Towards real-time geodemographic information systems: design, analysis and evaluation

Muhammad Adnan

Department of Geography

University College London (UCL)

Thesis submitted in the fulfillment of the requirement for the degree of:

Doctor of Philosophy (PhD)

September, 2011

Word Count: 60, 881

DECLARATION

I, Muhammad Adnan, confirm that the work presented in this thesis 'Towards real-time geodemographic information systems: design, analysis and evaluation' is exclusively my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis. The views expressed in this publication are those of the author and not necessarily of the University College London.

Signed:

Date:

ACKNOWLEDGEMENTS

I must thank, first of all, my supervisor Paul Longley, who has provided support, advice and guidance throughout this PhD. I will always be thankful to him for his constant encouragement, support, and help during this PhD. Paul's guidance and encouragement enabled me to complete this research successfully. Secondly, I am grateful to Alex Singleton, my second supervisor, for his support and guidance during the PhD. Paul and Alex have always been very kind and supportive during the course of the PhD research, and it wouldn't have been possible to complete the PhD without their support and guidance.

Thanks must go to Local Futures Group (LFG) who funded this PhD through a Knowledge Transfer Partnership. The team at LFG was very supportive throughout the research. Before the Knowledge Transfer Partnership, PhD was supported by a UCL's SPLINT (Spatial Literacy in Teaching and Learning) grant. Many people gave me very valuable ideas, support or materials through my research. I would like to thank Pablo Mateos, Maurizio Gibin, James Cheshire, Dan Lewis, Chris Gale, John Fisher, Liane Hoogland, Amreen Sandhu, Juliana Cipa, Amna Abdul Wahid and Muhammad Salman for their valuable ideas and moral support.

Finally, I would like to thank my friends, colleagues and family with whom I have shared different bits of the process of this PhD. It would have been very difficult to complete the PhD research without the moral support of my family in Pakistan, so special thanks goes to them.

ABSTRACT

Geodemographic classifications provide discrete indicators of the social, economic and demographic characteristics of people living in small neighbourhood areas. They have been regarded as products, which are the final 'best' outcome that can be achieved using available data and algorithms. However, reduction in the cost of geocomputation, increased network bandwidths and increasingly accessible spatial data infrastructures have together created the potential for the creation of classifications in near real-time within distributed environments. Current geodemographic classifications are said to be 'closed' in nature due to the data and algorithms used. This thesis is a step towards an open geodemographic information system that allows users to specify the importance of their selected variables and then perform a range of statistical analysis functions which are necessary to create classifications tailored to user requirements.

This thesis discusses the socio-economic data sources currently used in the creation of geodemographic classifications, and explains the work towards the creation of a non-conventional data sources arising out of the UCL's surname database. Such data sources are seen as key to the creation of tailor made classifications. The thesis explains and compares different cluster analysis techniques for the segmentation of geodemographic classifications. The development of an online information system employs an optimisation of *k*-means clustering algorithm. This optimisation uses CUDA (Computer Unified Development Architecture) for parallel processing of computationally expensive *k*-means on NVIDIA's graphics cards.

The concluding chapters of the thesis set out the architecture of a real-time geodemographic information system. The thesis also presents the results of the creation of bespoke local area classifications. The developmental work culminates in a pilot real-time geodemographic information system for the specification, estimation and testing of classifications on the fly.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
ASP	Active Server Pages
CCJs	Country Court Judgments
CLARA	Clustering Large Applications
CLG	Communities and Local Government
CPD	Canadian Postcode Directory
CPM	Counts Per Million
CPU	Central Processing Unit
CUDA	Computer Unified Device Architecture
DBMS	Database Management System
ED	Enumeration District
ER	Entity Relationship
ERD	Electoral Roll Data
EU	European Union
GA	Genetic Algorithm
GB	Great Britain
GBP	Great Britain Pound
GDAL	Geospatial Data Abstraction Library
GHz	Giga Hertz
GIS	Geographic(al) Information System
GNPC	Geonames Postcode Gazetteer
GNPL	Geonames Place name Gazetteer
GPS	Global Positioning System
GPU	Graphical Processing Unit
HTML	Hypertext Markup Language
IATA	International Air Transport Association
ICT	Information and Communication Technology
IMD	Index of Multiple Deprivation
JRI	Java to R Interface
KML	Keyhole Markup Language
LGDX	Local Government Data Exchange
LRBD	Land Registry Building Data
LRHPD	Land Registry House Price Data
LSOA	Lower Super Output Area
Ness	Neighbourhood Statistics
NSPD	National Statistics Postcode Directory
OA	Output Area
OAC	Output Area Classification

List of Abbreviations

ONS	Office of National Statistics
OS	OpenSpace
PAM	Portioning Around Mediods
PCA	Principal Component Analysis
RAM	Random Access Memory
SDRC	Social Disadvantage Research Centre
SQL	Sequential/Structured Query Language
TELDIR	Telephone Directory
TFL	Transport for London
UK	United Kingdom
USA	United States of America
URL	Uniform Resource Locator
UKER	United Kingdom Electoral Roll Address Table
XML	Extensible Markup Language

1 CHAPTER 1: INTRODUCTION

This is a thesis about geodemographics – the analysis of people by where they live (Sleight 2004). It seeks to advance the approach in a number of important respects. First, it develops new methods of delivering readily intelligible mapping of data at a wide range of scales to large numbers of end users. Second, it seeks to develop and apply improved methods for estimating and testing geodemographic models, in order to bring greater stability and consistency to the calibration process. And third, it introduces a flexible and interactive framework for specifying geodemographic representations using a range of open data sources, that offer greatly improved prospects for accommodating population dynamics (Birkin and Clarke 1988: 88) into geodemographic representations. It is argued that these different themes of specification, computation, visual evaluation and interaction render geodemographics more robust, transparent and applicable in an era of open data, faster computation, and enhanced methods for visual communication.

1.1 GEODEMOGRAPHICS

Geodemographics are small area classifications that provide summary indicators of the social, economic and demographic characteristics of neighbourhoods. They are based upon the fundamental premise that place and population are inextricably linked with one other. Knowing about where somebody lives, can reveal a lot of information about that person's identity Vickers & Rees (2007), a notion that is fundamental to the spirit and purpose of human geography. Singleton et al. (2010) describe how geodemographic modelling of neighbourhood conditions presents a successful and applied approach to understanding socio-spatial differentiation at the neighbourhood scale, that is widely used in the public as well as the private sector to predict the consumption of product, services and resources.

Geodemographics has a rich history, that has its origins in Charles Booth's studies of deprivation and poverty of London in 1889. This precedes the work of Marr (1904), who mapped housing conditions of Manchester and Salford based on their housing conditions, and established a tradition that can be traced through the work of the Chicago urban ecologists in the 1920s and 1930s, social area analysis of the 1950s and factorial ecologies of the 1960s and 1970s to the work of Webber (1975) and Webber and Craig (1978), who sought to identify deprivation in inner city areas of Liverpool. After this, many commercial companies contributed to this field (Experian, Nottingham: MOSAIC, CACI, London: ACORN) (Sleight, 2004). Today, geodemographics is even more popular with the emphasis on creating bespoke geodemographic classifications in health (Farr and Evans 2005; Shelton et al. 2006; Peterson et al. 2010), policing (Ashby and Longley, 2005), education (Singleton, 2010) and local government (Longley and Singleton, 2009). In their current form, the use of geodemographics is in both public and private sector projects (Birkin et al. 1996) (Peterson et al, 2010).

Geodemographic classifications continue to use census data as the principal data source for neighbourhood segmentation. Census data cover a wide range of attributes regarding demographics and socio-economic characteristics of the population. However, over the last decade or so, commercial companies have come to use other data sources in addition to census data. These data sources include lifestyle surveys, financial measures data, property characteristics data, County Court judgements, credit card data etc. This list of data sources to create geodemographic classifications is not exhaustive, but the motivation for using these sources has lain in the quest to devise ever richer and more relevant depictions of small area conditions.

1.2 MOTIVATION FOR THIS THESIS

The motivation for this PhD arises from the need to create open geodemographic information systems for improved public service delivery and decision making.

Hitherto, geodemographic classifications have been created from static data sources which do not necessarily reflect the dynamics of population change in modern cities. Census data have been the main source of creating geodemographic classifications: however, they are usually collected at infrequent (usually decennial) time intervals in different countries. Thus at the time of writing (September 2011), the most recent available census data for United Kingdom are for the year 2001. In view of the scale and rapidity of changes consequent upon regional, national and international migration, alongside a range of changes in household composition and built form, commentators (particularly with interests in commercial geodemographic systems) have argued that updating for inter-censal periods is in order. Data are increasingly available in near real-time and could be integrated to create more temporally responsive systems. With the advent of the ‘open data’ initiative in the United Kingdom (see http://www.data.gov.uk), many data sources have been made public, and it has became possible to create bespoke local area classifications in real-time. The ONS Ness Data Exchange (Office for National Statistics, 2009) and London Data Store (London Data Store, 2010) are examples of derived data sources and products in which live feeds of data have been available using innovative application programming interfaces (APIs). In addition, static data can be downloaded from a wide range of other web locations and could be used for various purposes. All of these possibilities enable us to create geodemographic information systems which could exploit these data sources in building geodemographic classifications.

A closely related issue is the openness of the classifications that can and have been produced. Many commercial companies create and sell their geodemographic classifications, but for obvious commercial reasons do not disclose the precise data sources used, their provenance, or the methods used to build them. The nature of these classifications is thus essentially closed and users are required to proceed without the knowledge of data and methods used. This makes public scrutiny very difficult for these classifications, and also raises issues of user accountability. While public consultation is becoming popular in various areas, it is also of demonstrable importance in geodemographics (Longley & Singleton, 2009). Together these issues raise the need to build classifications where the information regarding the data and methods used are open to public. Also there is a need of the verification of the final classifications produced either by expert users or by public consultation.

Each of these issues provides the motivations for this PhD research, with the aim to develop open geodemographic information systems which could produce free and open classifications, in ways that are sensitive to user requirements and which can be interpreted by the user at various stages in the classification process.

1.3 AIMS AND OBJECTIVES

The aims of the PhD are twofold. First, the aim is to explore different data sources for their inclusion in creating more open and responsive geodemographic classifications. Second, the aim is to design and develop an open software product for building local area bespoke geodemographic classifications. This software product will enable the users to control how they produce a geodemographic classifications.

These aims will be achieved by the following objectives:

- Investigate geodemographic classifications, and how they are produced

- Build a novel data source of family names that could provide an international benchmark for geodemographic analysis. This entails accessing diverse data sources with different data structures, and the creation of a large and complex dataset.

- Develop analytical visual methods for data aggregation, monitoring and exploration of the composite data source created in the previous objective.

- Review different data sources available in the public domain and investigate the potential for their inclusion in creating geodemographic classifications. This will also involve the development of a web-based service for extracting live XML feeds of data from an online data source.

- Investigate the utility of different cluster analysis techniques to build geodemographic classifications. This will be followed by technical work on the optimization of clustering algorithms (specifically *k*-means) to allow users to build open geodemographics quickly and efficiently.

- Design and develop a pilot software utility that will enable users to create geodemographic classifications. This software utility will be tested through creation of a number of bespoke geodemographic classifications.

Detailed specifications of each objective are developed in individual chapters. As such, the primary motivations for the PhD are both methodological and technical. It will investigate the development of new data sources from of view of creating geodemographic classifications. Data sources available in the public domain will also be investigated for their inclusion in the classifications, and the methods for building geodemographic classifications will be investigated and enhanced in order to allow classifications to be produced in real-time. The concluding chapters of this thesis will discuss the design, development and testing of a desktop software utility for creating local area bespoke geodemographics classifications.

1.4 METHODS AND OUTPUTS

As described above, this thesis is a mix of methodological discussion and technical software engineering concepts. The thesis starts (Chapter 2) with an overview of geodemographic classifications and the methodology that underpins their creation. Chapter 3 describes the creation of a very large data source of family names from 26 countries around the world. This data source is created by aggregating data from a number of telephone directories and electoral registers. Chapter 4 takes the international database of Chapter 3 a step further by devising a visualisation strategy for a huge geographical data source across a range of spatial scales. This chapter discusses different visualisation techniques and their advantages and disadvantages in different circumstances. To this end, the thesis emphasises use of a single indicator (in this case a surname) to pioneer deployment of a range of visualisation techniques across a full range of geographic scales.

While Chapters 3-4 emphasise database design and visualisation principles applied to a single indicator, Chapters 5 discusses a fuller range of data sources available in public domain for their possible inclusion in creating geodemographic classifications. This Chapter investigates the variables and spatial coverage of the data sources. A methodology is also devised for the creation of a web service for extracting live xml feeds of the data from an online data source. This is important if classifications are to be created in real-time using live feeds of data from a number of data sources.

Once the data sources have been investigated in sufficient detail, Chapter 6 emphasises on the methodology of creating a geodemographic segmentation. This

Chapter discusses cluster analysis in detail, and discusses different clustering algorithms and assesses their relative efficacy in real-time environments. Different optimisation of *k*-means clustering algorithms are devised and explained from the view of creating geodemographic classifications quickly and efficiently.

Chapters 7-8 implement the data and methods discussed earlier in the thesis. Chapter 7 emphasises the design and development of a public domain software utility, 'GeodemCreator'. This software utility enables users to create their own bespoke geodemographic classifications. Chapter 8 discusses the implementation of 'GeodemCreator' in the creation of a number of open geodemographic classifications.

Thus the main outputs of this research are:

- Interactive websites for visualising very large socioeconomic datasets across a full range of spatial scales.
- An optimised version of a clustering algorithm (*k*-means) for purposes of creating classifications in real-time environments.
- The design and architecture of 'GeodemCreator' which is a publicly available software utility for building geodemographic classifications
- The software ('GeodemCreator') itself.

The software utility 'GeodemCreator' will be made available to download from a web link with an outline user guide, so that users can download the software and create their own bespoke classifications.

Published work arising from these different activities is detailed and included in the Appendix 1 section.

1.5 THESIS STRUCTURE

The thesis is organised into nine chapters. Each of the seven core chapters addresses one or more of the six objectives specified in section 1.2.

A brief introduction of each of the chapters is given below:

CHAPTER 2: GEODEMOGRAPHIC CLASSIFICATIONS

This chapter introduces the concept of geodemographic classifications, and describes their rich and varied history. It continuous with discussion on the ways in which the current generation of geodemographic classifications are created by commercial companies. This chapter also discusses how geodemographic classifications differ from other social statistics, and provides a thorough critique of the diffusion and use of geodemographic classifications. It also emphasises how classifications are created with a thorough description of data and methods used. In the final part, this chapter introduces international geodemographics classifications and various problems relating to content and geographical coverage of the data used to create them. This chapter lays out a motivation for the creation of a data source of family names from around the world.

CHAPTER 3: CREATION OF A GEO-REFERENCED GLOBAL NAMES DATABASE

There are numerous and innovative sources of data which have hitherto not been exploited in the creation of geodemographic classifications. This chapter discusses the creation of a database of surnames from 26 countries around the world. This database contains useful information about the migration, social mobility, and ethnicity (Mateos, Webber et al. 2007). The chapter is divided into two parts.

The first part discusses the ways of extracting attribute data from different data sources and storing them in the database. It starts by discussing naming conventions around different countries of the world and discusses the sources of the data (telephone directories, electoral registers, etc) that can be used to generalise about them. This continues with the discussion of the geo-referencing of names and address data extracted from telephone directories in particular. This is important for visualisation of the names data. The chapter also explains how the data are saved in Oracle database in the form of database tables and entities. In the subsequent discussion, the overall structure of the 'Worldnames' database is described.

The second part discusses the extraction of attribute data from graduated map data. This part explains the process of extracting attribute data from scanned map images. Chinese maps were used as an example dataset. This part explains the algorithm used to extract the data from the maps and how the data are stored in the database.

To an end, this chapter describes the development of innovative procedures to establish a consistent international database which potentially provides the spatial data infrastructure to an international geodemographic system. Moreover, classification of the individual names data provides valuable information about the cultural, ethnic and linguistic patterning of different parts of the world, as well as evidence concerning the socio-economic and demographic characteristics of neighbourhood areas.

CHAPTER 4: GEO-VISUALISATION OF THE WORLD NAMES DATABASE

The research described in this chapter describes the development of multi-scale web tools for visualising the data sources created in chapter 3. It discusses and explains different geo-visualisation techniques and the visualisation of the 'Worldnames' database. The chapter begins by describing the different web mapping platforms and visualisation techniques that are in current use. The relative usefulness of static map renderers, slippy maps, API based map services and flash maps are all described in detail. This section also explains the advantages and disadvantages of each of the available techniques.

The Chapter goes on to explain the architecture and design of the 'Worldnames' flash website. This includes project specification and user requirements, justification of using flash over other web mapping techniques, web application design (using appropriate state-transition, class, and use case diagrams), and a data flow diagram of the web application. The screen shots of the 'Worldnames' beta website are given and explained, which then leads to the development of the final version of the 'Worldnames' website with a modification in the visualisation strategy (i.e. the choice between 'grid based flash maps' and 'administrative boundaries flash maps').

The final section of this chapter charts the success of the website success and its promotion in leading news outlets across the world.

CHAPTER 5: DATA SOURCES REVIEW

Chapter 4 documents the use of a single indicator variable (surname) to produce statistics and visualisation. Gedeomographic classifications are based on multiple indicators of socio-economic characteristics, and Chapter 5 discusses the data sources available in the public domains and could be used for building bespoke geodemographic classifications.

First part of the chapter discusses the creation of a web service for extracting data from the ONS NeSS (Neighbourhood Statistics) Data Exchange (Office for National Statistics, 2009). This API provides live XML feeds of data for a number of data sources available in the public domain. This web service is important from the point of view of creating real-time geodemographic classifications where data is integrated from a range of different online data sources. This section explains the architecture of the web service.

The second part of this chapter explains the data sources available in the public domain that might be used for creating geodemographic classifications. Variables in each data source are explained with the description of their spatial coverage.

The third section explains the creation of two data sources from the amalgamation of London's land registry house price data with the ethnic classification of the Great Britain Electoral Roll. Onomap (Mateos, 2007) was used to code names data to their corresponding ethnicities. The ethnicity information and London's Land Registry house price data were joined and two new data sources were created. These data sources could be very useful resources for exploring socio-economic distributions of different ethnic groups in London.

The final section of this chapter outlines the variables those could be used for the creation of geodemographic classifications.

CHAPTER 6: CLUSTER ANALYSIS FOR THE CREATION OF GEODEMOGRAPHIC CLASSFICATIONS

Geodemographic classifications are created using cluster analysis of socio-economic data. Cluster analysis combines/groups similar items into categories based upon an over-all measure of homogeneity. This chapter discusses relevant cluster analysis procedures in detail.

The first part of this chapter explains different clustering algorithms. These clustering algorithms include hierarchical clustering, *k*-means clustering, partitioning around medoids (PAM), and genetic clustering. This part explains the use of *k*-means for creating geodemographic classifications, and explains the instability of the basic *k*-means clustering algorithm. Later, a detailed comparison of *k*-means, Clara (Clustering Large Applications), and genetic clustering algorithms is given using OAC (the ONS Output Area Classification) (Vickers and Rees, 2007) data at three different spatial levels. This comparison is important from the view of evaluating different clustering algorithms for their relative performance in real-time environments.

The final part of the chapter sets out recent computational improvements in *k*-means clustering. This includes using principal components analysis (PCA) on the dataset before applying the *k*-means clustering algorithm, and a parallel implementation of *k*-means on Nvidia's graphics cards using CUDA (Computer Unified Device Architecture) (Harish & Narayanan, 2007).

CHAPTER 7: FUNCTIONAL AND TECHNICAL APPLICATION SPECIFICATION

Previous chapters have discussed the need to building real-time geodemographic classifications, where the methods and tools of building the classifications are open and known to the users and public scrutiny. This chapter provides the proof of concept that these kinds of tools and methods can be developed in practice. It discusses the functional and technical specification for the development of a desktop software utility 'GeodemCreator', that can be used to create bespoke classifications. The application specification defines the overall functionality included in the software in the form of textual information and flow charts. The technical application specification defines the intended audience for the software, the operation environment required, the software scope, and the technology that can be used to develop the software. This part also describes the design and overall structure of the software.

CHAPTER 8: A PILOT GEODEMOGRAPHIC DECISION SUPPORT SYSTEM

This chapter describes the use of 'GeodemCreator' for building bespoke geodemographic classifications. The use of the classification in practice is illustrated using screen shots of the software in action, along with detailed descriptions of how users can create their own classifications.

This chapter explains the creation of three different geodemographic classifications using 'GeodemCreator'. Output Area Classification (Vickers & Rees, 2007) data was used to create the first two classifications. Third classification was created by using

Output Area Classification (Vickers & Rees, 2007) data and an ancillary data source of ethnicity. This data sources was created by using the 'Worldnames' database. Chapter no. 3 describes the creation of the 'Worldnames' database. Thus the three classifications created in this chapter are:

- An output area classification of Greater London: This geodemographic classification is the output area classification of Greater London. This is created by using the output area classification (Vickers & Rees, 2007) data for Greater London.

- An output area classification of the United Kingdom: This geodemographic classification is the re-creation of the output area classification created by (Vickers & Rees, 2007). This demonstrates the creation of a national level classification using 'GeodemCreator'.

- An output area level 'socio-economic and ethnic' classification of Greater London: This case study described the creation of an output area level 'socio-economic and ethnic' classification of Greater London by using 'GeaodemCreatoer'. This case study gives a proof of the concept of building local area classifications using census data and other ancillary data sources.

2 CHAPTER 2: GEODEMOGRAPHIC CLASSIFICATIONS

2.1 AREA CLASSIFICATIONS

Area classifications provide a unique way of viewing zonal geographies that bring together a range of variables, in order to identify similarities between areas (Webber & Craig 1978) independent of their locations relative to one another. The idea of area classification is not a new one: the human mind tends to group different things into categories; we tend to live near people we are similar to ourselves, in terms of social structure, economic conditions, and cultural behavioral values. Figures 2.1 and 2.2 illustrate this idea. Figure 2.1 shows the concentration of Bangladeshi people living in London, while Figure 2.2 shows the concentration of Hindi people residents. There are clear groupings of people from different communities living in areas of London which are closer to each other. The source of Figures 2.1 and 2.2 is http://www.londonprofiler.org, and the shaded areas are unit postcodes.

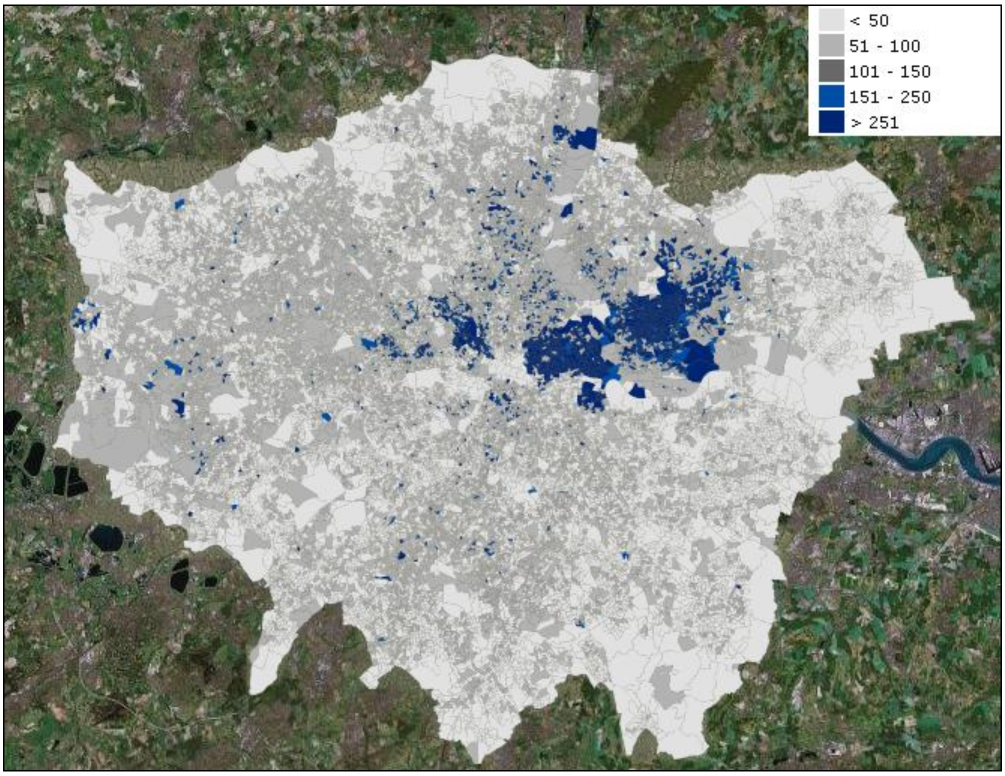


Figure 2.1: Concentration of Bangladeshis living in London

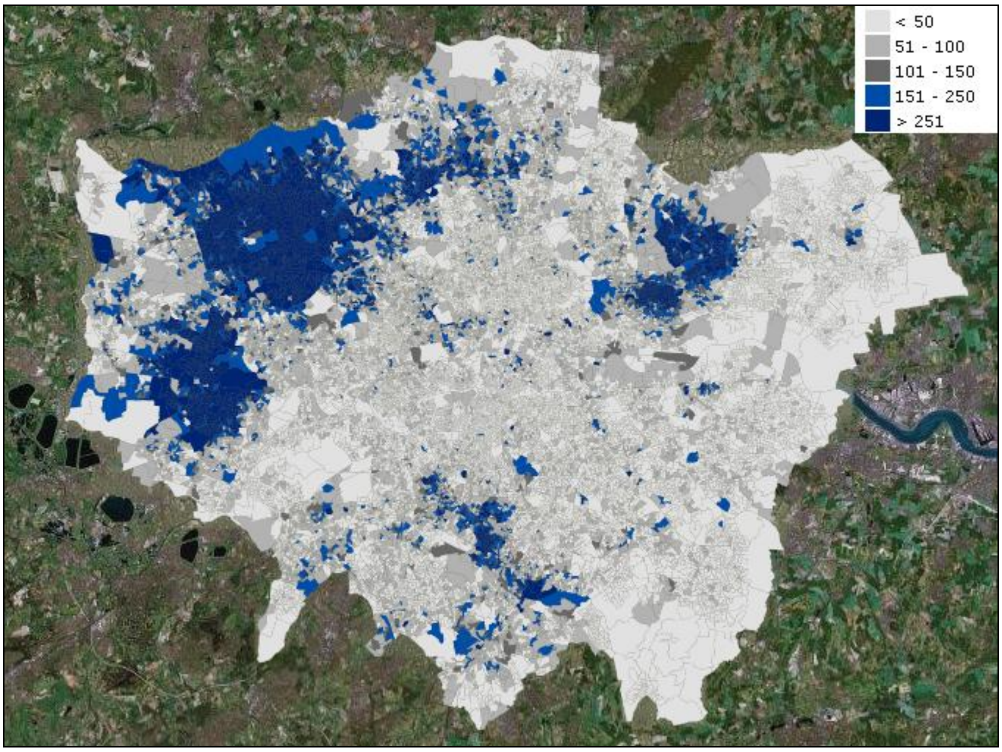


Figure 2.2: Concentration of Hindi people living in London

Group or Area Classifications can be seen everywhere. Figure 2.3 and 2.4 show a different example. These figures show the concentration of two surnames in Great Britain. Figure 2.3 shows the concentration of surname 'Longley', and figure 2.4 shows the concentration of surname 'Cheshire'. Despite the fact that both are Anglo-Saxon surnames, they are concentrated in different geographic areas. This shows another example of grouping. The source of figures 2.3 and 2.4 is http://gbnames.publicprofiler.org.

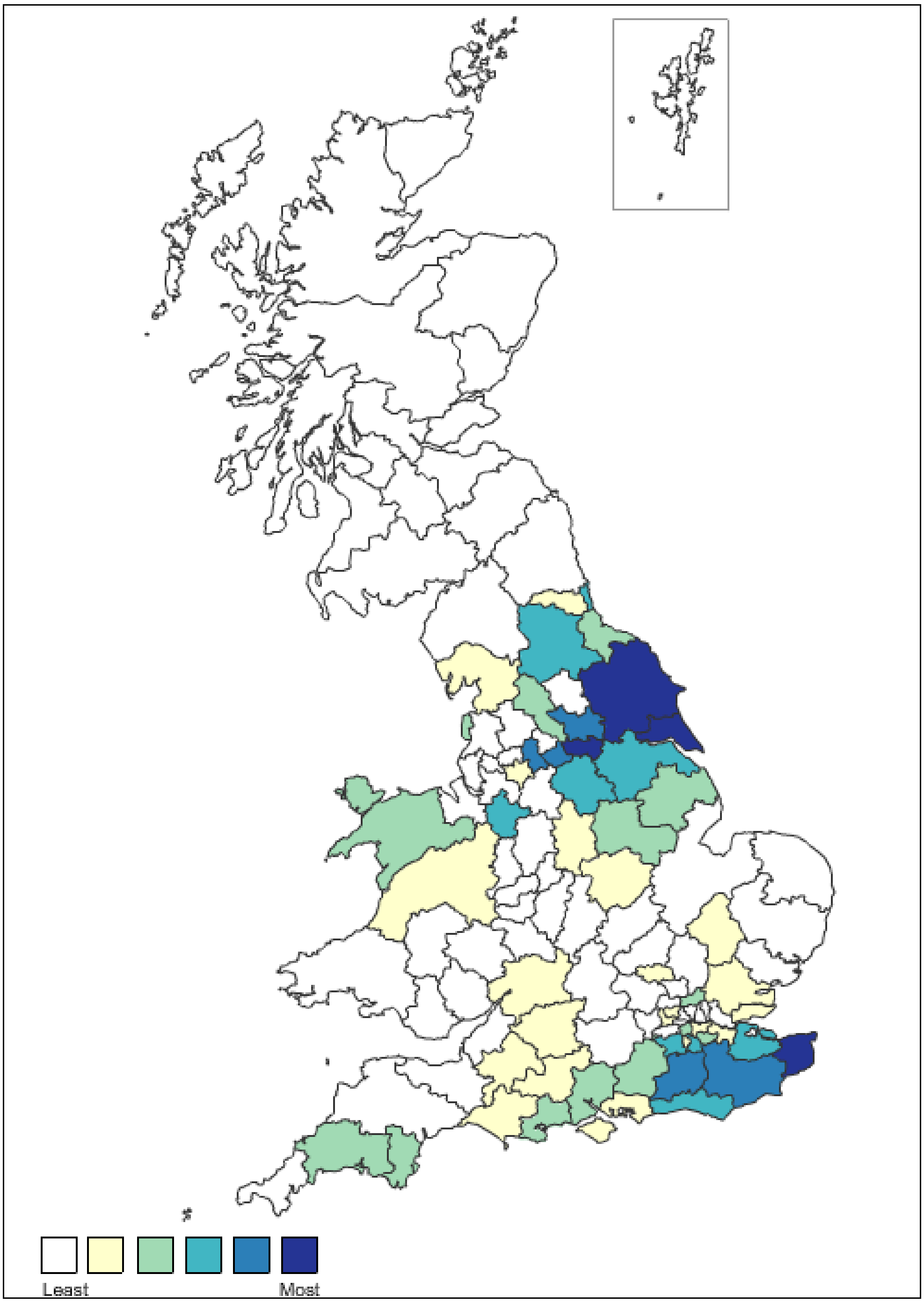


Figure 2.3: Concentration of Surname 'Longley' in the Great Britain

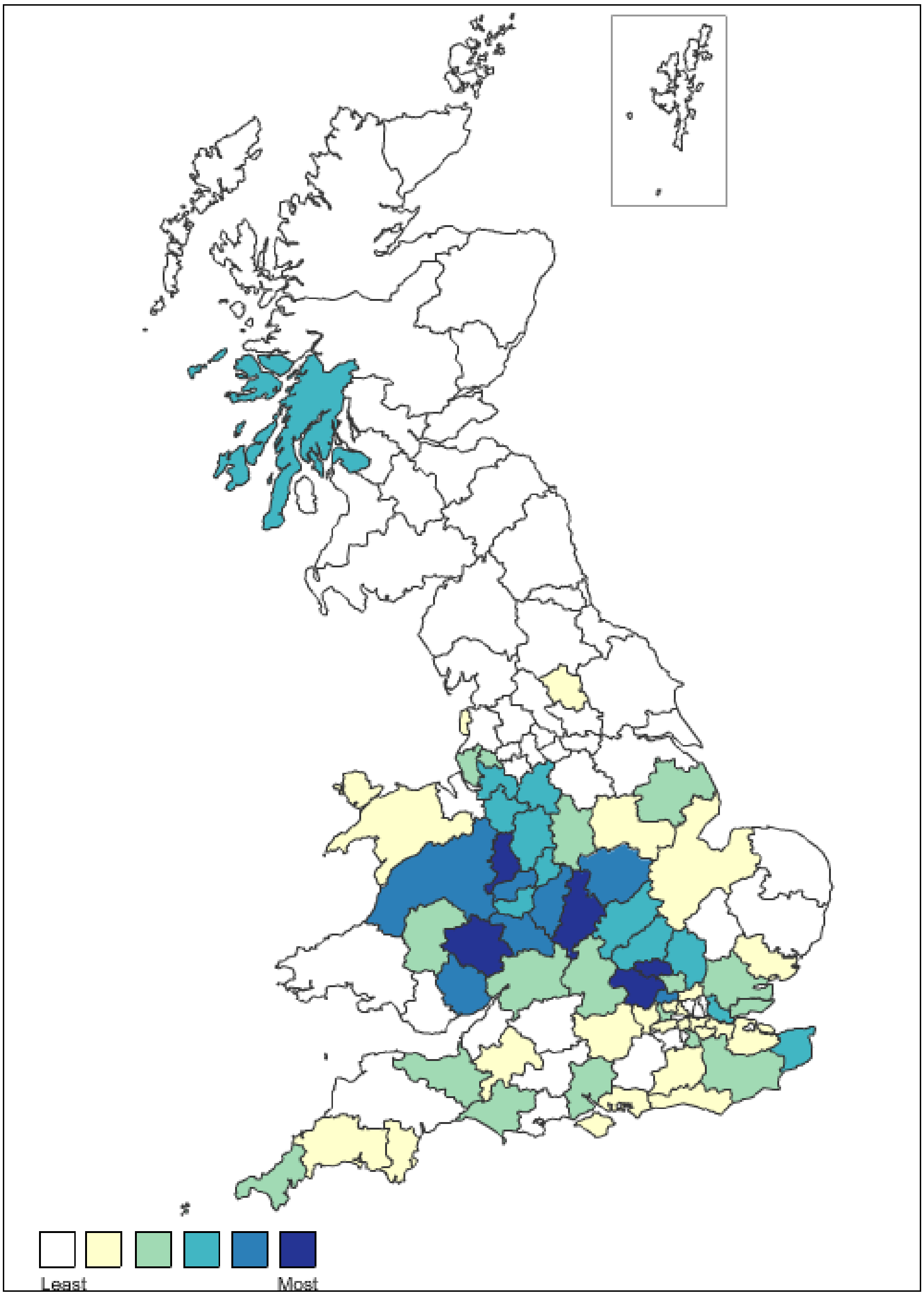


Figure 2.4: Concentration of Surname 'Cheshire' in the Great Britain

There are groups and clusters everywhere around us. The idea of area classification is a core part of human psychology, and it forms the basis for geodemographic classification.

