial diffusion of those individuals not resident within the UK. Such analysis could be disaggregated further such that nationality is considered. For example, individuals travelling into the UK via Stansted or Luton are likely to be predominantly European. Do these travellers exhibit different behaviour to those who fly into Heathrow who is most likely to have originated within the Americas?

7.2.4 Summary

In the preceding analysis, we have explored how Twitter data may be employed in place of conventional passenger data as a means to understand the demographic and mobility of individuals observed within four London airports. The analysis has considered the key demographic attributes of age, gender and ethnicity, and also identified airport catchments based on individuals believed home locations. The analyses have so far sought to replicate the types of outputs which are currently available in the public domain. This has not, however, exploited the temporal or textual information associated with the original data. Analysis of such data takes the potential of social media data beyond what can be achieved by conventional passenger survey data. This will be the focus of the subsequent analysis.

[2]Gatwick Airport publishes a range of general statistics which may be accessed at the following URL. http://www.gatwickairport.com/business-community/about-gatwick/company-information/gatwick-by-numbers/

7.3 Opportunities: New Forms of Data

As has been discussed previously, new forms of data offer a range of new capabilities previously not feasible with conventional datasets. Many these data are generated as a by-product of other processes such as energy monitoring devices and public WiFi provision. However, of the new forms of data, social media are easily the most accessible and ubiquitous. In this section, a range of potential insight is generated which would not easily be feasible with a conventional data source and more importantly, without access to internal or proprietary data.

7.3.1 Footfall and Activity Patterns

In the first instance, the objective is to analyse activity patterns across each airport. First, looking at the general trend of activity and subsequently across space. To begin, the raw Tweets associated with each airport were aggregated by the day of week and hour. Performed for each of the four airports, the results of the analysis are visualised in the following graph.

Figure 7.17, the temporal activity plots, provide an illustration of the daily rhythm of tweeting activity observed within each of the four UK airports. Inspection of the data suggests a typically bimodal distribution with a morning peak around 9 am and an afternoon peak around 7 pm. This pattern is somewhat distinct in the case of Heathrow, Gatwick and Stansted. However, is less evident in the case of Luton. For the purpose of reference, the limited operating hours at each airport are highlighted in red. During those times, strict quotas exist to limit noise disturbances to local residents. The quotas are most stringent in the case of Heathrow and least in the case of Luton. Details of the quotas and operating restrictions, published by the Department of Transport, can be accessed from the following URL (https://www.gov.uk/government/publications/night-flying-restrictions-at-heathrow-gatwick-and-stansted-airports).

In interpreting the above, the assumption is made that total Tweeting activity will correlate with the total number of individuals who are occupying the space at any given time. This assumption has been employed widely in the use of new forms of data for modelling human dynamics and has been suggested in the case of both

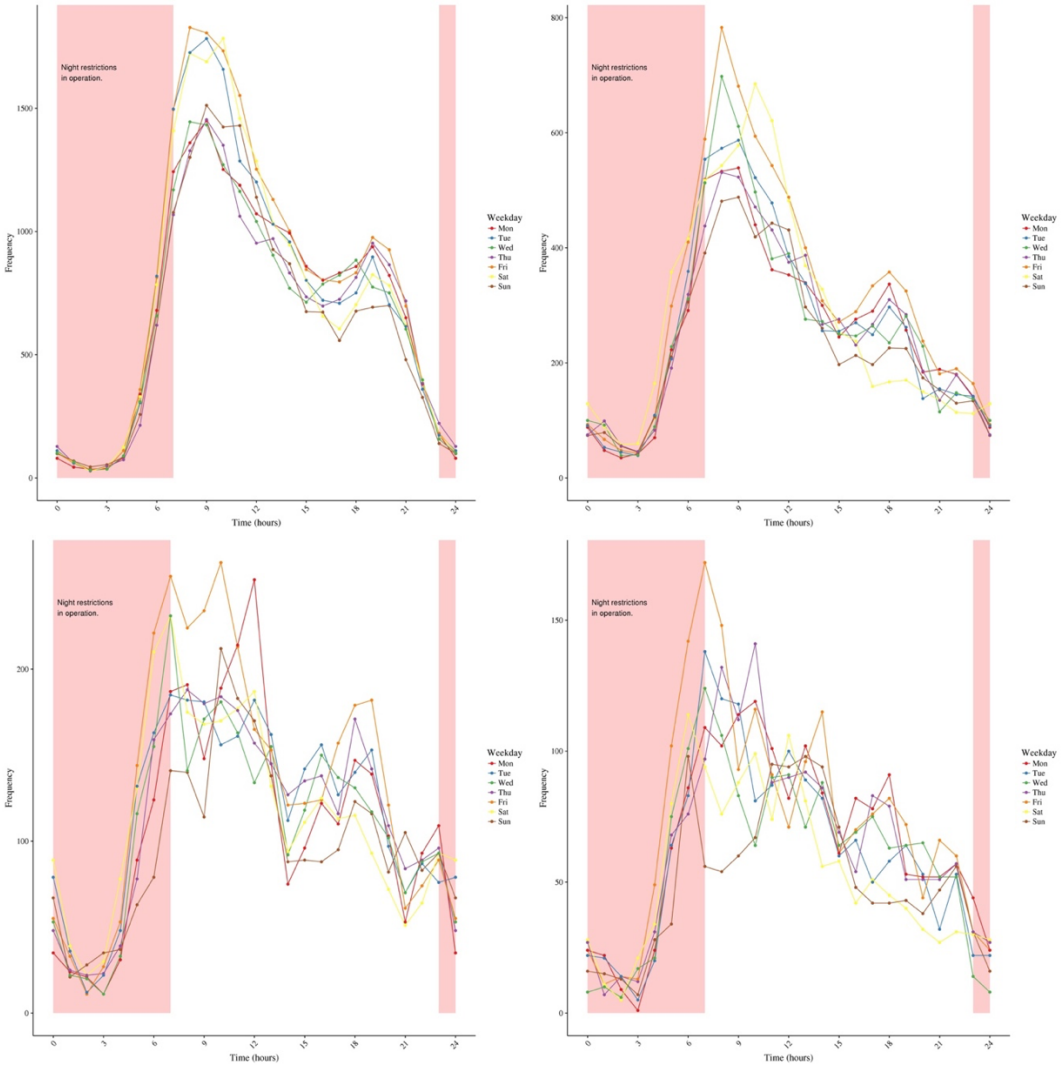


Figure 7.17: Time series plot showing the daily activity patterns based on Tweets for Heathrow (top left), Gatwick (top right), Stansted (bottom left) and Luton (bottom right). The typical hours of aircraft movement restrictions are shaded in red.

Twitter data mining and likewise in the analysis of mobile-phone data. For example, the 'Smart Steps' application developed by Telefonica's Smart Insights team predicts footfall based on the total number of phone calls made on the O2 phone network aggregated in space and time. Likewise, (Lansley and Longley, 2016b) suggest the footfall correlates with the total number of Tweets observed in a given space. Several key differences exist concerning the application of the two approaches to footfall estimation. First, in regards to accuracy and richness. Given the ubiquitous nature of mobile phones, the cell-tower data are rich in observations and ownership is relatively consistent across the population. Conversely, Twitter is used by a relatively youthful population with known gender bias. Second, Cell Tower Data, while rich in volume lacks spatial accuracy. It is not uncommon for cell tower data to have an accuracy in the region of 1 mile; a resolution which significantly limits fine-scale analysis of human mobility (Becker et al., 2013). Third, the ease in which data may be accessed. In the case of much academic research employing cell tower data, limited samples of data have been provided directly from the cell providers. Such data, while rich, constrain the potential applications of the research being performed. In contrast, legacy and present Twitter data are relatively accessible enabling research outcomes to be reproduced and actioned. Such is the limitation assumed through the use of cell-tower data that social media data pose an attractive solution to modelling footfall and human mobility more broadly.

Given that the analysis goes beyond the data that are presently in the public domain, there exist limited means by which the activity profiles may be validated. One possibility, though limited are the activity profiles published by Google as part of their Maps service. Using data collected during subsequent weeks, Google can report near real-time activity levels in specific locations. The data are, however, limited in that they are only accessible for a limited window of time. Thus, given that the Twitter data were collected three years previously, an explicit comparison cannot be made. Further, given that Google data do not provide an indicator of magnitude, it is not possible to make a direct comparison between days or locations.

Compare profiles for Heathrow and Gatwick. Note that the profiles were ac-

cessed on Tuesday the 21 March 2017. As yet, no means is provided by Google to access historical activity data.

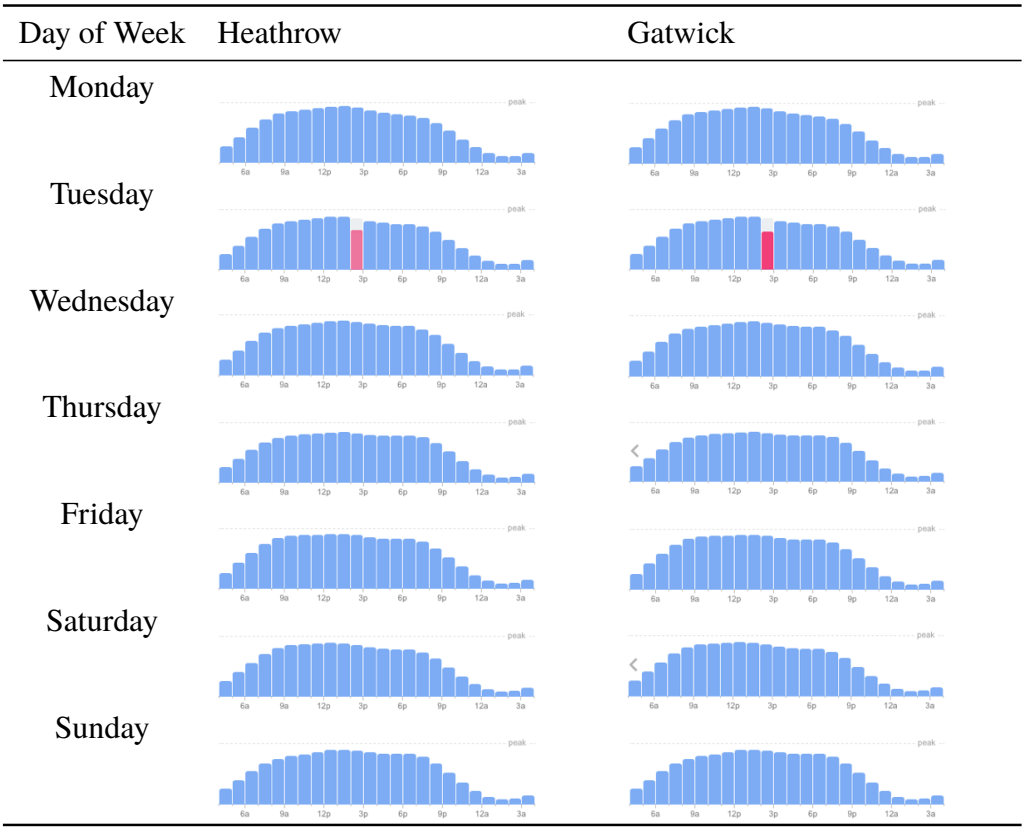


Figure 7.18: Plots showing the typical activity patterns at Heathrow and Gatwick Airports by day of week as determined by Google. The red denotes actual activity versus the typical day at the time of recording (21/03/2017)

Comparison of Figures 7.17 and 7.18 depicts largely similar patterns of behaviour with activity increasing between 3am and 9am before reaching a relatively consistent level of activity which continues until around 8pm. Activity then decreases to a low point around 1 am before beginning to rise gently. This pattern of behaviour provides only limited support to the Twitter-inferred activity profiles suggesting significant over-representation in the morning and under-representation in the afternoon. It may, however, be the case that given sufficient consideration, a temporal standardisation mechanism could be applied.

It may be observed that the data for the Tuesday includes a red bar at 3pm. A feature of Google's activity profiles is that it reports an estimate of real-time activity versus the typical activity over the measurement period. In the case of Figure 7.18,

the red measure indicates that on the 21/03/2017, activity within the airport was less that is typical for this time.

In possession of a general understanding of the temporal activity patterns of Twitter users, the next consideration is how this pattern evolves across space. Various methods exist for monitoring footfall ranging from manually counting the number of people to pass a particular point to the use of various technological counting solutions (Kobsa, 2014). New approaches include the use of aggregated and anonymised mobile phone data and the passive monitoring of WiFi enabled devices (Weppner et al., 2016).

Botta et al. (2015) investigated the utility of mobile phone data and Twitter data for inferring crowd size. Of particular relevance, Botta et al. (2015) conducted a comparison between Twitter activity and mobile phone activity at a football stadium (San Siro Stadium) and an airport (Linate) in Italy. In the case of the football stadium, exact numbers of people were known while passenger numbers at the airport were inferred based on the total number of flights. In the case of the stadium, a correlation of strong positive correlation was observed between the number of attendees and both the mobile phone and Twitter activity datasets. An R-squared value of 0.937 ($n=10$, $p < 0.001$) was observed for the mobile phones and an R-squared value of 0.855 ($n=10$, $p<0.001$) for the Twitter data. These coefficients were lower in the case of the airport case study, however, this is unsurprising given the crude means by which passenger numbers were inferred. Total Twitter activity was correlated with the total number of flights. Such an approach fails to account for aircraft capacity and fullness. One limitation with the analysis by Botta et al. (2015) was that the comparison was performed at the full day scale. Consequently, the differences in daily trends identified above would not have been identified and thus not considered.

For the sake of demonstration, the subsequent analysis is performed based on data recorded at Heathrow Airport. As before, in modelling footfall across space, the assumption is made that an increase in tweeting activity is indicative of an increased number of individuals occupying said space. In monitoring this, several data processing steps were performed. Notably, the Tweets were filtered such that those

containing URLs were omitted. In a large proportion of cases, the Tweets containing URLs were generated by applications linked to Twitter such as Foursquare or Instagram. In both cases, the services employ a form of location-based generalisation leading to apparent hot spots of activity. The applications can submit spatial coordinates through the Twitter API. For example, all Foursquare tweet checking into Heathrow Terminal 1 would be assigned identical latitudes and longitudes, irrespective of whether the sender was in that location or not.

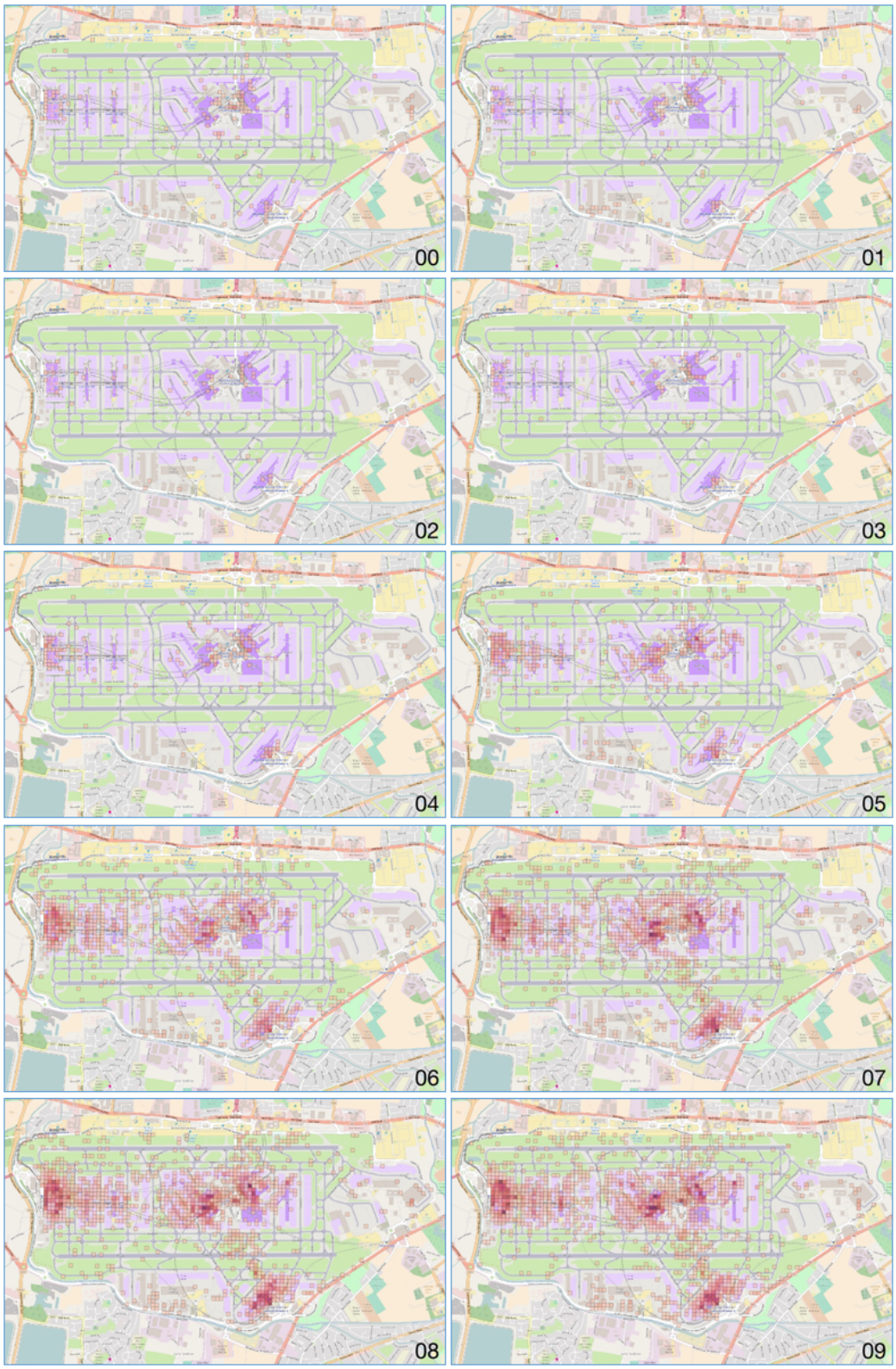
The decision to omit all Tweets containing URLs was due to the use of 'URL shortening' as a means to minimise the proportion of the tweet text taken up by the web address. Such shortening prevents the destination of the URLs being easily determined. It may be possible that the process of URL extension can be performed enabling a more precise form of filtering to be applied.