2.2 GEODEMOGRAPHIC CLASSIFICATION

Geodemographics can be described by more than one definition. (Sleight 2004: 18) described geodemographics as "the analysis of people by where they live" or "locality marketing". Geodemographics are small area classifications that provide summary indicators of the social, economic and demographic characteristics of neighbourhoods. Geodemographics works on the concept that place and population are linked with each other. Knowing about where somebody lives, can reveal a lot of information about that person (Vickers and Rees, 2007).

Geodemographics gets power from Tobler's First Law of Geography which states "Everything is related to everything else, but near things are more related than those far apart" (Tobler 1970). Geodemographics uses the same principle that two residential neighborhoods that are close to one another are most likely to be similar than the ones that are more geographically separated. Geodemographics analyses the socio-economic attributes of the population and groups the places in classes of different socio-economic structures.

Geodemographic classifications group areas into modest numbers of categories (typically between 10 and 50) and typically classify neighborhoods at two or more hierarchical levels: for example, the five Categories that make up the top level of the commercial ACORN classification (CACI Ltd., London) are divided into a total of 17 Groups at the secondary level, and these 17 Groups are divided into a total of 56 Types at the lowest, tertiary, level. The characteristics of each class within the typology are summarized by a label (e.g. "City Living"), a verbal "pen portrait" and other descriptive material such as photographs, montages and videos. These multimedia are used to give users of the classification a clearer understanding of the characteristics of the underlying population.

Geodemographics have a rich history. Their development continues a tradition of socio-spatial differentiation in human geography and urban sociology which extends over an 80 year period to formative work in urban ecology and social area analysis (see Batey and Brown 1995 and Harris et al. 2005 for reviews and interpretations). The term was first coined in the 1970s in work seeking to identify deprived inner city areas (Webber 1975; Webber and Craig 1978). The success of these public sector applications led to the commercialization of geodemographics during the 1980s when the approach became widely used in the private sector as a method of target marketing (Birkin et al. 2002). More recently, public sector applications have moved somewhat to the fore (Longley 2005), with applications in health (Farr and Evans 2005; Shelton et al. 2006), policing (Ashby and Longley 2005), education (Singleton 2010) and local government (Longley and Singleton 2009). In their current form, the use of geodemographics is in both public and private sector projects (Birkin et al. 1996, 68-70).

2.3 HISTORY OF GEODEMOGRAPHICS CLASSIFICATION AS A SOCIAL MEASUREMENT

Geodemographics classifications have a rich history. This section describes the developments in geodemographics classifications in three eras.

a) Pre 1981 developments

b) Post 1981 developments (1981 - 2005)

c) Recent developments

2.3.1 PRE 1981 DEVELOPEMENTS

Geodemographics can be thought of as having its origins as Charles Booth and his studies of deprivation and poverty of London in 1889 - 1899. Charles booth visited different streets of London and categorised the streets into 7 different groups, based on observed indicators of poverty and deprivation. This was the first classification of areas into homogeneous groups, and formed the foundation of geodemographics or area classification. Figure 2.5 shows part of Charles Booth’s poverty map of London, and Figure 2.6 shows the 7 categories of the map. This map provides a remarkable example of how areas having similar characteristics of population group together and form a classification.

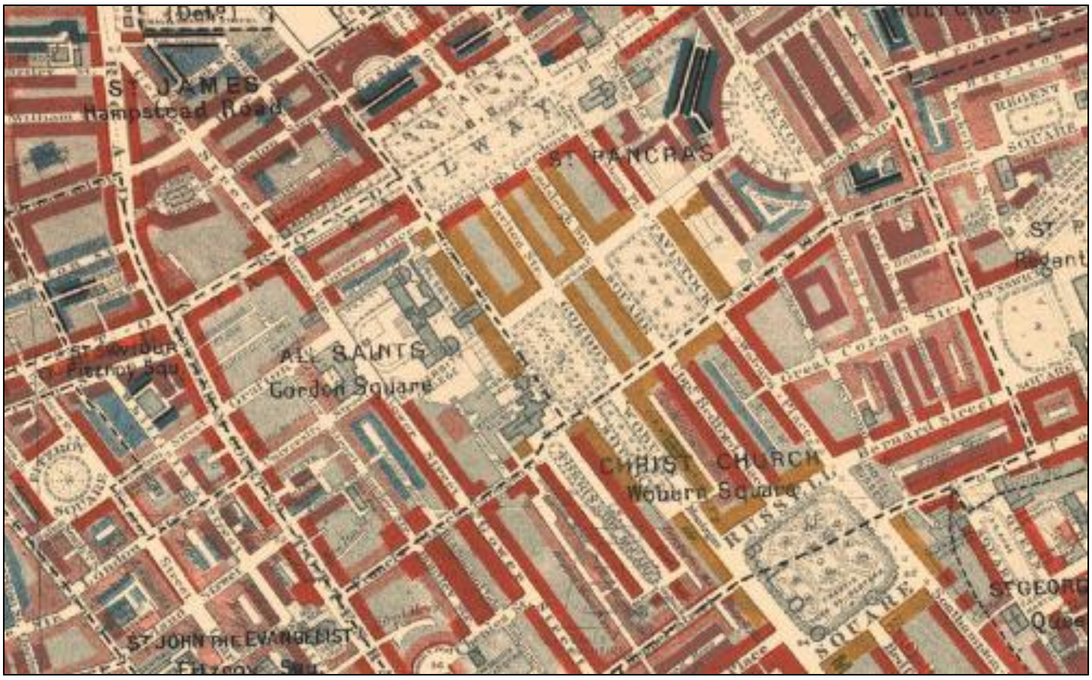
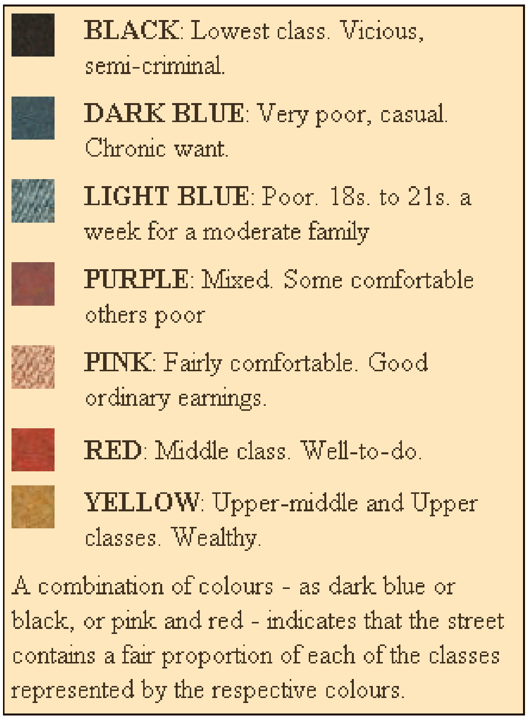


Figure 2.5: Charles Booth's poverty map of London

Source: (http://booth.lse.ac.uk/cgi-bin/do.pl?sub=view_booth_and_barth)



⬛ **BLACK**: Lowest class. Vicious, semi-criminal.

🟦 **DARK BLUE**: Very poor, casual. Chronic want.

▨ **LIGHT BLUE**: Poor. 18s. to 21s. a week for a moderate family

🟪 **PURPLE**: Mixed. Some comfortable others poor

▨ **PINK**: Fairly comfortable. Good ordinary earnings.

🟥 **RED**: Middle class. Well-to-do.

🟨 **YELLOW**: Upper-middle and Upper classes. Wealthy.

A combination of colours - as dark blue or black, or pink and red - indicates that the street contains a fair proportion of each of the classes represented by the respective colours.

Figure 2.6: Charles Booth's seven categories of London's population

Source: (http://booth.lse.ac.uk/static/a/4.html)

A second historical example of what would now be described as geodemographic classification is found in the work of T.R. Marr. Marr (1904) mapped housing conditions of Manchester and Salford. He created a map that divides the areas of Manchester and Salford into 10 different categories based on their housing conditions. Figure 2.7 show the Marr map of Manchester and Figure 2.8 shows the legend of the map.

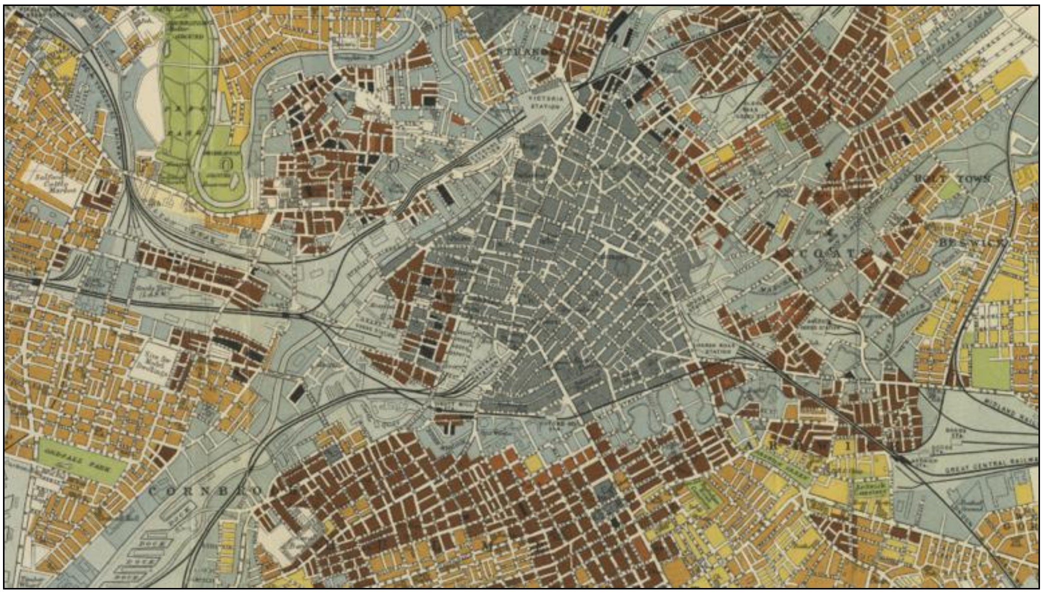
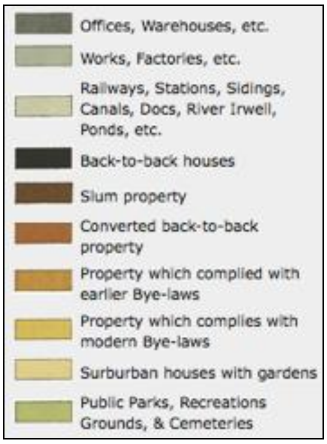


Figure 2.7: Part of the Marr map of Manchester

Source : (http://manchester.publicprofiler.org/marr/)



Offices, Warehouses, etc.

Works, Factories, etc.

Railways, Stations, Sidings, Canals, Docs, River Irwell, Ponds, etc.

Back-to-back houses

Slum property

Converted back-to-back property

Property which complied with earlier Bye-laws

Property which complies with modern Bye-laws

Surburban houses with gardens

Public Parks, Recreations Grounds, & Cemeteries

Figure 2.8: The 10 categories of the Marr map of Manchester

Source : (http://manchester.publicprofiler.org/marr/)

In 1970s Richard Webber contributed to geodemographics classifications by seeking to identify deprived inner city areas (Webber 1975; Webber and Craig 1978). Figure 2.9 show the map of the social area study for Liverpool city in 1971. This map categorises Liverpool into 5 groups based on the social status of the areas.

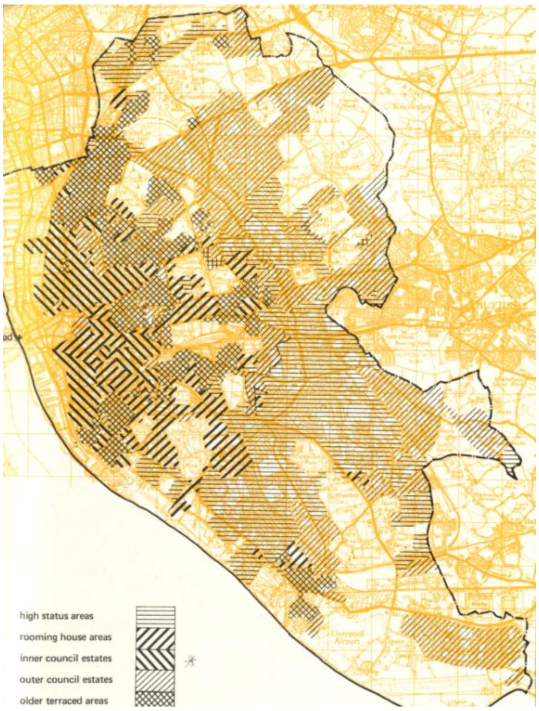


Figure 2.9: Liverpool Area Study (1971)

2.3.2 POST 1981 DEVELOPEMENTS (1981 - 2005)

After the dissemination of 1981 UK Census of Population results, area classification saw lots of progress and development in the private sector. Commercial companies who paid to be licensed users of 1981 Census data created categories of areas by using the socio-economic variables. ACORN, PiN, Mosaic, and Super Profiles developed the private sector market during mid 80s. Commercial companies also used new ancillary data sources for the creation of geodemographics classifications and it has taken geodemographics far from just being based on a census data source. Electoral Registers, In the UK, Country Court Judgements, credit reference agency data, vehicle registration data, and lifestyle surveys have been used as the other data sources for the creation of geodeomographic classifications (Harris et al 2005).

From 2001, UK Census data became free (there had hitherto been a license fee of £250,000 approx), and this further increased the popularity of geodemographics with an opportunity for smaller companies to create geodemographics classifications. With the increase in computing power and reduced costs of the data, it has now become possible for a single person to create a classification on his or her home PC. This does not require the resources of a whole company or state of the art computation devices to create the classifications. A large number of new companies have entered the market with their own geodemographics products. Table 2.1 lists the names of different organisations and their geodemographic classification systems available.

Organisation	Classification
CACI	ACORN
Experian	Mosaic
EuroDirect	CAMEO
Claritas	PRIZM
Acxiom	Personicx Geo
AFD Software	Censation
Allegran	Gnuggets
Beacon Dodsworth	P2 People and Places
Business Geographies	Locale
The Clockworkds/TRAC	SONAR
GeoBusiness	Locale
ISL	RESIDATA LIFETYPES
Streetwise Analytics	LIKEWISE

Table 2.1: Some Geodemographic systems available

2.3.3 CURRENT DEVELOPEMENTS

The use of geodemographic classifications is now becoming popular in different areas with applications in health (Farr and Evans 2005; Shelton et al. 2006), policing (Ashby and Longley 2005), education (Singleton 2010) and local government (Longley and Singleton 2009). General purpose and bespoke classifications are being created by public and private organisations and academic researchers. Census data have remained the core data source for creating geodemographics segmentations. But following ‘open data’ initiatives, many data sources have been made public, and it has in theory become possible to create bespoke local area classifications in real-time or on the fly. The ONS NeSS Data Exchange (Office for National Statistics, 2009) is an important open data source, where users can get feeds of census data by using the API. The London Data Store (London Data Store, 2010) has been created by the Greater London Authority as an initiative to make London's data free and accessible to all. Now programmers and data analysts can use thousand of data sources in addition to census data to create their own local area classifications. Crime Data have also been made public by UK Police in an initiative to the open data policy (http://www.police.uk). This enables general users, data analysts and programmers to map latest crime data either by downloading the data or getting live feeds from the http://www.police.uk website.

Computing equipment has seen major development with the increase in computation power almost fivefold in the last decade. Cloud computing is becoming cheaper and easy to use for the general public, which is bringing increased computation power to bear on fast transactions and real-time processing. Amazon (2010) has launched their EC2 cloud, which is a parallel computation architecture for running computationally intensive algorithms and processes.

Taken together, these data and computational power improvements are also creating the infrastructure for bespoke local area and real-time geodemographics classification systems.

2.4 HOW GEODEMOGRAPHIC CLASSIFICATION DIFFERS FROM OTHER STATISTICAL SOURCES

2.4.1 GEODEMOGRAPHICS VS IMD (INDEX OF MULTIPLE DEPRIVATION)

Created by the British Department for Communities and Local Government, the IMD (Index of multiple deprivation) is a measure of deprivation (hardship) made available at the small area level (Department of Communities and Local Government (2011)). Thus, the higher the deprivation index is, the more deprived is the area and vice versa. The Index was conceived by the Social Disadvantage Research Centre (SDRC) at the Department of Social Policy and Social Work at the University of Oxford. The IMD is based on the idea of distinct dimensions of poor physical and social conditions which can be recognized and measured separately. These are experienced by people living in the area. The overall IMD is conceptualized as a weighted area level aggregation of several constituent dimensions of deprivation. The new IMD is a lower layer super output area (LSOA) level measure of multiple deprivation and it has been developed by using seven constituent LSOA level indices. Each domain in turn comprises a number of variables to assess the level of deprivation in a LSOA area. Once a score is achieved for all of the constituent domains, a weighting procedure is used to calculate the average IMD of each LSOA. The following table lists the seven IMD domains with the corresponding weights that are assigned to them.

Domain	Weight (%)
Income Deprivation	22.5
Employment Deprivation	22.5
Health Deprivation and Disability	13.5
Education, Skills and Training Deprivation	13.5
Barriers to Housing and Services	9.3
Living Environment Deprivation	9.3
Crime	9.3

Table 2.2: The seven domains of the 2010 IMD

Source: http://www.communities.gov.uk/documents/statistics/pdf/1870718.pdf

The IMD is different to geodemographic indicators in that the specialised remit of the IMD is to record the level of deprivation in a particular area and has limited applications. The IMD has a number of uses in government policy, where there is a focus on the implications of high deprivation at the local level. The remit of geodemographics is usually somewhat broader, however, and is not usually focused upon the lowest socioeconomic echelons.

Another difference is the scale at which the different indicators are usually created. The IMD is created at the LSOA level. However, geodemographics classifications exist for much finer spatial scales, namely UK Output Areas and Unit Postcodes.

2.5 A CRITICAL REVIEW OF GEDEMOGRAPHIC CLASSIFICATIONS

Over time, the following critiques have been developed of geodemographic classifications.

2.5.1 DOES ONE SIZE FIT ALL? THE NEED FOR BESPOKE CLASSIFICATIONS

Most geodemographic classifications divide areas in United Kingdom into a pre-specified number of categories. This, for example, the ONS Output Area Classification (OAC) classifies each Census Output Area in the UK into one of seven broad categories. But question remains, do those seven categories adequately account for the characteristics of the population in the Output Areas in a sufficient level of detail for a given application? Singleton and Longley (2009) emphasize the need for local area bespoke classifications in their chosen application domain. Current classifications are created with the national coverage and they do not account for local level variations in the data. With the use of different data sources, application specific classifications can be devised for local areas. These classifications have been successfully demonstrated across a variety of areas, and there are many more sectors which could potentially benefit if the methods of construction and interpretation were more accessible. Decennial censuses of population have in the past been appropriate for creating geodemographics classifications, but the far-reaching and rapid changes that today characterize population characteristics and structure make it increasingly necessary to supplement census sources with data that are more timely and relevant to particular applications.

Better and more intelligent integration of a wider range of available data sources can open new horizons for depicting salient characteristics of populations and their behaviors. The art and science of creating geodemographic classifications has always been about much more than computational data reduction, and a key consideration in this quest is the availability of decision support tools to present areal data from a range of attributes in a format that is readily intelligible to the user. Thus, for example, in devising a local indicator of future risk of obesity, it might be appropriate to use data sources that variously measure demographic structure, school attainment, deprivation and existing health problems. In assembling such sources together, the analyst should be made aware of issues of data collection, normalization, weighting and data reduction methods.

Census data remain the main source for creating these classifications, and they are also used for academic study because they are the only spatially-detailed dataset freely available. But careful choice of census variables does not invariably account for the salient differences in the population of different areas. Also, geodemographic classifications assign an area to a single class or group. This also means that everyone living in a given small area has the same characteristics, which is rarely if ever the case. (Voas and Williamson, 2001) provide a critical review of geodemographic classifications by identifying that there are more differences found within particular classes than differences found between classes. Small areas are different in many different ways and a few dimensions do not provide enough information to describe an area fully. Geodemographic classifications can account for some of the diversity but do not account for the complete diversity of the areas.

2.5.2 OPEN METHODS

There has been considerable critique arising from the closed nature of commercial geodemographic classifications. In their current form, commercial geodemographic classifications are created in silos by expert producers, most prevalently with closed methods and little documentation of the data inputs; the weighting and normalization procedures are not disclosed; and neither is the exact method of clustering. Users only get the final classifications after areas have been grouped into different classes and they have to accept what they are given. Longley et al (2009) critique the closed nature of geodemographic classifications and form the view that there is a need for more open methods. These open methods are expected to be transparent in explaining all the procedures employed to build a geodemographic classification. Thus there is a need of a clear documentation about the method of selecting variables and their weightings, the normalization techniques employed, and the clustering algorithms used. Open methods will ensure that users have more confidence in the geodemographics classifications that they are using.

2.5.3 PUBLIC CONSULTATION

Another critique about geodemographics classifications is their closed nature and lack of openness to public scrutiny. In essence, users of the classification produced these days do not have any ability to give a feedback about the classification and alter the classification in any sense. Longley and Singleton (2009) addressed some concerns about geodemographic classifications being rarely transparent or open to scrutiny or challenge. They created a tool called E-society profiler (http://www.esociety.publicprofiler.org/). The aim of the project upon which this tool was based was to create a national classification of how people engage with new information and communication technologies (ICTs). The authors describe how this classification was opened to public scrutiny and challenge. Users can go on the website, see the classification, and provide feedback about the classification. The feedback may suggest a different class for areas for which users do not agree with what they see. They assessed 50,000+ searches and 3,952 responses during the consultation exercise. This method of open public scrutiny and challenge to a geodemographic classification make it more reliable and may also allow it to be updated.

2.6 BESPOKE VS STANDARD CLASSIFICATIONS

The increasingly complex, urbanized, and connected nature of human settlement is driving a demand for better contextual information to inform decisions about the needs and preferences of people and the places in which they live and work. Decennial censuses of population have in the past been appropriate for this task, but the far-reaching and rapid changes that today characterize population change are making it increasingly necessary to supplement census sources with data that are more timely and relevant to particular applications. Better and more intelligent integration of a wider range of available data sources can open new horizons for depicting salient characteristics of populations and their behaviors. The art and science of creating geodemographic classifications has always been about much more than computational data reduction, and a key consideration in this quest is the availability of decision support tools to present areal data from a range of attributes in a format that is readily intelligible to the user. Thus, for example, in devising a local indicator of school attainment, it might be appropriate to use data sources that variously measure demographic structure, school attainment, and deprivation. In assembling such sources together, the analyst should be made aware of issues of data collection, normalisation, weighting and data reduction method.

There is a need for the creation of more responsive and open geodemographic information systems. This would question the authority implied by classifications that purport to present 'best' solutions. There are a number of motivations for this. First, current classifications are created from static data sources that do not necessarily reflect the dynamics of population change in modern cities. Data are increasingly available at high temporal resolution and offer the potential to be integrated with other traditional sources to create more timely systems. For example, travel data recording the flow of commuters across a city network could be used to estimate daytime population characteristics. A further example could be extracting frequently updated patient registrations with doctors' surgeries in order to provide a more up-to-date picture of the residential composition of neighborhoods. A requirement for distributed and simple to use online classification tools arises from changes in the supply of socio-economic data and the potential that this creates for end users to create new intelligence on socio-spatial structures. In addition to Census data that are collected every 10 years in the United Kingdom, numerous supplementary data sources are becoming available, some of which are already updated in near real time. The availability of such resources will increase the potential to create more responsive and application specific geodemographic classifications that will make it less acceptable to uncritically accept the outputs of general-purpose classifications as received wisdom. Second, application specific classifications have been successfully demonstrated across a variety of domains, and there are many more sectors that could potentially benefit if the methods of construction and interpretation were more accessible and transparent. I argue here that there is a need for web-based applications that enable the creation of general purpose geodemographic classifications 'on the fly'. In these applications, the specification of (possibly real time) classification inputs should be guided to fulfill the objectives of the problem under investigation, with output from such analysis computed within a reasonable wait time. The task of creating real time geodemographics by integrating diverse and possibly disparate spatial databases raises a number of computational challenges concerning data normalization and optimization for fast transactions. As compared to the computational challenges posed in the past, real time computational solutions in an online environment are becoming possible and I anticipate that this will provide a stimulus to the development of real time application specific geodemographic classifications.

2.7 HOW GEODEGRAPHIC CLASSIFICATIONS ARE COMPUTED

Geodemographic classifications are computed by using the following four steps.

a) Specification

b) Normalisation

c) Clustering

d) Visualisation

This section gives a brief introduction of these steps. However, more detailed description can be found in other chapters.

2.7.1 SPECIFICATION

This is the first step in creating a geodemographics classification. The step concerns with selection of data sources, selection of variables from the data sources, and specification of weights for each variable selected.

The data used to create geodemographic classifications typically derive from any of a range of secondary data sources (Harris et al. 2005). Core to most classifications are census data and some classifications (such as the UK Output Area Classification: Vickers and Rees 2007) use no other source.

Commercial companies create their own classifications by using census data alongside other data sources. Other data may be derived from behavioral or attitudinal surveys (e.g. ‘lifestyle’ surveys from commercial sources), financial data (e.g. county court judgments, directorships) and property information (e.g. property tax bands).

Table 2.5 shows different classifications and data sources used in their creation.

Organisation	Classification System	Number of input variables	Number of clusters	Non-census data used?
CACI	ACORN	79	a) 6 b) 17 c) 54	No
Experian	Mosaic	87	a) 11 b) 52	Credit data, CCJs, electoral roll, postal address file, company directors, retail access
EuroDirect	CAMEO	48	a) 9 b) 50	No
	MicroVision	185	a) 11 b) 52 c) 200	Lifestyle data, company directors, share ownership, electoral roll, CCJs, risk indices, unemployment statistics
	DEFINE	146	a) 10 b) 50 c) 1050	Credit data, electoral roll, unemployment statistics, insurance ratings
Claritas UK	PRIZM	59 (+188)	a) 4 b) 19 c) 72	Lifestyle data, share ownership, company directors, unemployment statistics, postal address file, births and deaths
	SuperProfiles	120 (+130)	a) 10 b) 40 c) 160	Credit data, CCJs, TGI, Electoral Roll

Table 2.3: Geodemographic Classifications and data sources used

Source: Sleight (2004: 49)

Variables of other data sources (lifestyles surveys, credit data, unemployment statistics, etc.) differ from each other because of their availability from the professional organisations collecting data at different spatial levels.

After data sources have been selected, specification of a geodemographic classification includes the selection of variables and the weighting of their overall importance in the classification. This weighting insures that a variable having high weighting is more important in the classification than a variable having low weighting. Bespoke classifications tend to give more weight to those variables which are of importance in indicating local level variances in the target population. For example, one might be interested in creating a local level classification for Leicester to analyse the multidimensional characteristics of areas where most of the people are recent migrants. This classification might give more weight to the census variable 'Born Outside the UK' than any of the other variables.

A detailed review of commonly used data sources is given in Chapter 5 "Data Sources Review".

2.7.2 NORMALISATION

Once the variables have been selected for the classification, the next step is the standardisation of the input data. This step transforms the variables into rates or measures on the same scale enabling comparison. There are a number of normalisation techniques used to standardise data. The most commonly used techniques are z-scores, range standardisation, and principal components analysis.

Z-SCORES

Z-scores are the most commonly used technique for standardising variables when creating geodemographic classifications. Z-scores can emphasise the effect of outlying observations in the datasets, which may serve to highlight interesting patterns within the data: however, such observations can also adversely influence some clustering algorithms.

z-scores are defined by the following equation (2.1).

$$z_i = \frac{y_i - y_{mean}}{\sigma_y} \qquad (2.1)$$

Where y_i is the individual data value, y_{mean} is the mean of the data set, and σ_y is the standard deviation of the data values.

RANGE STANDARDISATION

Range standardization is another normalization technique for standardizing the variables. The ONS Output Area Classification (Vickers & Rees, 2007) uses range standardization for normalizing the variables. Range Standardization normalizes data values between the range of 0 and 1. Range standardization is less sensitive to outliers in the data set, but it can also obscure interesting patterns.

Range standardization is defined by the following Equation (2.2):

$$R_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \qquad (2.2)$$

Where y_i is the individual data value, $y_{\min}$ is the lowest data value, and $y_{\max}$ is the highest data value.

PRINCIPAL COMPONENT ANALYSIS

PCA (principal components analysis) transforms a number of correlated data inputs into a series of new uncorrelated principal component scores. PCA has previously been used in the creation of some geodemographic classifications, but is utilised less in contemporary classifications. A disadvantage of using PCA is that it can also potentially erase some interesting patterns from the dataset.

A detailed description of normalization techniques is given in Chapter 6 “Cluster Analysis for the creation of geodemographic classifications”.

2.7.3 CLUSTERING

The third step in creating a geodemographic classification is the creation of clusters or homogeneous groups.

The *k*-means clustering algorithm is used to create the finest level geodemographic classes. The *k*-means algorithm seeks to find the set of cluster centroids that minimises expression (2.3) below.

$$V = \sum_{x=1}^{n}\sum_{y=1}^{n}\left(z_x - \mu_y\right)^2 \qquad (2.3)$$

where *n* is the number of clusters, μ_y is the mean centroid of all the points z_x in cluster *y.* The *k*-means algorithm assigns a set of *n* seeds within the data set and then proceeds by assigning each data point to its nearest seed. Cluster centroids are then created for each cluster, and the data points are assigned to the nearest centroid. The algorithm then re-calculates the cluster centroids and repeats these steps until a convergence criterion is met (usually when switching of data points no longer takes place between the clusters).

A detailed description of different clustering techniques and their comparisons are given in Chapter 6 “Cluster analysis for the creation of geodemographic classifications”.

2.7.4 VISUALISATION

Clustering of data variables partitions data items (in the case of creating geodemographic classifications the “output areas” or “postcodes”) into homogeneous groups. Each area is assigned an identifier number representing its “geodemographics class”. Table 2.6 shows an example result of the cluster analysis.