The process for mapping footfall at each airport was as follows:

1. The extent of each airport was sourced from OpenStreetMap.
2. A 50-metre resolution grid for each airport was created.
3. Tweets were spatially joined to the grid.
4. Counts of Tweets for each grid cell during each hour were calculated.
5. Results were overlaid onto airport map for the purpose of contextualizations.

Midnight until 6am: plots 0 through 5

During the early hours of the day, it is clear that there is limited activity taking place. The activity which is occurring is predominantly observed in the main terminal buildings with very limited activity at the gates or around the runways. The behaviour remains consistent until around 4 am at which time activity begins to increase within the main terminal buildings. Likewise, this period sees the first arrivals from the Far East with passengers aiming to be in the UK for business. This rise in activity corresponds with those individuals arriving 2 hours before the first major flights are due to take off following the end of noise restrictions at 6 am. In the subsequent 2 hours, it is evident that a greater number of individuals are progressing


00
01
02
03
04
05
06
07
08
09

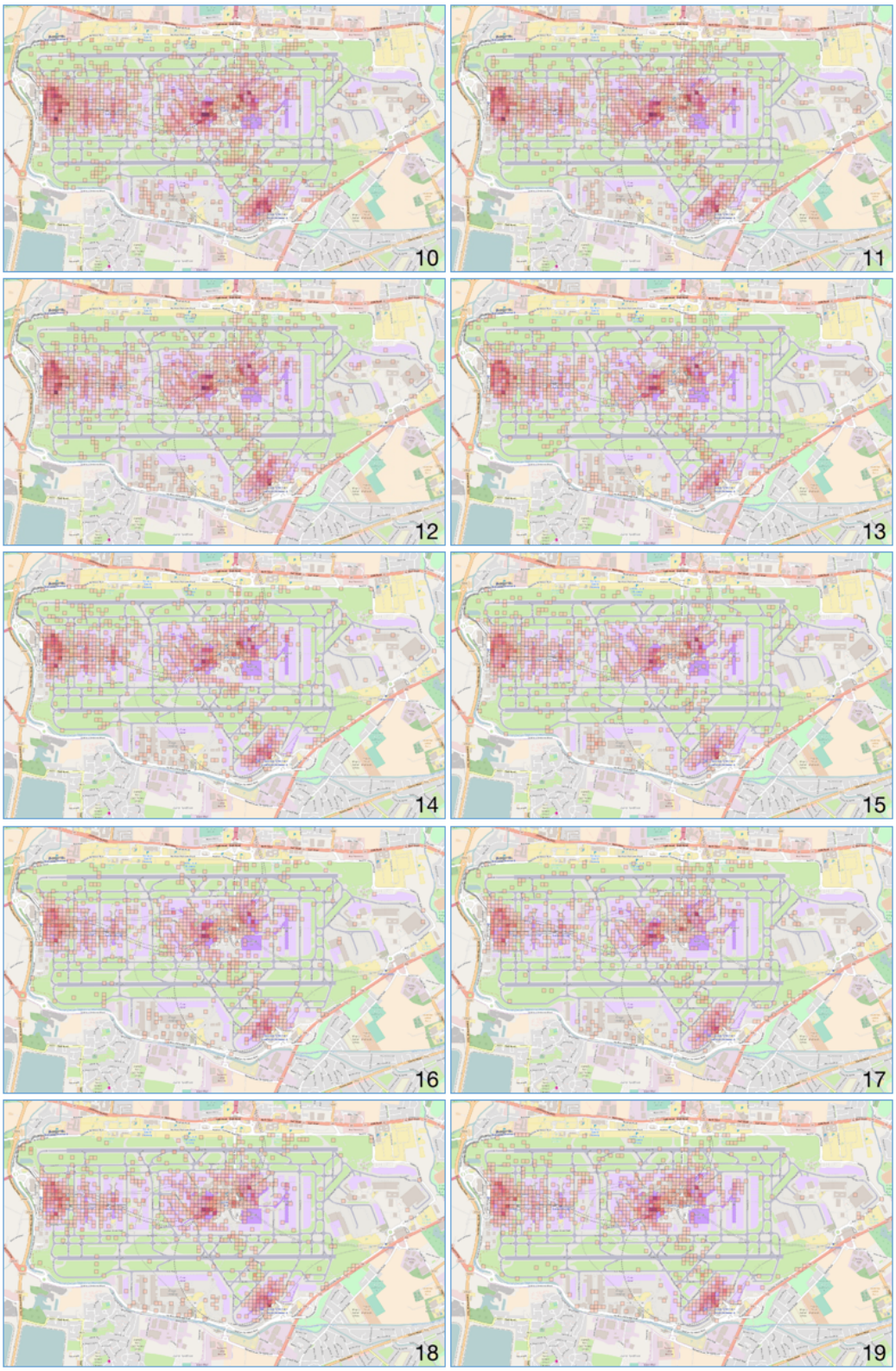

10
11
12
13
14
15
16
17
18
19

Figure 7.19: Maps showing Twitter activity across Heathrow Airport split by hour. The time marker indicates the beginning of the hour being shown.

through the airport with more activity around the main terminal and moving out to the various gates. Also, a greater amount of activity is evident on the various taxiways and runways suggesting aircraft movement are beginning. Also, activity in the baggage handling areas located to the SW of the airport begins around 5 am.

6am to 9pm: plots 6 through 20

With the airport operating with only daytime noise restrictions, activity is significant across the full airport campus. The highest density of activity remains in the main terminal areas and the main entrance and exit routes from the airport. Of note is the centrally located Heathrow Station on the London Underground. Likewise, the arrivals area or Terminal 5 indicates a large volume of activity. Activity in the baggage handling appears to decrease around 5pm suggesting that a proportion of those individuals employed work conventional 9am to 5pm hours.

9pm to midnight: plots 21 through 23

Between 9 pm and midnight, the overall level of activity sees a rapid decrease. Given that Heathrow's night restrictions come into operation around 11 pm, it is likely that fewer individuals are entering the airport campus resulting in the total stock of population within the airport decreasing. Between 11pm and 12pm, it is clear that

activity is predominantly within the main terminals and around the various airport access points.

Summary

Overall, the patterns of activity observed within Heathrow appear plausible based on prior knowledge regarding the structure and operation of the airport. While questions have been raised previously regarding the level of agreement between Twitter activity and observed footfall over time, this could potentially be addressed through undertaking a calibration exercise in much the same way as is employed in other footfall monitoring techniques.

7.3.2 General Patterns in Time

While the above analysis provides a useful means by which activity across space may be observed, it provides little detail regarding the demographic profile of those individuals being observed. Consequently, much of the value inherent in the data, specifically in regards to differentiating between users, is unrealised. In the subsequent analysis, the focus is placed on general patterns across each airport due to the issue of small numbers in certain demographic groups. In seeking to understand better the population being observed the composition of the passengers' nationalities is examined. The following plots illustrate said composition using an aggregation of countries based on the UN top-tier classification of regions. UK-residents are omitted from the Europe region and shown independently (Plot part A).

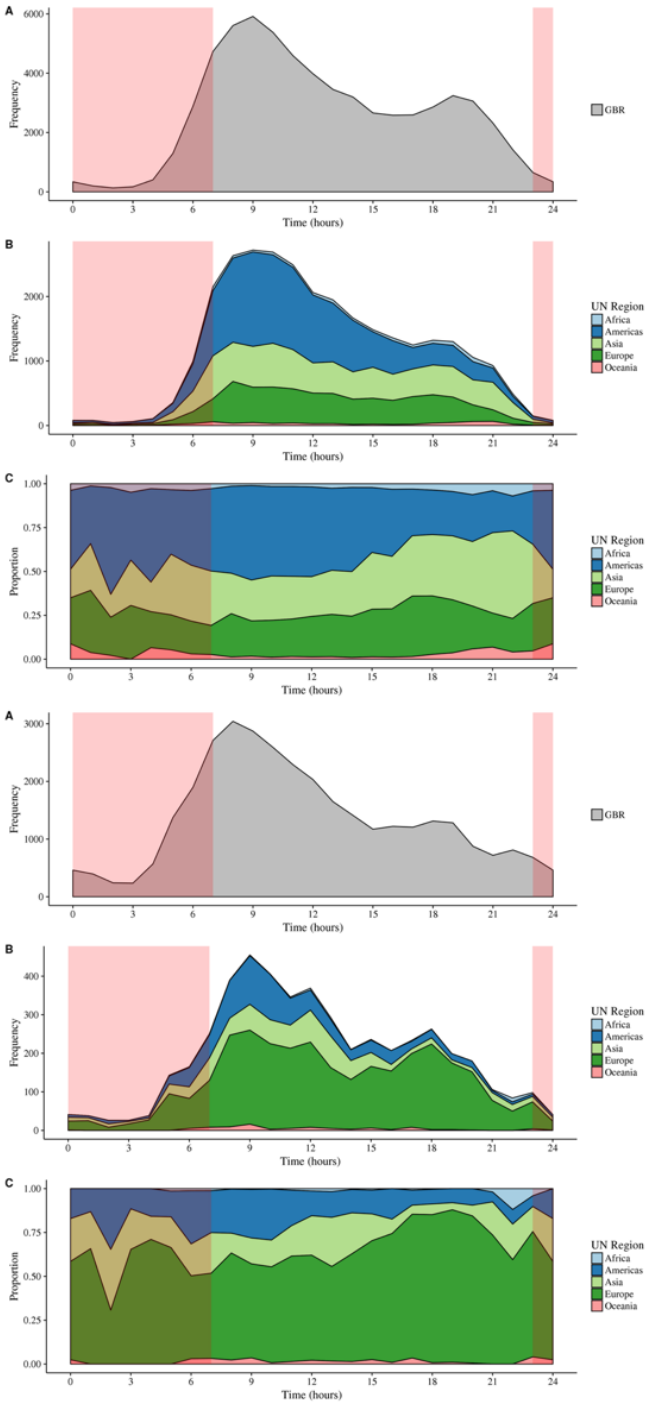


Figure 7.20: Temporal activity plot showing activity by region of residence for Heathrow (top) and Gatwick (bottom). Plot A depicts UK-residents only while Plots B and C indicate all other nationalities as raw counts in Plot B and as a total proportion in Plot C.

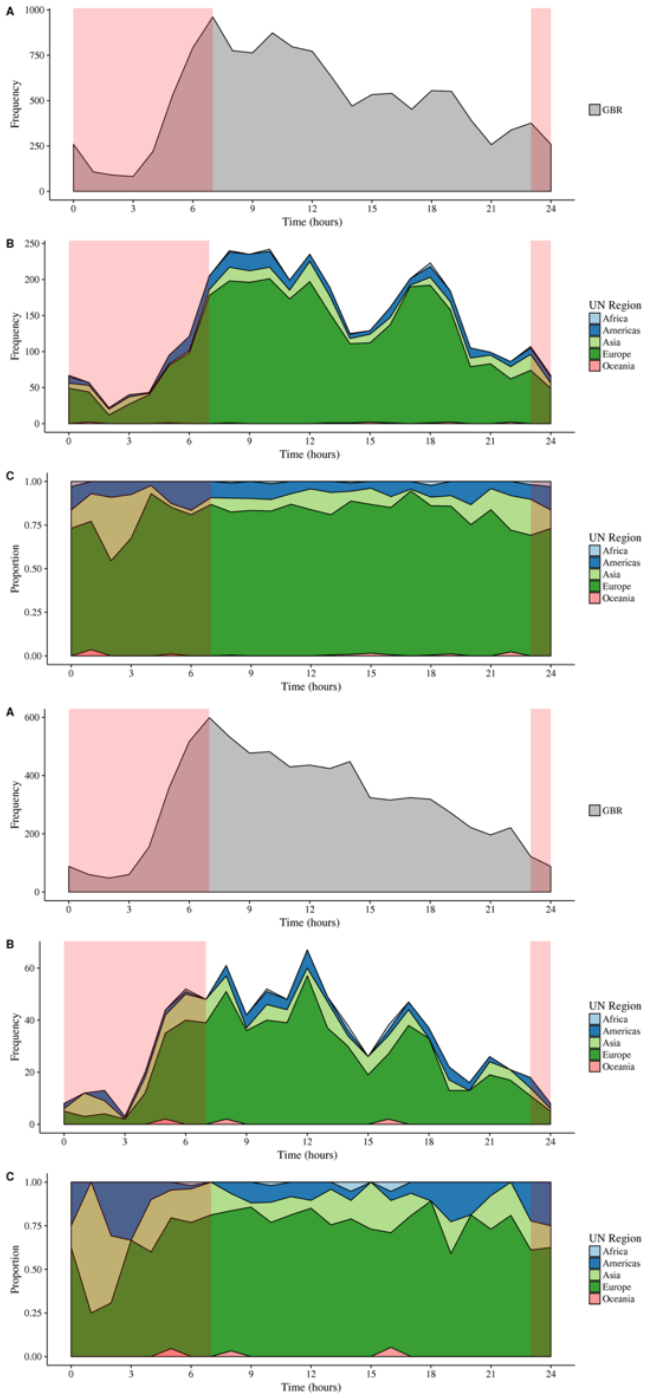


Figure 7.21: Temporal activity plot showing activity by region of residence for Stansted (top) and Luton (bottom). Plot A depicts UK-residents only while Plots B and C indicate all other nationalities as raw counts in Plot B and as a total proportion in Plot C.

Figures 7.20 and 7.21 depict the breakdown of activity patterns through time for each airport based on individuals' believed countries of residence. The regions are the most aggregate UN classification and are chosen as a balance between spatial resolution and the small numbers. For each airport, three graphs are shown. Subfigure A depicts the activity for the users believed to be UK residents, B depicts the activity patterns for the specified global regions (note that Europe excludes the UK) and C depicts the same information scaled such as to be indicative of ownership of space.

Part C of each plot provides a useful insight as to the dominant group within each airport at any given time. In some senses, this may be considered as a way to view ownership of space. In the case of Heathrow, can see how passengers from the Americas are dominant during the late morning and are then replaced by those from Asia as the dominant group during the evening. It should be highlighted that such analysis cannot differentiate between those arriving and those departing. A similar pattern, though less pronounced is observed for Gatwick. In the case of Stansted and Luton, the airports are both clearly dominated by European Travellers.

7.3.3 Analysis of Textual Content

In the preceding analysis, the insight generated has been derived based solely on the spatiotemporal component of the data along with the key personal identities extracted based on the analysis of personal names. The analysis has not yet, drawn on the potential insight which may be buried within the associate tweet texts, however. The analysis of text content offers an additional analysis dimension not possible using conventional sources of passenger data. Textual data mining is a broad subject area concerned with the extraction of useful insight from large volumes of unstructured and semi-structured text (Fan et al., 2006). Examples of standard text mining operations include sentiment analysis and topic modelling. Sentiment analysis is concerned with the identification and quantification of sentiment polarity while topic modelling is a technique for the identification of common themes within large corpora of text (Nikolenko et al., 2015). Both techniques have been widely applied to Twitter. One study, based on the same raw data as this thesis by Lansley and Lon-

gley (2016b), employed topic modelling for the purpose of characterising land use across Greater London.