Area	Cluster ID
Area 1	1
Area 2	2
Area 3	1
Area 4	3
Area 5	4
Area 6	3
Area 7	5
……	

Table 2.4: An example result of cluster analysis

Finally, each cluster id is given a colour code, and the colour code can be used to visualise the result of cluster analysis on a map. Figure 2.10 shows the Output Area Classification (Vickers & Rees, 2007) map around the area of Liverpool (OAC clusters output areas in the UK into 7 homogeneous groups).

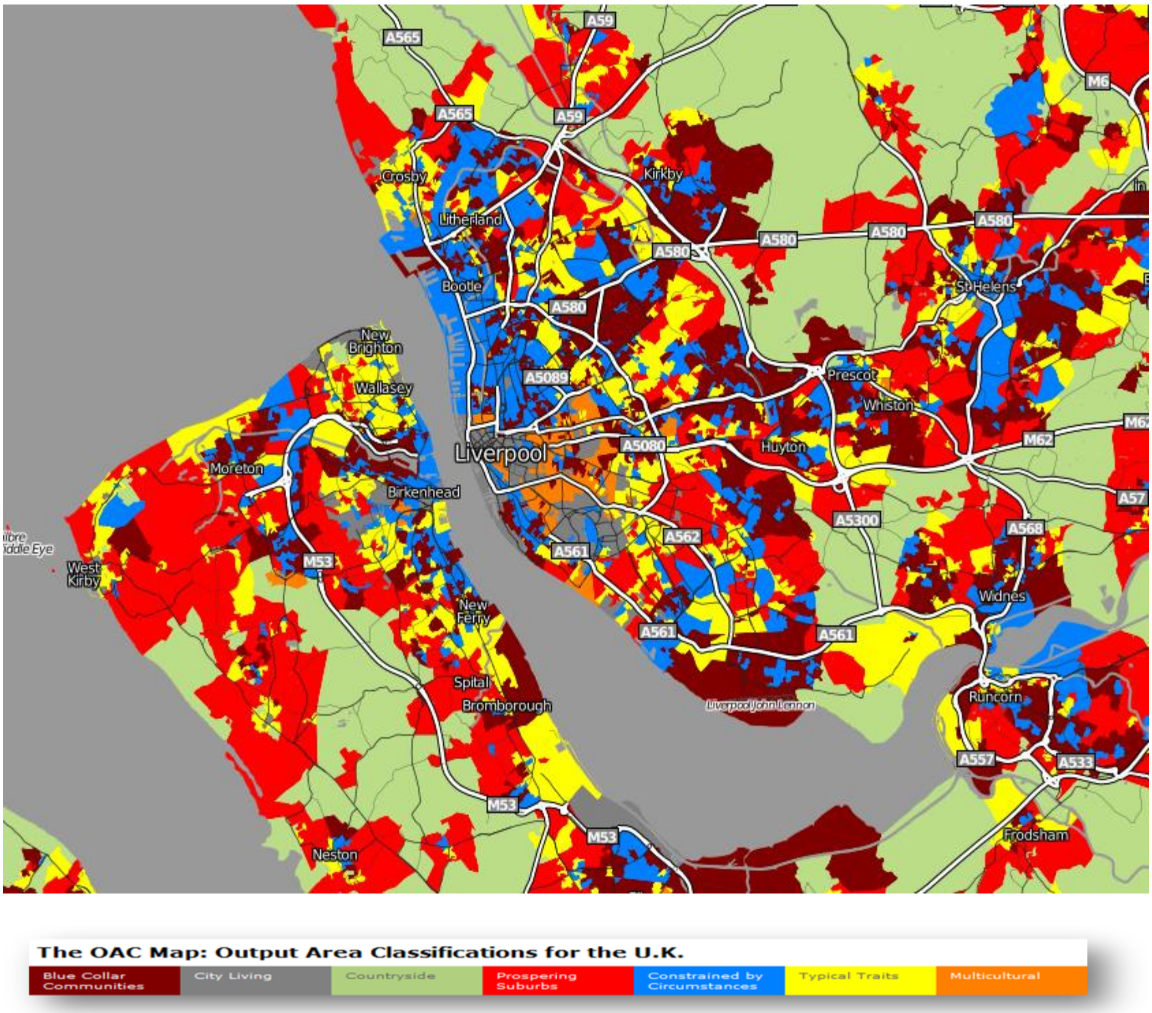


Figure 2.10: An example result of cluster analysis

Chapter 8 develop some case studies of creating different geodemographic classifications and explains the ways of visualizing the classification.

2.8 THE INTERNATIONAL DIMENSION

This chapter, so far, has focused on geodemographic classifications in the UK context. However, geodemographic classifications systems have also been implemented in other national settings. Burrows and Gane (2006) describe Jonathan Robbin pioneering work on computer based geodemographics in the USA, where he created PRIZM using funding from the US Department of Housing and Urban Development. PRIZM was focused on the allocation of housing grants between cities that have a history of rioting (Weiss, 2000: 142). PRIZM is now owned by a company called Claritas (Burrows and Gane, 2006). Harris et al (2006: 188) describe some of the geodemographic tools that have been developed in many countries of the world, including Australia, China, Denmark, Finland, France, Germany, Greece, Japan, Netherlands, New Zealand, Spain, Sweden, Norway, UK, and USA.

Geodemographic classifications also have an international dimension, with systems developed for multinational companies seeking common marketing strategies for targeting customers across many countries . According to Weiss (2000: 43):

"Throughout the world there's remarkable similarity in the way businesses are using the cluster technology - for analyzing trading areas, profiling customers, and driving media strategies. The increasing globalization of culture is also prompting multinational companies to look at clusters as a common marketing language to reach customers across many borders".

Thus, in an increasingly globalized world, there is a demand for classifications that transcend national boundaries. This is fraught with problems of deriving comparable data at comparable geographical scales of measurement, in order to derive classes that are themselves comparable between national areas. International geodemographic systems are also available (Brassington & Pettitt, 2006: 201). For example, Experian’s Mosaic Global covers 880 million people from many of the world's major economies, including North America, Europe, and Asia Pacific (http://www.uk.experian.com). Mosaic Global incorporates data from 24 national Mosaic classifications. However, it is not created by a single cluster analysis of data around different countries. Instead, it provides a framework into which the detailed geodemographic classes specific to each national market can be placed (Stroud, 2007: 142). A truly international geodemographic classification needs to arise out of a single cluster analysis on a data set which is consistent over the geographical areas of different countries.

There are many issues relating to the internationalisation of geodemographic classifications. Harris et al (2005: 193) describe the differences in census data attributes between different countries of the world. An international geodemographic system requires standardisation of data sources for building international classifications. This also applies to other non-census data sources. Another issue is the differences in neighbourhood geographies worldwide (Harris et al, 2005: 193). Any international geodemographic system should be based upon data aggregated at consistent geographical levels between different countries of the world.

There are some on-going developments in this quest. The EU INSPIRE (European Union INSPIRE) initiative (http://inspire.jrc.ec.europa.eu/) is an effort to establish an infrastructure for spatial information in Europe to support community environmental policies. This infrastructure seeks to archive information in consistent geographical levels for different countries across the EU. In the future, these spatial data structures could be useful in building transnational geodemographic classifications. Another development is the creation of consistent geographical boundaries around different countries of the world. Cheshire (2011) has created surname regions by using the 'Worldnames' database (discussed in Chapter 3 of this thesis) for 16 European countries. This example illustrates the creation of consistent boundaries independent of existing administrative geographies that could be useful for the aggregation of census and non-census data sources in geodemographic classifications.

However, this is just the beginning and there is a need of the creation of new data infrastructures, the coverages of which are not constrained by national or other essentially administrative boundaries. In the quest of the creation of new data sources which transcend national boundaries, the next chapter explains the creation of a data source of family names across 26 countries of the world. It sets out the issues of data collection from a number of data sources and how data may be standardised in a single format, even though it is assembled from multiple data sources across different countries and geographical areas. The next chapter also discusses important related issues of data aggregation and georeferencing in detail. In short, it is an effort to create a geographical data resource which is independent of zonal boundaries.

2.9 CONCLUSION

This chapter has provided an overview of the creation of geodemographic classifications. Area classification and geodemographics have a rich history and are useful for social measurement. However, there are problems with the use of commercial classifications – particularly in public sector applications in which life chances could be affected by misclassification. Therefore more intelligent systems are required. This chapter also has described the method to create geodemographic classifications. However, a more detailed description will be found in Chapter 6 "Cluster analysis for the creation of geodemographic classifications" and Chapter 8 "A pilot geodemographic decision support system".

The geodemographic classifications described and alluded to in this chapter are heavily (sometimes totally) reliant upon Census of Population data, but others are also reliant upon ancillary sources – whether available through open data government initiatives in the case of public sector data, or through lifestyles surveys in the case of some commercial solutions. Census data provide sound infrastructure for area classification in terms of their coverage – not least because of the legal requirement for households to complete Census enumeration forms – but lack content in terms of pertinent indicators of local social, economic and demographic structure, or adequate temporal updating. Ancillary data sources may fill these gaps, in terms of substantive relevance or temporal refresh.

However, the use of ancillary sources is not without problems. Open data may be available only for coarse areal units, or units that are incompatible with other data series. Commercial lifestyles data are of unknown provenance, as the characteristics of volunteer survey respondents may be very unrepresentative of non-respondents. These problems are compounded by the requirement to devise international geodemographic classifications that are inclusive, comparable and comprehensive. These provide severe impediments to the creation of geodemogrpahic classifications that are salient, timely and wide in geographical coverage. These challenges are not insuperable, however, as illustrated in the next Chapter in a discussion of the creation of an international names database that provides valuable indicators of population origins, migration patterns and cultural, ethnic and linguistic structure.

3 CHAPTER 3: THE CREATION OF A GEO-REFERENCED GLOBAL NAMES DATABASE

3.1 INTRODUCTION

As discussed in Chapter 2, the data inputs to geodemographic classifications typically comprise of numerous variables assembled from a range of sources including population census, survey data, government ‘open’ data and other transactional information such as credit scores, bad debts and directorships. However, there are numerous and innovative sources of data which currently remain underexploited either because of difficulty in their manipulation or lack of awareness of their existence or provenance. Previous research at University College London (Mateos et al. 2007) has demonstrate through the analysis of names that a variety of characteristics about migration, social mobility and ethnicity of populations can be extracted from (enhanced) registers of electors and the individual records of telephone directories. However, the geographic extent of this previous research was limited to Great Britain. In an increasingly complex and interconnected world (Wilson 1971), geodemographic classifications are being applied at international and indeed global levels. Another development in a different domain is demonstrated by Debenham et al. (2001) who indicate how workplace variables relating to a labour market can be integrated and used alongside residence based variables to create new classifications. Thus, a logical extension in the analysis of names is to consider their distribution at a global level. In this chapter, as an example of an innovative new data source the creation of a global names database is described.

In different parts of the world there are multiple ways in which family names are passed onto children or spouse which can be illustrated by a number of examples. A person’s name in Great Britain is usually composed of a forename and surname, and traditionally a female takes the male’s surname on marriage which is then typically inherited by any offspring they may have. However, in Spain, a child keeps the surname of both the father and the mother, so one name contains two surnames. For example “Aalquezar Felez Cristina” is a Spanish name with two surnames “Felez” and “Cristina”. Names in the Netherlands and Germany are sometimes composed with a prefix before the surnames. “A M Van Beek” is a Dutch name with surname “Van Beek”, where “Van” is a name prefix and “Beek” is the surname. Arab names also tend to include the name of father. “Abu bin Talib” is an Arab name where “Abu” is a given name and “Tablib” is the name of Abu’s father. Sometimes Arab names are very complex. “Fahr-ad-Din Abu Abdullah Muhammad Ibn Umer Ibn Al-Hasan Al-Hatib Ar-Razi” is such an Arab name where “Muhammad” is forename, “Ibn Umer Ibn al Hasan” is father’s or forefather’s name, and “Al-Hatib” denotes tribal descent.

There is an extended history of the study of names dating from an early example in 1874 when George Darwin investigated the frequency of marriages between cousins in England (Lasker 1985). This study further progressed to the use of names in many other fields. Colantonio et al. (2004) reflect how Crow and Mange in 1965, presented a formula for calculating an indicator of inbreeding from isonomy (a measure of the likelihood of partnering someone with the same name), which was a major development in the study of names and millions of surnames were analyzed by that formula in different fields.

Onomastics, or the study of the origins of names has an interdisciplinary range of applications in human genetics (Voracek and Sonneck 2007), anthropology (Lucchetti et al. 2005), public health/epidemiology (Degioanni et al. 1996) (Lagerberg et al. 2005), demography (Lewison and Kundra 2008), linguistics (Lucchetti et al. 2005), socioeconomic study (Lauderdale and Kestenbaum 2000), psychology (Dantzker and Freeberg 2003), computer science (Leino et al. 2003) and history (Burnard 2000). Within the field of Geography there have been numerous examples specifically looking at the onomastics of family names through the study of migration trends (Degioanni et al. 1996), cultural, ethnic and linguistic characteristics (Lauderdale and Kestenbaum 2000), and social mobility (Longley et al. 2007). Using the website http://gbnames.publicprofiler.org, much of this research is summarised in an online tool. For example, entering the surname Singleton reveals the concentration of the name in 1881 in the North West region on England, with a slightly more defuse distribution in 1998. The name appears very regional and has origins named after a village near Blackpool bearing the same name. The following figure (Figure 3.1) shows the distribution of surname Singleton in Great Britain for the years 1881 and 1998.

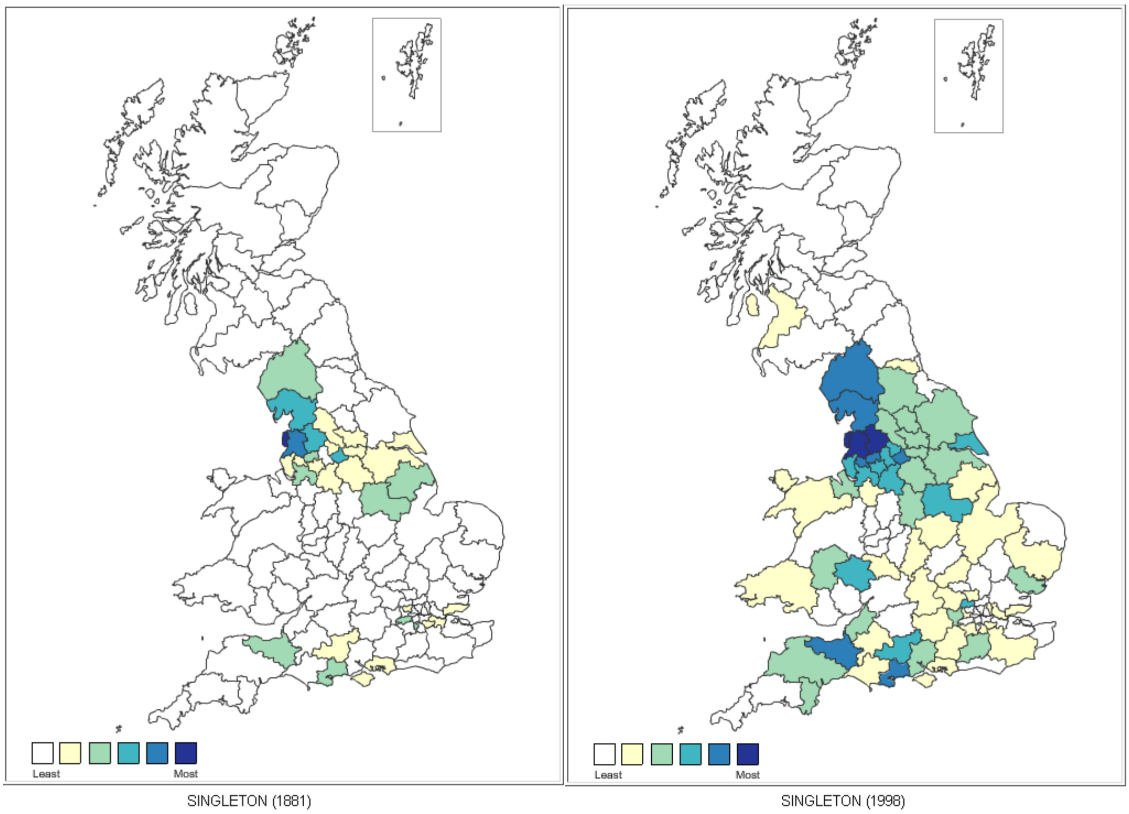


Figure 3.1: Distribution of surname Singleton for the years 1981 and 1998 (source: gbnames.publicprofiler.org)

A previously unexploited data source for spatial onomastics, and central to much of the research presented in the following chapters, is the increased availability of individual level digital data sources in the format of telephone directories and electoral rolls.

There are numerous providers of telephone directories that can be purchased digitally for a low cost. These data contain details of those people with telephone connections who agree for their information to be made public. These data are often at a household level, and will typically give details which include the family name along with address data such as postcode, city or state. Additionally, other sources of individual level data include registers of electors, again providing a variety of address data, typically at high spatial resolutions.

The large amount of different data sources and formats in which names data could be derived, creates a requirement for a standardised structure and database formats which makes for efficient information retrieval. This chapter explores the creation of a Worldnames database (containing 26 countries) created from a variety of different data sources. This chapter will explain the structure of the raw format of different data sources and how geographical variations in naming conventions were accommodated. It also explains in detail the process of geo-referencing the names using different data sources.

3.2 OVERVIEW OF DIFFERENCES IN NAMING CONVENTIONS

Individual level names data are available from a variety of different sources and this section provides an overview of those data which were used to create the 'Worldnames' Database. There are two major variations across the different unstructured data sources. Firstly is how names of people are represented and second is how address details are structured. We can divide these data sources into three categories:

- Telephone directories
- Electoral registers
- Surname Counts

3.2.1 TELEPHONE DIRECTORIES

Landline telephone directories were procured for the following countries:

- Austria
- Belgium
- Canada
- Denmark
- France
- Germany
- Hungary
- India
- Italy
- Luxembourg
- Netherlands
- Norway
- Poland
- Yugoslavia (Serbia, Montenegro, and Kosovo)
- Slovenia
- Spain
- Sweden
- Switzerland

- United States

Landline telephone directories typically include the names of householders who have subscribed to a landline service and who have consented to their inclusion in publicly available digital directories. As such, they are best viewed as representing total numbers of households, but exclude those that have opted out of this service as well, of course, as those that do not have landlines. Given the penetration of landline telephone services in most developed countries, it was assumed that these considerations were unlikely to result in bias in the distribution of names recorded, although this is less likely to be the case in developing countries. Specifically in India, where data was collected from 4 telephone directories each representing a city (Delhi, Mumbai, Hyderabad, Chennai).

Because each directory contained unstructured digital data these could not simply be entered into a standardized database. The various address and name encapsulations first had to be assessed. The following section presents these findings.

NAME ENCAPSULATION

Name encapsulation can be divided into three categories based on the way names were represented in telephone directory.

NAMES STARTING WITH FORENAMES (SLOVENIA, HUNGARY, AND DENMARK)

The telephone directories for these countries provided both forename and surname as part of a single full name string. Forename appeared at the start of the name string, and surname at the end. The full name string was divided into forename and surname using following algorithm:

1. Start from the end of full name. Traverse characters towards start. For example, if a name is “MAROLT ZVONE”, it will start traversing from character ‘E’ at the end.

2. When a space occurs, assign the word for surname. For example, in “MAROLT ZVONE” the space occurs after “MAROLT”, so “ZVONE” will be assigned as surname.

3. Keep traversing to left to see if remaining name contains prefixes. If yes, assign that prefix to surname as well. Rest will be forename. For example, in

“MAROLT van ZVONE” surname should be “van ZVONE” and part left i.e. “MAROLT” will be classed as a forename.

NAMES STARTING WITH SURNAMES (UNITED STATES, BELGIUM, LUXEMBOURG, ITALY, SERBIA & MONTENEGRO, AUSTRIA, GERMANY, AND NETHERLANDS)

The telephone directories for these countries again provided forename and surname as part of a name string. However in these data, the forename occurred at the end of the name string and surname at the start. In these names data, particular care had to be taken with prefixes to names such as Mr, Mrs or Dr. However, these obviously differed in each country. A sample of these prefixes is shown in Table 3.1.

Country	Example Prefix
United States	VAN, LES, LA, AL
Switzerland	DE, DA,DO, VAN, VON,DER
Canada	LESS, AL, DELLA, VON, DER, LI, ZUR
Belgium	VAN, DE, EL, VAN,DEN
Luxembourg	LA, DE, DOS, VAN, DEN
Italy	EL, DEL, LA, TEN, VON, DELLA
France	VAN, VON, DER, DE, LO, LA, AL, EL
Hungary	VAN, DER, DELLA, EL, LE
Denmark	DEL, DEN, AL, EL, VON, VAN
Yugoslavia (Serbia, Montenegro, & Kosovo)	DI, DEL, DER, LES, EL
Austria	ZUR, VAN, VON, DE, DER, DES, DEL,
Germany	VAN, VON, DER, DES, ZU
Poland	ZUR, VAN, VON, LE, EL, AL
Netherlands	AAN, VAN, DE, DER, AL, EL

Table 3.1: Example Country Name Prefix

The full name string was divided into forename and surname using the following algorithm:

1. Start at the beginning of the full name. Traverse characters towards end. For example, if a name is “Abbe Chad”, it will start traversing from character ‘A’ from the start.

2. When a space occurs, check whether first part is a prefix. If it is not, assign it as surname. If it is, keep traversing until the next space and assign the part as surname.
3. Assign the rest to forename.

NAMES ALREADY DIVIDED INTO FORENAMES AND SURNAMES (SWITZERLAND, SPAIN, CANADA, FRANCE, SWEDEN, NORWAY, POLAND, AND INDIA)

Telephone directories for these countries provided forenames and surnames as individual fields. So there was no need to separate forenames from surnames. Thus, these were entered directly into the database.

ADDRESS ENCAPSULATION

Address information varied between different telephone directories. Most of the telephone directories represented addresses by multiple fields, each of which corresponded to a spatial unit. For example address information for the United States was stored in the form of ‘State’, ‘City’, and ‘Postcode’ but the address information for Switzerland was stored as ‘City’ and ‘Postcode’. Because of the different address referencing conventions, a standard taxonomy was developed for address representation for the world names database. This taxonomy consists of the representation of each individual’s address by a combination of State, City, Area code, Postcode, and Street address. Each part of the address taxonomy is defined as follows:

STATE

A state represents the biggest spatial area inside a country. For example, states in United States of America represent 50 individual states. However, the 'state' field in Spain is taken to be the province in which the address lies. In each of these cases, a state contains multiple cities.

CITY

A city is an area which has administrative status and which has been classified as city according to its population or historical status. A city contains multiple postcodes.

AREA CODE

An area code is an administrative area representing a city/area and for which state and city are not specified in the telephone directory. An area code also contains multiple postcodes.

POSTCODE

A postcode is a spatial area possessing a general postcode (in the United Kingdom), zip code (United States) etc. A postcode is the collection of different houses (addresses).

ADDRESS/HOUSE NUMBER

An address/house number is the finest spatial area given in the telephone directories. An address represents an individual telephone subscriber.

Table 3.2 shows the address encapsulation in the world names database for the individual records taken from various telephone directories.

Country	State/ Province/ County	City/Area	Area Code	Postcode	Address/House Number	Smallest geography
Belgium	-	X	-	X	X	Address
Luxembourg	-	X	-	X	X	Address
Italy	-	X	-	X	X	Address
France	-	X	-	X	X	Address
Norway	-	X	-	X	X	Address
Denmark	-	X	-	X	X	Address
Austria	-	X	-	X	X	Address
Germany	-	X	-	X	X	Address
Canada	X	X	-	X	X	Address
United States	X	X	-	X	X	Address
Hungary	X	X	-	X	X	Address
India	X	-	-	-	-	State
Poland	-	X	X	-	X	Address
Slovenia	-	X	-	X	X	Address
Switzerland	-	X	-	X	X	House Number
Spain	X	X	-	X	X	Address
Sweden	-	X	-	X	X	Address
Yugoslavia	-	X	-	-	X	House Number

Table 3.2: Address Encapsulation for telephone directories

3.2.2 ELECTORAL REGISTERS

Electoral registers contain registration data for those people who are (or are about to become) eligible to vote. Access to publicly available versions of these data is possible for some countries, though public versions may exclude voters who opt out of inclusion in them. Electoral registers do, of course, exclude individuals who are not entitled to vote, and whilst the exclusion of minors is unlikely to bias names distributions, the absence of many recent immigrants from abroad is potentially more problematic. Electoral register data were obtained for the following countries:

- Great Britain
- New Zealand
- Argentina

In the case of Great Britain, a commercially enhanced version of the Electoral Roll was purchased in order to include non-electors and 'opt out' electors. (Electoral Roll data for Northern Ireland have not been made available for many years.) All electoral rolls were supplied with names data as individual fields for both forenames and surnames. Thus, there was no need to apply any name separation algorithms. Following table (Table 3.3) outlines the address encapsulation of electoral registers.

Country	State/ Province/ County	City/Area	Area Code	Postcode	Address/House Number	Smallest geography
United Kingdom	-	X	-	X	X	Address
New Zealand	-	X	-	X	X	Address
Argentina	-	X	-	-	X	Address

Table 3.3: Address Encapsulation for Electoral registers

3.2.3 SURNAME COUNTS

For some countries where individual level data were not available as either telephone directories or electoral rolls, surname counts were obtained at a number of levels of geographic aggregation. Thus, these count data represent the number of occurrences of a particular Surname within a geographical region.

Surname Counts were obtained for the following countries:

- Ireland
- Australia
- Japan

These data sources contained only surname counts and therefore there was no need to apply a name separation algorithm. Furthermore, the address references were singular and related to specific areal aggregations within each country. For Ireland and Australia these aggregations were Postcode Area level, and for Japan the administrative aggregation was the prefecture. The following table shows the number of geographical regions and total number of surnames for each of these countries. Following table (Table 3.4) outlines the number of geographical regions and their corresponding surnames for the three countries (Ireland, Australia, and Japan).

Country	No. of geographical regions	Total number of surnames
Ireland	34	46507
Australia	197	12269
Japan	47	49746

Table 3.4: No. of Geographical regions and Surnames

3.3 ACCOMMODATING GEOGRAPHIC VARIATIONS IN NAMING CONVENTIONS:

There are multiple variations in naming conventions across the globe which need to be accommodated prior to entering in the database. This requires a process of cleaning which standardises the format of the names. A name cleaning algorithm was designed and is outlined in the following steps.

STEP 1 – SEPARATING THE NAME INTO PARTS BASED ON SPACES

The algorithm starts by separating the name into parts based on the occurrence of spaces in the name. There are 3 kinds of spaces which this algorithm takes into account. The Unicode values for those spaces are U+0020, U+00A0, and U+2000.

A string of “Paul Longley” would be separated in two parts “Paul” and “Longley”. “De Ad-John” would be separated into “De” and “Ad-John”. However, “John-Cine” would remain as it is because this name does not contain any spaces, only hyphens.

STEP 2 – REMOVING PREFIXES AND SUFFIXES

The second step of the algorithm involves removing those parts of the name which are prefixes and suffixes. Two tables for prefixes and suffixes were used for this purpose.

The table of prefixes is given in the Appendix 2 section, and table of suffixes is given in the Appendix 3 section.

The prefix is the part of a name found at the start of a name string and the suffix is found at the end. However, this algorithm checks for typical prefix and suffix values that are found anywhere within the name string. This is due to the fact that many German, French, Dutch, and US names have prefixes at the end of the name string. Additionally, in some of the data, suffixes occurred in the middle of the name string. This normally occurred in the data of Netherlands and Denmark.

For example, the name “Mr. A Singleton” would be constructed of three parts after Step 1 in the algorithm: “Mr.”, “A”, and “Singleton”. Step 2 of the algorithm extracts from these three parts known prefixed and suffixes. Thus, after Step 2 the name will only have two parts left: “A” and “Singleton”. In the same way “John Jr.” will have one part left “John”. And “Longley Mr.” will have only “Longley” left.

STEP 3 – STANDARDIZING THE NAME

This step combines the parts of the name into a single variable and then standardises them. Standardisation is required due to the irregular positioning of spaces, apostrophes and dashes. An example of this irregularity would be in the name “O’brien” and “O’ brien”, where both are clearly the same name, but the second version has a space after the apostrophe. Three kinds of apostrophes need to be considered, which are `, ´ and '.

After this step, all of the names are in the same format. For example: "O' Neil", "O 'Neil", and "O'Neil" all became "O'Neil". And, "A –Singleton", "A- Singleton", and "A – Singleton" all became "A-Singleton".

STEP 4 – SEPARATING THE STANDARDIZED NAME INTO PARTS BASED ON SPECIAL CHARACTERS

This step of the algorithm separates the standardized name into parts, based on a series of special characters. A sample java source code of this process in tabulated in the Appendix 5 section. Thus "A-Singleton" will be separated into two parts "A" and "Singleton". In the same way, "DE APPRIAH-OFFORE-BAWE" will be separated in four parts "DE", "APPRIAH", "OFFORE", and "BAWE".

STEP 5 – REMOVING NAME PARTS COMPRISING 1 OR 2 CHARACTERS

This step builds on the parts of the name obtained from the previous step.

If all parts of a name have length less than or equal to 2 characters, then this step won't do anything and goes to step 6. For example, if a name is "M A", then both parts ("M" and "A") have length which is less than 2, so this algorithm will go to step 6. In the same way, "M,A," will have two parts "M" and "A" and this step won't change anything and goes to step 6.

If the length of any name part is greater than 2, the following operations are performed:

1. Check all parts and see if any part does not contain any vowels. Delete the parts from memory which do not have vowels. So, "John GB" will become "John" and "MIAN-PO" will remain "MIAN-PO" because both parts have vowels.

2. Delete those parts which have a length less than or equal to two, and which are connected to the whole name using a space. So, "De Appriah-Offore-Bawe" will become Appriah-Offore-Bawe. This part had De which does contain a vowel, but it is of length 2 and it is connected to the name via a space so it is deleted.

STEP 6 – COMBINE PARTS AND REMOVE LEADING AND TRAILING CHARACTERS

This step combines the remaining name parts into one single name and removes leading and trailing special characters. So "&Singleton" would become "Singleton".

STEP 7 – REPLACING UNICODE CHARACTERS WITH THEIR CORRESPONDING ASCII CHARACTERS

This final step checks the occurrence of any Unicode characters in the name and then replaces the Unicode characters with their corresponding ascii characters. Replacement of Unicode characters with their corresponding ascii characters is done using a conversion table which holds an ascii character entry for each Unicode character. This table is specified in the Appendix 4 section.

3.4 CHOOSING AN APPROPRIATE DBMS (DATABASE MANAGEMENT SYSTEM)

There are many data structures and database systems that can be used to store the data in table format. Two relational database management systems are in widespread use: Oracle (www.oracle.com) and SQL Server (www.microsoft.com/sqlserver). Each has particular properties in the ways in which it stores, manages, retrieves and displays data to users. Oracle (2005) suggests that Oracle outperforms SQL Server in performance, using its Concurrency model, Indexing, Partitioning, and Parallel execution of processes. This system was thus chosen to manage the names data.

Oracle has versions of the DBMS which offer different functionality. The most recent version of Oracle is 11g, which was used in this study. Data were loaded into individual tables of an Oracle database: Table 3.5 shows the names of the tables used to store data for different countries.

Country	Oracle table name
Belgium	TBL_BE_TEL_DIR_CLEANED
Luxembourg	TBL_LU_TEL_DIR_CLEANED
Italy	TBL_IT_TEL_DIR_CLEANED
France	TBL_FR_TEL_DIR_CLEANED
Norway	TBL_NO_TEL_DIR
Denmark	TBL_DK_TEL_DIR_CLEANED
Austria	TBL_AT_TEL_DIR_CLEANED
Germany	TBL_DE_TEL_DIR_CLEANED

Canada	TBL_CA_TEL_DIR_CLEANED
United States	TBL_US_TEL_DIR_CLEANED
Hungary	TBL_HU_TEL_DIR_CLEANED
India	TBL_IN_TEL_DIR_CLEANED
Poland	TBL_PL_TEL_DIR_CLEANED
Slovenia	TBL_SI_TEL_DIR_CLEANED
Switzerland	TBL_CH_TEL_DIR_CLEANED
Spain	TBL_ES_TEL_DIR_CLEANED
Sweden	TBL_SE_TEL_DIR
Yugoslavia (Serbia, Montenegro, & Kosovo)	TBL_YU_TEL_DIR_CLEANED
United Kingdom	TBL_GB_CACI_2007
New Zealand	TBL_NZ_ER
Argentina	TBL_AR_ER_CLEANED
Ireland	TBL_IE_SURNAME_COUNTS
Australia	TBL_AU_SURNAME_COUNTS
Japan	TBL_JP_SURNAME_COUNTS
Netherlands	TBL_NL_TEL_DIR_CLEANED

Table 3.5: Table names for different countries in Oracle database

3.5 GEO-REFERENCING NAME ADDRESS DATA

Geo-referencing is the assignment of locations to atoms of information (Longley et al. 2011: 124). It is required in order to uniquely differentiate places from one another. There are millions of addresses in names database, and it is required that each address should be uniquely identified so that the names data could be plotted on maps easily. Geo-referencing has a rich history and giving place names to areas is a fundamental human trait. Beyond assigning place names to areas, Postal Addresses and Postal Codes (Longley et al. 2011: 127) have been developed since the 19th Century in order to expedite mail delivery. Nowadays, geo-referencing is not just limited to place names or postal codes: new geo-referencing systems have been developed so that information may be easily linked with a particular location. The National Grid of Great Britain is a geo-referencing system that assigns every point in England, Scotland and Wales to a cell of 100 km grid cell (see Figure 3.2).