As an initial step in understanding the text data within each airports' Tweets, the commonality cloud is constructed using the 'wordcloud' package in R (see: Fellows, 2014). The purpose of the cloud is to visualise words common to each of the four airports. This is both useful for understanding the data, and for providing a reality check. Before constructing the cloud, several data preparation techniques were applied. Tweets containing URLs were omitted as these are predominantly associated with other web services such as Instagram and Twitter. Further, usernames and non-alpha characters are deleted.

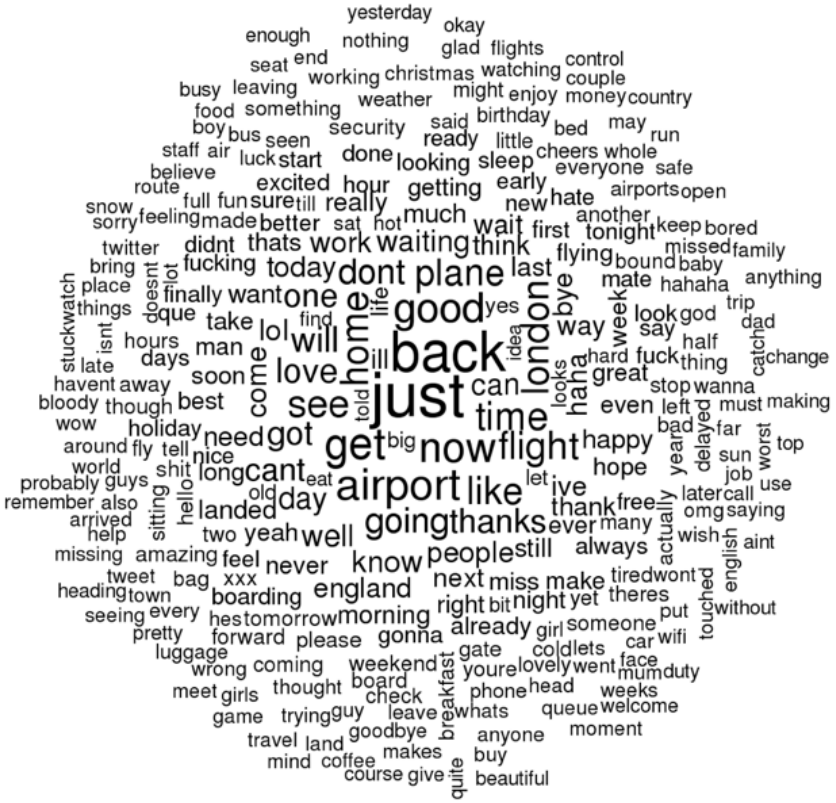


Figure 7.22: Common terms across Heathrow, Gatwick, Stansted and Luton airports.

Figure 7.22, the commonality cloud provides initial insight into the terms which are most common to the four London airports. The size of the words is indicative of how often they occur. Note that many of the keywords identified are clearly associated with individuals' experiences and travel. Examples of such terms include just,

back, home and airport. A limitation of this approach, and various other text mining techniques is the focus on single word tokens. Consequently, paired terms are often missed. An example being the phrase 'just back.' In isolation, these two words provide no useful information. In combination, they provide a clear indication as to the sender's circumstance.

In contrast to identifying those terms common to the four airports, it can be valuable to identify terms which are more strongly associated with each of the four. This may be achieved through construction on a comparison cloud.

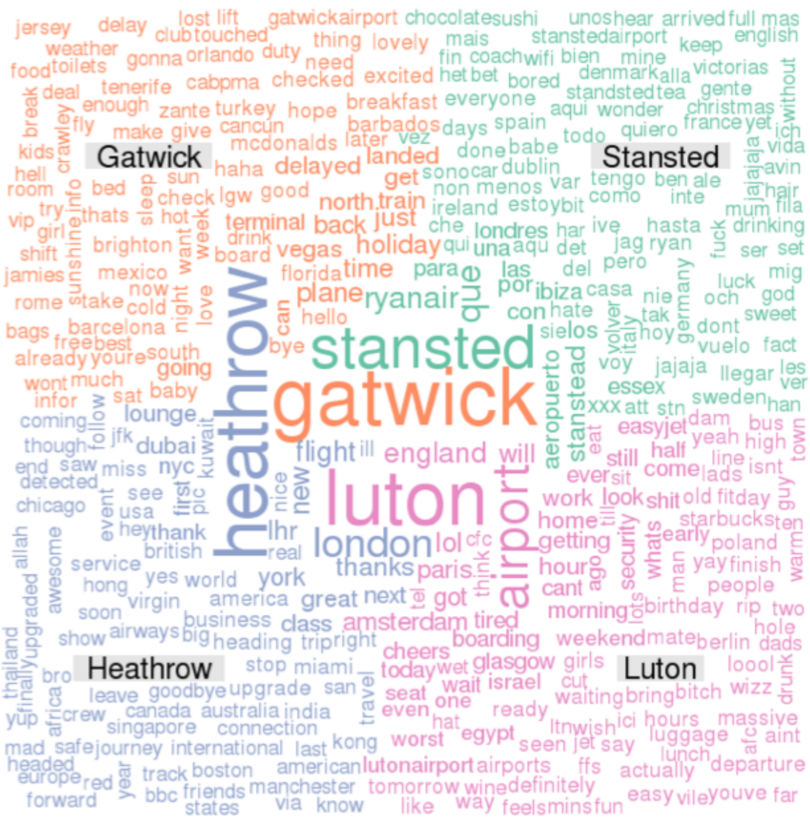


Figure 7.23: Comparison cloud contrasting word use between Heathrow, Gatwick, Stansted and Luton.

Figure 7.23, the comparison cloud, provides an starting point for the identification of terms and topics unique to each of the four London airports. Within each set of words various key terms may be observed. These include reference to specific airports and locations, various airport carriers, common terms related to air travel and also, a range of terms in languages other than English. The clear evidence of key

terms suggests that there may be opportunities to employ topic modelling techniques as a means to better understand the activity within each of the four airports. Likewise, the presence of emotive words such as 'yay', 'awesome' and 'tired' suggest the potential for the use of sentiment analysis.

7.3.4 Sentiment

Beyond knowledge of what is being discussed, it may also be useful to understand the collective mood of individuals as they progress through each airport. It is possible to analyse the general mood through the application of sentiment analysis. Sentiment analysis encompasses a broad range of techniques designed to determine and quantify the degree of happiness or anger (Wilson et al., 2005). In this analysis, we employ a lexicon-based approach to determining the degree of sentiment within each Tweet.

The lexicon-based approachs may be divided into those concerned with individual words and those which analyse words within their original context (Taboada et al., 2011). In the case of the latter, the typical approach involves the use of a dictionary or lexicon of specific words and associated polarity scores. To determine the overall sentiment within a document, the sum of sentiment scores within the text is calculated. Such an approach is efficient. However, it fails to account for the more nuanced nature of language. More specifically, the effect of valence shifters, terms which alter the effect or interpretation of specific words (Polanyi and Zaenen, 2006). Examples include the use of specific terms which might negate or amplify the polarity of a term. In practice, it is important to consider the effect of valence shifters; terms which alter or nullify the polarity of a term. Specific valence shifters include negators, amplifiers, deamplifiers and adversative conjunctions. A negator is a term which reverses the emphasis of a specific word. For example, by placing the word 'not' before the word happy, the sentiment is shifted from being positive to negative. Amplifiers and deamplifiers either increase or decrease the magnitude of sentiment within a phrase. Adversative conjunctions are phrases which alter the polarity of previous phrases.

In this analysis, the 'sentimentr' package is developed by Rinker (2016). This

package is written in such a way as to identify the occurrence of negators, amplifiers, deamplifiers and also, the more complex adversative conjunctions. Further, while not employed in this analysis, the package contains methods for dealing with emoticons and various ratings methods. The approach applied is as follows:

1. Polarised terms are identified based on a pre-existing dictionary of terms.
2. Positive words initially scored 1, negative words -1, non-polar words as 0.
3. The four preceding and two super-ceding word tokens are subset and checked for amplifiers, deamplifiers, negators and adversative conjunctions.
4. Based on the above, each Tweet is assigned an overall sentiment score.

Having applied the sentiment classification algorithm, various possible aggregations of the data may be implemented for the purpose of understanding opinion within each of the four airports.

Table 7.6: Mean sentiment by airport versus 2017 Google Review Scores.

Airport	Mean Sentiment	Google Review Score
Heathrow	0.037	3.9*
Stansted	0.027	3.9*
Gatwick	0.025	3.1*
Luton	0.014	2.8*

In the initial assessment, the mean sentiment within each London airport was calculated. The results of this assessment are shown in in Table 7.6. It may be observed that those individuals within Heathrow express the greatest overall positive sentiment while the least positive sentiment was observed at London Luton. As a point of reference, the Google review score of each airport was sourced. It is interesting to observe some degree of ordinal correlation between the two scores, though, it is acknowledged the Google Review Scores have been collected a number of years after the Twitter data. Seeking to understand the sentiment across the four airports better, the data were disaggregated such the sentiment by hour could be observed.

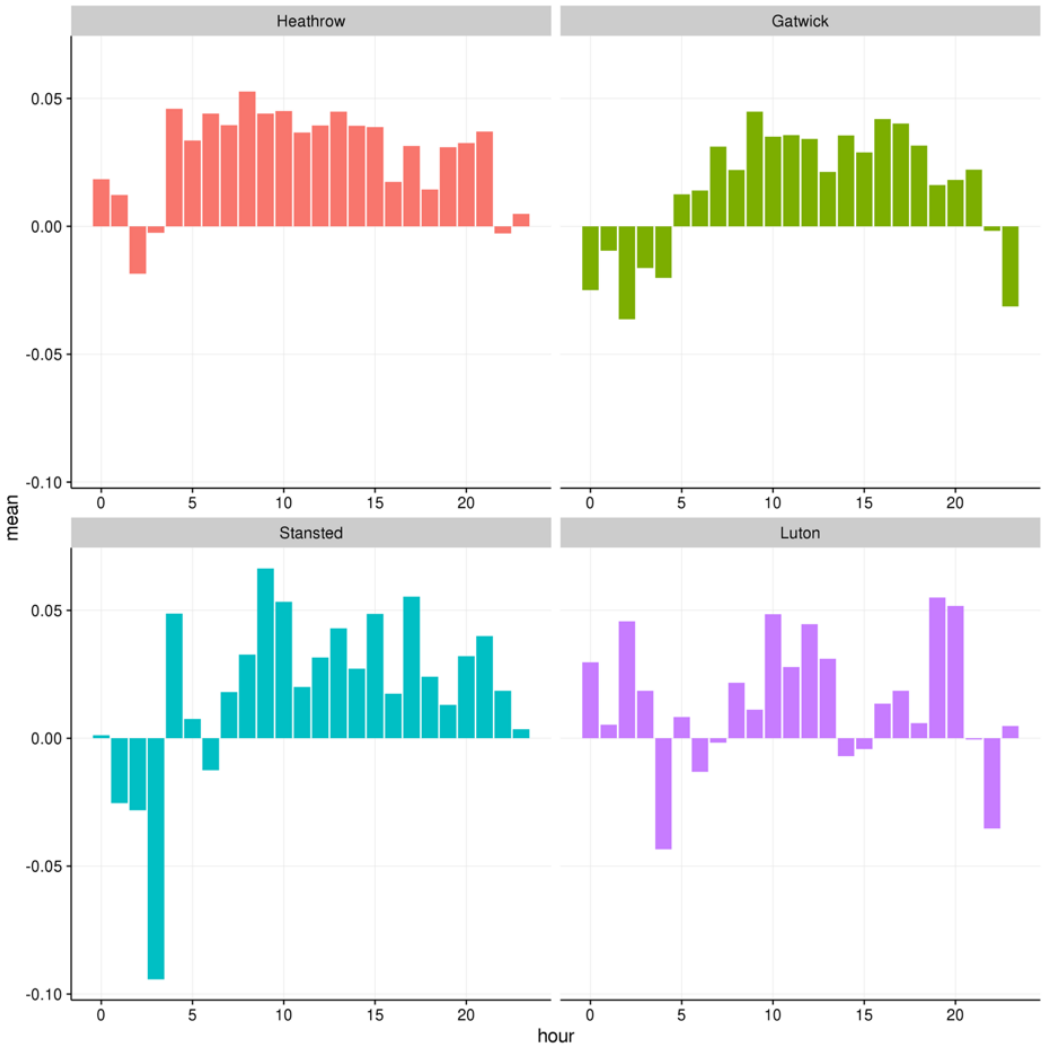


Figure 7.24: Plot showing the mean Tweet sentiment for each airport during each hour.

Figure 7.24 provides an illustration of the mean sentiment across a typical day at each of the four London airports. The four plots illustrate a range of patterns of sentiment. In the case of Heathrow, the sentiment is generally positive throughout the day decreasing during the night hours. In the case of Gatwick, a similar pattern of positive sentiment is observed during the day, however, during the night, the sentiment is varying degrees of negative. At a similar daily trend is observed, however, the variability between hours is more significant. The issue of variability is further evident for Luton which depicts a relatively volatile profile. Except for Luton, it would appear that during the day individuals are generally happier and during the night are less happy. The effect of time of day is acknowledged in the literature and

is associated with individuals' circadian pattern and normal activities (Csikszentmihalyi and Hunter, 2003). The effect of time of day is somewhat more complex in the airport context in which the individuals present may have originated in multiple different time zones and also, due to the variations in individuals purpose for being at the airport: for their work or, business or leisure travel.

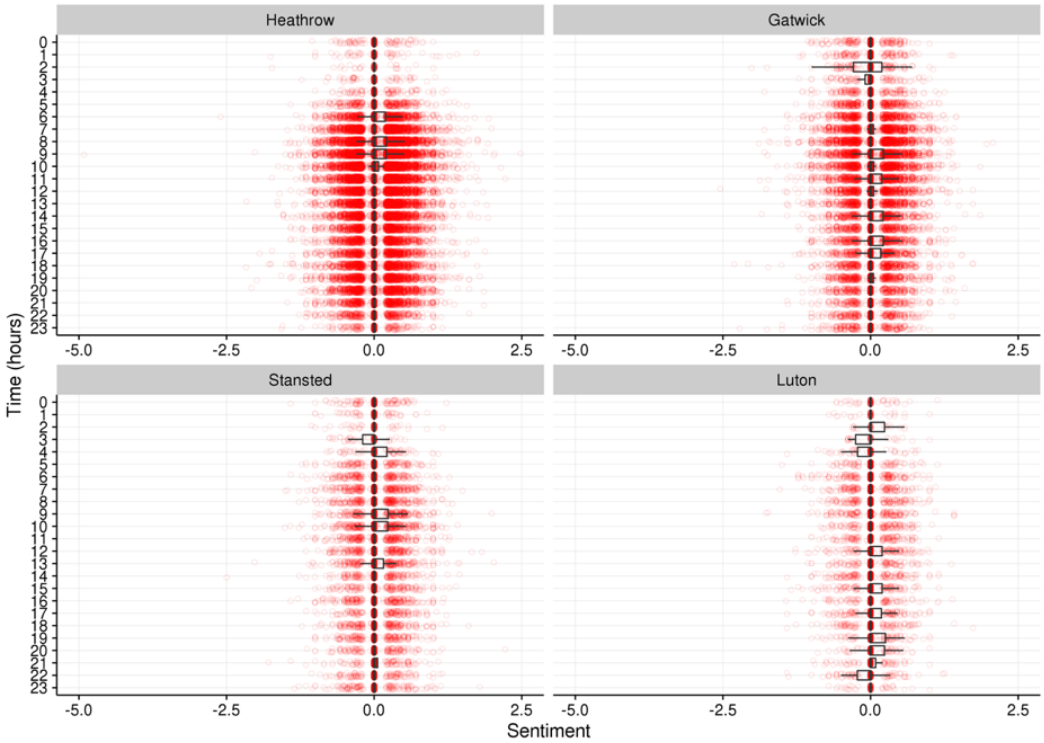


Figure 7.25: Box plot showing the distribution of sentiment split by airport and hour. All data points are shown.

While Figure 7.24 provides a useful depiction of sentiment, it fails to demonstrate the volume of activity during each hour. Consequently, while 3 am – 4 am at Stansted appears to be negatively orientated, the amount of activity is relatively small compared to the rest of the day. One alternative is the use of a combined dot plot and box plot. Figure 7.25 illustrates the hourly distribution of sentiment for each of the four London airports. The graph provides both an indication of the range of sentiment values recorded each hour and also, by virtue of transparency, an indication as to the number of Tweets being submitted. The minimal size of the box plots in the majority of hours' highlights that the vast majority of Tweets contain very limited sentiment either positive or negative.

A potential extension to this analysis would be to map sentiment across the same grid employed in the footfall analysis. Such an application may provide an effective means by which stress points within each airport could be identified. In partnership with an understanding of the internal airport layout, this may provide a useful tool in the case of retail, security and planning. Further, greater linkage could be made between the analysis of sentiment and personal identities. Such a link may enable more effective provision of internal airport security as it could help target those individuals who are expressing the greatest degree of negative sentiment.

7.4 Discussion

In the preceding analysis, the objective was to showcase the potential for richly attributed social network data for the purpose of understanding human mobility and behaviour within a constrained environment. A case study was delivered for four of the six major London airports in two parts. In the first, the focus was on replicating what could already be achieved using conventional forms of data generated by or for UK airports and aviation. In the second, the focus shifted to forms of analysis and insight which could not be achieved with data that were previously available. Through the course of the above, a range of strengths, weaknesses and opportunities have been raised. In many cases, these observations are either associated with the rich spatiotemporal nature of the data or the issue relating to the representativeness of the data given that they are sourced from a self-selecting sample of the population.

A key advantage of the Twitter-based approach is that the data are readily available and may be collected in real-time or purchased retrospectively via data-warehousing services such a gnip (see: http://www.gnip.com). Not only does the availability of data facilitate analysis at custom timescales, it means that the forms of insight generated may easily be applied to a range of other locations such as retail locations, sports venues or alternative travel hubs. In effect, Twitter provides a consistent data source across a huge range of application. This is particularly important regarding developing an analysis framework making insight generation fast and effective. However, it is also important that limitations in the data are recog-

nised. Notably, while we often consider Twitter data as being rich in volume, once the window of analysis has been constrained, for example in the case of an airport, the number of individuals who are sending spatially referenced Tweets is surprisingly small. This problem is exacerbated by the inclusion of automated location reporting applications such as Foursquare and Instagram which can be challenging to distinguish from other Tweets. The issue is of most concern in the smaller airports in which the total number of recorded Tweets is already low.

Given the sparseness of Tweets once processed, and the reality that the number of messages may not easily be improved upon, the most practical option available is some form of spatial or temporal aggregation. Fortunately, given the nature of airports and the routine patterns of activity, it is possible to perform various temporal aggregations. The most useful of which, used in the above, is the day of the week. However, unlike much traditional analysis of population data, the analyst retains a degree of control over the modifiable areal unit problem. In practice, whatever decision made would involve some form of compromise. That said, it is a common fallacy within quantitative geography to place too great an emphasis on achieving a high spatial/temporal resolution when a more general measure may be more appropriate. Examples of possible aggregations include zoning based on functional zones or aggregation of specific activity times. The issue of spatial aggregation becomes more complex when we consider that while it is possible to access the horizontal position of each Tweet, no information is available on altitude. Consequently, when faced with an environment with multiple overlapping functional zones across several floors, it will not be possible to differentiate between them.

The issue of aggregation may be considered more broadly regarding individuals' personal identities. While it may be desirable to know the specific nationalities of individuals', it could be argued that such precision is unnecessary. The issue is discussed by Mateos et al. (2009) in the context of ethnicity who highlight that both the aggregation and granulation of ethnicity classifications is inherently unstable. This issue is compounded in the sense that the ethnicity classification associated with each user is purely a prediction and may not necessarily represent the individ-

uals' true ethnic type. It may, therefore, be more accurate to employ a lower level of precision.

Other concerns include the availability of data and challenges in differentiating between those employed within the airport and those which are passengers. As has been noted previously, the quality of analysis is dependent on the volume of data available. Given that in practice the data observed are quite sparse, it would suggest the techniques demonstrated are best suited to the identification of general patterns of behaviour rather than investigating specific circumstances. In effect, pushing a nomothetic agenda in which we seek to identify laws rather than model the behaviour of individuals'. The second challenge in such analysis is differentiating between those individuals who are transient within the airport and those who are employed in some regard. This specific question is not easily answered. Lacking attribution associated with employment, it would be necessary to analyse either individuals' spatiotemporal tweeting patterns or look at options to mine each users' Tweets for employment indicators.

The last major concern is the ethics associated with the use of social media for the observation of human mobility patterns and behaviour. As with all forms of public data collection, there are various issues relating to privacy and consent. In the context of Twitter, such concerns are largely ignored by many using the justification that the data are publicly available and individuals have consented by virtue of agreeing to the stated terms and conditions (Zimmer and Proferes, 2014). In further support, in comparison to other forms of population monitoring data, one might argue that the use of Twitter is the least intrusive and most controlled by the user. Where individuals make an active decision to share content on social media and are conscious that what is being submitted may be viewed by others. Relatively few are aware that mobile phone providers are using their data for the purpose of monitoring population dynamics. The ethics are increasingly concerning when we consider the use of WiFi monitoring of footfall. WiFi monitoring relies on individual's phones interacting with strategically placed WiFi receivers. These receivers do not broadcast their presence, and thus public awareness of their use is limited. While the

output of such data collection should be anonymised and aggregated, it remains the case that such forms of data collection are intrusive. Such is the concern associated with the ethics of WiFi monitoring that the UK Information Commissioner Office published guidance on the appropriate use of such technologies (ICO, 2016).

Beyond ethics, a further debate surrounds the issue of data ownership. This is a major concern of individuals who are often unaware that data generated by-product of their activities may be employed in a broad range of applications. Many of the social media platforms maintain the stance that any data submitted by the user remains the property of the user. However, the platforms maintain the right to employ or commercialise the data as and how they choose. Given the ethical concerns associated with the use of cell-tower data and WiFi monitoring, the justification for the use of social media in the inference of general patterns of human mobility and insight generation may be considered quite positive. With sufficient consideration given to the biases inherent in the data, and maintaining the appropriate degree of privacy, Twitter appears to be a valuable and useful meaning of generating high-quality insight into the stocks and flows of populations.

7.5 Conclusions

In this chapter, the objective was to demonstrate the potential of Twitter as a viable alternative to conventional forms of demographic data. To showcase this, an analysis of the four largest London airports was performed. In the first instance, conventional insight, such as demographic profiles and catchment analysis were performed. Building upon this, it was demonstrated how the rich spatiotemporal attribution associated with each Tweet facilitated the generation of previously inaccessible insight. Notably in regards to modelling footfall and tracking sentiment. In appreciating the above, it must be recognised that the insight demonstrated has been creating using data that are freely available. The significance of this is that the analyst maintains full control over the data and the processes which are subsequently applied. Versus alternative crowd monitoring solutions, such as Google Places or Telefonica's Smart Steps, the methods employed are fully transparent and reproducible. As such,

the framework may be readily tailored to the task at hand. Further, while the analysis is performed in the UK, it may easily be reproduced in other situations with the only significant limitation being the popularity of Twitter within the study area. Consequently, the benefits of employing Twitter as an alternative, or in support of conventional population data is increasingly evident.

Chapter 8

Conclusions

8.1 Introduction

At the outset, the aim of this thesis was to explore the potential of new forms of data to address limitations in current demographic profiling practices. By means of a critical evaluation of current approaches, it was identified that the main constraint was the data by which the various human behaviours are studied. Conventional social survey data and secondary sources are collected on an infrequent basis, lack consistency between years and regions and are inherently cross-sectional in nature. A case in point is the three regional Censuses of Population collected across the UK. While collected concurrently, the specific questions are inconsistent, limiting the ease with which UK-wide analysis may be performed. Further, Scotland and Northern Ireland employ a different recording geography to England & Wales for some data sources, such as the Census of Population. Further, given the infrequency of their publication (every ten years), the data are increasingly uncertain over time (Singleton and Longley, 2009). In light of this, it was proposed that new forms of data may provide an improved means by which the stocks and flows of the population may be observed and analysed. The concept of new forms of data is relatively recent and has arisen alongside the development of Web 2.0 and the Internet of Things. Increasingly, people's offline and online identities are entwined, with vast volumes of data being generated as a result of their activities. Such data range from energy consumption records collected via smart meters to online communication data shared within on-

line social networks. These data, while not collected for the purpose of observing populations, provide a new and exciting means by which population insight may be generated.

While new forms of data are relatively broad in scope, this thesis is focused predominantly on data generated by the Twitter Social Network. The use of Twitter data in the generation of population insight is not in itself new, however, as was discussed in Chapter 3, such analysis has often been beset with limitations. Chief amongst these and a recurring theme throughout this thesis is the issue of representativeness. By their very nature, Twitter users are a small self-selecting sample of the population. However, while it is well recognised that the Twitter population is biased, little has been done to address or acknowledge the issues which this introduces. In effect, research conclusions are being made in which the results are generalised to the population as a whole, disregarding the true demographic of the Twitter user base. This is well demonstrated in the literature modelling flu trends based on the analysis of Twitter data. While it may well be possible to track general patterns in flu, the population considered most at risk of contracting the Flu, the young and old, are not necessarily those who are represented by the data. There is thus a risk that lacking recognition of the bias inherent in the data that resources are incorrectly allocated and the target population missed.

In light of the above, this thesis set out an ambitious plan to assess the representativeness of Twitter at a range of spatial scales from the local to the global. A key feature of this assessment, outlined in Chapter 4, was the approach employed for the construction of functional population inventories based on the analysis of geographically referenced Tweets. Guided by the United Nations (2001) definition for population registers, the framework transformed the raw Twitter data into a global population inventory containing distinct ids, extracted forenames and surnames, and also inferred the probable place of residence at a range of scales. In Chapter 5 the analysis sought to demonstrate how representative Twitter users are of the population at the global scale while in Chapter 6 the issue of representativeness in the UK was explored in more detail.

In Chapter 5, the analysis measured the compositional similarity of surnames across 22 countries drawing on data from the UCL Worldnames Database for reference. In turn, these data were used as a basis for modelling social media penetration across the globe providing both insights on the global geography of Twitter, and also a valuable point of reference in regards to the interpretation of international travel behaviour.

Following the global assessment, a UK based benchmarking exercise was performed. Unlike the preceding analysis, the objective was to determine the extent to which bias existed within the Twitter population in regards to a series of imputed key demographic attributes: age, gender, ethnicity and geographic distribution. Based on individual users' personal names, the analysis employed a range of novel data-mining and heuristic techniques to infer characteristics of those posting Twitter messages. Not only does this attribution facilitate comparison against conventional forms of data, but it also provides a means to differentiate between different groups within the population when performing analysis; a feature which may have a broad and significant range of applications.

Although aspects of the analysis shine a somewhat negative light on the potential for Twitter data in the generation of population insight, Chapter 7 sought to remedy this through demonstration of the potential opportunities of such an approach. Through completion of a case study on London's four largest airports, it was demonstrated how the enriched Twitter data might be employed in place of conventional data, and subsequently, the additional insight that they may provide. The ability to differentiate between different users based on their inferred identities enables more extensive insight than previously possible. What is more, this insight was generated without direct observation or access to privileged data. Thus, the techniques may readily be applied in a broad range of contexts from retail and marketing to security and crime.

8.2 Reflection on Methods

A key feature of this thesis has been the extraction of insight based on the novel application of name-based heuristics and spatio-temporal data mining. Giving the unstructured nature of the Twitter data and the lack of explicit personal attribution it has been necessary to develop and employ novel algorithms and heuristics for the inference of individuals' key personal identities. Of these, the two most fundamental processes were the identification of individuals probable place of residence and the extraction of their personal names.

While several possible approaches exist for the extraction of location information, it was decided to use the geographic coordinates embedded within the geographically referenced Tweets. The full process for this is detailed in Section 5.3.3. This approach was based on using the least-ambiguous source of geographic reference information. The obvious alternative to this approach would have been to determine individuals nationalities and places of residence based on either their language, reference to locations within their Tweets, or the location field associated with the account. Potentially, such an approach would allow the incorporation of a greater number of users at the national scale. However, while it may be possible to process individuals in such a manner, the real barrier to this approach is associated with the ability to access sufficient data from Twitter. Such analysis would be better suited should the original data have been harvested from the Firehouse rather than Sample Stream. An additional consideration is the purpose for which the data were to be employed. In terms of the creation of dynamic population data there is a requirement for high precision observations in time and space; a feature not readily possible if manually geo-coding individuals' locations.

In terms of personal names extraction, Twitter unlike some alternative social networks, does not require individuals to report 'real' names and rather allows any combination of alpha-numeric characters to be recorded. It is a fortunate feature that individuals often use this field to report their actual name. This data is, however, unstructured in that there is no distinction as to the structure of the name. The process employed in the extraction of individuals' personal names is discussed in

Section 5.3.3. As was noted, the approach employed was based on the assumption of a western naming order in which the given name precedes the family name. Given that the large proportion of countries where Twitter is popular employ this naming convention, it could be considered fit for purpose. However, in future there exists the potential to develop the algorithm such that names in both western and eastern order may be processed. Given that individuals of different cultural, ethnic and linguistic groups are typically co-located, it may be necessary to employ Onomap on the raw data as a first step in determining an appropriate strategy for name processing.

Beyond location and personal names, a range of pre-existing heuristic techniques were employed through the course of the thesis. Developed in academia, many have seen limited application beyond their initial development. Consequently, efforts were made to assess the utility of these tools prior to their application such that a degree of confidence could be assumed in terms of their classification ability. Case in point being the assessment of the Onomap classification against the 2013 Consumer Register and Ethnicity data from the 2011 UK Census of Population. Here is was possible to identify apparent systematic bias in the Onomap classification and thus make informed decisions as to the validity of the Twitter-inventories in the UK. Lastly, a key process in the assessment of each inventory was the Morisita-Horn index of overlap. Drawn from the ecology literature, the Morisita-Horn index was designed to measure overlap in population structure based on samples of species. The approach was readily interpretable in the context of individuals names where each surname is considered as a separate species. A discussion on the merits of the Morisita-Horn index is given in Section 4.3.2.3. In their analysis, Wolda (1981) discusses the use of various similarity measures and the effects that these may have on the observed similarity scores. The use of the Morisita-Horn index was critical given the variation in size and diversity observed within the various population inventories.

Given the relative infancy of social media data and name-based heuristics for the creation of demographic insight, it is as important to focus on the application and utility of the tools as it is to establish new knowledge. As, without confidence in the

methods, any insight generated is of very limited value. In terms of the above, it is important to consider replication and generalisation. Replication is concerned with the recreation of existing research or academic findings to ascertain the reliability of the results. Often replication is conducted by third parties and includes their own interpretation of the literature, methods and outcomes. Successful replication of findings serves to strengthen the outcome of both analyses and conversely, unsuccessful replication suggests that one or both of the approaches are flawed. Generalisation is concerned with the application of pre-existing methods beyond the context within which they were developed. The purpose of generalisation is to ascertain the scope of a method and consequently its broader applicability. In the case of generalisation, new data may be required for validation.

Recognising the importance of the above, this thesis has sought to establish the provenance of both the methods and data being employed prior to their application. In particular, in Chapter 6, focus was placed on the use of the name-based heuristics for the allocation of key personal identities. by establishing the accuracy and precision of the methods at the correct scale of analysis, it was possible to assess each methods efficacy and consequently, interpret the analysis outcomes appropriately. This validation is exemplified in the case of the Onomap CEL classification tool. Reported in Section 6.5.2, it was found that the classification tool was biased toward the 'White' CEL group and failed to classify individuals into categories such as 'other'. In effect, the assessment may be considered as replication, while the application of the method to the Twitter data may be considered as generalisation. In both cases, these analysis enable us to have increased confidence in the methods and observed results.

Moving forwards and looking to establish the use of social media there is clearly a need that these methods be assessed further. In particular, it would be beneficial to employ the location and name-based heuristics in a range of alternative countries such that their provenance is further established. The framework for benchmarking, employed in Chapter 6, provides an effective means by which such an assessment could be structured. The main challenge in achieving the goal of generalisation is

the availability of comprehensive individual level and aggregate population data to be used for reference. A potential first candidate for assessment would be the USA given the popularity of Twitter and the availability of detailed demographic data. However, in the long term, it would be beneficial to target countries based on the dominant cultural, ethnic and linguistic group such that the transferability of the methods may be established. It is probable that collaboration with other global institutions would be required such that knowledge of local trends and conventions may be incorporated into the analysis and interpretation of the methods.

8.3 Summary of Findings and Limitations

As a result of this thesis, a broad range of insights has been generated – not only in terms of whom the Twitter data best represent, but also with regards to better practice in the collection, analysis and interpretation of Twitter data. Considering the issue of for whom the data represent, the global scale analysis delivered a unique viewpoint on Twitter and its global reach. Until now, only data pertaining to a select few countries has been analysed. The production of a global assessment provides both a reference for practitioners wishing to explore the use of Twitter in specific regions and also a means of standardisation when considering the activities of international travellers. Interestingly, it was observed that Twitter was most popular in English and Hispanic speaking countries.