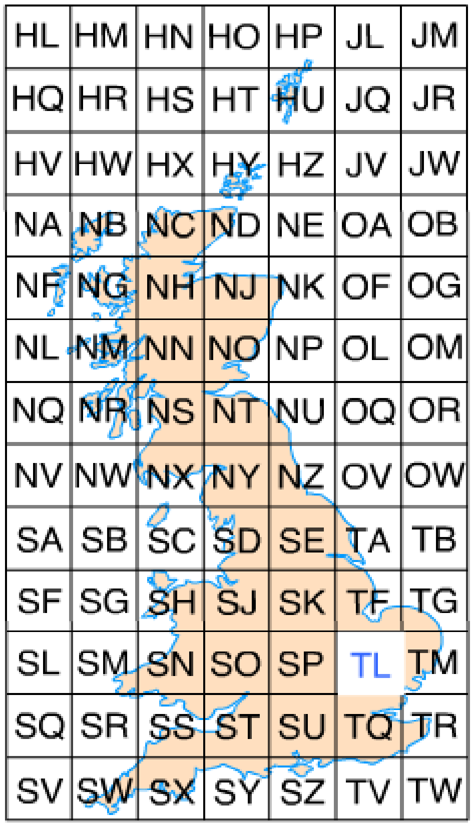


Figure 3.2: The National Grid of Great Britain

Each point in any geo-referencing system may be identified by latitude and longitude units of geo-processing. Latitude is measured from the Equator, where positive values increase northwards and negative values increase southwards, and longitude is measured east or west from Greenwich England to the International Date Line.

Once the names data had been cleaned and loaded into the database, a geo-referencing process was required to assign an *X* (latitude), *Y* (longitude) coordinate to each individual address. This process used the Geonames (www.geonames.org) free gazetteer to assign (*X,Y*) coordinates to the names data. Geonames is a geographical database containing eight million geographical names. It runs under a creative commons license (www.creativecommons.org), allowing the free use of data. Data from Geonames can either be downloaded as a CSV file or using an XML service. If data are downloaded as a CSV file, they can be uploaded to a database directory: however, if an XML service is used to obtain live XML feeds of data from Geonames, then coordinates may be archived to a file and then uploaded to a database. Data quality of the Geonames database varies between different data sources because of the update of data by general public. Geonames allows the updating of data through a wiki interface. Users are invited by Geonames to improve the data manually by editing names of places, postcodes, latitudes, longitudes, and elevation information.

For the geo-referencing process, geonames data were downloaded from the website and uploaded in following Oracle tables:

GNPC: GEONAMES POSTCODE GAZETTEER

The Geonames postcode gazetteer contains the latitude and longitude information for the postcodes of different countries. It contains an entry of latitude (*X*) and longitude (*Y*) for postcodes in each country. However, this table does not contain all the postcodes for all the countries. The structure of this table is shown below (Table 3.6):

Field name	Description
Country Code	Two digit IATA code for the representation of a country
Postcode	Postcode entry for a country
Place Name	Name of the place that the postcode lies in. These places names represent cities or towns
Admin name 1	Name of the biggest geographical unit in a country. For example, in United States biggest geographical unit is the State
Admin name 2	Name of the second biggest geographical unit in the country. For example in United states second biggest geographical unit is County
Latitude	Latitude of the postcode
Longitude	Longitude of the postcode

Table 3.6: Structure of Geonames postcode gazetteer table

GNPL: GEONAMES PLACENAMES GAZETTEER1

This gazetteer contains latitude (*X*) and longitude (*Y*) information for places names in a country. This table was required because the Geonames postcode gazetteer does not contain the entries of all postcodes for all countries. The structure of this table is shown below (Table 3.7):

Field name	Description
Country Code	Two digit IATA code for the representation of a country
Place name	This field represents a Place name entry. A place name could be a county, city, or town
Latitude	Latitude of the postcode

Longitude	Longitude of the postcode

Table 3.7: Structure of Geonames postcode gazetteer1 table

GSPL: GEONAMES PLACENAMES GAZETTEER2

This gazetteer also contains latitude (*X*) and longitude (*Y*) information for place names in a country. This table was required because it contains many more entries of place names than the Geonames placename gazetteer1. The structure of this table is shown below (Table 3.8):

Field name	Description
Country Code	Two digit IATA code for the representation of a country
Place name	This field represents a Place name entry. A place name could be a county, city, or town
Latitude	Latitude of the postcode
Longitude	Longitude of the postcode

Table 3.8: Structure of Geonames postcode gazetteer2 table

Additionally the following non geonames tables were used for UK and Canadian addresses:

NSPD: NATIONAL STATISTICS POSTCODE DIRECTORY

The National Statistics Postcode Directory contains the entries of WGS84 geographical coordinates for the postcode of United Kingdom. This table is a repository of United Kingdom postcodes with the reference to their WGS84 geographical coordinates. The structure of this table is shown below (Table 3.9):

Field name	Description
Postcode	This field shows an entry of postcode in United Kingdom
East Ref	WGS84 East reference. This could be converted to Latitude.
North Ref	WGS84 North reference. This could be converted to Longitude.

Table 3.9: Structure of National Statistics Postcode Directory table

UKER: UK ELECTORAL ROLL ADDRESS TABLE

This table contains the Electoral Roll data for the Great Britain. Electoral Roll data comprise the names and addresses of people who are registered to vote in United Kingdom, but are not available for Northern Ireland.

CPD: CANADIAN POSTAL DIRECTORY

The Canadian postal directory contains the latitude (*X*) and longitude (*Y*) information for the postcodes of Canada. This table is a repository of Canadian postcodes with linkage to their latitudes and longitudes. The structure of this table is shown below (Table 3.10):

Field name	Description
Postcode	This field shows an entry of postcode in Canada
City	Name of the city containing the postcode
Province	Name of the province containing the postcode
Latitude	Latitude of the postcode
Longitude	Longitude of the postcode

Table 3.10: Structure of the Canadian Postal Directory table

3.5.1 CREATION OF THE POSTAL MASTER TABLE

Prior to geocoding the addresses in the surnames database, a Postal Master table was created with the following fields:

- Location ID
- Country code
- Postcode
- City
- State
- Latitude
- Longitude

The first five of these fields were populated by aggregating the TELDIR table for each country with a unique combination of postcode, city, and state (state for the countries where it was available). Location ID is a unique identifier generated with an Oracle Sequence Object, and it uniquely identifies a combination of state, city, and postcode in a telephone directory. The Postal Master table was then geo-coded in three phases.

- Phase1: Geo-coding by Geonames Postcode Gazetteer, which applies to: USA, Netherlands, Germany, Belgium, Luxemburg, France, Denmark, Hungary, Slovenia, Austria, Italy, Switzerland, Spain, and Sweden.
- Phase2: Geo-coding by Geonames and GNS Placename Gazetteer (GSPL & GNPL) which applies to: USA, Netherlands, Germany, Belgium, Luxemburg, France, Denmark, Hungary, Slovenia, Austria, Italy, Switzerland, Spain, and Sweden.
- Phase3: Geo-coding of the rest of the countries, which applies to: Yugoslavia (Serbia, Montenegro, and Kosovo), Romania, Canada, United Kingdom, and Poland.

PHASE1: GEO-CODING BY GEONAMES POSTCODE GAZETTEER

The first stage of the geo-coding process was to match the Postal Master table against the Geonames postcode gazetteer (GNPC). The following table (Table 3.11) shows the format of the postcodes in both the Geonames postcode gazetteer (GNPC) and the postal master table.

Country	Postcode formats in GNPC	Postcode formats in postal master table
Belgium	4 digits format (e.g. 2660)	4 digits format (e.g. 2660)
Luxembourg	4 digits format (e.g. 4247)	4 digits format (e.g. 4247)
Italy	5 digits format (e.g. 18038)	5 digits format (e.g. 18038)
France	5 digits format (e.g. 38960)	5 digits format (e.g. 38960)
Norway	4 digits format (e.g. 3183)	4 digits format (e.g. 3183)
Denmark	4 digits format (e.g. 8870)	4 digits format (e.g. 8870)
Austria	4 digits format (e.g. 8330)	4 digits format (e.g. 8330)
Germany	5 digits format (e.g. 40789)	5 digits format (e.g. 40789)
Canada	6 digits format (e.g. T1C1R3)	6 digits format (e.g. T1C1R3)
United States	5 digits format (e.g. 96941)	5 and 7 digits formats (e.g. 74525, or 73401-1034)
Hungary	4 digits format (e.g. 2626)	4 digits format (e.g. 2626)
Poland	2+3 digit format (e.g. 00-001, 00-002)	2 digits format (e.g. 71, 72, 74)
Slovenia	4 digits format (e.g. 8293)	4 digits format (e.g. 8293)
Switzerland	4 digits format (e.g. 8404)	4 digits format (e.g. 8404)
Spain	5 digits format (e.g. 28037)	5 digits format (e.g. 28037)

Sweden	5 digits format (e.g. 23044)	5 digits format (e.g. 23044)
Netherlands	4 digits format (e.g. 9400)	4+2 digit format (e.g. 3356CP, 3137RR)

Table 3.11: Postcode formats for different countries

The process works by matching ‘city’ and ‘postcode’ fields of the Postal Master table with the ‘place_name’ and ‘postcode’ fields of GNPC table. If this match does not work, then latitudes and longitudes in the Postal Master table are updated using the averaged latitudes and longitudes for the matches performed on some different fields. E.g. for the records of United States, the process requires a match to be performed between the ‘state’ and ‘postcode’ fields of the Postal Master table and the ‘admin_name1’ and ‘postcode’ fields of GNPC. Because this match is at much higher geography and it can return multiple records, an average of the resultant latitudes and longitudes are assigned to the record in Postal Master Table. The following table (Table 3.12) shows this process of multiple results returned by the match of the ‘state’ and ‘postcode’ fields of Postal Master table with the ‘admin_name1’ and ‘postcode’ fields of the GNPC for United States and the conversion of the returned latitudes and longitudes to averages.

For, Country= United States, State=Ohio, postcode=44334, GSPL returns following records:

Country Code	Postcode	Admin1 Code	Admin1 Name	Place Name	Latitude	Longitude
US	44334	OH	Ohio	Akron	41.129000	-81.54000
US	44334	OH	Ohio	Fairlawn	40.397000	-83.328000

Table 3.12: Records returned by GNPC

By averaging latitudes and longitudes of the returned records, we get the following fields (Table 3.13):

Country Code	Postcode	Admin1 Code	Admin1 Name	Latitude	Longitude
US	44334	OH	Ohio	40.763	-82.434

Table 3.13: Averaged latitudes and longitudes

The following is a description of this phase when applied to different countries.

USA

The GNPC table contains postcodes for the USA with a five digits format (e.g.96941). However TELDIR for the USA contains postcodes with five and seven digits formats (e.g. 74525, or 73401-1034). Furthermore, the fields 'admin_code1', and 'place_name' fields of GNPC respectively contain 'US state code' and 'US city name'. Using these fields, this step of geo-coding for the USA is completed by matching on state, city and the first five digits of the postcode from the Postal Master table with the admin_code1, place_name, and postcode fields from the GNPC table.

For the records in the Postal Master that remained ungeo-coded after this process, a further match was performed between state and the first five digits of the postcode from the Postal Master table with the 'admin_code1' and 'postcode' fields of the GNPC table. The average latitude and longitude of the matches found in GNPC were then assigned to that record in the Master table.

The following tables (3.14 & 3.15) show the record in the postal master table and record returned by GNPC.

Country Code	Postcode	State	City
US	99553	AK	Akutan

Table 3.14: Record in Postal Master Table

Country Code	Postcode	Admin1 Code	Admin1 Name	Latitude	Longitude
US	99553	AK	Alaska	55.431000	-162.558000

Table 3.15: Record returned by GNPC

GERMANY

The Geo-coding of Germany was the most complex because of the large volume of entries without postcodes in the TELDIR table. First, a temporary table was created by grouping the TELDIR table for Germany by city, postcode, and area_code (extracted from the digits in telephone directory contained within brackets of length 2-5 digits). This table was called tbl_temp_DE. Then, the following two steps were taken:

Firstly, each instance of ‘city’ and ‘postcode’ in the table tbl_temp_DE was matched against the ‘place_name’ and ‘postcode’ fields of GNPC. The table tbl_temp_DE was then updated with the latitude and longitude of the matches. Secondly, for the records in tbl_temp_DE which remained un-geocoded, each instance of ‘postcode’ was matched with the postcodes from the GNPC and the averaged latitude and longitude of the matches were updated in tbl_temp_DE.

The following tables (3.16 & 3.17) show the record in the postal master table and the record returned by the GNPC.

Country Code	Postcode	State	City
DE	01945	-	Lindenau

Table 3.16: Record in Postal Master Table

Country Code	Postcode	Admin1 Code	Admin1 Name	Latitude	Longitude
DE	01945	BB	Brandenburg	51.400000	13.733000

Table 3.17: Record returned by GNPC

NETHERLANDS

The GNPC table contains postcodes for the Netherlands in 4 digit format (e.g. 9400). However the Netherlands TELDIR table contained postcodes with a 4+2 digit format (e.g. 3356CP, 3137RR) – that is, at a higher resolution than the postcodes of the GNPC. Furthermore, the 'place_name' field of GNPC represents 'city name'. Therefore geo-coding for Netherlands was done by matching the fields 'city', and 'first 4 digits of postcode' from Postal Master with 'place_name' and 'postcode' fields of GNPC table. For the records in the Postal Master table that remained un-geocoded, a match on the 'first 4 digits of postcode' from the Postal Master table was performed against the 'postcode' field of GNPC table. The average of the latitude and longitude of these matches was then stored in the Postal Master table.

The following tables (3.18 & 3.19) show the record from the postal master table and the record returned by GNPC.

Country Code	Postcode	State	City
NL	9408	-	Assen

Table 3.18: Record in Postal Master Table

Country Code	Postcode	Admin1 Code	Admin1 Name	Latitude	Longitude
NL	9408	01	Provincie Drenthe	52.997000	6.562000

Table 3.19: Record returned by GNPC

ELEVEN STANDARD POSTCODES COUNTRIES: BELGIUM, LUXEMBURG, FRANCE, DENMARK, HUNGARY, SLOVENIA, AUSTRIA, ITALY, SWITZERLAND, SPAIN, SWEDEN

The format of the postcodes in the GNPC table for the above twelve countries is the same as the format of the postcodes found in the individual TELDIR tables of each country. Furthermore, the 'place_name' field in GNPC contains the 'city name' for each of the above listed countries. Therefore, geo-coding of these countries was done by matching the 'city', and 'postcode' fields of the Postal Master table with the 'place_name' and 'postcode' fields of the GNPC table. For the records in the Postal Master that remain non-geocoded, a match between the 'postcode' in the

Postal Master table was done against the 'postcode' field of GNPC table. The average of the latitude and longitude of the matches was then stored in the Postal Master Table.

The following tables (3.20 & 3.21) show the record in the postal master table and record returned by GNPC, for Belgium (BE).

Country Code	Postcode	State	City
BE	1060	-	Brussels

Table 3.20: Record in Postal Master Table

Country Code	Postcode	Admin1 Code	Admin1 Name	Latitude	Longitude
BE	1060	BEBRU	-	50.842000	4.353000

Table 3.21: Record returned by GNPC

PHASE2: GEOCODING BY GEONAMES PLACENAME GAZETTEERS (GSPL & GNPL)

This section explains the second phase of the geo-coding process where the remaining non-geocoded records in the Postal Master table are assigned a latitude (*X*) and longitude (*Y*) using the place name gazetteers (GNPL or GSPL), as opposed to a postcode gazetteer.

These steps apply only to records in the Postal Master table with a 'multi-city postcode', that is, a postcode apparently lying in two different cities. So it might be possible that those cities are far away from each other, but they have the same postcode. This step is necessary to further refine the geocoding process, because it is possible in PHASE1 that the same latitude and longitude is assigned to 'multi-city postcodes' if those cities are in same 'state' (e.g. In United states where matches were also performed based on State and the first 5 digits of postcodes). In this phase, postcodes which lie in multiple cities and which have a corresponding entry of their cities in the Geonames placename gazetteers were assigned the latitude (X) and longitude (Y) of the matched city in the Geonames Placename gazetteer.

For each country, a series of steps were implemented. These are described below:

USA

'State' and 'city' were matched with the 'admin_cod1' and 'ascii_name' fields of the GNPL. The following procedures were invoked.

For each city for that has an exact match found in GNPL, take the difference between the latitude (*X*) and longitude (*Y*) already present in the Postal Master table (assigned in Phase 1) and the latitude (*X*) and longitude (*Y*) of the exact match in the GNPL. If the difference is less than 1 decimal degree, then update the Postal Master table with latitude (*X*) and longitude (*Y*) of the exact match in GNPL, otherwise keep the existing latitude (*X*) and longitude (*Y*). For the records in the Postal Master table that remain without a georeference, self join the table based on 'state', 'city', and first '2 digits of postcode'. Calculate the average of the latitude and longitude of the matches and update them in the Postal master table. For each such postcode in the Postal Master table, select 'state' and 'city' and match it with the 'admin_cod1' and 'ascii_name' fields of (GNPL). If an exact match is found, update the latitude and longitude.

THIRTEEN COUNTRIES: GERMANY, NETHERLANDS, BELGIUM, LUXEMBURG, FRANCE, DENMARK, HUNGARY, SLOVENIA, AUSTRIA, ITALY, SWITZERLAND, SPAIN, SWEDEN

Select 'city' in the Postal Master table and match it with the 'short_name' field in the GSPL. For each city where there is an exact match found in GSPL, take the difference between the latitude (*X*) and longitude (*Y*) of the Postal Master table (as assigned in Phase 1) and the latitude (*X*) and longitude (*Y*) in GSPL. If the difference is less than 1 degree, update the Postal Master with latitude (*X*) and longitude (*Y*) of the exact match in GSPL, otherwise keep the existing latitude and longitude. This process is applied to all countries except Germany.

For Germany, self join the table based on 'city' and the first 2 digits of the postcode. Calculate the average latitude (*X*) and longitude (*Y*) of the matches and update them in the Postal Master table. For the records in Postal Master that remain ungeocoded, self join the table based on 'city' and calculate the average latitude (*X*) and longitude (*Y*) of the matches and update them in the Postal Master table.

PHASE3: GEO-CODING THE REMAINING COUNTRIES

For a number of countries, alternative methods of geocoding were available. These are detailed in this following section.

UNITED KINGDOM

Geo-coding of the UK is done by joining the UKER table (tbl_gb_caci_2007_addresses) to the National Statistics Postcode Directory NSPD table (tbl_uk_nsp_wgs84) by standardizing the format of the unit postcodes in both tables through removal of the internal spaces.

YUGOSLAVIA (SERBIA, KOSOVO AND MONTENEGRO)

The Yugoslavia TELDIR table did not contain postcodes, although the telephone book areas to which the entries related could be extracted. Thus, the country is divided in 9 telephone book areas, 7 for Serbia, one for Kosovo and one for Montenegro. These will be referred to here as 'State'. For the geo-coding of Yugoslavia, a temporary table was created by grouping state, city, and area_code (this area_code comes from the first two digits of telephone number). This table was called tbl_temp_YU.

Then the following steps were applied sequentially:

1. Each instance of 'city' field in tbl_temp_YU is matched with the 'city' field of GSPL to update the latitudes and longitudes of those entries in tbl_temp_YU. If more than one match is found, do nothing at this step.

2. For the records which are still left non-geocoded in tbl_temp_YU, apply the following steps sequentially.

    a) Join the table tbl_temp_YU with itself using the 'state' and 'city' fields. If there are some records found, average the latitudes and longitudes of the matches and update the non-geocoded entries in tbl_temp_YU.

       The result of this step was that 1,002 records were updated.

    b) For the records still left ungeocoded, join the table tbl_temp_YU using the 'state' and 'area_code' fields. If there are some matching records found, average the latitudes and longitudes of the matches and update the ungeocoded entries in tbl_temp_YU.

The result of this step was that 36 records were updated.

c) For the records still left ungeocoded, join the table tbl_temp_YU with itself based on the 'state' field only. Average the latitudes and longitudes of the matches and update the non-geocoded entries in tbl_temp_YU.

   The result of this step was that 3 records were updated.

This latter step involved manual correction of some of the records. This step was needed because some of the cities were displaced from their known 'state' (i.e. telephone book area) to a different one after completing the first two steps above. Therefore, all the records were imported into ArcGIS and locations were assigned a colour based on their 'state' (i.e. telephone book area). Cities in the wrong state location were identified using Internet searches for their exact latitudes and longitudes, and were corrected manually. All the geocoded records from tbl_temp_YU were transferred to the Postal_Master table. Tbl_temp_YU was deleted afterwards.

CANADA

For Canada, data from ZipWorld, a company that sells zip codes databases (www.zipcodeworld.com), was uploaded in a table CPD (tbl_ca_postal_code). Postcodes in Canada's TELDIR were of the same format as the postcodes found in the CPD file. In this way, the process involved matching 'state' and 'postcode' fields of the Postal Master table with the 'state' and 'postcode' fields of CPD.

For those locations still remaining un-geocoded (which were around 40 records out of 200,000), a self join on the Postal Master for Canada based on the 'state' and 'city' fields was performed and the average latitudes and longitudes of the matches were calculated. Finally, these were used to update the averaged latitudes and longitudes.

POLAND

GNPC table contained postcodes for Poland with 2+3 digit format (e.g. 00-001, 00-002). However, TELDIR for Poland contained postcodes with a 2 digits format (e.g. 71, 72, 74), because it only contained information about the postal area. Furthermore, the 'place_name' field in GNPC represents 'city name'. Geocoding of Poland was therefore completed by matching the 'city name' and 'postal area' fields of Poland’s TELDIR against the ‘place_name’ and ‘first 2 digits of the postcodes’ fields from GNPC.

For the records in Postal Master that remained un-geocoded, a self join was performed on the Postal Master table (for Poland) based on the ‘city’ field alone, in order to calculate the average latitudes and longitudes.

For the records in the Postal Master that remained un-geocoded, a match on the ‘area code’ field from Poland’s TELDIR was performed with the ‘postcode’ field of GNPC table. Again the average latitudes and longitudes of these matches was calculated and updated in the table.

3.6 GEO-REFERENCING NAME DATA

3.6.1 AGGREGATING GEO-REFERENCED NAMES DATA TO GEOGRAPHIC AREAS

This section describes the assignment of geo-referenced names data to geographical areas and levels. A digital map consists of different polygons which combine to represent any geographical area. Once data have been aggregated to areas of different geographies, they can be visualised. The preceding discussion identifies that the following four geographical levels are amenable to names mapping:

- Country Level
- Region Level
- County Level
- City Level

For the purpose of this aggregation, the following Oracle tables were created.

TBL_MASTER_NAMES_FINAL (MASTER NAMES)

This table contains all names with their surname and forename frequencies as sourced from telephone directories and electoral registers. This is a global representation of all the names in the database. This table has following fields:

Row_Id Name Name_Eng Surname_Freq Forename_Freq Total_Freq

Where 'Row_ID' is the unique id for a name, 'Name' is the surname we want to aggregate data for, 'Name_Eng' is the standardised English name (because names were in different Unicode formats, so they needed to be in the same standard English format), 'Surname_Freq' represents the number of occurrences in telephone directories and electoral registers where this name has been classified as surname, 'Forename_Freq' represents the number of occurrences in telephone directories where this name has been classified as forename, and 'Total_Freq' represents the number of occurrences of the name in the whole database.

TBL_COUNTRY_MASTER_NAMES (COUNTRY MASTER NAMES)

This table contains all names at country level and it provides the snapshot of data for each country. It has the following fields:

Row_Id Country_Code Surname_Freq Forename_Freq

Here 'Row_Id' represents a link from this table to (Master Names). This link is required so that data can be linked between two tables, so this key is a foreign key in this table. 'Country_Code' represents 2 letter IATA country codes for different countries; 'Surname_Freq' is the number of occurrences within a country where the name has been classified as surname, in the same way 'Forename_Freq' represents the number of occurrences within a country where the name has been classified as forename.

TBL_POSTAL_MASTER_ FINAL (POSTAL MASTER)

This is the postal masters table created in Section 3.5 of this Chapter. It contains a reference to latitude and longitude for each location. This table will be used to link names data with latitude and longitude information which are necessary to plot any data on the map. This table has the following fields

Location_Id Latitude Longitude

'Location_Id' is a unique identifier of locations in the database i.e. it represents a single occurrence of (country code, state, city, and postcode). 'Latitude' and 'Longitude' represent latitude and longitude values of the 'location_id'.

TBL_LOCATION_ NAMES_FINAL (LOCATION MASTER NAMES)

This table contains the location level names data. A location is represented by 'location_id' in TBL_POSTAL_MASTER_FINAL, and this table contains the surname and forename frequencies of each surname for each 'location_id'. It has the following fields:

Location_Id Row_Id Surname_Freq Forename_Freq

Here 'Location_Id' represents a unique geographical region in the (Postal Master) table. This 'Location_Id' is a foreign key in this table, where its primary key is in (Postal Master) table. 'Row_Id' represents a unique identifier of a name in (Master Names) table, so this field is foreign key as well. 'Surname_Freq' is the number of occurrences in a geographical region (Country, Region, County, or City) where the name has been classified as a surname. In the same way, 'Forename_Freq' is the number of occurrences in a geographical region (Country, Region, County, or City) where the name has been classified as forename.

Names were aggregated to geographical areas by using the following procedure:

COUNTRY LEVEL

Country level geographical areas are the regions which represent a country. Transformation at this level involved creating a new table from the (Master Names) and (Country Master Names) tables. The new table has the following fields:

Country_Code Name_Eng Surname_Freq Cpm Rank

'Country_Code' is the 2 digit IATA code for countries, 'Name_Eng' is the standardised English surname, and 'Surname_Freq' is the number of occurrences in a country where the name has been classified as surname, 'Cpm' is the 'Counts Per Million' of a name in a country, and 'Rank' represents the rank of any name in a country. These fields were used in displaying data in maps at country level.

REGION LEVEL

This level corresponds to different geographical regions which are smaller than countries. The population size and areal extent of these regions vary from country to country. In the United States and Canada, regions are equivalent to the constituent States. However, in the United Kingdom, Government Office Regions represent the equivalent geographical areas. Data at this level were produced by combining (Postal Master), (Master Name), and (Location Master Names). The result was a table possessing the following fields:

Country_Code Admin1_Code Admin1_Name Name_Eng Surname_Freq Cpm Rank

'Country_Code' is once again the 2 digit IATA code for a country, 'Admin1_Code' is the code of the geographical area, 'Admin1_Name' is the English name of the geographical area, 'Name_Eng' is the standardised English name, 'Surname_Freq' is the number of occurrences in a geographical area where the name has been classified as surname, 'Cpm' is the counts per million in a geographical area, and 'Rank' represents the rank of surname in this geographical area. These fields were used to display data in maps at region level.

COUNTY LEVEL

This level corresponds to county level geographical areas. They are smaller than 'Region Level' geographical areas. For example, counties within the United States, Canada, and United Kingdom represent this type of geographical area. Data for this level were produced by combining (Postal Master), (Master Name), and (Location Master Names). The result was a table having the following fields.

Country_Code Admin1_Code Admin2_Code Admin2_Name Name_Eng Surname_Freq Cpm Rank

Where 'Country_Code' is the 2 digits IATA code for a country, 'Admin1_Code' is the code of the 'Region Level' geographical area in which 'Admin2_Code' lies, 'Admin2_Code' is the code of geographical area, 'Admin2_Name' is the English name of the geographical area, 'Name_Eng' is the standardised English name, 'Surname_Freq' is the number of occurrences in a geographical area where the name has been classified as surname, 'Cpm' is the counts per million in a geographical area, and 'Rank' represents the rank of surname in this geographical area.

CITY LEVEL

This level corresponds to city level geographical areas which are smaller than 'Region Level', and 'County Level' geographical areas. Examples include cities of the United States, Canada, and the United Kingdom. Data for this level were produced by combining (Postal Master), (Master Name), and (Location Master Names). The result was a table having the following fields.

Country_Code Admin1_Code Admin2_Code City_Code City_Name Name_Eng Surname_Freq Cpm Rank

'Country_Code' is once again the 2 digit IATA code for a country, , 'Admin1_Code' is the code of 'Region Level' geographical area in which the 'City_Code' lies, 'Admin2_Code' is the code of 'County Level' geographical area in which the 'City_Code' lies, 'City_Code' is the code of the city level geographical areas, 'City_Name' is the English name of the geographical area, 'Name_Eng' is the standardised English name, 'Surname_Freq' is the number of occurrences in a geographical area where the name has been classified as surname, 'Cpm' is the name count per million in a geographical area, and 'Rank' represents the rank of surname in the geographical area.

3.6.2 DATABASE STRUCTURE

Once data have been stored in the database, and aggregated to appropriate geographical levels, it is important to devise an appropriate structure of the database. The names data were stored in an Oracle 11i database. The database structure defines how different entities (tables) interact with each other to make relationships. A database structure can be represented using a database diagram. The database diagram shows the interaction of different database tables with each other.

DATABASE DIAGRAM

Following figure (Figure 3.3) shows the diagram of the database.

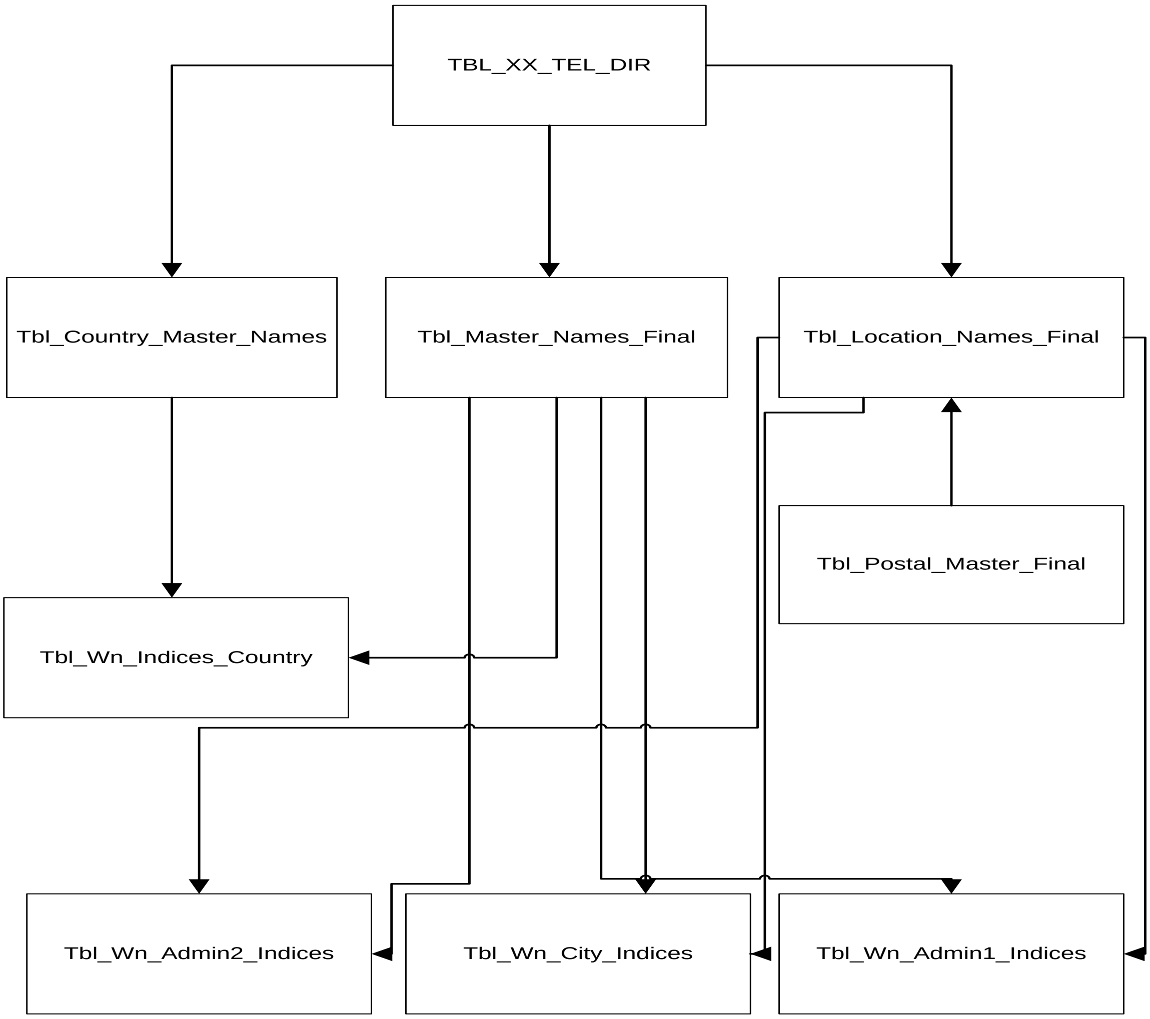


Figure 3.3: Database diagram

3.7 CONVERTING MAPS DATA TO ATTRIBUTE DATA

This chapter, so far, has described the creation of the world names database from text (telephone directories or electoral registers) sources. Whilst the coverage of the 'Worldnames' database is extensive, there are nonetheless some countries where text data sources are hard to come by. In the case of China, text sources were not available to the project, but we were able to source maps of the 300 most common names that had been compiled from official sources. This section is aimed at describing the conversion of maps data to attribute data. It is important in the sense that how we can extract information from maps and store them in the database. Graduated shading maps of the 300 surnames were scanned, showing the concentration of each surname in different regions of the country. Figure 3.4 illustrates the source data for the surname 'Li' across the different regions of China.