Considering the analysis at the national scale, Chapter 6 delivers an assessment of how representative the Twitter data are of the UK population. The analysis confirmed various anecdotal beliefs concerning of whom the Twitter data are representative. A first issue was the presence of gender bias. It was found that within the sample studied, a notable male bias was present. In terms of age, Twitter users were found to be predominantly of a younger age group than the observed population as recorded in the Census of Population. This behaviour was consistent across the UK with the male bias increasingly evident as age increased. In terms of ethnicity, the results were inconclusive. While efforts were made to infer individuals' ethnicities based on the Onomap classification tool, when applied to the 2013 Consumer Reg-

ister, the most comprehensive list of personal names publicly available, the level of agreement with the Census of Population was poor.

Based on the analysis performed, a case can be made for the use of the processed Twitter data as a population inventory. However, it should be noted that there are clear limitations in terms of applicability. These limitations include the proportion of the population for which the data are considered representative and the degree of spatial and temporal granularity with which the data may be considered sufficient. It may be that the data are better described as a sample population inventory such that the description is not considered to be misleading. As new forms of data continue to be incorporated into population analysis, it will be necessary that further consideration be given to the nomenclature used in the description of individuals identified and profiled within the data. In many respects, greater parallels may be drawn between Telefonica's SmartSteps application and the Twitter population inventories than conventional records of population and traditional demographics.

A core feature of this thesis was the adoption of a data-rich approach to analysing individuals' identities. Unlike much conventional analysis using Twitter data, the raw data employed in the study were collected at the global scale. Consequently, the activities of individuals could be examined beyond the specific study area. The merits of such an approach were demonstrated in the case of the London airports in Chapter 7. In possession of the global data, it was possible to infer individuals' nationalities and effectively differentiate between residents and non-residents and also determine airport catchments. Such an extension not only increases the range of insight that may be generated, but it also ensures that the various name-based heuristics and nationality-dependent analysis are only applied where they are appropriate.

The second core feature of this analysis is its transferability. When considering new forms of data, much excitement surrounds the use of smart sensors, consumer data and cell phone tracking. These forms of data do provide a range of novel insight, however, are severely limited in regards to their application beyond academia. While it may be possible to demonstrate a correlation between crime rates and footfall based

on cell-tower activity, it is necessary that further data be procured for any subsequent analysis. Conversely, Twitter provides a range of methods through which data may be harvested in real-time or purchased retrospectively. The attractiveness of Twitter is further enhanced due to its maintenance of a consistent data format enabling the data to be employed efficiently in a range of scenarios.

A key assumption within the analysis was that individuals' online and offline identities are inextricably linked as a result of physical and cultural anchors. This feature of the analysis is critical for the inference of individuals' personal identities which is reliant on the analysis of individuals' online identities. In Section 2.3.2 it was reported how two main views exist regarding the association between individuals' online and offline identities: The extended real-life hypothesis and the idealised-self hypothesis. It was argued that the nature of modern social networks is such that individuals are relatively constrained in the degree to which they may embellish their online identities, however, that this varied in degree based on the specific identity being represented. Within the course of this thesis, the primary means of determining an individual's identity was their personal name which is a key anchor between an individual's identities. In addition to personal names, the location at which individuals tweeted was fundamental to determining their identity. Unlike a users name, a user may have greater interaction with the locations which they report and this may provide an opportunity for embellishing personal identity. For example, including location information where it may be deemed 'cool' to be seen. However, this is likely to effect only a small proportion of users with many enabling location sharing by default and consequently providing a more passive indication as to the locations visited. Further, in the context of this thesis, the scale of analysis - typically at the national or regional scale - is such that localised location reporting bias is unlikely to have a significant impact on the conclusion drawn. The exception to this would be in the completion of micro-scale analysis, such as in Chapter 7, when analysing population dynamics within the London airports.

Moving forwards, it is advisable the the likelihood of self idealisation and behavioural biases in the use of online social media are considered in the context that

the analysis are being performed. When using highly anchored attributes such as personal names, a certain degree of confidence may be assumed. Conversely, if seeking to mine opinion or sentiment, it is necessary that any systemic biases in individual observations is identified and adjusted for.

As with any analysis, this thesis has some limitations. These include data uncertainty, availability of suitable reference data and data sparsity. In regards to uncertainty, there are various points of this thesis whereby uncertainty is a factor. This includes the completeness of the raw Twitter data and the attribution of essential identity characteristics to individuals based on their personal names. Regarding data completeness, there are various concerns which cannot easily be addressed. First, it is not possible to determine exactly what proportion of the full Twitter stream is being collected by the filtered streaming API. We know from the work of Morstatter et al. (2013) that when the API is set to return only geotagged Tweets that the sample data are largely consistent with what could have been achieved using the full stream. Further, uncertainty is encountered in the attribution of key identities based on individuals' personal names. As has been discussed, the attributes assigned to each user are based on our collective understanding of people bearing the name. A case in point being the inference of gender. For example, a user with the forename 'Sam' would be predicted male, however, not every user named 'Sam' will be of the male gender. In much the same way as the ecological fallacy, the attributes are indicative of all people called 'Sam' and are not necessarily applicable to each person within that group. In recognition of this uncertainty, it is highlighted throughout that it is the collective behaviour of the population which should be examined and not the activities of individuals.

A further challenge in conducting this thesis has been the availability of consistent and relevant reference data. The issue is evident both in terms of comparing national level static populations and also local level population dynamics. Throughout the course of this thesis, efforts have been made both to identify suitable sources of data and also to provide insight on data quality, applicability and overall usefulness. In particular, a focus has been placed on the UCL Worldnames Database.

While it may be argued that the database remains the best source of global names data, the diversity and completeness of data range considerably. Efforts have been made throughout to quantify and qualify these issues and factor the results into all interpretation and suggestion of implications. In regards to the availability of national level data, the UK is something of an exception with small area demographic data readily available. However, the ease in which UK data may be procured is not the norm. Rather, many countries do not collect or make publicly available such comprehensive data. Seeking to address this, this thesis explores a novel means of comparison based upon the composition of personal names within each country's population. Considering next the issue of local level dynamics data, there are no non-proprietary available sources of such information. It is for this reason that both government, industry and academia are exploring means by which such data may be generated. As noted in Chapter 7, current approaches are often limited to manual or electronic counting at specific locations.

The final significant challenge is related to data sparseness. While Twitter is often grouped under the umbrella of Big Data, when considered in a geographic context, cleaned and split by space and time the data are in fact quite sparse. Given that the study dataset contains the majority of geographically attributed Tweets, the options to enrich the dataset are limited. Therefore, this thesis employs spatial and temporal aggregation as a means to increase observation density. In doing so, it was necessary to strike a balance between the desire to increase granularity while maintaining sufficient data density to identify stable patterns. The challenge of sparseness should also be considered regarding the proportion of the population who use Twitter and submit data. As has been discussed, the sample of Twitter users is neither random nor stratified. Rather, as was demonstrated in Chapter 6, it is strongly skewed towards the younger age groups with a slight male bias. Some might use this point to the detriment of Twitter-based analysis. However, this is somewhat of a glass-half-empty mind-set. In practice, while the data do not depict the full demographic spectrum, the younger cohort are well represented. In fact, in the case of some applications, this bias may be considered advantageous.

8.4 Applications and Implications

While the bulk of this thesis is concerned with whom the data are representative, consideration is also given to potential applications and implications. Considering first the implications, this thesis provides a valuable point of reference for those wishing to employ Twitter data in the study of population stocks and flows. The provision of detailed demographic profiles will enable practitioners to better plan, analyse and interpret Twitter-derived insight. This output alone is a valuable contribution to the literature.

The potential of the demographically attributed Twitter data is showcased in Chapter 7 in which London's four largest airports were examined. The analysis demonstrates both the replication of conventional insight and also, new forms of insight not previously attainable. The significance of the above is that this insight has been generated without direct observation or access to any proprietary or privileged data. Obviously, as has been noted, there are issues of calibration which may need to be addressed on an ad-hoc basis. However, this is a relatively trivial task in relation to procuring data which are not available in the public domain.

There is also a case to be made that the Twitter data may be employed by practitioners for the purpose of exploratory analysis. Consequently, sufficient insight may be drawn and used as justification for the acquisition and analysis of more comprehensive data. Given the challenges associated with the acquisition of proprietary data, this may prove valuable to a broad range of practitioners. This potential and obvious extension to the work would entail development of an end-to-end toolkit which enabled practitioners to lever the potential of Twitter data for the purpose of observing stocks and flows of populations. Such a toolkit could be applied in a broad range of contexts from retail and marketing to security and crime.

As is often the case with new technologies, the pace of developments exceeds the rate at which legislation and guidance can be implemented. Discussed in detail by Zimmer and Proferes (2014), little recognition is given to the ethical implications of the use of new forms of data. From an ethical perspective, the use of Twitter-based analysis is an attractive option. Given that the data collected are within the public

domain and consent, albeit assumed, is given, the data may be employed without the stringent controls associated with personal or proprietary data. Further, the analysis of Twitter, unlike other forms of hyper-local monitoring is far less intrusive. This is especially the case when we consider the alternatives such as cell-tower and WiFi monitoring.

While the overarching focus of this thesis has been the analysis of social media data, it should be remembered that name-based analysis has formed a critical conduit by which the social media and demographic data have been linked. Approaches from the names literature have provided novel means by which the representativeness of the Twitter inventories could be assessed. Consequently, this thesis has made various contributions which could further extend the names literature. Key amongst these is the novel implementation of the Morisita-Horn Index of Overlap, a similarity measure commonly employed in ecology as a means to quantify the similarity between collections of species. Compared to other methods, the Morisita-Horn technique has several advantages. The measure is more easily interpreted using a linear scale between 0 and 1, and also the ability to handle populations of significantly different sample sizes. Such a feature is particularly valuable when the populations being compared are orders of magnitude different in size. Beyond providing a means to standardise between populations, the measure, due to its standardisation, can be employed to draw out more nuanced aspects of population structure.

In many respects, the social sciences are only just beginning to scrape the surface of social media and its potential for the study of human populations. To date, much of the academic interest has been on the extraction of novel insight or the application of cross-disciplinary tools. Yet, so far there been limited attention on the establishment of fundamental principles or truths. Case in point is the ongoing publication of analysis which fails to account for the demographic and geographic biases inherent within the source data. This observation is evidenced by the lack of literature on the factors effecting social media uptake and also the lack of recognition of any biases which are manifest within the data.

Throughout the course of this thesis, numerous applications of social media data

have been discussed in regards to their strengths, weaknesses and opportunities. These applications have spanned a diverse range of topics drawing on techniques and concepts from across the sciences. This observation alone is testament to the potential of social media data to be transformative, providing a new medium for analysis. However, while these data offer unparalleled richness in both space and time, the data lack the formalised recording and structured nature of much traditional data. This lack of formalised recording introduces significant uncertainty with key concerns including: the proportion of the population represented by each platform, Any discrepancies in the geography of each platforms users; and the extent to which the data are fit for purpose. As has been noted, while social media data are often considered under the umbrella of 'Big Data', in reality, once the data are subset they are often particularly sparse. In many respects this contradicts the supposed panacea of Big Data and statements regarding the end of theory in the 'Big Data' era.

In much traditional social science, the biases in the data are both recognised and understood with protocols in place to measure and address the issues which they might present. In regards to social media data, the key challenge is a lack of consistency and transparency in the current measurement and reporting of social media demographics and geography. While the biases are increasingly recognised, there does not yet exists a common interpretation or formalised framework for their quantification. Moving forwards two specific areas of research would benefit from a greater amount of focus. These are the production and implementation of a formal framework for the effective benchmarking and application of each social media platform and, research into the the effective data linkage of multiple social media platforms. In this thesis we take a step forwards in establishing the provenance of the Twitter online social network which may be considered as a blueprint for future benchmarking exercises. The provision of such information in a accessible and standard format should provide academics and industry with a consistent point of reference whereby the aforementioned biases may be recognised and addressed in future analysis. Not only would a standard for benchmarking contribute to more robust use of social media data within the social sciences, the provision of said standards could potentially

improve the integrity of research outputs addressing many of the common criticisms levelled against social media analysis. While this focus is on overall coverage, it should be reiterated that a demographic bias is not in itself a problem, rather it constrains the scope of analysis to specific portion of the population. This portion may in fact be better represented than the population as a whole.

Second, considering the synthesis of data from multiple social media platforms, the nature of social media is such that not all platforms appeal to all parts of the population. For example, it is discussed in Section 3.2 how many social media platforms target specific portions of the population for a range of different purposes. Given that a large proportion of the population use some form of online social media, there exists significant opportunities to develop cross-platform analyses designed to best capture the activities of specific portions of the population. In looking to exploit such data a number of challenges must be addressed. These include the relative size of each platform and the effective standardisation of scale, effective linkage such that biases are understood and controlled rather than magnified and, the effective standardisation of data formats.

In looking to the future of social media and the social sciences we must also remain aware of issues of data ownership, privacy and ethics. Given the range of personal insight potentially available, we must be mindful of the risks of intrusion which may occur inadvertently through the analysis and linkage of new data. A further consideration is the sharing of data. While Twitter data are relatively easily accessible, restrictions are placed on the subsequent sharing and distribution of the data. In terms of disseminating the explicit research outputs this is a potential challenge. It should be noted that this issue is evident with many new forms of data and that there exists a growing interest in developing pathways by which the products of research using such data may be made publicly available. In the interim, it is important that the methods and general findings are made available such that future analysis can gain the maximum advantage.

Beyond the direct impacts which social media data will have within the social sciences, many of the challenges and opportunities raised throughout this thesis may

be observed more broadly across new forms of data. Over the past 10 years there has been an explosion in the rate at which data are generated with data increasingly being considered as a raw data resource and a business asset with increasing reference to a new data economy. In parallel, there has been a dramatic growth in the publication and availability of open data and new forms of data more generally. The availability of these data has seen a shift in who provides access to data, what data are available and subsequently the emergence of many innovative new data-driven products.

While many positives may be observed as a consequence of new data being available, there are also a number of challenges which remain to be faced. Two examples of such are the pressures on organisations to provide open-data products and also, the shift to dependence on commercial data providers. In the case of some open data, increased financial pressure has been placed on the data providers. Where historically there data could be marketed as a product, often these organisations are now required to provide access to a large proportion of the data at zero or limited cost. Second, with the availability of new data an increase in the use of commercially sourced data has been observed such as cell-tower data which is discussed in Section 7.3.1. The use of such data, while often effective, raises questions as to the usefulness of the research beyond academia. Analysis using proprietary data, whether from cell-towers or alternative sources, is intrinsically dependent upon the cooperation of the data provider to allow future access to their data. Another concern relates to the volume of data that are now available. This can be considered under the umbrella of 'Big Data'. The volume and variety of data is such that many of the tools and applications traditionally relied upon in academia are no longer fit for purpose. Increasingly need to employ new bespoke applications. For example, Hadoop, Apache Spark and Google's Big Query. These changes require social scientists, among others, to develop a new quantitative skill set.

A further consideration for the analysis of new forms of data is the issue of data standardisation and the meta-data reporting. Within the context of social media data it has been discussed how demographic biases within the population exist

and may impact upon the conclusions drawn when making data-based inference. It was suggested that consistent assessment of said bias and subsequent reporting could improve data provenance improving both the quality of analysis and also the degree of confidence which may be placed on the data. The creation of such 'meta-data' will be an important step in the transition from the use of traditional population data to administrative and consumer data. A noteworthy mention in the Consumer Data Research Centre (CDRC) based at University College London (see: https://www.cdrc.ac.uk). The CDRC has developed a portal for consumer data, similar to the UK Government's open data portal, providing a new standard for access to consumer data.

In closing, the use of new forms of data and 'Big Data' for the generation or population insight are still very much in their infancy. While the potential of such data is huge, it is, as with the case of social media data, that caution is exercised in the application and interpretation of analysis. Further, given the range of data now available, individuals' privacy should not be sacrificed in the pursuit of greater insight.

In regards to the role which social media data and Twitter more specifically have in the study of demographics, it is clear that considered independently, these data are insufficient to address the full spectrum of requirements in regards to the effective description of the population. However, such is the variety of new data that significant potential exists in the formation of novel data linkages. Through an improved understanding of the relative merits of each data source, and through cooperation with data providers their exists tangible opportunities to incorporate new forms of data in support of existing demographic datasets. In particular, the use of Twitter may provide a means to interpolate between paired static datasets such as the Census of Population Output Areas and Workplace Zones geography.

8.5 Future Work and Closing Remarks

While this thesis has covered significant ground, there are undoubtedly opportunities to extend the work. As a first contribution, it might be possible to develop an end-to-end toolkit that uses Twitter data to monitor population stocks and flows.

The delivery of such a toolkit would assist practitioners in delivering high quality and consistent insight. In addition, it would be beneficial to develop a series of case studies in which such a toolkit was applied to a series of specific scenarios. For example, in determining the optimum allocation of policing resources at a football stadium, the Twitter data could be employed to determine footfall, catchment and potential stress points. Such analysis could draw on historical data as a means to enrich the analysis and provide further insight for planning. This extension, in particular, would benefit from engagement with the relevant agencies and organisations to provide better tailored solutions and address any misconceptions in the flaws of social media-based analysis.

A further extension to this thesis is to explore the potential insight manifest within the Tweets text through the use of text-mining techniques. Text-mining offers an additional avenue of analysis and has the potential to extend the insight beyond what may be possible with alternative New Forms of Data. The potential for such analysis was briefly explored in Chapter 7. Within text mining, there are a number of major themes including topic modelling, sentiment analysis and event detection. Considering first topic modelling, various techniques exist to identify common themes within large collections of text data. Such techniques offer a means by which we may examine place and space providing an improved understanding of who is occupying space and for what purpose. Sentiment analysis provides a means of determining the general mood within a string of text. As discussed in Chapter 7, sentiment analysis may be employed for a range of purposes. In the airport case study, it was proposed that it could facilitate the identification of stress points in space and time, so that the airport authority at Stansted knows when to deploy street entertainers. Event detection is a powerful tool for the identification of unusual or isolated behaviour. While some of the proposed ideas have already been explored, they have not previously accounted for the population bias. Further, text mining offers an additional opportunity to differentiate between individuals based on their activities answering questions such as: How can we differentiate between those who are permanent vs. those who are transient? In the airport case study, how can we fil-

ter out those individuals who are employed on the airport campus versus those who are passing through? In effect, there is an opportunity to further attribute individual users based on their typical activity patterns.

Considering new forms of data more broadly it is evident that each has its own unique strengths and weaknesses. Considered in isolation, each form of data is subject to various limitations including issues of privacy and ethics, demographic representativeness and spatial coverage. Seeking to address these issues may necessitate the development of a framework or standard by which various new forms of data may be linked. One potential area of research involves the calibration of Twitter temporal profiles based on the samples collected using WiFi or other motion sensors. A UK-wide network of WiFi footfall sensors, suitably stratified across business sectors and land use, could provide a means to calibrate Twitter activity patterns more broadly. In effect, addressing the limitations of both the Twitter-derived footfall and WiFi monitoring techniques simultaneously.

This thesis may thus conclude on a positive note. Clearly, the analysis of Twitter data is by no means a perfect substitute for conventional population survey data and other secondary sources. However, it does provide an exciting new frontier in how we may investigate and observe the stocks and flows of populations.

Appendix A

Character Substitutions

Table A.1: List of Special Characters and the substitutions employed.

Character	Substitution	Character	Substitution
Á	A	Ñ	N
À	A	Ň	N
Â	A	Ņ	N
Ä	AE	Ŋ	N
Ã	A	Ø	O
Ă	A	Ó	O
Å	AA	Ò	O
Ā	A	Ô	O
Ą	A	Ö	OE
Ć	C	Õ	O
Ĉ	C	Ő	O
Č	C	Ō	O
Ċ	C	Ŏ	O
Ç	C	Ŕ	R
Đ	D	Ř	R
Ď	D	Ŗ	R
É	E	Ś	S
È	E	Ŝ	S
Ê	E	Š	S
Ë	E	Ş	S
Ě	E	~~Ŧ~~	T
Ė	E	Ť	T
Ē	E	Ţ	T
Ę	E	Ú	U
Ĕ	E	Ù	U
Ĝ	G	Û	U
Ğ	G	Ü	UE
Ġ	G	Ū	U
Ģ	G	Ů	U
~~Ħ~~	H	Ű	U
Ĥ	H	Ų	U
Í	I	Ū	U
Ì	I	Ụ	U
Î	I	Ẃ	W
Ï	I	Ý	Y
Ĩ	I	Ŷ	Y
İ	I	Ÿ	Y
Ī	I	Ź	Z
Į	I	Ž	Z
Ĭ	I	Ż	Z
Ĵ	J	Þ	TH
Ķ	K	Æ	AE
Ł	L	IJ	IJ
Ĺ	L	Œ	OE
Ľ	L	ß	SS
Ļ	L		
Ŀ	L		
Ń	N		

Appendix B

Audit of the UCL Worldnames Database

B.1 Introduction

The Worldnames Database is a collection of over 20 publicly available individual-level inventories of personal names. The data are largly sourced from electoral roll and telephone directory datasets. Combined, the data are representative of some two billion of the Earth's population. The objective here is to audit the data such that future work, which relies upon the data as an accurate reference, may be performed in confidence. As part of the audit process, meta-data will be generated pertaining to data source, time of collection, and representative capability. The motivation for this audit is to determine the provenance of the component datasets, as this is not always immediately clear or well documented.

B.2 Method

The approach to the audit is as follows. for each of the Worldnames countries, a series of tests will be performed. The results of the test and data sources used in the case of each validation will then be reported. The representativeness of the data will be tested on a country-by-country basis. In each case, the most common names as recorded by the Worldnames datasets will be compared against any alternative sources available.

Two tests will be used to measure how representative the data are. In the first

case, the degree of overlap between the alternative data sources will be measured. Overlap will be determined on the number if names shared in the top 10, 20, 50 and 100 where sufficient data are available. The second test will use the Spearman's rank statistic to measure the rank correlation between the most top 10, 20, 50 and 100 names where available. Where possible the validation data will be sourced from national statistic agencies, however, this is not always possible. The key premise for each test is that the frequency distribution of names is generally very distinct within the most commonly occurring names.

B.3 Results

Argentina

The Worldnames data for Argentina are sourced from the Electoral Register though no year is available. The data contain records for 24.9 million individuals, representing 60% of the 2013 population. Three sets of validation data were identified, neither of which specified a source or year. The data are from Forebears.io, Wikipedia (allegedly sourced from the Worldnames Database) and Behindthename.com.

Table B.1: Validation of Argentine names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	100	80	84	96	0.648	0.903	0.92	0.897
p-value					0.049	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table B.2: Validation of Argentina names versus Behindthename.com and Wikipedia data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	70	68	79	0.55	0.736	0.788	0.702
p-value					0.133	0.004	5.50E-07	< 2.2e-16

- http://forebears.io/argentina
- http://surnames.behindthename.com/top/lists/argentina/2006
- http://en.wikipedia.org/wiki/List_of_most_common_surnames_in_South _America

The overlap in names is fairly high for all four samples for each validation dataset. Overlap ranged from 80-100% in the case of the Forebears data and 68-90% in the case of the BehindTheName.com data. Further, the data appear to exhibit strong positive relationships in all but the top-10 sample sets. The analysis suggests that the Worldnames data for Argentina are a good representation of the true population.

Australia

The Worldnames data for Australia are sourced from the 2002 Telephone Directory. The data contain records for 7.8 million individuals, which represents 34% of the population in 2013. The directory has been cleaned to keep only residential addresses. The data, used for validation are sourced from IP Australia, a portion of the Australian Government devoted to intellectual property.

Table B.3: Validation of Australia names versus IP Australia data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	100	98	98	0.983	0.971	0.996	0.997
p-value					5.0E-05	6.6E-06	< 2.2e-16	< 2.2e-16

- http://pericles.ipaustralia.gov.au/atmoss/Falcon_Search_Tools.Main?pSearch=Surname&pWord=*&pCommand=Search

The overlap between names is very high in each of the four selections with 90-100% or names common to both populations. Further, there is a high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames data for Australia are a very good representation of the true population.

Austria

The Worldnames data for Austria are sourced from the 2007 Telephone Directory. The data contain records for 2.9 million individuals, which represents 35% of the 2013 population. The directory has been cleaned to keep only residential addresses. Two sets of validation data were identified. The data are from Forebears.io and the Wiener Sprachblätter newspaper (a periodical focused on language). The Forebears data claims to be from 2014 though lists no source. The Sprachblätter data claims to be from the 2005 telephone directory.

Table B.4: Validation of Austria names versus Sprachblätter data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	100	100	97.5		0.903	0.976	0.981	
p-value					0.001	6.60E-06	< 2.2e-16	

Table B.5: Validation of Austria names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	95	88	88	0.667	0.895	0.963	0.884
p-value					0.059	< 2.2e-16	< 2.2e-16	3.90E-30

- http://www.sprache-werner.info/WSBAlles_Gruber_in_Oestereich.10287.html
- http://forebears.io/austria

The overlap between names is very high in each of the four selections with 90-100% and 88-95% of names common to both populations respectively. Further, there is a high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames data for Austria are a very good representation of the true population.

Belgium

The Worldnames data for Belgium are sourced from the 2007 Telephone Directory. The data contain records for 3.52 million individuals, which represents 31% of the 2013 population. The directory in questions has been cleaned to keep only residential addresses. The data, used for validation are sourced from Behindthename.com which attribution suggests are from the Belgium Statistics Authority. It is believed the data are from 2001.

Table B.6: Validation of Belgium names versus Behindthename.com data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	90	92	90	0.967	0.992	0.965	0.978
p-value					2E-04	1.00E-05	< 2.2e-16	< 2.2e-16

- http://surnames.behindthename.com/top/lists/belgium/2001

The overlap between names is very high in each of the four selections with 90% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames data for Belgium are a very good representation of the true population.

Brazil

The Worldnames data for Brazil is the Telephone Directory for an unknown year. The data contain records for 0.28 million individuals, which represents 0.13% of the 2013 population. The data have been cleaned to remove businesses addressed and keep residential addresses. It is important to note that the data represent only 5 cities: Sergipe, Salvador, Curitiba, Vitoria and Alagos. The data used for the validation are sourced from Forebears.io, which gives the impression the data are from 2014.

Table B.7: Validation of Brazil names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	60	60	68	72	0.886	0.552	0.561	0.552
p-value					0.033	0.067	0.001	8.20E-07

- http://forebears.io/brazil

The overlap between names is low in each of the four selections ranging from 60-72%. Further, there is a low degree of correlation between the two sets of ranked data. Whilst the source of the validation dataset is uncertain, it suggests that the Worldnames data for Brazil are a poor representation of the true national population. The poor performance is almost certainly associated with the limited geographic coverage of the data.

Bulgaria

The Worldnames data for Bulgaria is the Telephone Directory for an unknown year. The data contain records for 0.74 million individuals, which represents 10% of the 2013 population. The data used for the validation are sourced from Forebears.io, which gives the impression the data are from 2014. The analysis is performed using both raw and gender standardised surnames.

Table B.8: Validation of Bulgarian names versus Forebears.io Unstandardised data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	70	50	58	50	0.571	0.81818	0.922	0.801
p-value					0.2	0.007	2.40E-07	< 2.2e-16

Table B.9: Validation of Bulgarian names versus Forebears.io Standardised data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	50	55	52	55	0.9	0.961	0.852	0.579
p-value					0.083	2.40E-06	3.40E-08	3.70E-06

- http://forebears.io/bulgaria

In the case of both comparisons the level of agreement between the two datasets is fairly poor. Notably, the level of agreement is the most common names varies between 50-70% and 50-55% for the unstandardised and standardised data respectively. Whilst the sources of the validation dataset is uncertain, the results suggest that the Worldnames data for Bulgaria are a poor representation of the true national population.

Canada

The Worldnames data for Canada is sourced from the Telephone Directory. The data contain records for 4.75 million individuals, which represents 14% of the 2013 population. In this case, the year of collection is unknown. The directory has been cleaned to remove businesses addressed and keep residential addresses. The data used for reference are sourced from the Wikipedia, however, are attributed to Statistics Canada and appears to be correct as of 2006. The data available only referred to the top 20 most common names.

Table B.10: Validation of Canadian names versus Wikipedia data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	10	20			n/a	-0.224		
p-value					n/a	0.537		

Table B.11: Validation of Canadian names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	20	50	46	55	1	0.079	0.202	0.309
p-value					1	0.838	0.355	0.022

- http://en.wikipedia.org/w/index.php?title=List_of_most_common_surnames_in_North_America&oldid=367685342
- http://forebears.io/canada

The overlap between names is very low in both cases with only 50% overlap in the top 20 and 10% in the top 10. Further, there is a very poor degree of correlation between the two sets of ranked data, though these are not statistically significant. The names that were not matched in the top 20 were Lee, Lam, Roy, Tremblay, Lee, Gagnon, Wilson, Williams, Cote and Chan. The high proportion of ethnic names suggests systematic under-reporting. The analysis suggests that the Worldnames data for Canada are a poor representation of the true population.

Denmark

The Worldnames data for Denmark are sourced from the Telephone Directory. The data contain records for 3.29 million individuals, which represents 59% of the 2013 population. In this case, the year of collection is unknown. The directory has been cleaned to keep only residential addresses. The data used for reference are sourced from the 'My Danish Roots Website' and appear to be correct as of 2014. This data was validated against data published by the Danish Statistics Authority that provide the top 20 names.

Table B.12: Validation of Danish names versus MyDanishRoutes.com data

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	100	95	98	95	1	1	0.988	0.988
p-value					< 2.2e-16	8.4E-06	1.5E-39	5.2E-77

- http://www.dst.dk/da/Statistik/emner/navne/navne-i-hele-befolkningen
- http://www.mydanishroots.com/surnames-meaning-and-origin/the-100-most-common-surnames-in-denmark.html

The overlap between names is very high in each of the four selections with 95% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames data for Denmark are a very good representation of the true population.

France

The Worldnames data for France are sourced from the Telephone Directory. The data contain records for 20.36 million individuals, which represents 31% of the 2013 population. In this case, the year of collection is unknown. The directory has been cleaned to keep only residential addresses. The data used for reference are sourced from Le Journal des Femmes. The data are based on data for 11 million users of Copains d'avant, a website similar to Friends Reunited.

Table B.13: Validation of French names versus Le Journal des Femmes data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	95	94	93	0.533	0.909	0.918	0.942
p-value					0.148	< 2.2e-16	< 2.2e-16	< 2.2e-16

- http://www.journaldesfemmes.com/nom-de-famille/noms/1/2/france.shtml

The overlap between names is very high in each of the four selections with 90% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data though the correlation in the top 10 is not statistically significant. The analysis suggests that the Worldnames data for France are a very good representation of the true population.

Germany

The Worldnames data for Germany are sourced from the 2007 Telephone Directory. The data contain records for 32.54 million individuals, which represents 40% of the 2013 population. The directory has been cleaned to keep only residential addresses. The data used for reference are sourced from Wikipedia.

Table B.14: Validation of German names versus Wikipedia data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	100	95	98	97	1	1	0.999	0.999
p-value					<2.2e-16	8.4E-06	<2.2e-16	<2.2e-16

- http://de.wikipedia.org/wiki/Liste_der_h%C3%A4ufigsten _Familiennamen_in_Deutschland

The overlap between names is very high in each of the four selections with 95% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data though the correlation in the top 10 is not statistically significant. The analysis suggests that the Worldnames data for Germany are a very good representation of the true population.

Hungary

The Worldnames data for Hungary are sourced from the Telephone Directory. The data contain records for 0.28 million individuals, which represents 3% of the 2013 population. In this case, the year of collection is unknown. The directory has been cleaned to remove businesses addressed and keep residential addresses. The data used for reference are sourced from the Hungarian 'Administrative and Public Services Central Office' and appears to be correct as of 2011.

Table B.15: Validation of Hungary names versus the Hungarian Administrative and Public Services data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	70	78	71	0.433	0.776	0.735	0.805
p-value					0.25	0.002	4.90E-07	2.80E-17

- http://infolux.uni.lu/familiennamen/grundstrukturen/#haeufigste

The overlap between names is fairly high in each of the four selections with 70% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. However, the correlation is considered statistically insignificant when only the top 10 names are compared. The analysis suggests that the Worldnames data for Hungary are a good representation of the true population.

India

The Worldnames data for India are from the India Telephone directory for 2005/2006. The data contain records for 3.73 million individuals, which represents 0.2% of the 2013 population. In the case of the Indian data, it is important to note that it represents only 3 cities: New Delhi, Mumbai and Hyperabad. The data have been cleaned to keep only residential addresses. The data used for reference are sourced from Forebears.io. The data claim to be from 2014 though no formal source or year is recorded.