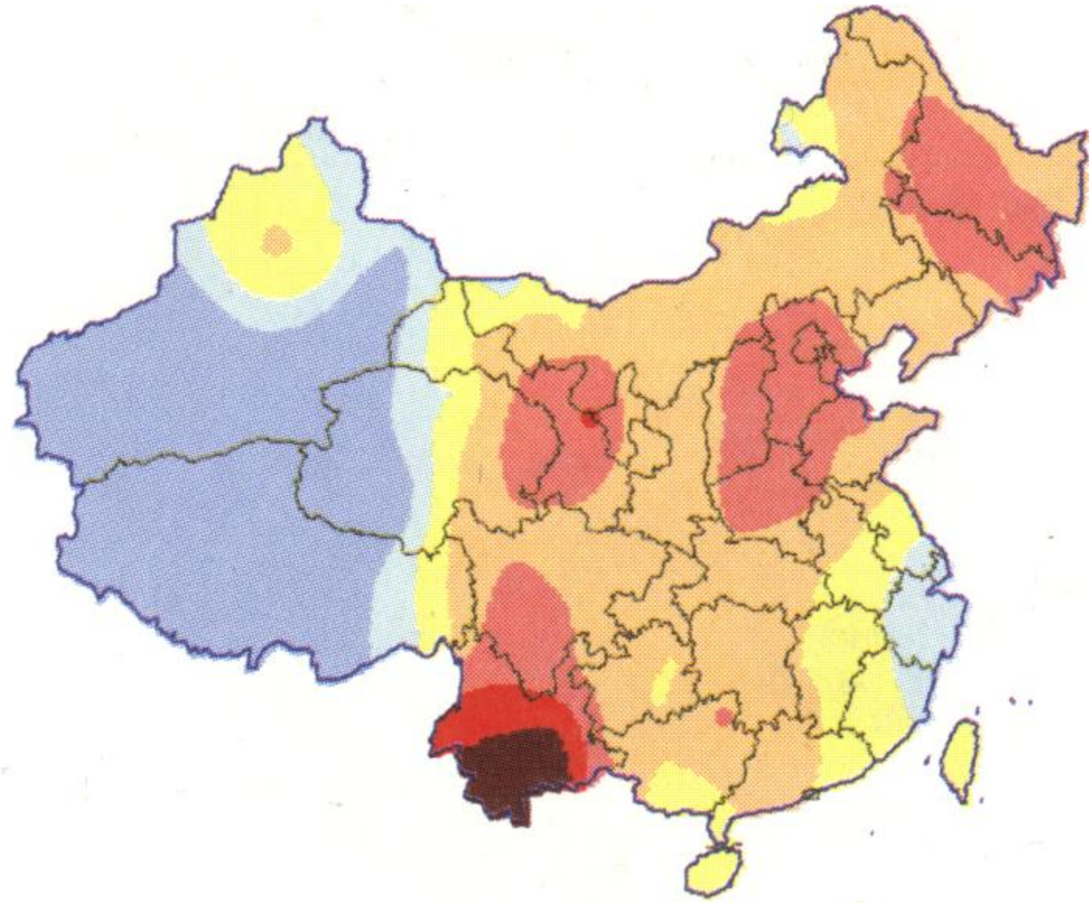


Figure 3.4: Concentration of Surname 'Li' across different regions of China

Source: Yuan (2007)

The following figure describes the colour ramp used on the map.

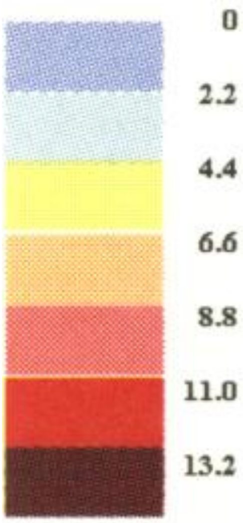


Figure 3.5: Color ramp used in Figure 3.4

Source: Yuan (2007)

The following section describes the extraction of the data from scanned maps/images by using different programming techniques.

3.7.1 ALGORITHM FOR EXTRACTING DATA FROM SCANNED MAPS/IMAGES

Java was used as the programming language for extracting data from scanned maps/images. The algorithm works in a series of following steps, that were each repeated for each scanned map/image.

a) Read the scanned map/image pixel by pixel, and identify regions (There are 30 regions in the map).

b) For each region in the map, repeat the following steps.

    1) Read each pixel value from the region, and create a matrix of different colours.

    2) Count the number of pixels for each colour.

3) Weight the number of pixels by their ‘colour concentration score’ and ‘population of the region’.
4) The final weighted score is the frequency of people living in the region with a particular surname.
5) Store the frequency of people in the matrix, and repeat steps 1-5 for other regions.

3.7.2 STORING INFORMATION IN THE DATABASE

Once all the scanned maps had been processed by the above algorithm (2.7.1), the data were stored in the database. The following tables were used to store the data for China.

TBL_MASTER_NAMES_FINAL

TBL_POSTAL_MASTER_FINAL

TBL_WN_ADMIN1_INDICES

TBL_COUNTRY_MASTER_NAMES

TBL_WN_INDICES_COUNTRY

3.8 CONCLUSION:

This chapter began by setting out the case for using names as cultural, ethnic and linguistic markers that had important geodemographic connotations. In an era of globalisation, it also set out the case for using names as a key element of socio-economic spatial data infrastructure that is comparable between national settings and that can be extracted at a full range of spatial scales from publicly available telephone directories and registers of electors. The grounding of any geodemographic representation is ideally at the level of the individual, but this chapter has also set out the usefulness of household data, as well as the imputation procedures that may be used to generate small area counts from previously published areal data. All of these procedures have underpinned the creation of the 'Worldnames' database after overcoming a number of procedural difficulties.

This chapter has presented an overview of the differences in naming conventions used in different parts of the world and has explained how different data sources are used in the creation of this database. Subsequent sections have described how geographical variations in naming conventions can be accommodated in the database and also explained the procedures of name cleaning. This chapter also explained the detailed processes of how the address data appended to the names data were georeferenced and then used in order to georeference the actual names data. The final section of the chapter describes the conversion of maps data to attribute data, using the example of China.

In the next chapter, the utility of this innovative new data source will be discussed, specifically by outlining the visualisation challenges in describing the geographic distribution of such a large number of names. Effective visualisation is central to evaluation of the provenance of the 'Worldnames' database, and hence its suitability for inclusion as constituents of the geodemographic classifications.

4 CHAPTER 4: GEO-VISUALISATION OF WORLDNAMES DATABASE

4.1 INTRODUCTION

The previous chapter has demonstrated how a unique international names database could be created using a range of innovative heuristics and data processing algorithms, allied with novel georeferencing procedures and creative common resources. Following data cleaning, aggregation and weighting into a form suitable for analysis, the next step is to create a method of visualisation. Longley et al. (2011: 327) described geovisualization as a 'technique to explore, analyze, synthesize, and present spatial data'. Tile based mapping applications have grown in popularity as a method of disseminating spatial data over the web, driven by the recent developments of a number of different mapping providers (Google Maps (http://maps.google.com), Yahoo Maps (shttp://maps.yahoo.com), OpenStreetMap (http://www.openstreetmap.org/) etc).

Google Maps is one of the widely used applications, which allow users to 'mash up' their own data on top of the base maps provided by Google maps API. OpenLayers provides an open source solution for the creation of rich Internet applications and is compatible with any data source.

Flash maps are another alternative to mapping as they are based on vector rather than raster images. A vector image is better than raster image as it does not distort with changing dimensions. This allows a smoother, clearer and better view in the resultant maps. However, flash maps have disadvantages in terms of the number of objects they can handle.

This chapter discusses the geo-visualisation process of the 'Worldnames' database. It comprises the following sections:

a) In the first section this chapter evaluates a range of different web mapping platforms. This discussion is based on static map renderers, slippy maps, API based map services, and a new way to map data in the form of flash maps.

b) The second part describes the geo-visualisation techniques employed. This section describes the different geo-visualisation techniques available for the creation of rich Internet mapping applications along with their relative advantages and disadvantages.

c) The third part of this chapter also explains the creation of the 'Worldnames' Flash website and the design requirements and project specification that guided its creation. It also discusses the operation and design of the website from a software engineering perspective.

d) Finally, this chapter explains the development of the beta 'Worldnames' website and how usability and visualisation issues in the beta website led to the development of the final version of 'Worldnames' website.

4.2 EVALUATION OF DIFFERENT WEB MAPPING PLATFORMS

Once data have been classified into different geographical areas, they can be visualised in maps. Visualisation of data in maps has became increasingly popular in recent years, with hundred of websites representing different data in geographical context. This popularity of web based mapping applications is due in part to the activities of GIS vendors, who are making it easy for users and developers to publish their data on the web in the form of geographical context. Also, usability is another cause of the increased popularity of visualisation techniques. Aoidh et al. (2008) described how, given the increasing popularity and growing choice of spatial information interfaces in recent years, it has become increasingly important to provide an efficient user friendly interface. In the same way Caschera et al. (2008) demonstrated that the growing interest of visualisation and analysis of social networks has led to the development of several methods of structural analysis in order to analyse individual and group behaviours. As such, developments in visualisation are not only restricted to the display of conventional data sources in maps but are also opening up the representation of some non conventional data sources in geographical context. One example is the heat map of the users of Second Life (http://www.lab.kathar.in/heatmap/), which is quite a non-conventional data source. However, this could explain the extent of the use of technologies in different countries.

The following is a discussion of the different geo-visualisation techniques that are currently available.

4.2.1 STATIC MAP RENDERERS

Static map renderers have been heavily used in the creation of maps. They render maps in the form of images. Static map renderers work by receiving a user input, rendering the map to an image (preferably a PNG image), and then sending the image back to the user. The user application then shows the image in the web or desktop application. Static map renderers do not facilitate any interaction in the maps because they are just static images, so they do not allow users to pan or zoom in/out around the map. On the one hand, the produced map is very simple and easy to interpret, because there are no requirements of the user on the client side, but on the other hand it poses a flexibility issue. Because a user cannot interact with the maps they are not suitable for the rich Internet applications of Web 2.0.

Figure 4.1 shows an example of static map renderer, with a map produced by the old National Trust Names website (www.nationaltrustnames.org.uk).

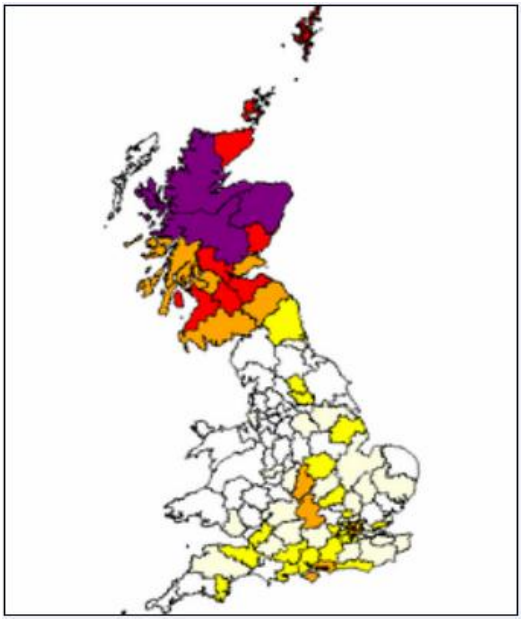


Figure 4.1: A map produced by a static map renderer

Static map renderers are fast as far as the computational time is concerned but they are not a flexible solution for representing maps because of the non-interactive nature of the final maps produced. The solution to these deficiencies is the use of slippy or tile-based maps, which are explained in the next section.

4.2.2 SLIPPY (TILE-BASED) MAPS

One important and widely used geovisualisation technique is tile-based or slippy maps. This technique works by dividing the map into a discrete number of tiles at zoom levels, so that each zoom level has identical number of tiles. Instead of accessing the whole map at once, a tile-based client builds the map by accessing individual tiles and then assembling them to form a map. Google maps (http://maps.google.com), Bing maps (http://www.microsoft.com/maps/), Yahoo Maps (http://maps.yahoo.com), OpenStreetMap (http://www.openstreetmap.org) are all based on tile based or slippy maps. Slippy maps were introduced as an alternate to static maps because they are dynamic and faster to load and the only specific tiles are loaded on user requests.

The following figure (Figure 4.2) shows the example of division of large tile into discrete number of small tiles.

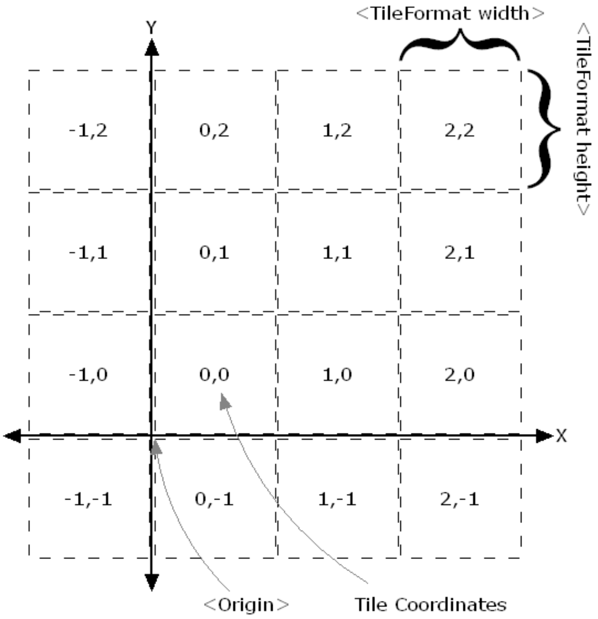


Figure 4.2: Tiling in Slippy maps

The tiling process works by dividing the whole map into small equally sized tiles. The coordinates that the user clicks to view at finest level of detail acts as the origin of the tiling divisions and all the tiles are numbered as shown in the Figure 4.2. Finally, the user is presented with the 'origin tile'. This process of dividing a map into tiles continues if the user clicks on the map again.

What follows is a discussion of OpenLayers, ArcGIS Server, and Mapnik which can be used to provide the facility of slippy maps in web or desktop GIS applications.

OPENLAYERS

OpenLayers provides a mapping solution in the form of slippy maps. OpenLayers is a solution to embed dynamic maps into websites. It can be used to show map tiles from any data source. MetaCarta (http://www.metacarta.com/) was the first developer of OpenLayers and then they made it available to the public. It is based on a JavaScript library to show the map tiles in the websites.

Following figure 4.3 shows a simple overlay using OpenLayers.

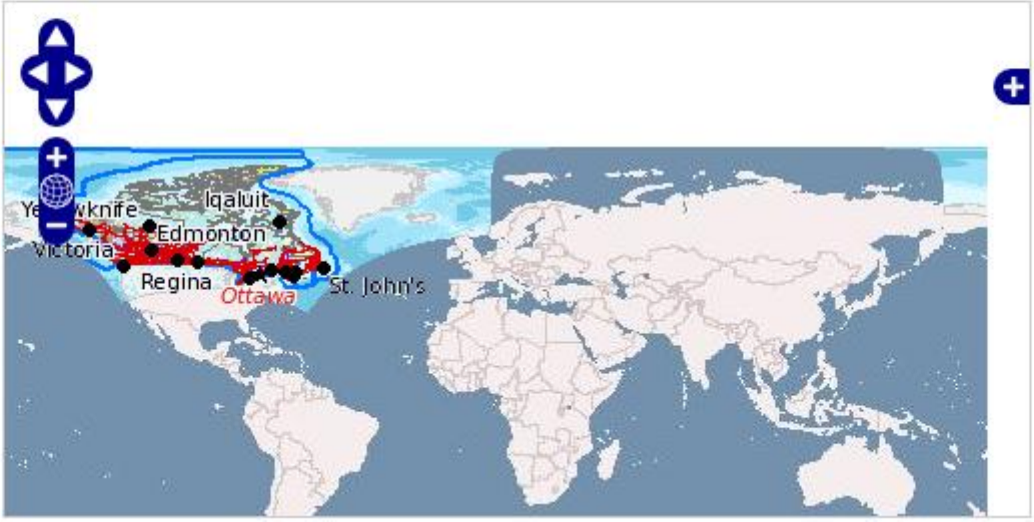


Figure 4.3: A simple overlay with OpenLayers

An advantage of using OpenLayers is that it can map any data source. Other GIS servers (for example Google Maps, Yahoo Maps etc), use only one data source.

ARCGIS SERVER

ArcGIS Server serves tiles in the form of slippy maps. ArcGIS Server is a Geographic Information System developed by ESRI Inc. (Redlands, CA: www.esri.com) and is used to serve map services to clients. ArcGIS server provides a set of APIs used to develop rich internet applications in different platforms. It provides .NET, Java, JavaScript, Flex, REST, and Silverlight APIs for the development of web and GIS mapping applications. Applications developed using any of these technologies can access maps from an ArcGIS server and show them to users.

The mapping solution works by making a service in ArcGIS Server, which contains the actual map. Any of the client applications created using the above mentioned APIs are required to send a request to the ArcGIS server and in turn ArcGIS server provides a response to the client with the actual map requested. Client applications can either show the map as it is or it can overlay user data on top of the map. This provides a powerful architecture for the creation of dynamic and data oriented maps based on a 3 tier web architecture.

Following figure 4.4 shows a simple map produced from ArcGIS server with population data overlaid on the base map.

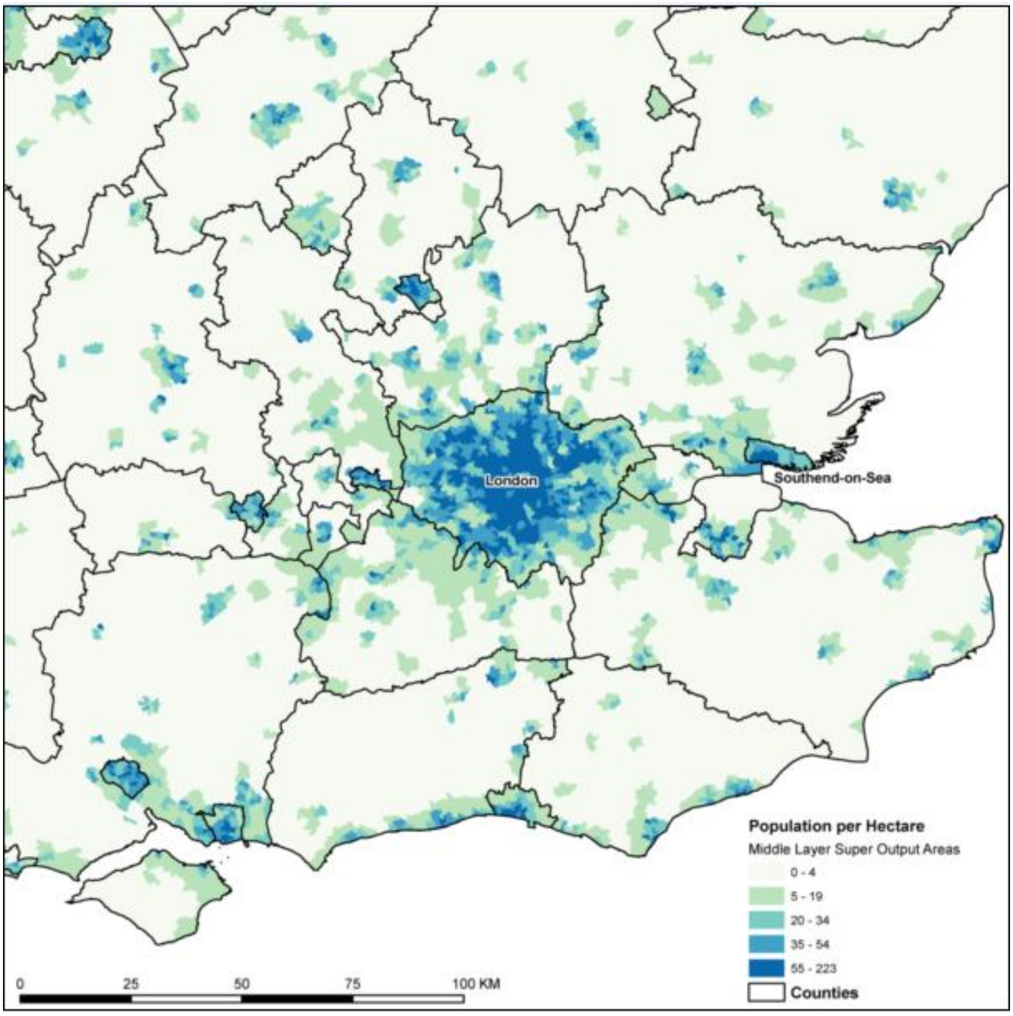


Figure 4.4: A map produced by ArcGIS Server

Once a user interacts with the map, for example by zooming into an area or clicking on any area, another request goes to the ArcGIS server, which in turn sends another response to the user in the form of a map.

MAPNIK

Mapnik (www.mapnik.org) is a free toolkit for developing mapping applications. Users can use Mapnik to produce thematic overlays of their own data sources. These thematic overlays can be overlaid on base maps (e.g. Google Maps, OpenStreetMaps etc). It is developed in C++ with the support of Python. Mapnik can read all raster and vector formats that are supported by GDAL (Geospatial Data Abstraction Library). GDAL provide an abstract model for the representation of spatial data. Mapnik also has ESRI shape file readers, which make it a flexible platform when using any data source. Thus Mapnik can use data from different data sources and show them as a part of map. Figure 4.5 shows the use of Mapnik to show a custom layer using the Google Maps API. The source of this map is http://mapnik.org/tiling/oxford/.

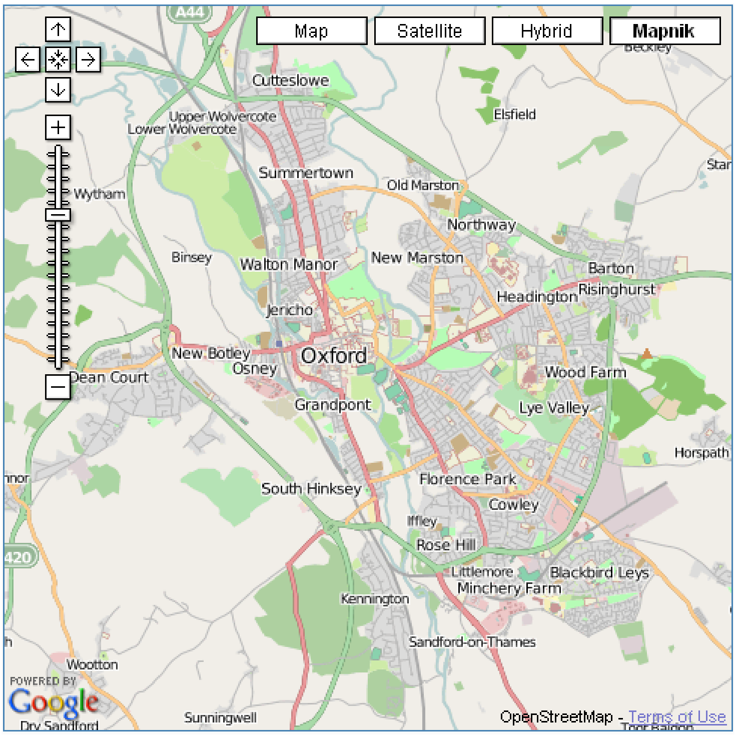


Figure 4.5: Mapnik showing a custom layer using Google Maps API

ADVANTAGES OF SLIPPY MAPS

Slippy maps are better than static maps because tiles are served to users based on requests rather than serving the whole map at once. Slippy maps are dynamic and faster to load because only the tiles that a user wants to see load up and it is not necessary to load up all the tiles at once. This saves computation and memory resources both on the client and on the server that provides the mapping service. In addition, it provides a powerful architecture for the creation of rich internet GIS web applications.

DISADVANTAGES OF SLIPPY MAPS

Slippy maps are not a good solution for the development of on the fly rendering applications. On the fly rendering requires a lot of memory and computational resources and this makes a slippy maps solution non-flexible to overlay data on top of all the tiles of the base map. Taking the example of mapping surname data on top of Google maps, if a user searches for a name ‘Longley’ and this needs to be visualised on 3 levels of geography (Country, Region, and County), this requires the facility to overlay the data for all levels on the tiles of Google maps. This may be fine for one level, but to do it for all the tiles of all the spatial levels is time consuming. Data could be overlaid on Google Maps in form of KML as well, but KML files have restrictions on the size of the data. So in the environment where data changes dynamically based on user requests and it requires on they fly rendering, slippy maps are not a suitable option.

An alternative as explained by Gibin et al. (2008) is based on pre-rendering the tiles using a tool called GMAPCreator (http://www.casa.ucl.ac.uk/software/gmapcreator.asp), storing the pre-rendered tiles on disk, and serving the tiles to users from disk rather than rendering them on the fly. While, on one hand, this technique solves the problem of computation time, it poses another problem of storing the tiles on disk. There could be millions of tiles and this needs a considerable disk space to store all the tiles, which is not efficient.

4.2.3 API BASED MAP SERVICES

There are a number of API based map services which users can use to develop their own mapping websites. Google Maps, Yahoo Maps, and Ordnance Survey OS (OpenSpace) APIs are examples of API based map services. This section discusses them in detail:

GOOGLE MAPS API

The Google Maps API allows users to embed Google maps into their websites. Google maps API is a free service from Google and it just needs an API key to be used in a website. Once a user gets an API key for a particular website, that can be used to access Google Maps data from Google and a user can show their own data on top of Google maps as well. For creating a customized interface a user is required to use JavaScript.

In 2006, Google also launched a mobile phone version (http://www.google.com/mobile/default/maps.html) of the Google Maps API which can run on any Java enabled mobile phone or hand held device. This provides the functionality of using Google Maps on mobile devices (mobile phone, PDA etc) as well.

GOOGLE MAPS API AND KML

KML (Keyhole mark-up Language) is an XML based mark-up language used to express geographical visualisation on 2 or 3 dimensional maps. Initially, it was developed to be used with Google Earth, but it is also used heavily with Google Maps. It is a standard put forward for the representation of geographical data (point, poly lines, polygons, place marks, images, and 3D models etc) in XML based format, which could be displayed in conjunction with Google Maps or Google Earth. KML represents geographical features in terms of their latitude, longitude, and altitude. Latitude and longitude are identified using WGS84 (World Geodetic System of 1984) format, whereas altitude information is based on the WGS84 EGM96 Geoid vertical datum. KML has some limitations as far as the size of KML file is concerned. , KML limitations are as shown in Table 4.1. The source of this table is http://code.google.com/apis/kml/documentation/mapsSupport.html.

Maximum fetched file size	3MB
Maximum uncompressed file size	10MB
Maximum number of network links	10
Maximum number of total-document-wide features	1,000
Maximum number of features visible in any give viewport	80

Table 4.1: Limitations of KML

YAHOO MAPS API

Yahoo Maps API also allows users to add maps to their web sites with the choice of platforms. At the moment, the Yahoo Maps API comes with Flash, Ajax, and Map Image platform support. Flash API for Yahoo maps works by using DHTML and java script to show the data on top of base maps. Thus users can develop their web applications which are dynamic in the way they show graphics. An Ajax API could be used to extract maps by asynchronous web requests to the Yahoo Maps server. Ajax requests work by sending the user requests to the server and showing the results without refreshing the web page. So, this works as a background process to retrieve data from the server and then displays the data. The map Image API works on a request-response process. The user specifies latitude (X) and longitude (Y) for the content and then the Map Image API returns the static image in PNG format, which a user can show in the website as an overlay of the maps.

ORDNANCE SURVEY OS (OPENSPACE) API

Ordnance Survey (GB) is the National mapping agency of Great Britain and has launched an OpenSpace API. A user needs to get an API key from Ordnance Survey before using this API. This API is based on JavaScript, so the user needs to implement requests in JavaScript. The resultant map is from the Ordnance Survey map series and a user can add points, lines, and polygons on top of the map using JavaScript. However, there is a limit on the number of user requests per day. A user can only access 30,000 tiles and 1,000 place name look-ups per day, which makes this API non-flexible. However, it is useful for a website that has small number of users. If user requests increase, this API ceases to be a suitable mapping solution.

OPENSTREETMAP API

OpenStreetMap is the Wikipedia of maps. It is a collaborative project that aims at creating a free editable map of the world. OpenStreetMap represents maps created by portable GPS devices, aerial photography, and other free sources or simply from local knowledge. The OpenStreetMap API allows users to use JavaScript to embed

OpenStreetMap maps to the web applications. There is no limitation on the number of requests from the user web application to the OpenStreetMap server. This provides a great flexibility in the creation of rich internet applications.

4.2.4 FLASH MAPPING

Flash maps are based on vector images. Busselle (2008) describes a raster image as a collection of dots called pixels, and a vector image as a collection of connected lines and curves that produce objects. Raster images are resolution dependent, so their shape distorts if the image is made bigger than the resolution for which they were intended. Vector images have an advantage over raster images that they are resolution independent. Thus their shape does not distort if the image's size changes. This is because a vector image is drawn dynamically by using mathematical calculations that do not allow the map to distort and keeps it in the right shape.

Flash has become popular as a mapping solution during the last few years. It has been incorporated as an API in ArcGIS (ArcGIS Flex API) (http://help.arcgis.com/en/webapi/flex/index.html). It is also possible to integrate Flash with the Google maps API (http://code.google.com/apis/maps/documentation/flash/) . The increasing popularity of Flash arises out of its rich Internet and dynamic visual contact generation capability, which is easy to integrate with other technologies, and this is making it possible to consider Flash as a mapping solution.

FLEX

Flex (http://flex.org/) encapsulates Flash by making it easy for users to develop interactive animations without the need to learn Flash Action Script in detail. Flex has become popular due to the additional functionality it provides to the users. Based on a Flash technology, it encapsulates the complexity posed in Flash and provides an easy to use architecture for applications. ArcGIS's Flex API (http://resources.esri.com/arcgisserver/apis/flex/) shows the increasing popularity of Flex in the web GIS development world. This API is used to develop rich internet applications by overlaying user's data on top of the base maps produced by the ArcGIS Server.

INTEGRATION OF FLEX WITH GOOGLE MAPS AND ARCGIS SERVER

Google maps API for Flash and the Flex API for ArcGIS server allow users to use Flash and Flex to embed the functionality of Google Maps or ArcGIS maps in the web applications. Both APIs hide complex details of connectivity to the GIS servers and they allow an easy to use interface for the developers so that Flash can be used to develop more interactive and rich Internet applications without knowing Flash Action Scripts in great detail. This opens up a new way of developing GIS applications where the GIS server produces the map and then Flash is used to add some more information to the map. This information could entail adding a Point, Line, Polygon or could entail adding some more complex graphics to the map e.g. custom navigation Flash controls.

4.3 THE WORLDNAMES FLASH WEBSITE

4.3.1 PROJECT SPECIFICATION AND USER REQUIREMENTS

The names profiler for Great Britain (www.nationaltrustnames.org) was launched in January 2006 and it became very popular following extensive media coverage. It has been promoted by media channels and newspapers including the Daily Mail, the Times, and BBC. This success led to the development of a 'Worldnames' website which can profile the names of the residents of 26 countries across the world.

We can define user requirements or project scope in terms of three functional areas of the website which are ‘Name Search’, ‘Area Search’, and ‘Ethnicity Search’. These requirements were defined before the work was started on the 'Worldnames' website. The following is a description of each of these functional areas.

NAME SEARCH

a) 'Worldnames' profiler should have the ability to search for a surname in 26 countries (as mentioned in Chapter 3) of the world and it should show the concentration of names using a suitable mapping technique.

b) In addition to showing the concentration of names in different countries, the worldnames profiler should be able to show some statistics about the name. These statistics are the ‘Ethnic roots of a surname’, ‘Top countries of a surname’, ‘Top regions of a surname’, ’Top cities of a surname’, and ‘Top forenames of a surname’.

c) The user should be able to provide feedback on ‘Ethnic roots of a surname’ and ‘Top forenames of a surname’, if results are not as expected by the user.

AREA SEARCH

a) Area search allows the ability for a user to search for an area and see the top forenames and surnames associated with that particular area. So the website should enable a search for ‘Top forenames’ and ‘Top surname’ in any part of the 26 countries. This will allow users to see the popular forenames and surnames in different areas around the 26 countries.

ETHNICITY SEARCH

a) Ethnicity search is based on searching for one of the ethnicities as described by Mateos (2007).

b) 'Worldnames' website should have the ability to show the concentration of the searched ethnicity around 26 countries of the world using a suitable mapping solution.

c) In addition to showing the concentration of an ethnicity in different countries, Worldnames website should be able to show some statistics for the ethnicity. These statistics are the ‘Top countries of an ethnicity’, ‘Top regions of an ethnicity’, ’Top cities of an ethnicity’, and ‘Top surnames of an ethnicity’.

Sub section 4.3.5 describes the development of the beta version of the website that is only based on ‘Name Search’ facility, and subsequently Sub Section 4.3.6 describes the development of the final version of Worldnames website which is based on this project specification. The choice of mapping solution for the website is explained in the next sub section.

4.3.2 CHOOSING FLASH OVER A TILE BASED APPROACH

Section 4.1 of this chapter discussed different mapping options in detail. Opting for one of them from the available options was based on the requirement of the time that elapses for large numbers of users to submit queries and receive results. On the fly rendering for every level of Country, Region, and County was very difficult in this situation due to the long computational time that is experienced. Previous experience from National Trust Names website demonstrates that on the fly rendering cannot cope with high user load. This means that when a large number of users accessed the website at the same time, the National Trust Names website crashed. So Flash mapping was chosen as the resultant mapping solution for the 'Worldnames' website because flash files are easy to load in the browser and they do not require the map to be rendered over and over again for each user request. So once a user sends a request, the web application application retrieves the information for a searched surname from the database and writes it to the XML file and then loads a particular Flash file in the browser. While the flash file is loading, it reads the XML file and paints the objects (in this case the polygons of a map) of the flash file with the colour values given in the XML file. Thus, the processing in this case involves reading from the database and painting the polygons which is much quicker that producing the maps on the fly.

4.3.3 WEB APPLICATION DESIGN

Designing software or web applications before development is an important part of Software Engineering process. Software Design makes it possible to show graphically the entities in the system, how these entities change states for various inputs, and how users interact with the entities. Reeves (2005) describes how most current software development processes try to segregate the different phases of software development into separate pigeon-holes. The top level design must be completed and frozen before any code is written.

This section explains the design of web application in the form of a:

- State-Transition diagram
- Class diagram
- Use case diagram

STATE-TRANSITION DIAGRAM

Copeland (2008) describes how State-Transition diagrams describe all of the states that an object can have, the events under which an object changes state (transitions), the conditions which must be fulfilled before the transition will occur (guards), and the activities undertaken during the life on an object (actions). State transition diagrams are useful to describe the behaviour of objects (i.e. entities that make up the system). The behaviour of all objects in the system describes the overall functionality of any software or web application.

In terms of the 'Worldnames' website, we can divide the web application into three major objects or functions i.e. Surname Search, Ethnicity Search, and Area Search. Each represents a different search facility in the web application. What follows are the state transition diagrams of each object.

The following figure 4.6 shows the state transition diagram for ‘Surname Search’.

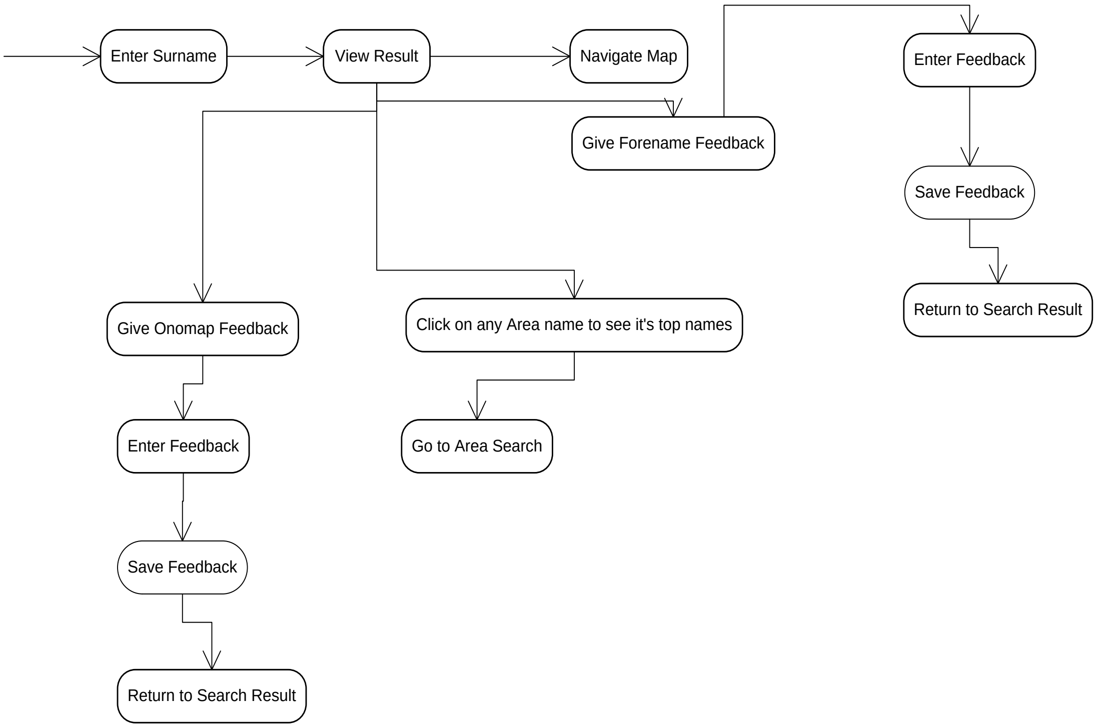


Figure 4.6: State Transition Diagram for Surname Search

When a user enters a surname, results of the query are shown as a map alongside associated statistics. A user can navigate around the map to view the concentration of surname in different geographical regions of the map; this is done by hovering he mouse over any geographical area. A user can also click on any geographical area to view the map of the next spatial level. For the 'Ethnic roots of a surname' and 'Top forenames of a surname' a user can also provide feedback, which takes the user to different pages of the website to record the feedback to the database.

The following figure 4.7 shows the state transition diagram for 'Ethnicity Search'.

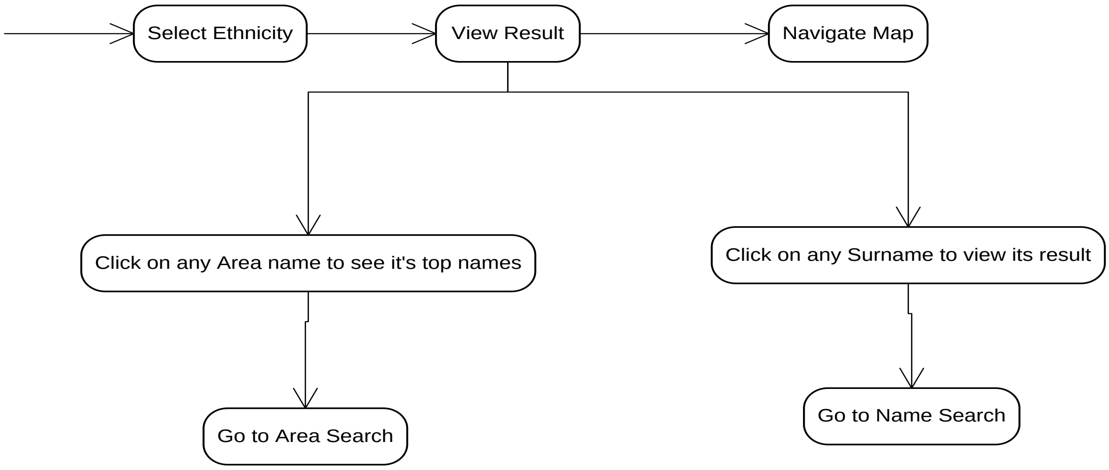

Select Ethnicity
View Result
Navigate Map
Click on any Area name to see it's top names
Click on any Surname to view its result
Go to Area Search
Go to Name Search

Figure 4.7: State Transition Diagram for Ethnicity Search

When a user searches for an ethnicity, results of the query are shown as a map and statistics. A user can navigate around the map to view the concentration of surname in different geographical regions of the map. The user can also select any area from 'Top countries of an ethnicity', 'Top regions of an ethnicity', and 'Top cities of an ethnicity' to go to 'Area Search' for that particular clicked area. In addition, a user can click on any surname from 'Top Surname of an ethnicity' to go to 'Name Search'.

The following figure 4.8 show the state transition diagram for 'Area Search'.

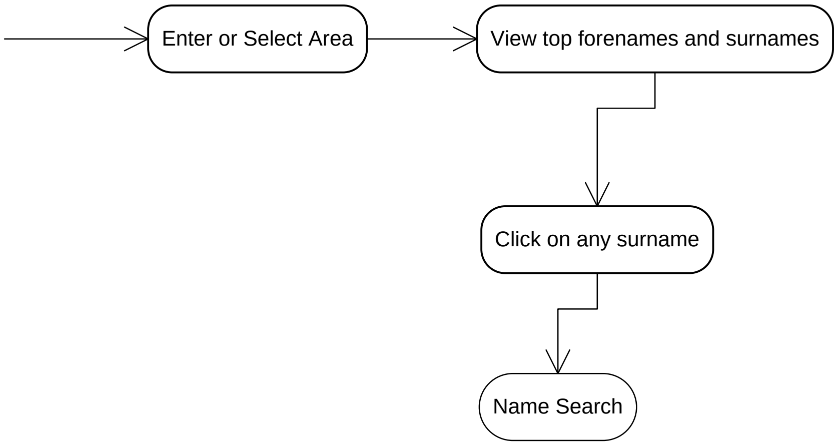


Figure 4.8: State Transition Diagram for Area Search

When a user types in an area name and searches, top (most common) forenames and surnames of the searched area will be presented to the user. The user can click on any surname from ‘Top Surnames’ to go to the name search.

CLASS DIAGRAM

Copeland (2008) described that a class diagram shows the classes that make up a system and the static relationships between them. Classes are defined in terms of their name, attributes (or data), and behaviours (or methods). The static relationships are association, aggregation, and inheritance.

Classes can be considered as components or individual entities of the system and a class diagram shows all the components and the relationship they have between them. A class diagram reflects the overall picture of the software or web application, because it shows all the entities constituting a system.

The structure of the classes should be finalized before actual coding has started for the software or web application. This helps the software development team or individual to identify all the objects of the system in advance and it shows a clear picture of the software or web application under development. The 'Worldnames' website was divided into 13 major classes, each representing a separate entity in the web application.

Following figure 4.9 shows the final class diagram of the Worldnames website.

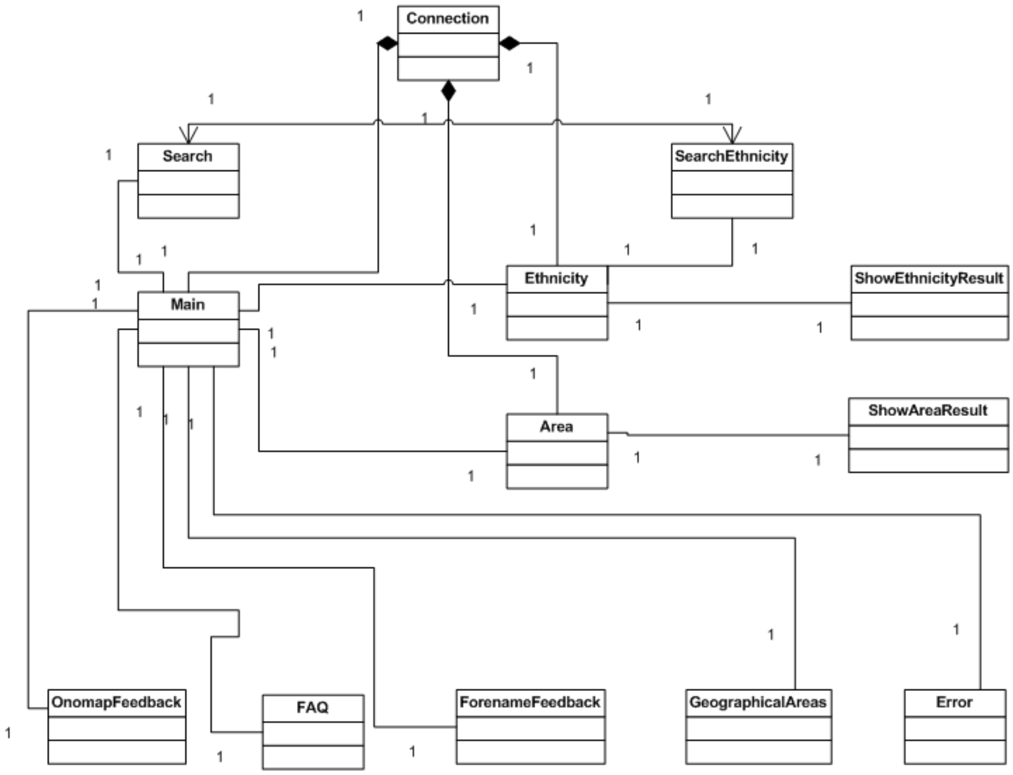


Figure 4.9: Class diagram of Worldnames website

In the class diagram, a diamond represents an aggregation (containment) of a class in another, while an arrow shows the association (linkage) of one class with another. On the top of the hierarchy 'Connection' is the class which represents a database connection with the Oracle database. This class is aggregated into 'Main', 'Area', and 'Ethnicity' classes. 'Main' class encapsulates the functionality of name search, 'SearchEthnicity' class encapsulates the functionality of ethnicity search, and 'Area' class encapsulates the functionality of area search. 'OnomapFeedback', 'FAQ', 'ForenameFeedback', 'GeographicalAreas', 'Error' are the utility classes which act as the helping classes of others.

USER CASE DIAGRAM

Copreland (2008) described that a use case is a scenario that describes the use of a system by an actor to accomplish a specific goal. An actor is a user playing a role with respect to the system. Actors are generally people although other computer systems could be actors as well. A scenario is a sequence of steps that describe the interaction between an actor and the system. The use case model consists of the collection of all actors and all use cases.

From a software engineering perspective, use cases are used for the following purposes:

a) They describe what functionality a user needs in the system

b) They can help in identifying components of the system and the relationships of different components

c) They can help to design test cases

The following figure 4.10 shows the use case diagram of the 'Worldnames' website.

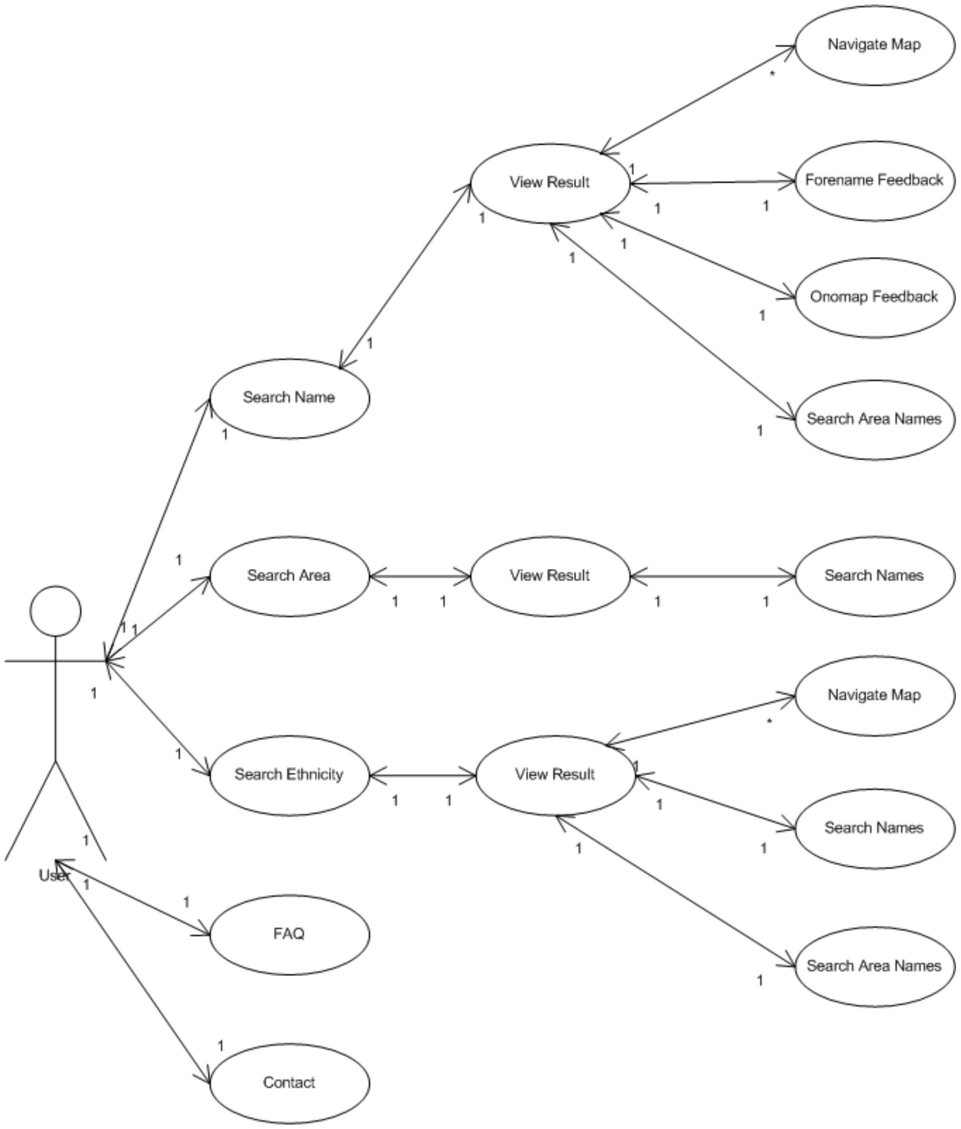


Figure 4.10: Use Case diagram of Worldnames website

A user can search a name, an area, and an ethnicity, see FAQs and click on ‘contact us’. While searching a name, the user can view the results on a navigable map. The user can also provide 'forename' and 'ethnicity' feedback if the results are not according to what might be expected, and a user can click on any area to see its top forenames and surnames. While searching an area, a user can view the results in the form of top forenames and top surnames. The user can also click on any surname to view its map and statistics. While searching an ethnicity, a user can view the results on a navigable map. The user can also click on any surname to view its map and statistics. Also, the user can click on any area to see its top forenames and surnames.

4.3.4 DATA FLOW OF WEB APPLICATION

The following figure 4.11 shows the data flow diagram of the 'Worldnames' website. The figure shows the integration of all the components of the website and this demonstrates the functionality of ‘Name Search’, ‘Ethnicity Search’, and ‘Area Search’ in the web application.

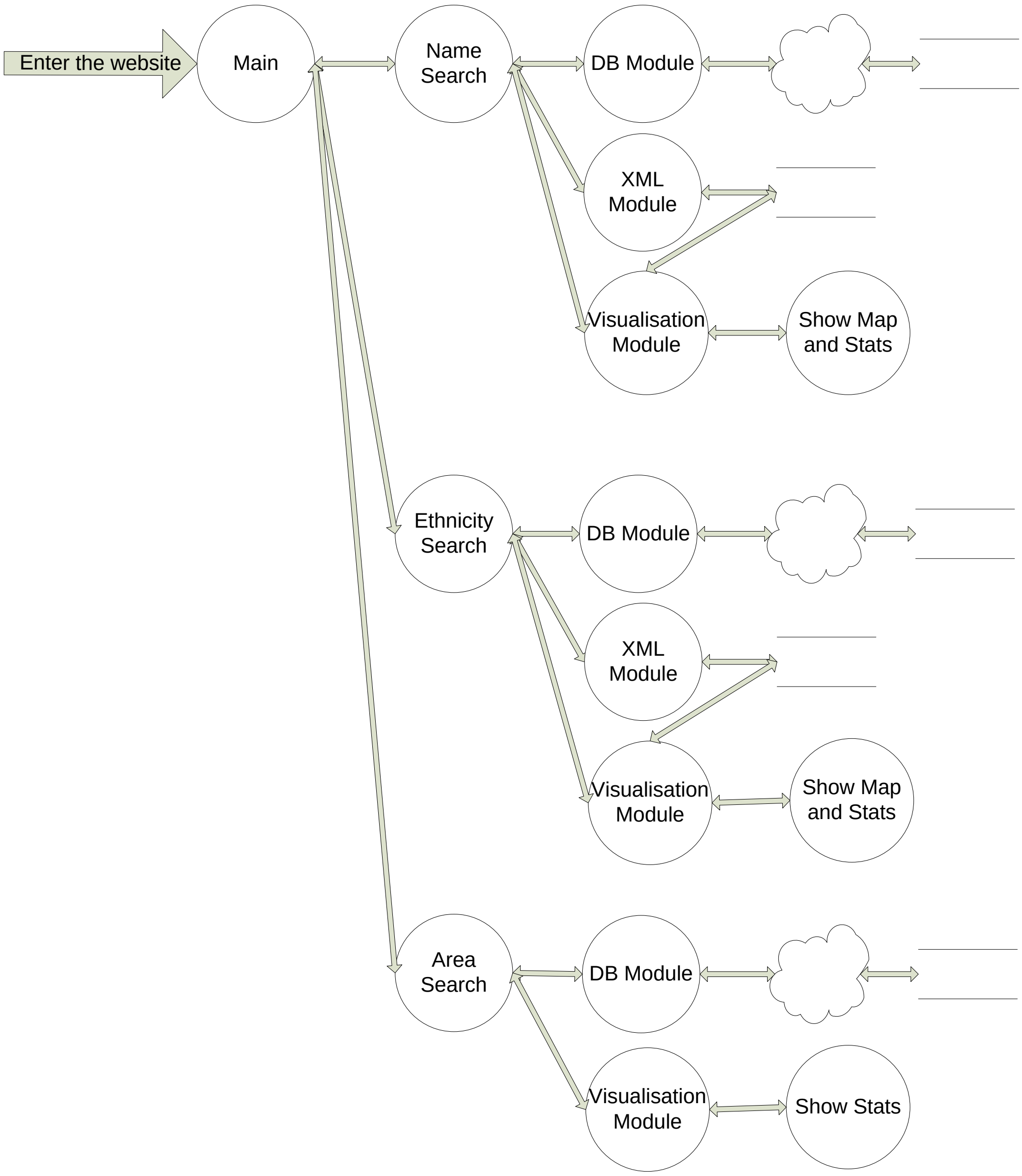


Figure 4.11: Flow diagram of the web application

This figure shows the different functional components stitching together to form the overall functional part of the website. Functional components can be divided into three main parts.

NAME SEARCH

This functional component implements the functionality of Surname search in the website. When a user searches for a surname, the 'Name Search' component passes the searched surname to 'DB Module'. 'DB Module' interacts with the Oracle database, which is on a remote server, and passes the information to 'XML Module'. 'XML Module' writes the information retrieved from the database to XML files. Finally, 'Visualisation Module' reads the XML files and displays information as maps. This module also shows the statistics for a searched surname.

ETHNICITY SEARCH

This functional component implements the functionality of Ethnicity search in the website. When a user searches for an ethnicity, 'Ethnicity Search', the component passes the searched surname to 'DB Module'. 'DB Module' interacts with the Oracle database, which is on a remote server, and passes the information to 'XML Module'. 'XML Module' writes the information retrieved from the database to XML files. Finally, 'Visualisation Module' reads the XML files and displays information in maps. This module also shows the statistics for a searched surname.

AREA SEARCH

This functional component implements the functionality of Area search in the website. When a user searches for an area, 'Area Search', the component passes the searched surname to 'DB Module'. 'DB Module' interacts with the Oracle database, which is on a remote server, and passes the information to 'Visualisation Module'. 'Visualisation Module' shows the results in the form of statistics.

4.3.5 WORLDNAMES BETA WEBSITE

In the development process of the 'Worldnames' website the first step was to create a beta version. A grid level technique was chosen for showing data in map form. This technique was aimed at providing a grid square on each country of 26 countries. There were two grid levels opted, 40Km and 5Km. It was decided that users would be able to navigate between the 40Km and 5Km levels by using the zoom option.

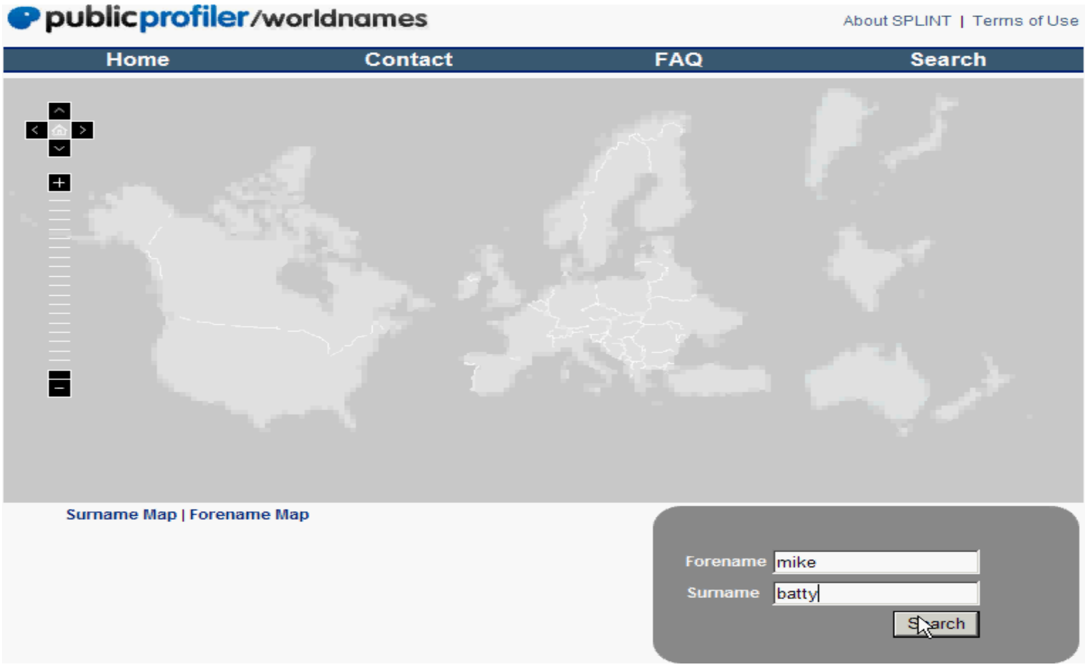
By using ESRI ArcGIS, a point in polygon operation was completed for all geo-referenced occurrences of names data. This enabled 40km and 5km grids to be overlain on top of the country level geography of the world. The idea was to paint grid cells dynamically when a user searches for a name, where painting would show the concentration of the searched surname. After completing the point in polygon operation in ArcGIS, a flash file was created for the two zoom levels, where each grid cell was represented a flash object. However, flash could not handle a 'movie file' where number of objects exceeds 8000. So, Flash files were divided into three more levels representing North America, Europe, and Asia.

SCREEN SHOTS OF THE BETA WEBSITE

Forename=Mike, Surname=Batty were used to show the screen shots of the beta website. Screen shots of the website are divided into three sections:

- Country level
- 40km grid level
- 5km grid level

The following figure 4.12 shows the main screen of the beta website. A user can input forename and surname.


publicprofiler/worldnames
About SPLINT | Terms of Use
Home
Contact
FAQ
Search
Surname Map | Forename Map
Search

Forename: mike

Surname: batty

Figure 4.12: Main Screen for input

COUNTRY LEVEL SCREEN SHOTS

The following figure 4.13 shows the concentration of surname 'Batty' in the world.

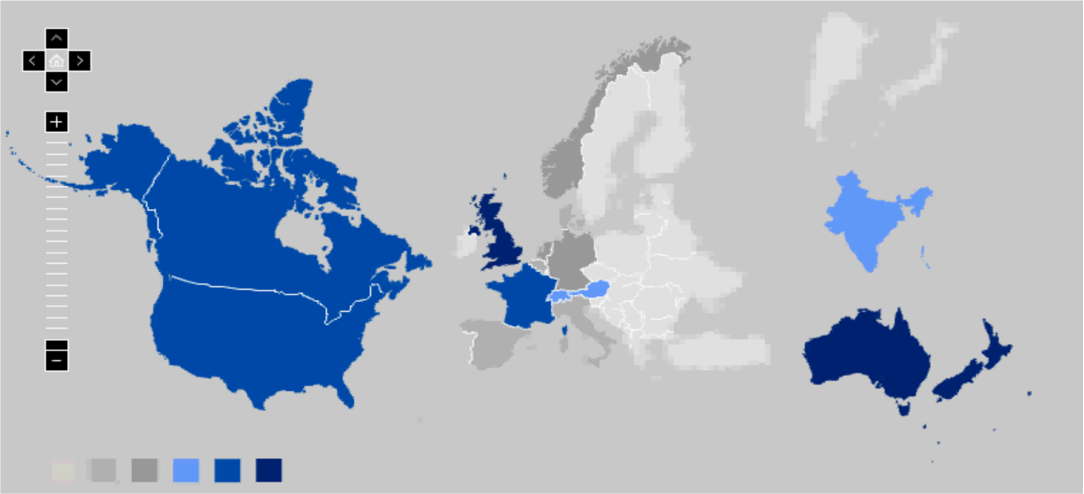


Figure 4.13: Concentration of surname 'Batty' in the world

The following figure 4.14 shows the concentration of forename 'Mike' in the world.

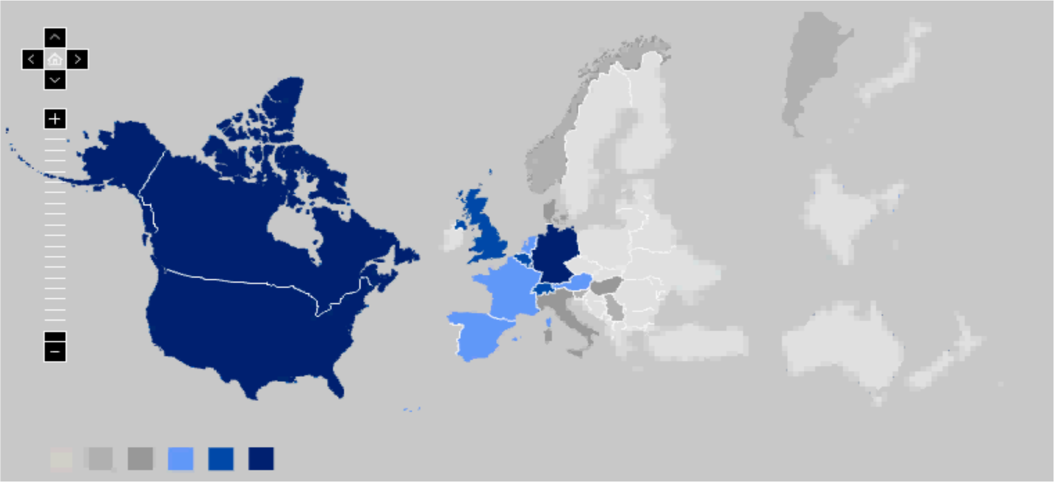


Figure 4.14: Concentration of forename 'Mike' in the world

40KM GRID LEVEL SCREEN SHOTS

The following figure 4.15 shows the 40km grid level screen shot for surname 'Batty' in North America.