Table B.16: Validation of Indian names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	10	35	26	29	n/a	0.143	0.571	0.562
p-value					n/a	0.783	0.045	0.002

Table B.17: Validation of Indian names versus Low Chen Australia data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	10	15	20	23	n/a	1	0.345	0.403
p-value					n/a	0.333	0.331	0.058

- http://forebears.io/india
- http://www.lowchensaustralia.com/names/popular-indian-names.htm

Based on the forebears.io data, the overlap between names is very low in each of the four selections with less that 30% of names hared between the two populations. Further, there is a very low degree of correlation between the two sets of ranked data. From the alternative source, the overlap between names in the second reference is also very low in each of the four selections with less that 23% of names hared between the two populations. Further, there are no significant correlations between either of the two sets of ranked data. Whilst the sources of the validation datasets are uncertain, the results suggest that the Worldnames data for India are a very poor representation of the true national population.

Ireland

The Worldnames data for Ireland are from an unknown source in an unknown year. The data contain records for 2.92 million individuals which suggests they may be from a telephone directory or similar, which represents 63% of the 2013 population. The data have been cleaned to keep only residential addresses. The data used for the validation are sourced from Forebears.io that gives the impression the data are from 2014. This information is not verified, however.

Table B.18: Validation of Irish names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	90	88	89	0.833	0.913	0.943	0.883
p-value					0.01	< 2.2e-16	< 2.2e-16	< 2.2e-16

- http://forebears.io/ireland

The overlap between names is very high in each of the four selections with 88% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. Whilst the source of the validation dataset is uncertain, it still suggests that the Worldnames data for Ireland are a good representation of the true population.

Italy

The Worldnames data for Italy are from the Italian Telephone directory for 2005/2006. The data contain records for 15.94 million individuals, which represents 27% of the 2013 population. The data have been cleaned to remove businesses addressed and keep residential addresses. The data used for reference are sourced from Cognomix which is an Italian genealogical website. The data were published in 2010 though no formal source or year is recorded.

Table B.19: Validation of Italian names versus Cognomix data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	80	100	88	93	0.929	0.925	0.968	0.955
p-value					0.002	4.40E-06	< 2.2e-16	< 2.2e-16

- http://www.cognomix.it/top100_cognomi_italia.php

The overlap between names is very high in each of the four selections with 80% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. Whilst the source of the validation is uncertain, it still suggests that the Worldnames data for Italy are a good representation of the true population.

Japan

The Worldnames data for Japan are from an unknown source and the year of collection is unknown. The data contain records for 44.99 million individuals, which represents 35% of the 2013 population. The data have been cleaned to remove businesses addressed and keep residential addresses. The data used for reference are sourced from the 'The National Same Family Name Investigation' by the Meiji Yasuda Life Insurance Company and appears to be correct as of 2008. The data are based on customer details for 6.1 million individuals.

Table B.20: Validation of Japanese names versus Wikipedia data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	95	92	95	0.9	0.809	0.933	0.857
p-value					0.002	3.10E-05	4.10E-21	1.60E-28

- http://en.wikipedia.org/wiki/List_of_most_common_surnames_in_Asia#Japan

The overlap between names is very high in each of the four selections with 90% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. However, the correlation is considered statistically insignificant when only the top 10 names are compared. The analysis suggests that the Worldnames data for Japan are a good representation of the true population.

Luxembourg

The Worldnames data for Luxembourg are sourced from the national Telephone Directory. The data contain records for 0.12 million individuals, which represents 23% of the 2013 population. In this case, the year of collection is unknown. The directory has been cleaned to remove businesses addressed and keep residential addresses. The data used for reference are sourced from data shared by the University of Luxembourg.

Table B.21: Validation of Luxembourg names versus infolux data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	70	75	72	68	1	0.993	0.84376	0.90297
p-value					4E-04	< 2.2e-16	5.10E-08	< 2.2e-16

- http://infolux.uni.lu/familiennamen/grundstrukturen/#haeufigste

The overlap between names is fairly high in each of the four selections with 70% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames data for Luxembourg are a fair representation of the true population.

Malta

The Worldnames data for Malta are the Electoral roll in an unknown year. The data contain records for 0.33 million individuals, which represents 79% of the 2013 population. The data have been cleaned to remove businesses addressed and keep residential addresses. The data used for the validation are sourced from Forebears.io which gives the impression the data are from 2014. This information is not verified, however.

Table B.22: Validation of Malta names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	100	98	93	0.983	0.983	0.974	0.987
p-value					< 2.2e-16	< 2.2e-16	6.5E-06	5.0E-05

- http://forebears.io/malta

The overlap between names is very high in each of the four selections with 90% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. Whilst the source of the data is uncertain, it still suggests that the Worldnames data for Malta are a very good representation of the true population.

Netherlands

The Worldnames data for the Netherlands are the Telephone Directory for an unknown year. The data contain records for 4.87 million individuals, which represents 29% of the 2013 population. The data have been cleaned to remove businesses addressed and keep residential addresses. The data used for the validation are sourced from Forebears.io which gives the impression the data are from 2014.

Table B.23: Validation of Malta names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	80	90	82	86	0.762	0.944	0.953	0.937
p-value					0.037	2.20E-06	< 2.2e-16	< 2.2e-16

- http://forebears.io/netherlands

The overlap between names is very high in each of the four selections with 80% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. Whilst the source of the validation is uncertain, it still suggests that the Worldnames data for the Netherlands are a good representation of the true population.

New Zealand

The Worldnames data for New Zealand are recorded as being from the Electoral Roll, though the specific year is unclear. The data contain records for 2.84 million individuals, which represents 63% of the 2013 population. However, on inspection, the data are clearly a synthesis of datasets spanning almost 100 years. The earliest records pertain to 1897 and the most recent to 1992. This theory has been confirmed though investigation of a specific user, a Mr Austin Edgar Andrews, who is recorded in the Worldnames for the period 1912 to 1917. The individual was also matched to a news report about the history of the Thomas's department store (http://www.thomass.co.nz/page/about-us.aspx) and referenced by mentions of his wife. A further observation of the data was that multiple individuals were recorded at the same address over an extended period. In one instance, 4 different individuals from 3 different families were recorded at one address (26 Francis Street, Blenhiem). When a subset was created for only the most recent residents (1992) only 122297 individuals remained of the previous 2.84 million. The data used for reference are sourced from the New Zealand Department for Internal Affairs and contain data for 2015. In this case, only the top 20 names are made available.

Table B.24: Validation of Malta names versus Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	70	65			0.964	0.989		
p-value					0.003	< 2.2e-16		

- http://www.dia.govt.nz/press.nsf/d77da9b523f12931cc256ac5000d19b6/738cf2a0e2ef4d2ecc257d340013fc3b!OpenDocument

The overlap between names is medium in each of the two selections with 65-70% of names shared between the two populations. There is a very high degree of correlation between the two sets of ranked data. It should be noted that the Worldnames data appears to under-represent ethnic minorities. From the top 20, the 7 names not evident in the Worldnames data were Singh, Wang, Li, Chen, Pa-

tel, Zhang, and Kumar. The analysis suggests that the Worldnames data for New Zealand are a poor representation of the true population.

Norway

The Worldnames data for Norway are sourced from the Telephone Directory. The data contain records for 3.58 million individuals, which represents 70% of the 2013 population. In this case, the year of collection is unknown. The directory has been cleaned to keep only residential addresses. The data used for reference are sourced from the Norwegian National Statistics Authority and pertain to 2014.

Table B.25: Validation of Norwegian names versus the Norwegian National Statistics Authority data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	100	95	94	97	0.976	0.996	0.997	0.982
p-value					< 2.2e-16	8.4E-06	< 2.2e-16	< 2.2e-16

- http://www.ssb.no/en/befolkning/statistikker/navn/aar/2015-01-27 ?fane=tabell&sort=nummer&tabell=216083

The overlap between names is very high in each of the four selections with 94-100% of names common to both two populations. Further, there is a very high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames data for Norway are a very good representation of the true population.

Poland

The Worldnames data for Poland are sourced from the national Telephone Directory. The data contain records for 8.17 million individuals, which represents 21% of the 2013 population. In this case, the year of collection is unknown. The directory has been cleaned to keep only residential addresses. The data used for reference are sourced from the Polish Ministry for Interior and contain data for 2014. The data are available split by gender and as such must be combined. In this, it is important to note that the orders differ potentially meaning a count of either the feminine, or masculine form of the name is not available. In this situation, the name is omitted.

Table B.26: Validation of Poland names versus the Polish Interior Ministry data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	50	60	68	62	0.2	0.378	0.543	0.622
p-value					0.783	0.227	0.001	1.40E-07

- https://www.msw.gov.pl/pl/aktualnosci/12891,100-najpopularniejszych-polskich-nazwisk.html?search=96137220

The overlap between names is quite poor in each of the four selections with no observation over 70% shared between the two populations. Further, there is a very low degree of correlation between the two sets of ranked data. Furthermore, the correlations in the case of both the top 10 and 20 are considered insignificant. The analysis suggests that the Worldnames data for Poland are a poor representation of the true population.

Serbia

The Worldnames data for Serbia are the Yugoslavia Telephone Directory for an unknown year. The data are a subset of this for Serbia. The year is estimated as being between 1992 and 2006 based on the breakup of Serbia and Montenegro. The data appear to represent key cities, which include Pristina, Belgrade, Uzice, Kragujavac, Nis and Novi Sad. The dataset also include records for Podgorica, which is now the capital of Montenegro. The data contain records for 1.59 million individuals, which represents 22% of the 2013 population. The data have been cleaned to keep only residential addresses. The data used for the validation are sourced from Forebears.io, which gives the impression the data are from 2014. This information is not verified, however. It should be noted that the Worldnames uses the Æ character in place of Ć. For the comparison the Æ character is pre-converted to Ć.

Table B.27: Validation of Serbian names versus the Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	75	84	89	0.767	0.882	0.752	0.889
p-value					0.021	< 2.2e-16	1.10E-07	< 2.2e-16

- http://forebears.io/serbia

The overlap between names is fairly high in each of the four selections with 75-90% of names shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. Whilst the source of the validation is uncertain, it still suggests that the Worldnames data for Serbia are a good representation of the true population. However, it would appear that the character substitution would be necessary.

Slovenia

The Worldnames data for Slovenia are the Telephone Directory for 2006. The data contain records for 0.35 million individuals, which represents 23% of the 2013 population. The data have been cleaned to keep only residential addresses. The data used for the validation are sourced from Forebears.io that gives the Slovenian Statistics Authority. The data include the 200 most common surnames. It is evident that there issues with the use of specific characters. Specifically, the È character is used in the telephone directory where it should be a Ć. Once changed, the results are drastically improved.

Table B.28: Validation of Slovenia names versus the Slovenian Statistics Authority data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	80	80	85	0.967	0.95	0.958	0.886
p-value					2E-04	< 2.2e-16	< 2.2e-16	< 2.2e-16

- http://www.stat.si/ImenaRojstva/sl/FamiliyNames/ExpandFamiliyNames

The overlap between names is high in each of the four selections ranging from 80-90% between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames for Slovenia are a very good representation of the true population. However, for this to be the case the character issue must be resolved.

Spain

The Worldnames data for Spain are sourced from the 2004 Telephone Directory. The data contain records for 10.4 million individuals, which represents 22% of the 2013 population. The directory has been cleaned to remove businesses addressed and keep residential addresses. The data used for reference are sourced from the Spanish National Statistics Authority and contain data for 2014.

Table B.29: Validation of Spanish names versus the Spanish National Statistics Authority data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	90	95	86	80	0.933	0.988	0.987	0.972
p-value					7E-04	8.40E-06	< 2.2e-16	< 2.2e-16

- http://www.ine.es/apellidos/formGeneralresult.do?vista=1

The overlap between names is very high in each of the four selections with 80% or higher shared between the two populations. Further, there is a high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames for Spain are a very good representation of the true population.

Sweden

The Worldnames data for Sweden are sourced from the 2006 national Telephone Directory. The data contain records for 0.79 million individuals, which represents 8% of the 2013 population. The directory has been cleaned to remove businesses addressed and keep residential addresses. The data used for reference are sourced from Statistics Sweden and contain data for 2014.

Table B.30: Validation of Swedish names versus the Statistics Sweden data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	100	85	86	74	0.99	0.99	0.95	0.96
p-value					< 2.2e-16	1.1E-05	< 2.2e-16	< 2.2e-16

- http://www.scb.se/sv_/Hitta-statistik/Statistik-efter-amne/Befolkning/Amnesovergripande-statistik/Namnstatistik/30898/30905/Samtliga-folkbokforda–Efternamn-topplistor/31063/

The overlap between names is very high in each of the four selections with 85% or higher shared between the two populations. Further, there is a very high degree of correlation between the two sets of ranked data. The analysis suggests that the Worldnames data for Sweden are a very good representation of the true population.

Switzerland

The Worldnames data for Switzerland are the Telephone Directory for an unknown year. The data contain records for 1.87 million individuals, which represents 23% of the 2013 population. The data have been cleaned to remove businesses addressed and keep residential addresses. The data used for the validation are sourced from Forebears.io that gives the impression the data are from 2014.

Table B.31: Validation of Switzerland names versus the Forebears.io data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	60	50	60	58	0.714	0.6	0.407	0.407
p-value					0.136	0.073	0.026	0.002

- http://forebears.io/switzerland

The overlap between names is very low in each of the four selections ranging from 50-60% between the two populations. Further, there is a very low degree of correlation between the two sets of ranked data. Whilst the source of the validation is uncertain, it suggests that the Worldnames data for Switzerland are a poor representation of the true population.

UK

The Worldnames data for the UK are an enhanced version of the 2011 public electoral register. The data, provided by CACI Ltd, are enhanced to account for those individuals who have chosen to opt out of inclusion in the public register. The data contain records for 54.29 million individuals, which represents 85% of the 2013 population. The data have been cleaned to remove businesses addressed and keep residential addresses. The data used for the validation are sourced from Forebears.io, which gives the impression the data are from 1991. Further, it appears that the data do not include Scotland and Northern Ireland. This information is not verified, however. Surname Frequency data are not published by the national statistics agency.

Table B.32: Validation of UK names versus the Behindthename.com data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	100	100	90	94	0.976	0.982	0.96	0.963
p-value					< 2.2e-16	6.5E-06	< 2.2e-16	< 2.2e-16

- http://surnames.behindthename.com/top/lists/england-wales/1991

The overlap between names is very high in each of the four selections ranging from 90-100% between the two populations. Further, there is a very high degree of correlation between the two sets of ranked data. Whilst the source of the validation data is uncertain, it suggests that the Worldnames data for the UK are a very good representation of the true population.

USA

The Worldnames data for the USA are sourced from the national Telephone Directory. The data contain records for 78.46 million individuals, which represents 25% of the 2013 population. In this case, the year of collection is unknown. The directory has been cleaned to remove businesses addressed and keep residential addresses. The data used for reference are sourced from US Census and contain data for 2000. The data contain all surnames and counts for names with greater than 100 owners.

Table B.33: Validation of USA names versus the US Census data.

	Overlap (%)				**Correlation (rho)**			
	Top 10	Top 20	Top 50	Top 100	Top 10	Top 20	Top 50	Top 100
Value	70	75	80	82	0.964	-0.261	0.511	0.859
p-value					0.003	0.35	0.001	< 2.2e-16

- http://www.census.gov/topics/population/genealogy/data/2000_surnames.html

The overlap between names is medium in each of the four selections with 70% or higher shared between the two populations. However, the correlation analysis suggests that: the internal ranks do not conform particularly well. The analysis suggests that the Worldnames data for the USA are a good representation of the true population.

Bibliography

Abbas, J., Ojo, A., and Orange, S. (2009). Geodemographics–a tool for health intelligence? Public Health, 123(1):e35–e39.

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2011). Predicting flu trends using Twitter data. In Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on, pages 702–707. IEEE.

Adnan, M., Leak, A., and Longley, P. (2014). A geocomputational analysis of Twitter activity around different world cities. Geo-spatial Information Science, 17(3):145–152.

Adnan, M., Longley, P. A., Singleton, A. D., and Brunsdon, C. (2010). Towards real-time geodemographics: Clustering algorithm performance for large multidimensional spatial databases. Transactions in GIS, 14(3):283–297.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired magazine, 16(7):16–07.

Aspinall, P. J. (2012). Answer formats in British census and survey ethnicity questions: Does open response better capture 'superdiversity'? Sociology, 46(2):354–364.

Austin, P. C. and Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. Journal of Clinical Epidemiology, 68(6):627–636.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B.,

and Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. Psychological Science, 21(3):372–374.

BBC (2014a). Social media 'at least half' of calls passed to front-line police. http://www.bbc.co.uk/news/uk-27949674.

BBC (2014b). Twitter 'misplaces' Taliban official. BBC. http://www.bbc.co.uk/news/world-asia-29490997.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. (2013). Human mobility characterization from cellular network data. Communications of the ACM, 56(1):74–82.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. In Proceedings of the 16th International Conference on Multimodal Interaction, pages 427–434. ACM.

Bollier, D. and Firestone, C. M. (2010). The promise and peril of Big Data. Aspen Institute, Communications and Society Program Washington, DC.

Botta, F., Moat, H. S., and Preis, T. (2015). Quantifying crowd size with mobile phone and Twitter data. Royal Society Open Science, 2(5):150162.

Boulos, M. N. K., Hetherington, L., and Wheeler, S. (2007). Second Life: An overview of the potential of 3-D virtual worlds in medical and health education. Health Information & Libraries Journal, 24(4):233–245.

Bowers, K. (1999). Exploring links between crime and disadvantage in north-west england: An analysis using geographical information systems. International Journal of Geographical Information Science, 13(2):159–184.

Briggs, D. and Baker, S. A. (2012). From the criminal crowd to the 'mediated crowd': The impact of social media on the 2011 English riots. Safer Communities, 11(1):40–49.

Burnap, P., Gibson, R., Sloan, L., Southern, R., and Williams, M. (2016). 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. Electoral Studies, 41:230–233.

CAA (2011). CAA market power assessments catchment area analysis working paper.

CAA (2015). CAA passenger survey report 2015.

Casilli, A. A. and Tubaro, P. (2011). Why net censorship in times of political unrest results in more violent uprisings: A social simulation experiment on the UK riots. Available at SSRN 1909467.

Chainey, S. and Tompson, L. (2012). Engagement, empowerment and transparency: Publishing crime statistics using online crime mapping. Policing, page pas006.

Chandra, K. (2012). Constructivist theories of ethnic politics. Oxford University Press.

Cheshire, J., Mateos, P., and Longley, P. A. (2011). Delineating Europe's cultural regions: Population structure and surname clustering. Human Biology, 83(5):573–598.

Cheshire, J. A. and Longley, P. A. (2012). Identifying spatial concentrations of surnames. International Journal of Geographical Information Science, 26(2):309–325.

Cheshire, J. A., Longley, P. A., and Singleton, A. D. (2010). The surname regions of Great Britain. Journal of Maps, 6(1):401–409.

Cheshire, J. A., Longley, P. A., Yano, K., and Nakaya, T. (2014). Japanese surname regions. Papers in Regional Science, 93(3):539–555.

Cockings, S., Martin, D., and Harfoot, A. (2015). A classification of workplace zones for England and Wales (COWZ-EW). Technical report, Tech. rep., University of Southampton, Southampton, UK.

Cornish, D. B. and Clarke, R. V. (1987). Understanding crime displacement: An application of rational choice theory. Criminology, 25(4):933–948.

Crow, G., Wiles, R., Heath, S., and Charles, V. (2006). Research ethics and data quality: The implications of informed consent. International Journal of Social Research Methodology, 9(2):83–95.

Crump, J. (2011). What are the police doing on Twitter? social media, the police and the public. Policy & Internet, 3(4):1–27.

Csikszentmihalyi, M. and Hunter, J. (2003). Happiness in everyday life: The uses of experience sampling. Journal of happiness studies, 4(2):185–199.

Darwin, G. H. (1875). Marriages between first cousins in England and their effects. Journal of the Statistical Society of London, 38(2):153–184.

De Smith, M. J., Goodchild, M. F., and Longley, P. (2011). Geospatial analysis: A comprehensive guide to principles, techniques and software tools. Matador.

Dewdney, J. (1981). The UK Census of Population, 1981. Geo Books, Norwich.

Diaspora (2016). About - the Diaspora project. https://diasporafoundation.org/about.

Ellison, N. B. et al. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1):210–230.

eMarketer.com (2015). Social network ad spending to hit $23.68 billion worldwide in 2015 - eMarketer. http://www.emarketer.com/Article/Social-Network-Ad-Spending-Hit-2368-Billion-Worldwide-2015/1012357.

ESOMAR (2011). Esomar Guideline on social media research. Technical report, European Society for Opinion and Marketing Research.

Facebook (2015). Facebook company info.

Fan, W., Wallace, L., Rich, S., and Zhang, Z. (2006). Tapping the power of text mining. Communications of the ACM, 49(9):76–82.

Fellows, I. (2014). wordcloud: Word Clouds. R package version 2.5.

Frederickson, H. G. and LaPorte, T. R. (2002). Airport security, high reliability, and the problem of rationality. Public Administration Review, 62(s1):33–43.

GADM (2012). Gadm database of global administrative areas, version 2.0.

Gale, C. G. (2014). Creating an open geodemographic classification using the UK Census of the Population. PhD thesis, UCL (University College London).

Gale, C. G. and Longley, P. A. (2013). Temporal uncertainty in a small area open geodemographic classification. Transactions in GIS, 17(4):563–588.

Gallagher, A. C. and Chen, T. (2008). Estimating age, gender, and identity using first name priors. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE.

Genç, K. (2014). When one door closes: Turkey turns east or west? Index on Censorship, 43(2):102–106.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232):1012–1014.

GlobalWebIndex (2015). Internet users have average of 5.54 social media accounts.

Goldin, C. and Shim, M. (2004). Making a name: Women's surnames at marriage and beyond. The Journal of Economic Perspectives, 18(2):143–160.

Goss, J. (1995). We know who you are and we know where you live: The instrumental rationality of geodemographic systems. Economic Geography, pages 171–198.

Graham, M., Hale, S. A., and Gaffney, D. (2014). Where in the world are you? geolocation and language identification in Twitter. The Professional Geographer, 66(4):568–578.

Grotevant, H. D. (1992). Assigned and chosen identity components: A process perspective on their integration, chapter 5, pages 73–90. Sage Publications, Inc.

Guppy, H. B. (1890). Homes of family names in Great Britain. London, Harrison & Sons.

Harrell, F. E. (2001). Regression Modeling Strategies. Springer Series in Statistics. Springer New York, New York, NY.

Harris, R., Sleight, P., and Webber, R. (2005). Geodemographics, GIS and neighbourhood targeting, volume 7. John Wiley and Sons.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41(3):260–271.

Hawn, C. (2009). Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. Health Affairs, 28(2):361–368.

Hermida, A., Fletcher, F., Korell, D., and Logan, D. (2012). Share, like, recommend: Decoding the social media news consumer. Journalism Studies, 13(5-6):815–824.

Hogg, M. A., Terry, D. J., and White, K. M. (1995). A tale of two theories: A critical comparison of identity theory with social identity theory. Social Psychology Quarterly, pages 255–269.

Holloway, S. M. and Sofaer, J. (1989). Coefficients of relationship by isonymy within and between the regions of Scotland. Human Biology, pages 87–97.

Horn, H. S. (1966). Measurement of 'overlap' in comparative ecological studies. American Naturalist, pages 419–424.

ICO (2016). Wi-fi location analytics.

International Telecommunication Union (2013a). Internet users (per 100 people). Technical report, International Telecommunication Union.

International Telecommunication Union (2013b). Mobile cellular subscriptions (per 100 people). Technical report, International Telecommunication Union.

Ipsos Mori (2015). How Britain voted in 2015: The 2015 Election – Who voted for whom?

Jan, M. T., Abdullah, K., and Momen, A. (2015). Factors influencing the adoption of social networking sites: Malaysian muslim users perspective. Journal of Economics, Business and Management, 3(2):267–270.

Jenkins, R. (2014). Social identity. Routledge.

Jobling, M. A. (2001). In the name of the father: Surnames and genetics. Trends in Genetics, 17(6):353–357.

Jobling, M. A. and Tyler-Smith, C. (2003). The human Y chromosome: An evolutionary marker comes of age. Nature Reviews Genetics, 4(8):598–612.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivariate Behavioral Research, 35(1):1–19.

Junglas, I. A. and Watson, R. T. (2008). Location-based services. Communications of the ACM, 51(3):65–69.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., and Newth, D. (2015). Understanding human mobility from Twitter. PloS one, 10(7):e0131469.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. Business Horizons, 53(1):59–68.

Kelsey, D. and Bennett, L. (2014). Discipline and resistance on social media: Discourse, power and context in the Paul Chambers 'Twitter Joke Trial'. Discourse, Context & Media, 3:37–45.

Khondker, H. H. (2011). Role of the new media in the Arab Spring. Globalizations, 8(5):675–679.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). Social media? get serious! understanding the functional building blocks of social media. Business Horizons, 54(3):241–251.

King, T. E. and Jobling, M. A. (2009). What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends in Genetics, 25(8):351–360.

Kobsa, A. (2014). User acceptance of footfall analytics with aggregated and anonymized mobile phone data. In International Conference on Trust, Privacy and Security in Digital Business, pages 168–179. Springer.

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24):8788–8790.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In Proceedings of the 19th international conference on World Wide Web, pages 591–600. ACM.

Lansley, G. and Longley, P. (2016a). Deriving age and gender from forenames for consumer analytics. Journal of Retailing and Consumer Services, 30:271–278.

Lansley, G. and Longley, P. A. (2016b). The geography of Twitter topics in london. Computers, Environment and Urban Systems, 58:85–96.

Lasker, G. W. (1977). A coefficient of relationship by isonymy: A method for estimating the genetic relationship between populations. Human Biology, pages 489–493.

Li, N. and Chen, G. (2009). Analysis of a location-based social network. In International Conference on Computational Science and Engineering, 2009. CSE'09., volume 4, pages 263–270. Ieee.

Li, N. and Chen, G. (2010). Sharing location in online social networks. IEEE Network: The Magazine of Global Internetworking, 24(5):20–25.

Lieshout, R. (2012). Measuring the size of an airport's catchment area. Journal of Transport Geography, 25:27–34.

Liu, Y., Chen, L., Yuan, Y., and Chen, J. (2012). A study of surnames in China through isonymy. American Journal of Physical Anthropology, 148(3):341–350.

Lloyd, A. and Cheshire, J. (2017). Deriving retail centre locations and catchments from geo-tagged Twitter data. Computers, Environment and Urban Systems, 61:108–118.

Longley, P. and Singleton, A. (2014). London Output Area Classification: Final report.

Longley, P. A., Adnan, M., and Lansley, G. (2015). The geotemporal demographics of Twitter usage. Environment and Planning A, 47(2):465–484.

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

Lumley, T. (2009). leaps: regression subset selection. R package version 2.9, Using Fortran code by Alan Miller.

Lupton, D. (2013). Understanding the human machine [Commentary]. IEEE Technology and Society Magazine, 32(4):25–30.

Maguire, D. J. (1991). An overview and definition of GIS. Geographical Information Systems: Principles and applications, 1:9–20.

Mallows, C. L. (1973). Some comments on Cp. Technometrics, 15(4):661–675.

Manago, A. M., Graham, M. B., Greenfield, P. M., and Salimkhan, G. (2008). Self-presentation and gender on MySpace. Journal of Applied Developmental Psychology, 29(6):446–458.

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

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

Mateos, P., Singleton, A., and Longley, P. (2009). Uncertainty in the analysis of ethnicity classifications: Issues of extent and aggregation of ethnic groups. Journal of Ethnic and Migration Studies, 35(9):1437–1460.

McConnell-Ginet, S. (2003). What's in a name? 'social labeling' and gender practices. The handbook of language and gender, pages 69–97.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. ICWSM, 11:5th.

Mitchell, A. (2005). The ESRI guide to GIS analysis, Volume 2: Spatial Measurements and Statistics.

Model, S. and Fisher, G. (2002). Unions between blacks and whites: England and the US compared. Ethnic and Racial Studies, 25(5):728–754.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? comparing data from Twitter's streaming API with Twitter's firehose. Proceedings of ICWSM.

Nikolenko, S. I., Koltcov, S., and Koltsova, O. (2015). Topic modelling for qualitative studies. Journal of Information Science, page 0165551515617393.

O'Brien, O. and Cheshire, J. (2015). Interactive mapping for large, open demographic data sets using familiar geographical features. Journal of Maps, pages 1–8.

OECD (2013). New Data for Understanding the Human Condition: International Perspectives. Technical report, OECD.

Ofcom (2015). The UK is now a smartphone society.

Omand, D., Bartlett, J., and Miller, C. (2012). Introducing social media intelligence (SOCMINT). Intelligence and National Security, 27(6):801–823.

ONS (2013). Travel trends: 2013.

ONS (2016). Statistical bulletin: Internet users in the UK: 2016. Technical report, Office for National Statistics.

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. Communications & Strategies, 65(1):17.

Oxford University Press (2010). Oxford Dictionary of English. OUP Oxford.

O'Malley, L., Patterson, M., and Evans, M. (1997). Retailer use of geodemographic and other data sources: an empirical investigation. International Journal of Retail & Distribution Management, 25(6):188–196.

Papacharissi, Z. (2009). The virtual geographies of social networks: A comparative analysis of Facebook, LinkedIn and ASmallWorld. New Media & Society, 11(1-2):199–220.

Paul, M. J. and Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. ICWSM, 20:265–272.

PeerReach (2014). 4 ways how Twitter can keep growing.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In Computing attitude and affect in text: Theory and applications, pages 1–10. Springer.

Poston Jr, D. L. and Micklin, M. (2006). Handbook of population. Springer Science & Business Media.

Procter, R., Vis, F., and Voss, A. (2013). Reading the riots on Twitter: Methodological innovation for the analysis of Big Data. International Journal of Social Research methodology, 16(3):197–214.

Quan, H., Wang, F., Schopflocher, D., Norris, C., Galbraith, P. D., Faris, P., Graham, M. M., Knudtson, M. L., and Ghali, W. A. (2006). Development and validation of a surname list to define Chinese ethnicity. Medical Care, pages 328–333.

Rinker, T. W. (2016). sentimentr: Calculate text polarity sentiment. University at Buffalo/SUNY, Buffalo, New York. version 0.5.3.

Shagrir, I. (2003). Naming patterns in the Latin kingdom of Jerusalem, volume 12. Iris Shagrir.

Silver, N. and McCann, A. (2014). How to tell someone's age when all you know is her name. https://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name/.

Singleton, A. D. and Longley, P. (2015). The internal structure of Greater London: a comparison of national and regional geodemographic models. Geo: Geography and Environment, 2(1):69–87.

Singleton, A. D. and Longley, P. A. (2009). Geodemographics, visualisation, and social networks in applied geography. Applied Geography, 29(3):289 – 298.

Smith, M. A. and Kollock, P. (1999). Communities in cyberspace. Psychology Press.

Statista (2016). Distribution of EU referendum votes by age and gender UK. https://www.statista.com/statistics/567922/distribution-of-eu-referendum-votes-by-age-and-gender-uk/.

Stone, G. P. (1990). Appearance and the self: A slightly revised version. Life as theater: A dramaturgical sourcebook, pages 141–62.

Suler, J. (2004). The online disinhibition effect. Cyberpsychology & Behavior, 7(3):321–326.

Sykes, B. and Irven, C. (2000). Surnames and the y chromosome. The American Journal of Human Genetics, 66(4):1417–1419.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2):267–307.

Tajfel, H. and Turner, J. C. (1979). An integrative theory of intergroup conflict. The Social Psychology of Intergroup Relations, 33(47):74.

Twitter (2014). 80% of UK users access Twitter via their mobile. https://blog.twitter.com/en-gb/2014/80-of-uk-users-access-twitter-via-their-mobile.

Twitter (2015). The streaming APIs. https://dev.twitter.com/streaming/overview.

Twitter (2016). Twitter | Company | About. https://about.twitter.com/company.

United Nations (2001). Population registers.

Vicente, C. R., Freni, D., Bettini, C., and Jensen, C. S. (2011). Location-related privacy in geo-social networks. Internet Computing, IEEE, 15(3):20–27.

Vickers, D. and Rees, P. (2006). Introducing the area classification of output areas. POPULATION TRENDS-LONDON, 125:15.

Wade, T. and Sommer, S. (2006). A to Z GIS.

Weppner, J., Bischke, B., and Lukowicz, P. (2016). Monitoring crowd condition in public spaces by tracking mobile consumer devices with wifi interface. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pages 1363–1371. ACM.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., and Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology, 75(5):1182–1189.

Wickham, H. (2016). rvest: Easily harvest (scrape) web pages. R package version 0.3.2.

Williamson, T. (2008). The handbook of knowledge based policing: Current conceptions and future directions. John Wiley & Sons.

Willis, F. N., Willis, L. A., and Gier, J. A. (1982). Given names, social class, and professional achievement. Psychological Reports, 51(2):543–549.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 347–354. Association for Computational Linguistics.

Wolda, H. (1981). Similarity indices, sample size and diversity. Oecologia, 50(3):296–302.

Zhao, S., Grasmuck, S., and Martin, J. (2008). Identity construction on Facebook: Digital empowerment in anchored relationships. Computers in Human Behavior, 24(5):1816–1836.

Zimmer, M. and Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. Aslib Journal of Information Management, 66(3):250–261.