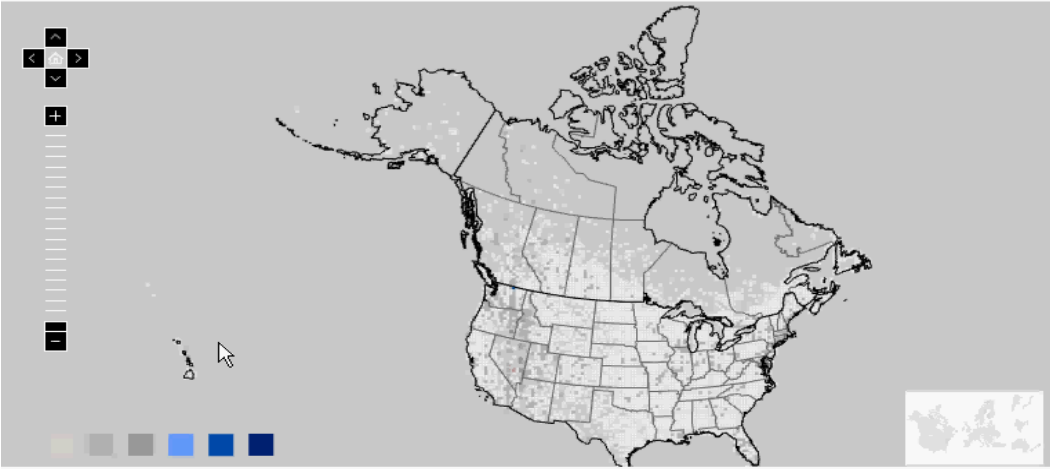


Figure 4.15: Concentration of Surname 'Batty' in North America (40km grid)

The following figure 4.16 shows the 40km grid level screen shot for surname 'Batty' in the United States.

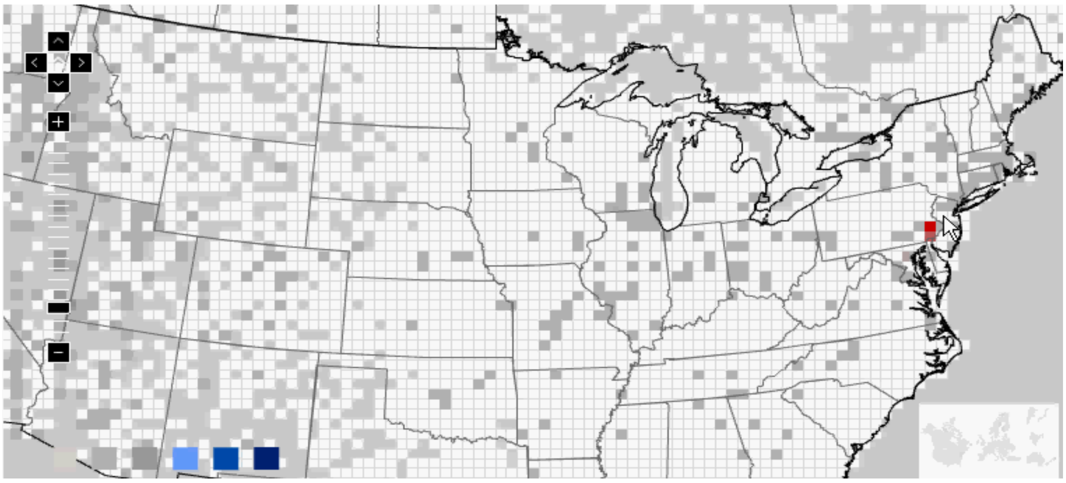


Figure 4.16: Concentration of Surname 'Batty' in United States (40km grid)

The following figure 4.17 shows the 40km grid level screen shot for surname 'Batty' in Europe.

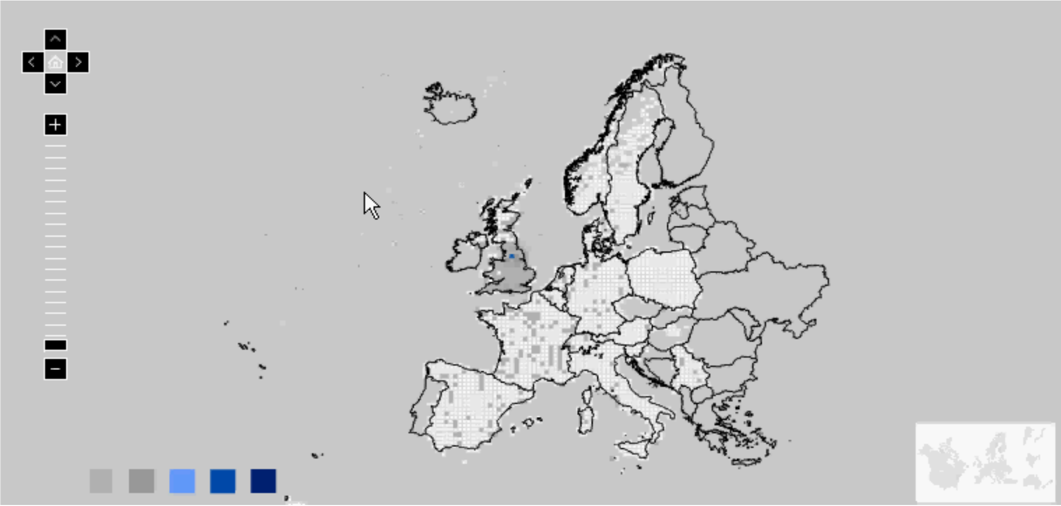


Figure 4.17: Concentration of Surname 'Batty' in Europe (40km grid)

The following figure 4.18 shows the 40km grid level screen shot for surname 'Batty' in Europe.

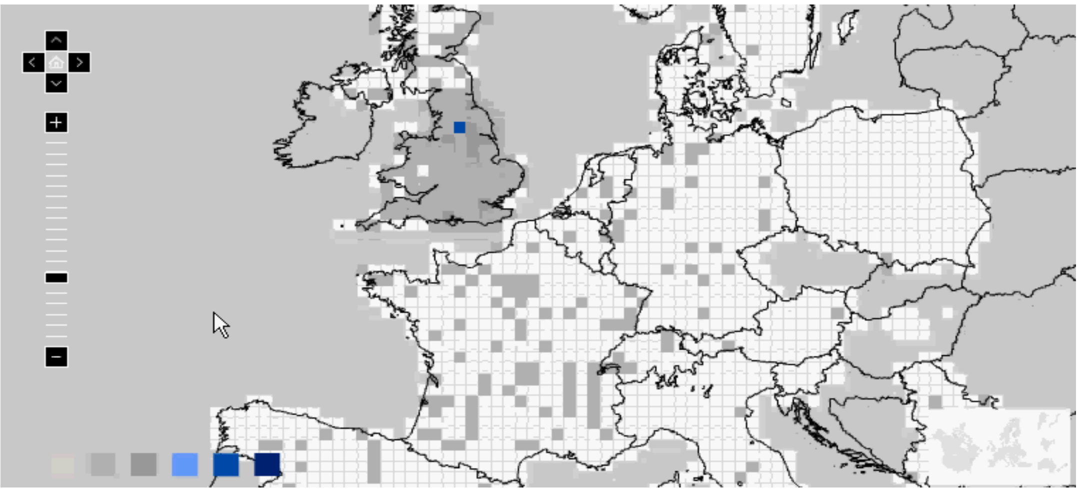


Figure 4.18: Concentration of Surname 'Batty' in Europe (40km grid)

The following figure 4.19 shows the 40km grid level screen shot for forename 'Mike' in United States.

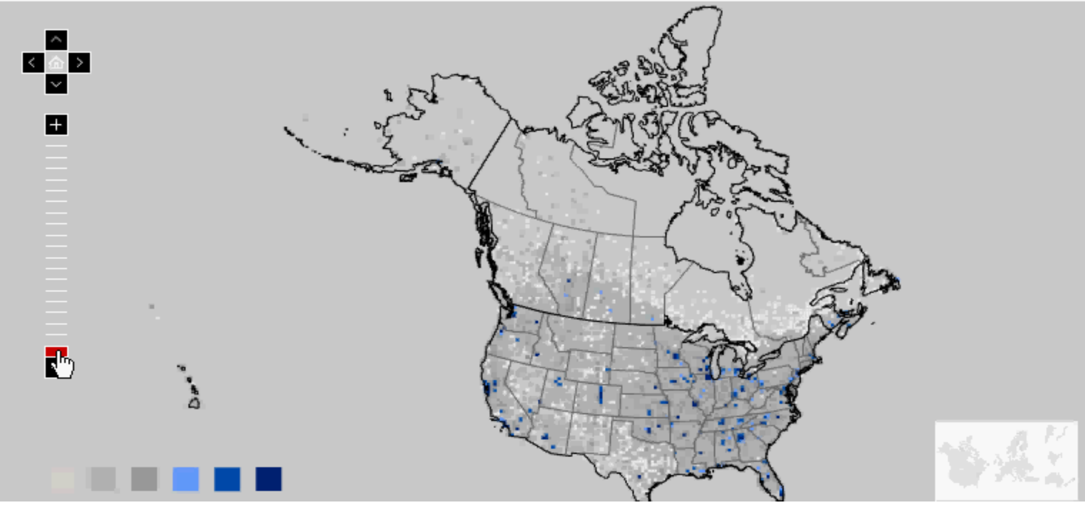


Figure 4.19: Concentration of Forename 'Mike' in North America (40km grid)

The following figure 4.20 shows the 40km grid level screen shot for forename 'Mike' in Europe.

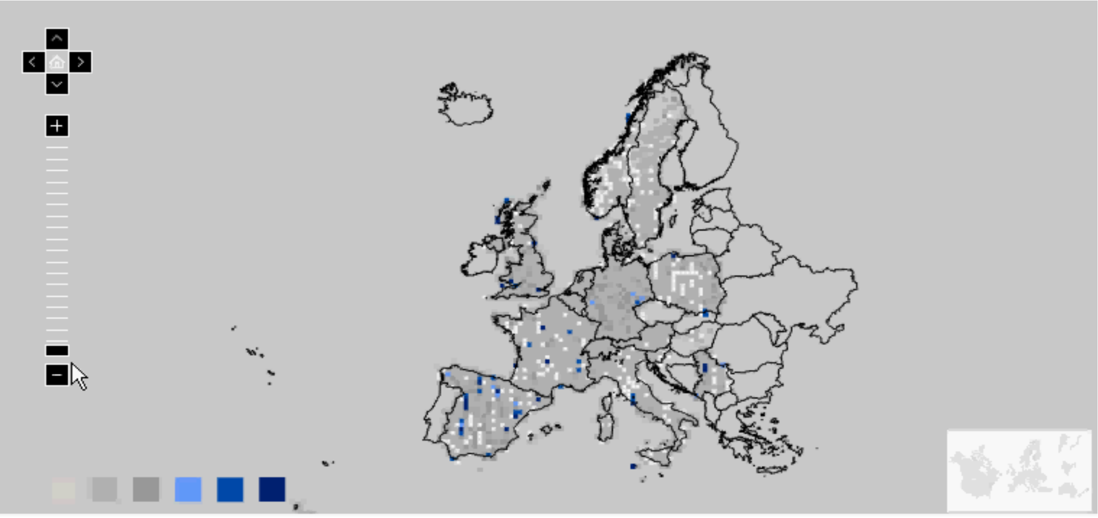


Figure 4.20: Concentration of Forename 'Mike' in Europe (40km grid)

The following figure 4.21 shows the 40km grid level screen shot for forename 'Mike' in Europe.

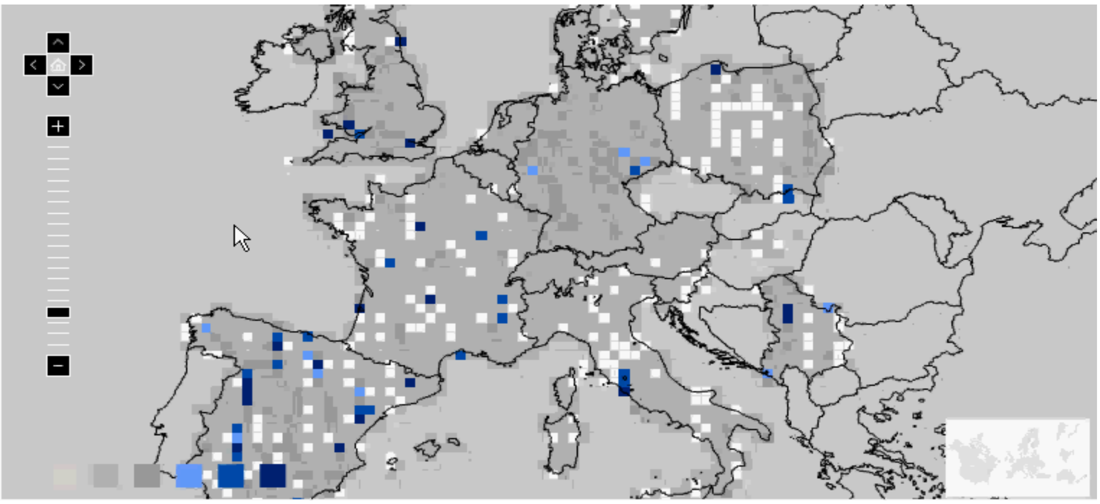


Figure 4.21: Concentration of Forename 'Mike' in Europe (40km grid)

5KM GRID LEVEL SCREEN SHOTS

The following figure 4.22 shows the 5km grid level screen shot for surname 'Batty' in west of the United States.

Figure 4.22: Concentration of Surname 'Batty' in the west of United States (5km grid)

5KM GRID LEVEL SCREEN SHOTS

The following figure 4.23 shows the 5km grid level screen shot for surname 'Batty' in the east of the United States.

Figure 4.23: Concentration of Surname 'Batty' in the east of the United States (5km grid)

DISADVANTAGES OF GRID BASED MAPPING APPROACH

There were some processing, visualisation, and usability issues associated with the grid based mapping approach.

Firstly, the web application was slow to run in the browser because of a large number of objects in the flash files. As a property of vector images, because they are drawn dynamically, it makes rendering of the final output slow if the number of objects becomes large. Thus even if objects at 5km grid level were divided into three parts (North America, Europe, and Asia) , the application was still slow to run in the browser because of the large number of objects in the flash file. From a usability perspective, it was very difficult for a novice user to understand the location of grid cells. There was no labelling of the grid cells, which made it difficult to understand the underlying geographical areas. From a visualisation and usability point of view, there was a need that the maps should show the labelling of the places clearly.

These disadvantages led the move to another alternative of flash maps which might be characterised as the "Google maps version of flash maps". This version of flash maps works in the same way as grid based maps, but it uses country, admin1, and admin2 areas as the polygons. This means that there is smaller number of flash objects. This enabled better management of the flash objects and it increased the speed at which they are rendered. In terms of usability, it was decided that this version of flash maps would show the labelling of flash polygons when a user brushed the mouse over it. Thus it was proposed that the revised web application would offer the following enhancements over the prototype:

- The new version would be based on a 'Google maps version of flash maps' with no grid cells. This makes the number of polygons (flash objects) much smaller and allows the web application to run faster in the browser.
- The new version of the web application should label every map polygon, so that a user can identify the areas and feel confident using the application.
- Zoom in/out between different levels of geographies should be done by clicking the polygons rather than using a scroll bar.

4.3.6 WORLDNAMES FINAL WEBSITE

Based on the usability input from the beta version, the final version of the 'Worldnames' website was developed to account for the processing, usability, and visualisation issues identified in the beta version. Maps for this version of the web application have three zoom levels: these are Country Level, Region Level, and County Level. A country level map is the representation of countries around the world, a region level map shows the geographies which are smaller than countries (e.g. states in United States represents this level), and a county level map shows the geographies which are even smaller than region level geographies (e.g. counties in United States and United Kingdom are represented at this level).

It was required that a user could navigate between these levels, fulfilling the need for an interactive mapping application. By using ESRI ArcGIS, a point in polygon operation was performed for all geo-referenced points for names data. This enabled the creation of maps for all countries at three different levels (Country, Region, and County) or geography. Maps from ArcGIS were then exported to Flash files, which were used by the ASP.NET application (web application) to plot the surnames data. So, ASP.NET web application retrieves data from an Oracle database, converts it to a particular format as needed by flash files, and writes that to XML files. Flash files, then, read XML files and show the data by painting corresponding polygons according to the values in XML files.

From a usability perspective, these maps are easy to use and interpret. They allow the users to click on any part of the map to zoom into the next finest level of geography. In this way these maps allow to build up an interface which is interactive and fast to run.

In addition to showing the data painted in maps, this website also shows different statistics for the searched ‘Surname’. These statistics are:

- Ethnic roots of the surname as classified by (Mateos 2007).
- Top countries, which show the ranked top ten countries for the searched surname where the concentration of the surname (in terms of share of the total population) is higher than all others.
- Top regions, which show the ranked top ten regions for the searched surname where the concentration of surname is higher than all others.

- Top cities, which show the ranked top ten cities of the searched surname where the concentration of surname is higher than others.
- Top Forenames, which show the ranked top ten forenames associated with the searched surname, where these are known.

SCREEN SHOTS OF THE WEBSITE

Screen shots of the website are shown below with their textual description. Screen shots are divided into three sections:

- Name Search
- Area Search
- Ethnicity Search

NAME SEARCH

The following screen shot (Figure 4.24) shows the concentration of surname 'Singleton' in different countries.

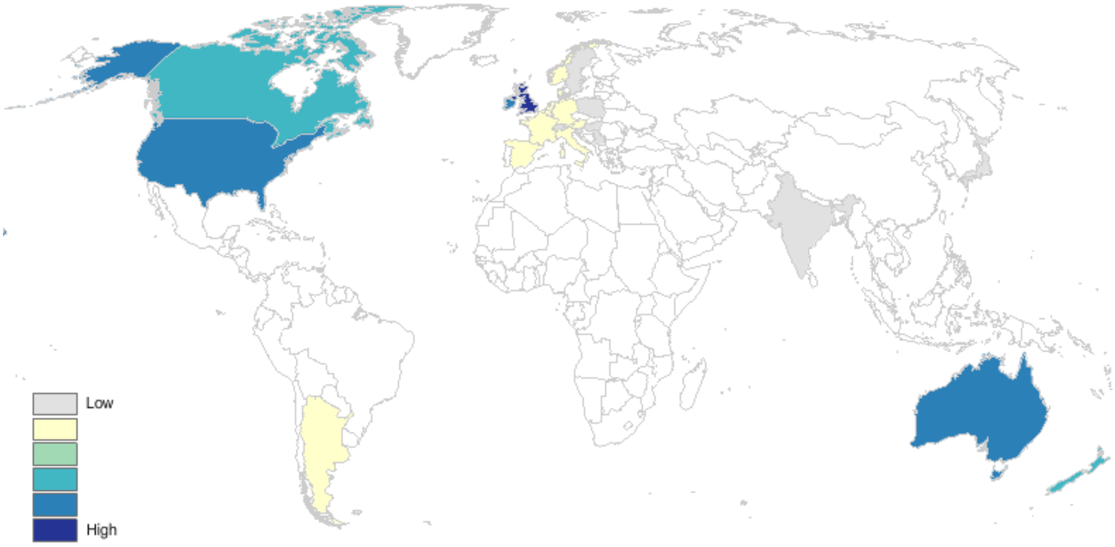


Figure 4.24: Concentration of Surname 'Singleton' in different countries

The following screen shot (Figure 4.25) shows the concentration of surname 'Singleton' in Europe.

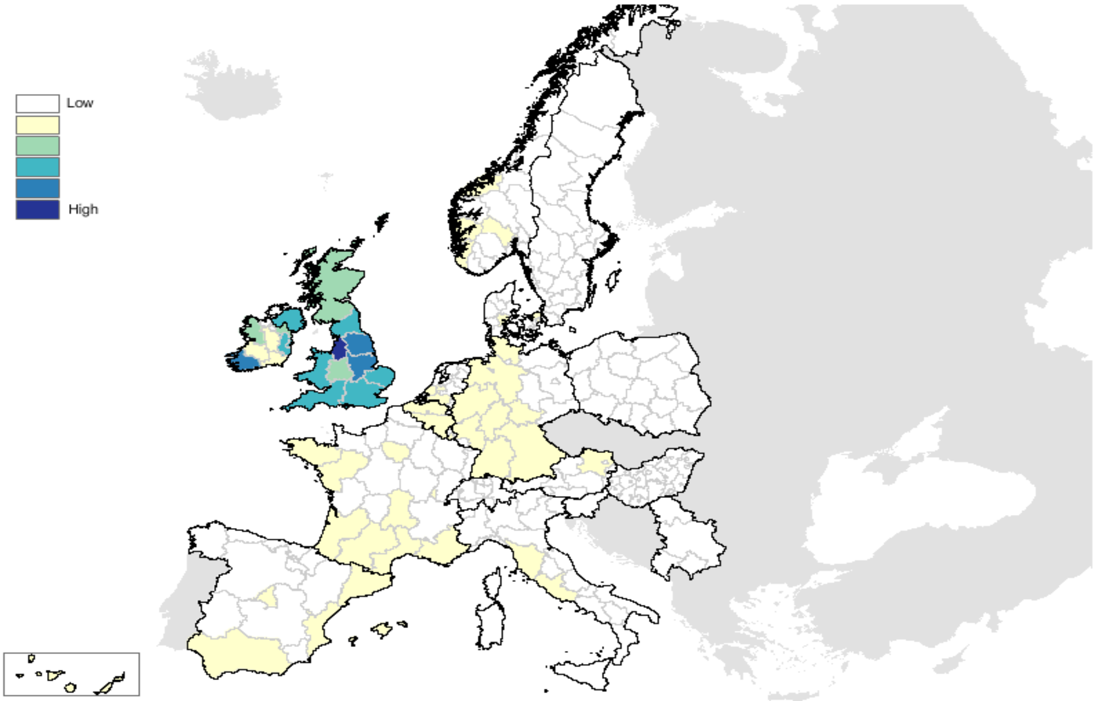


Figure 4.25: Concentration of Surname 'Singleton' in Europe

The following screen shot (Figure 4.26) shows the concentration of surname 'Singleton' in North America.

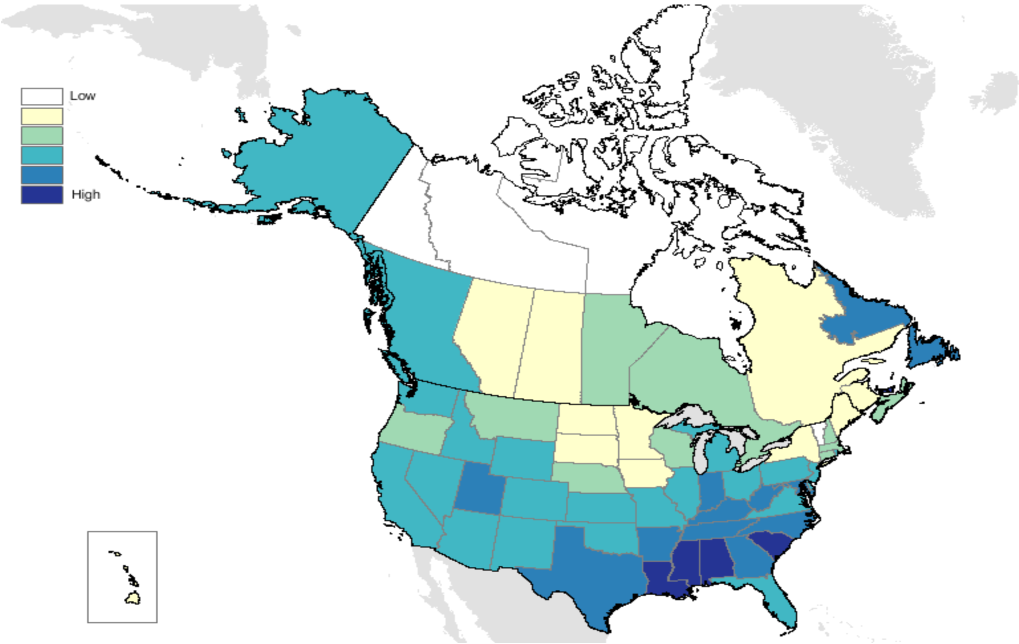


Figure 4.26: Concentration of Surname 'Singleton' in North America

The following screen shot (Figure 4.27) shows the concentration of surname 'Singleton' in the counties of Alabama, United States.

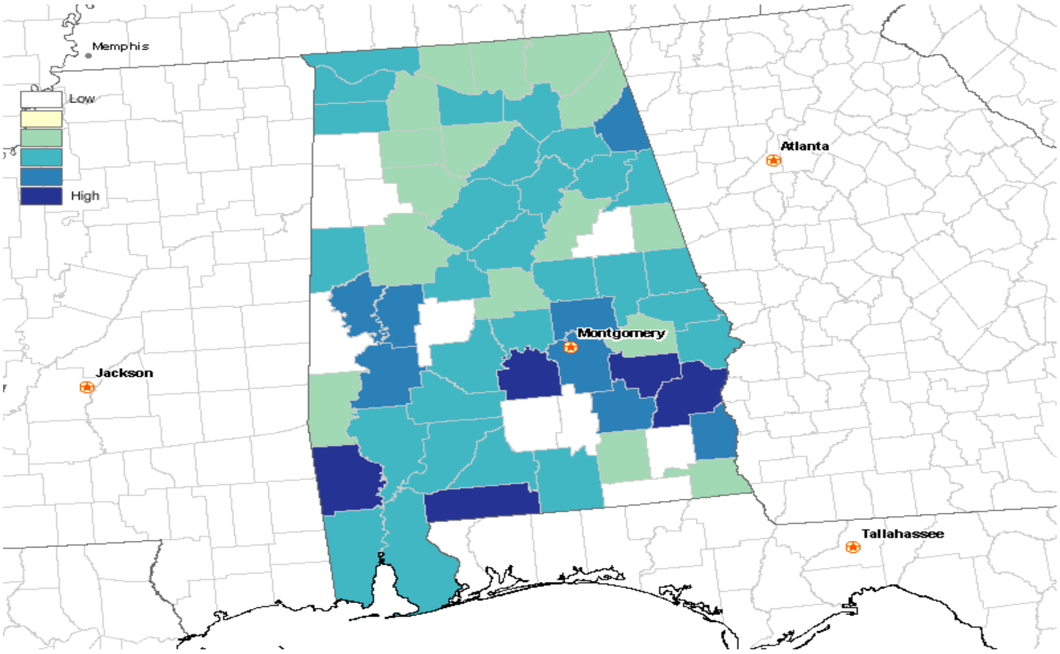


Figure 4.27: Concentration of Surname 'Singleton' in counties of Alabama, United States

The following screen shot (Figure 4.28) shows the concentration of surname ‘Singleton’ in the United Kingdom.

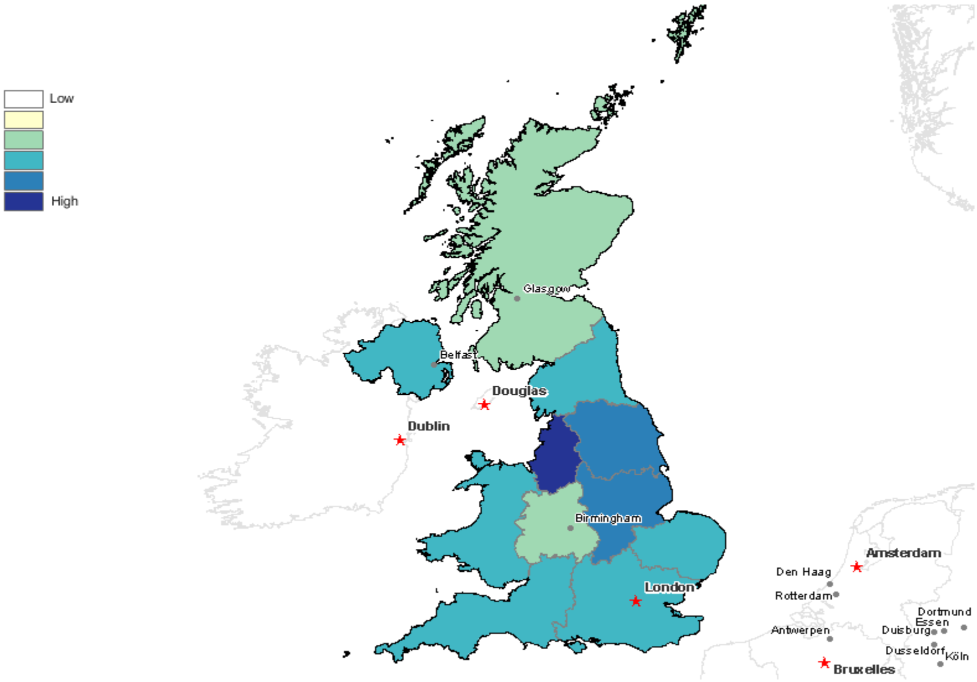


Figure 4.28: Concentration of Surname ‘Singleton’ in the United Kingdom

The following screen shot (Figure 4.29) shows the concentration of the surname ‘Singleton’ in the counties of United Kingdom.

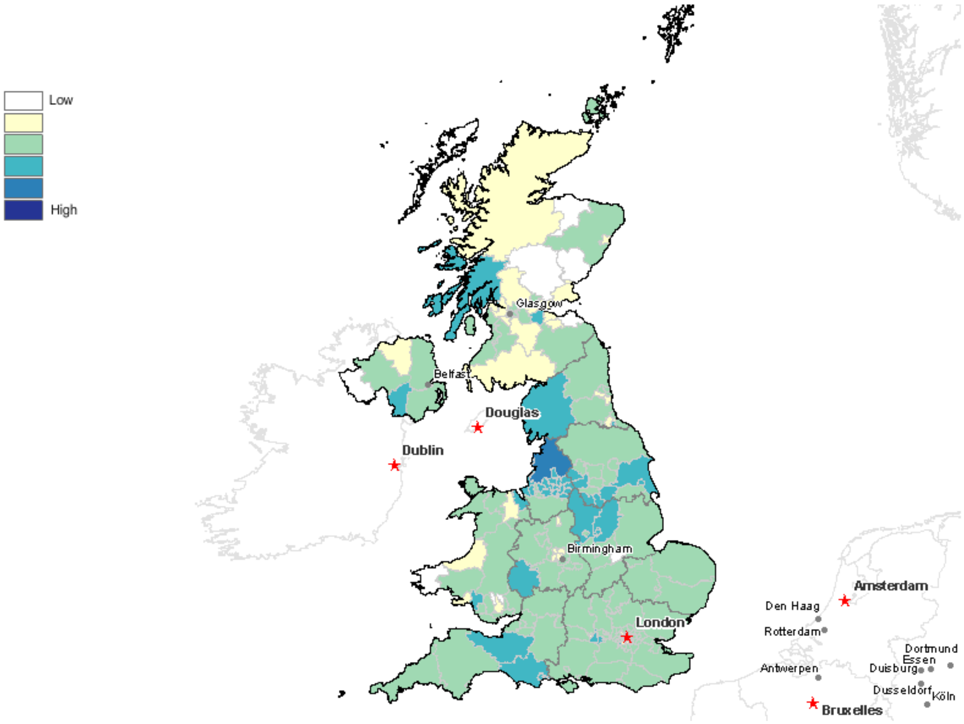
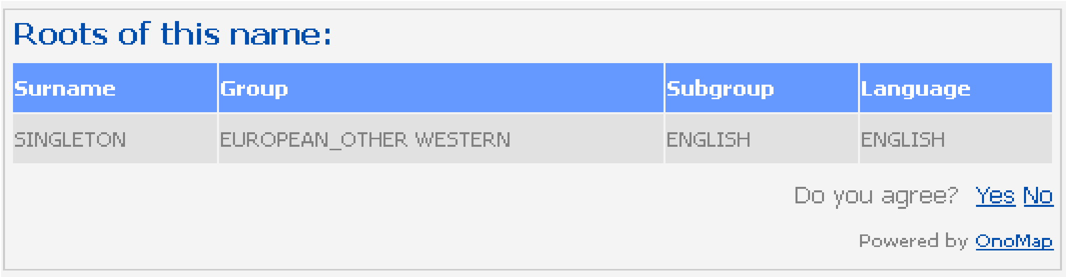


Figure 4.29: Concentration of the surname 'Singleton' in the counties of United Kingdom

The following screen shot (Figure 4.30) shows the Onomap roots of surname 'Singleton'.



Roots of this name:

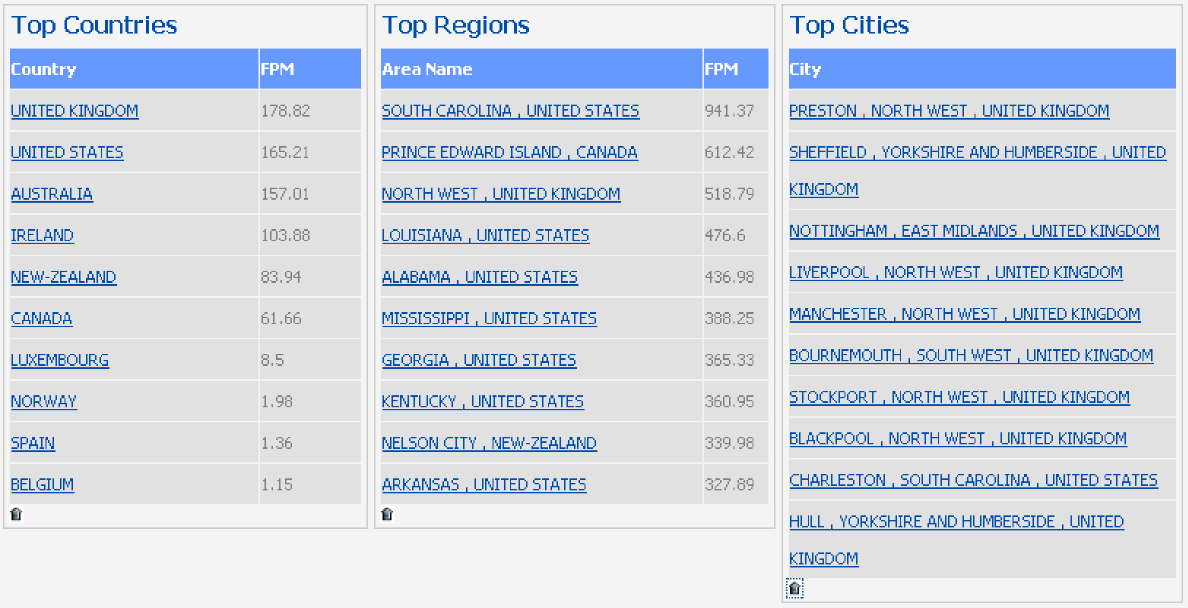
Surname	Group	Subgroup	Language
SINGLETON	EUROPEAN_OTHER WESTERN	ENGLISH	ENGLISH

Do you agree? Yes No

Powered by OnoMap

Figure 4.30: Onomap roots of Surname 'Singleton'

The following screen shot (Figure 4.31) shows the top countries, top regions, and top cities of the surname ‘Singleton’.



Top Countries

Country	FPM
UNITED KINGDOM	178.82
UNITED STATES	165.21
AUSTRALIA	157.01
IRELAND	103.88
NEW-ZEALAND	83.94
CANADA	61.66
LUXEMBOURG	8.5
NORWAY	1.98
SPAIN	1.36
BELGIUM	1.15

Top Regions

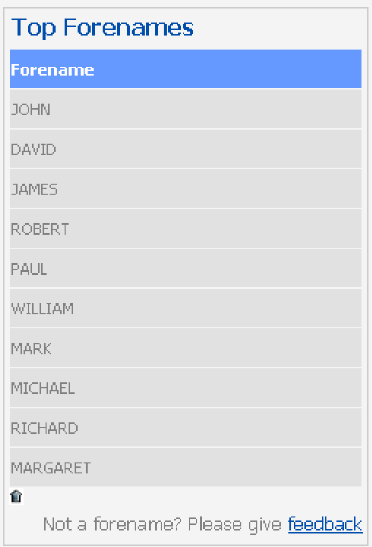
Area Name	FPM
SOUTH CAROLINA , UNITED STATES	941.37
PRINCE EDWARD ISLAND , CANADA	612.42
NORTH WEST , UNITED KINGDOM	518.79
LOUISIANA , UNITED STATES	476.6
ALABAMA , UNITED STATES	436.98
MISSISSIPPI , UNITED STATES	388.25
GEORGIA , UNITED STATES	365.33
KENTUCKY , UNITED STATES	360.95
NELSON CITY , NEW-ZEALAND	339.98
ARKANSAS , UNITED STATES	327.89

Top Cities

City
PRESTON , NORTH WEST , UNITED KINGDOM
SHEFFIELD , YORKSHIRE AND HUMBERSIDE , UNITED KINGDOM
NOTTINGHAM , EAST MIDLANDS , UNITED KINGDOM
LIVERPOOL , NORTH WEST , UNITED KINGDOM
MANCHESTER , NORTH WEST , UNITED KINGDOM
BOURNEMOUTH , SOUTH WEST , UNITED KINGDOM
STOCKPORT , NORTH WEST , UNITED KINGDOM
BLACKPOOL , NORTH WEST , UNITED KINGDOM
CHARLESTON , SOUTH CAROLINA , UNITED STATES
HULL , YORKSHIRE AND HUMBERSIDE , UNITED KINGDOM

Figure 4.31: Top countries, top regions, and top cities of the surname ‘Singleton’

The following screen shot (Figure 4.32) shows the top forenames known to be associated with the surname ‘Singleton’.



Top Forenames

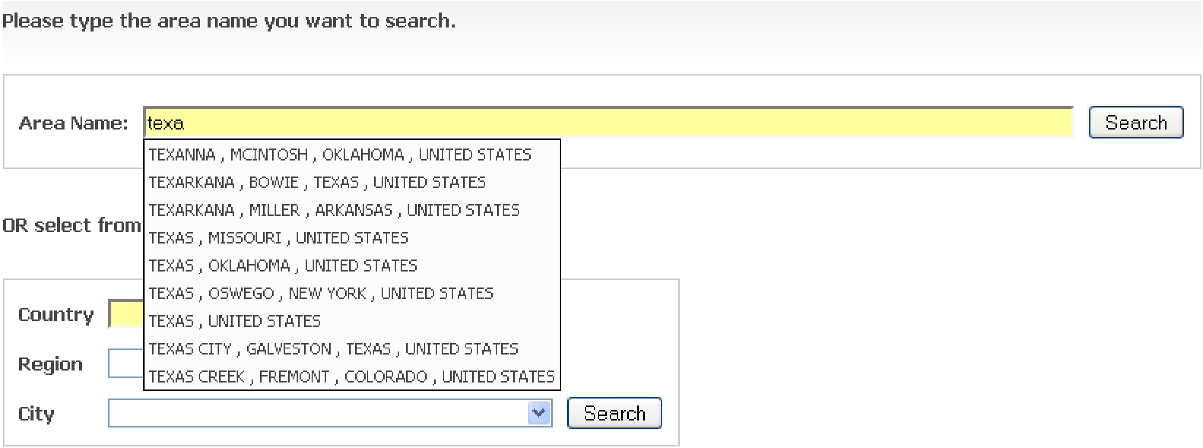
Forename
JOHN
DAVID
JAMES
ROBERT
PAUL
WILLIAM
MARK
MICHAEL
RICHARD
MARGARET

Not a forename? Please give feedback

Figure 4.32: Top Forenames associated with the surname ‘Singleton’

AREA SEARCH

The following screen shot (Figure 4.33) shows the area search facility.



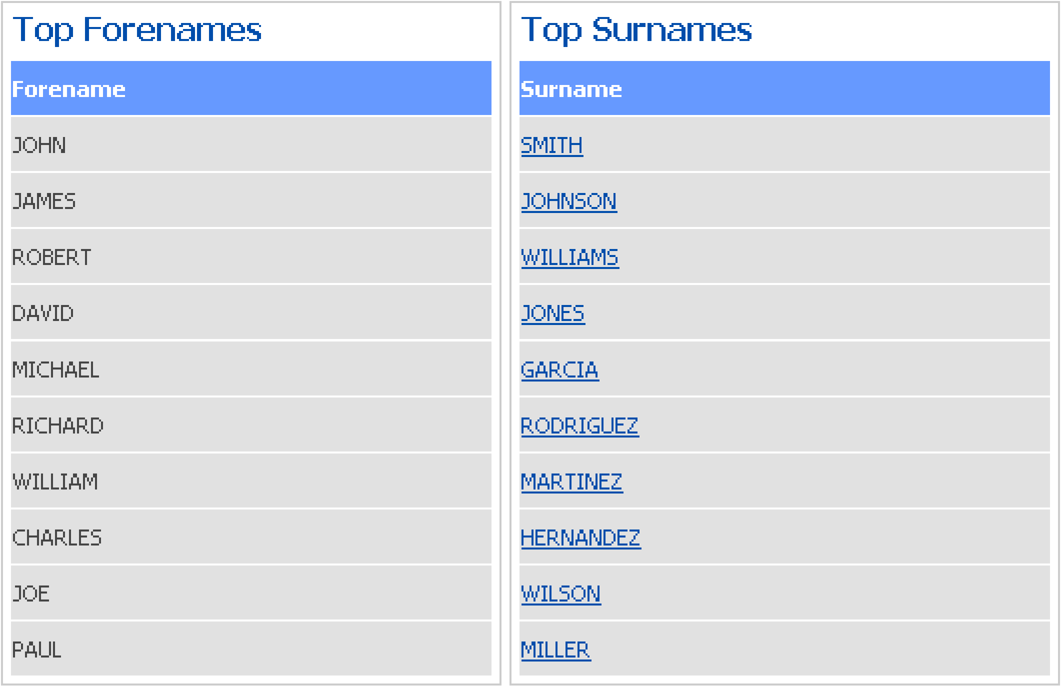
Please type the area name you want to search.

Area Name: texa

TEXANNA , MCINTOSH , OKLAHOMA , UNITED STATES
TEXARKANA , BOWIE , TEXAS , UNITED STATES
TEXARKANA , MILLER , ARKANSAS , UNITED STATES
TEXAS , MISSOURI , UNITED STATES
TEXAS , OKLAHOMA , UNITED STATES
TEXAS , OSWEGO , NEW YORK , UNITED STATES
TEXAS , UNITED STATES
TEXAS CITY , GALVESTON , TEXAS , UNITED STATES
TEXAS CREEK , FREMONT , COLORADO , UNITED STATES

Figure 4.33: Area search

The following screen shot (Figure 4.34) shows top forenames and top surnames for the area name 'TEXAS, UNITED STATES'.



Top Forenames

Forename
JOHN
JAMES
ROBERT
DAVID
MICHAEL
RICHARD
WILLIAM
CHARLES
JOE
PAUL

Top Surnames

Surname
SMITH
JOHNSON
WILLIAMS
JONES
GARCIA
RODRIGUEZ
MARTINEZ
HERNANDEZ
WILSON
MILLER

Figure 4.34: Top Forenames and top Surnames for the area name 'TEXAS, UNITED STATES'

ETHNICITY SEARCH

The following screen shot (Figure 4.35) shows the concentration of 'ANGLO-SAXON: ENGLISH' ethnicity in different countries of the world, as classified by the Onomap software (Mateos 2007).

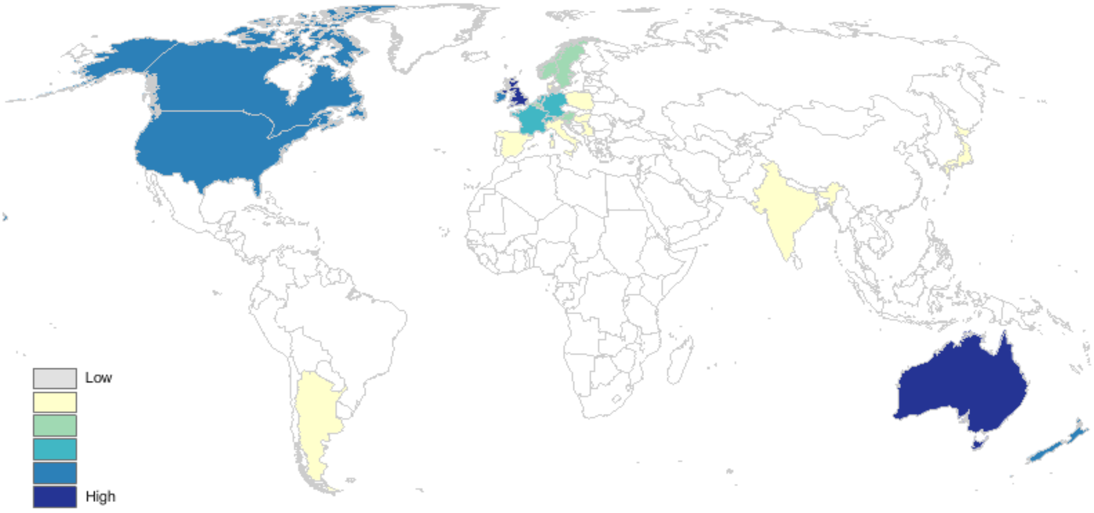


Figure 4.35: Concentration of 'ANGLO-SAXON: ENGLISH' in different countries

The following screen shot (Figure 4.36) shows the concentration of the 'ANGLO-SAXON: ENGLISH' ethnicity in North America.

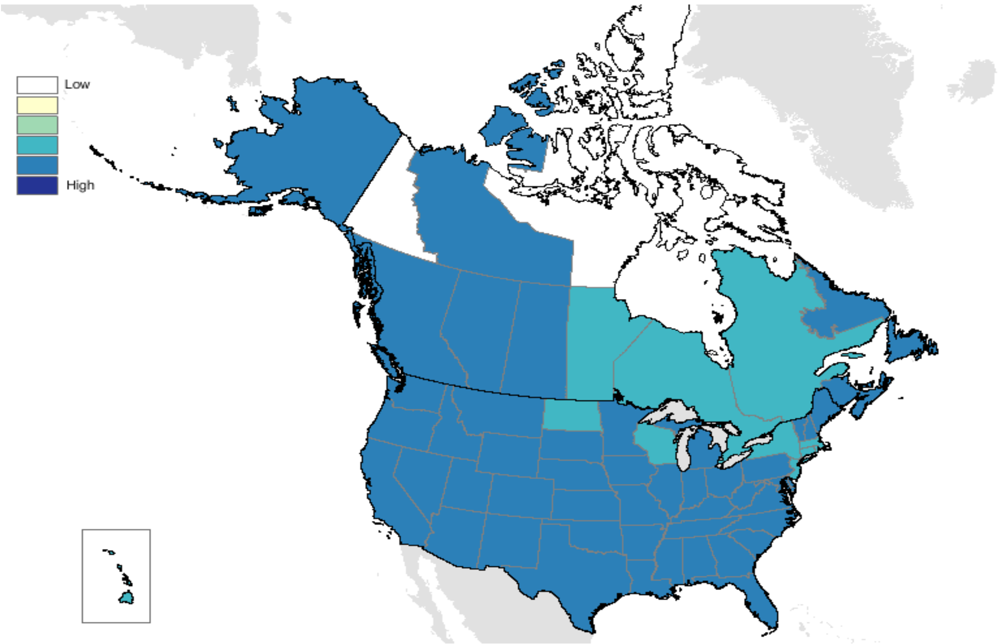


Figure 4.36: Concentration of the 'ANGLO-SAXON: ENGLISH' in North America

The following screen shot (Figure 4.37) shows the concentration of the 'ANGLO-SAXON: ENGLISH' ethnicity in Europe.

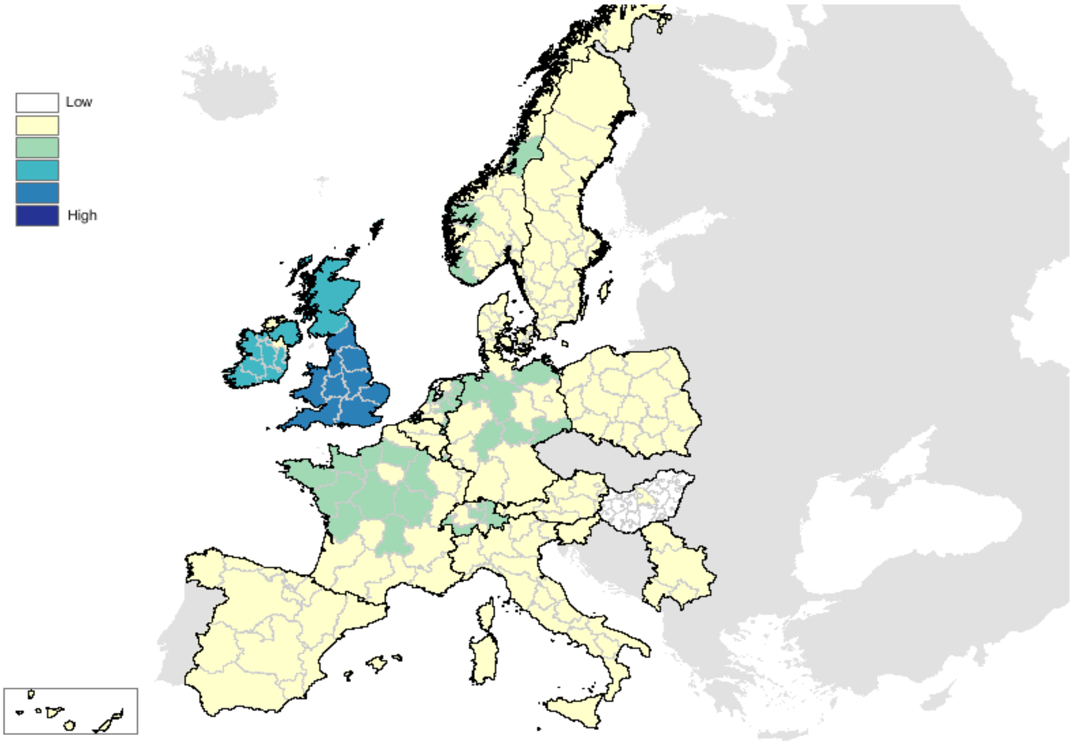
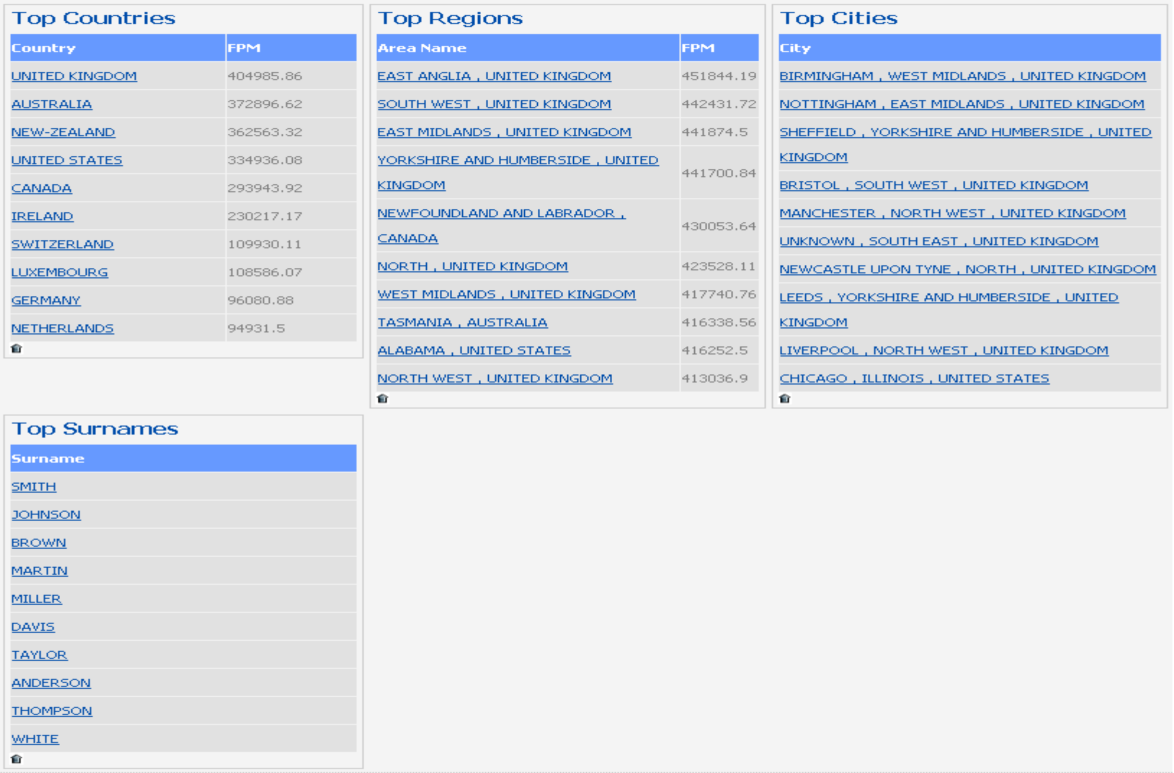


Figure 4.37: Concentration of the 'ANGLO-SAXON: ENGLISH' in Europe

The following screen shot (Figure 4.38) shows the top countries, regions, cities, and surnames of the 'ANGLO-SAXON: ENGLISH' ethnicity in Europe.



Top Countries

Country	FPM
UNITED KINGDOM	404985.86
AUSTRALIA	372896.62
NEW-ZEALAND	362563.32
UNITED STATES	334936.08
CANADA	293943.92
IRELAND	230217.17
SWITZERLAND	109930.11
LUXEMBOURG	108586.07
GERMANY	96080.88
NETHERLANDS	94931.5

Top Regions

Area Name	FPM
EAST ANGLIA , UNITED KINGDOM	451844.19
SOUTH WEST , UNITED KINGDOM	442431.72
EAST MIDLANDS , UNITED KINGDOM	441874.5
YORKSHIRE AND HUMBERSIDE , UNITED KINGDOM	441700.84
NEWFOUNDLAND AND LABRADOR , CANADA	430053.64
NORTH , UNITED KINGDOM	423528.11
WEST MIDLANDS , UNITED KINGDOM	417740.76
TASMANIA , AUSTRALIA	416338.56
ALABAMA , UNITED STATES	416252.5
NORTH WEST , UNITED KINGDOM	413036.9

Top Cities

City
BIRMINGHAM , WEST MIDLANDS , UNITED KINGDOM
NOTTINGHAM , EAST MIDLANDS , UNITED KINGDOM
SHEFFIELD , YORKSHIRE AND HUMBERSIDE , UNITED KINGDOM
BRISTOL , SOUTH WEST , UNITED KINGDOM
MANCHESTER , NORTH WEST , UNITED KINGDOM
UNKNOWN , SOUTH EAST , UNITED KINGDOM
NEWCASTLE UPON TYNE , NORTH , UNITED KINGDOM
LEEDS , YORKSHIRE AND HUMBERSIDE , UNITED KINGDOM
LIVERPOOL , NORTH WEST , UNITED KINGDOM
CHICAGO , ILLINOIS , UNITED STATES

Top Surnames

Surname
SMITH
JOHNSON
BROWN
MARTIN
MILLER
DAVIS
TAYLOR
ANDERSON
THOMPSON
WHITE

Figure 4.38: Top countries, regions, cities, and surnames of the 'ANGLO-SAXON: ENGLISH'

4.3.7 WEBSITE SUCCESS AND USER LOAD MANAGEMENT

This website was launched on 28th August, 2008 under the URL: www.publicprofiler.org/worldnames. It became very popular in few days from its launch, with user hits reaching to 1 million during the first 13 days of launch. The following graph (Figure 4.39) shows the number of hits on the website from 30th August to 11th September, 2008.

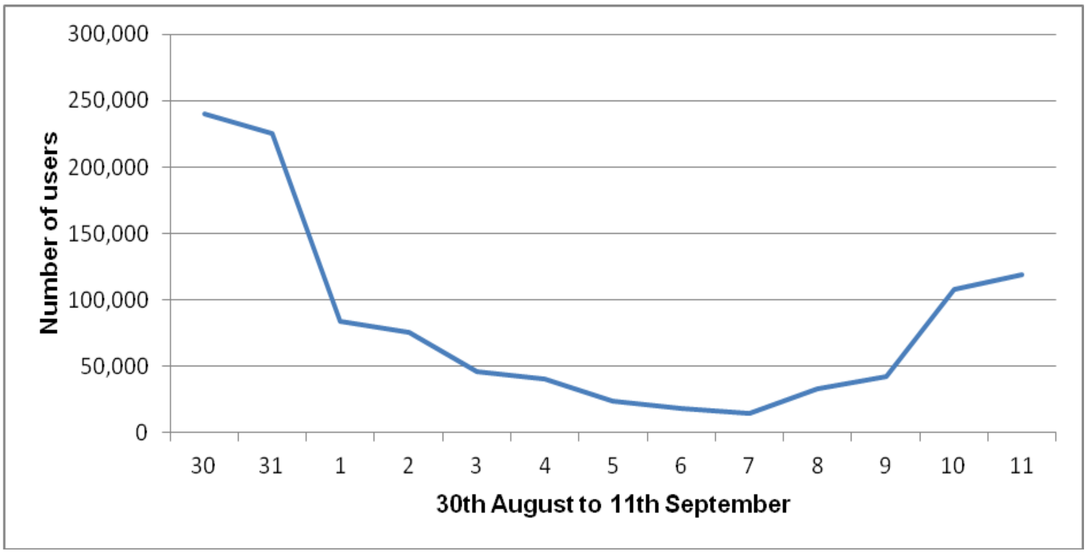


Figure 4.39: Number of user hits for the period 30th August to 11th September, 2008

Under this enormous user load, the website performed well as far as the load management and concurrent user access is concerned. Using 3-tier ASP.NET and Oracle architecture, it successfully served the concurrent user requests without errors. This website also attracted attention from the international media of different countries. The following table shows some of the media links that drove traffic to the website.

Media Name	Website link
BBC Online	http://news.bbc.co.uk/1/hi/uk/7588968.stm
BBC Radio 4	http://www.spatial-literacy.org/wp-content/uploads/Video/today.mp3
The Independent	http://www.independent.co.uk/news/science/putting-you-on-the-map-the-website-that-pinpoints-where-your-name-is-in-the-world-913312.html
Channel4 News	http://www.channel4.com/news/articles/science_technology/global+surname+website+launched/2438237
Telegraph	http://www.telegraph.co.uk/news/uknews/2648378/Whats-in-a-name-A-great-deal-say-researchers.html
The Scotsman	http://news.scotsman.com/uk/New-family-history-website-Where.4442462.jp
New Zealand Herald	http://www.nzherald.co.nz/technology/news/article.cfm?c_id=5&objectid=10529791

SG.HU (Hungary)	http://www.sg.hu/cikkek/62405/nevegyezeseket_kutat_fel_egy_kereso
ZDNet.de (Germany)	http://www.zdnet.de/news/wirtschaft_telekommunikation_suchmaschine_findet_weltweit_namensverwandte_story-39001023-39195547-1.htm
Telekom Presse (Austria)	http://www.telekom-presse.at/channel_internet/news_34072.html
Presstext (Germany)	http://pressetext.de/news/080901002/suchmaschine-findet-weltweit-namensverwandte/?phrase=publicprofiler
Globo (Brazil)	http://oglobo.globo.com/mundo/mat/2008/08/30/site_localiza_sobrenomes_ao_redor_do_mundo-548019784.asp
Yahoo News Hong Kong	http://hk.news.yahoo.com/article/080831/4/7z90.html
Saigon Giai Phong Online (Vietnam)	http://www.sggp.org.vn/thegioi/2008/8/163741/
News.com.au (Australia)	http://www.news.com.au/story/0,23599,24269490-401,00.html
Mbl.is (Iceland)	http://mbl.is/mm/frettir/taekni/2008/08/30/vefur_kortleggur_aettarnofn_i_heiminum/
Computerwoche.de (Germany)	http://www.computerwoche.de/knowledge_center/web/1872583/
WebUser	http://www.webuser.co.uk/news/266673.html
Worthing Herald	http://www.worthingherald.co.uk/latest-london-news/Family-trees-researched-on-website.4443307.jp
Portugal Diario	http://diario.iol.pt/tecnologia/public-profiler-sobrenome-historia-internet-lugares-tecnologia/986289-4069.html
The Dominion Post	http://www.stuff.co.nz/dominion-post/archive/national-news/606280

Table 4.2: Media links where Worldnames website was featured

4.3.8 LESSONS LEARNED FROM THE VISUALISATION OF 'WORLDNAMES' DATABASE

There are a number of lessons that were learned from the visualisation of 'Worldnames' database. The first was the management of a large number of concurrent user requests. The web application proved to be flexible enough to manage a huge user load. As set out in the previous section, the huge number of user requests during the first few days of the launch, provides ample evidence that the website performed really well in different circumstances. From the point of view of creating geodemographic classifications on the fly, this knowledge and experience was recognised as important in the design of a system to manage potentially large numbers of concurrent user requests.

The second lesson learned was the importance of choosing a suitable visualisation technology for different circumstances. Because slippy maps would have taken longer for serving maps of individual 'surnames', flash mapping technology was chosen instead. This technology serves maps on the fly faster than slippy maps. This technology was thus subsequently chosen for the web application in which users can create their own geodemographic classifications on the fly.

4.4 CONCLUSION

In conceptual terms, geovisualisation can be thought of as an inherently important step in the conception, representation and analysis of geographic phenomena (Longley et al 2011: 327). There are vagaries inherent in assembling, concatenating and conflating diverse data sources prior to creating geodemographic indicators, and these problems are multifaceted in an era in which there is demand to create consistent indicators that transcend nation states and other administrative divisions. In this context, this chapter has discussed the visualisation aspects for creating the 'Worldnames' web application. It discussed different visualisation technologies and options available for a very large socioeconomic dataset, and evaluated the advantages and disadvantages of each of them.

An additional consideration in creating this particular dataset was the huge public interest that it was likely to generate, resulting in a requirement that the website be accessible to very large numbers of users at any one time. This chapter also justified the decision to adopt Flash mapping as the solution for this website. In the second part of this chapter, a detailed description of the user requirements and web application design were explained. The 'Worldnames' beta website was explained with its usability and visualisation issues and those issues were used for the development of the final version of 'Worldnames' web application.

In the next chapter, this experience of using different techniques to display a large and complex dataset will be extended to a broader consideration of the sources of data to develop geodemographic information systems.

5 CHAPTER 5: DATA SOURCES REVIEW

5.1 INTRODUCTION

Previous chapters have demonstrated how through the integration of a range of innovative new data sources a database of the distribution of names around the world can be created and visualised. The emphasis in these chapters was upon standardisation of data across space, when they pertained to different political or national jurisdictions. Chapters 3 and 4 described the creation of 'Worldnames' database and web application. 'Worldnames' website was an attempt to devise fast on the fly rendering of a single variable extracted from a very large database. The last chapter also dealt with important issues of visualisation and user interaction. A grid based flash mapping technique was devised for the visualisation, but it had disadvantages. Thus it was replaced by a better visualisation technique.

Thus, previous chapter emphasised on using a single indicator (‘surname’) extracted from a large database. But, geodemographic classifications are the result of intensive cluster analysis on a large number of variables e.g. Output Area Classification (Vickers & Rees, 2007) was created by using 41 census variables. These variables accounted for the socio-economic characteristic of the population living in different output areas of the United Kingdom. Some commercial geodemographic classifications are created at postcode level e.g. Mosaic (by Experian, UK) is created at postcode level in the United Kingdom (Experian 2011).

Choosing appropriate data sources when building geodemographic classifications plays an important role in the classification procedure. Currently the core to most geodemographic classifications remains census data, since they pertain to a wide range of social, economic and demographic circumstances of people at small area (neighbourhood) level. There are different methodologies for creating geodemographic classifications but these differ based on the proprietary solution offered. The difference in methods is based on the data set used, the number of socio-economic variables incorporated, the data normalisation technique applied to the data set and the weightings applied to the constituent variables of the classification. All of the neighbourhood classifications developed for the UK market to date are the result of performing variants of a basic cluster analysis methodology. Purely census-based classifications involve clustering census ‘neighbourhoods’. Past (pre 2001) clustering techniques in the UK were applied to enumeration districts (EDs), but since 2001 the neighbourhoods in question are census output areas (OAs). EDs averaged about 170 households, while OAs are smaller at an average of 124 households each.

As Sleight (2004) suggests, the most useful data for producing neighbourhood classifications remain census area statistics. 2001 Census area statistics are available at OA (Output Area) level and can be used to create geodemographic classifications. Some companies (e.g. Experian UK, CACI UK) also use credit reference agency data, County Court Judgments (CCJs), the Electoral Roll, postcode address file, and retail access data as ancillary sources in the creation of geodemographic classifications. Promotional material for the Mosaic classification by Experian (Nottingham, UK) states that it is based on 60% Census area statistics and 40% data from credit data, CCJs, the Electoral Roll, the postal address file, the register of company directors, retail access, and lifestyle data. This enables the classification to cover a wider range of socio-economic variables and is deemed to create a more responsive classification.

The following table 5.1 shows the classifications created by different companies and the data sources used in creating those classifications.

Organisation	Classification System	Number of input variables	Number of clusters	Non-census data used?
CACI	ACORN	79	d) 6 e) 17 f) 54	No
Experian	Mosaic	87	c) 11 d) 52	Credit data, CCJs, electoral roll, postal address file, company directors, retail access
EuroDirect	CAMEO	48	c) 9 d) 50	No
	MicroVision	185	d) 11 e) 52 f) 200	Lifestyle data, company directors, share ownership, electoral roll, CCJs, risk indices, unemployment statistics

	DEFINE	146	d) 10 e) 50 f) 1050	Credit data, electoral roll, unemployment statistics, insurance ratings
Claritas UK	PRIZM	59 (+188)	d) 4 e) 19 f) 72	Lifestyle data, share ownership, company directors, unemployment statistics, postal address file, births and deaths
	SuperProfiles	120 (+130)	d) 10 e) 40 f) 160	Credit data, CCJs, TGI, Electoral Roll

Table 5.1: Geodemographic Classifications and data sources used

Source: Sleight (2004: 49)

The rest of this chapter is focused on investigating the different data sources available for the creation of new and innovative geodemographic classifications. In the first section, this chapter explains the creation of a software service to extract live XML feeds of data from the ONS (Office for National Statistics) NeSS (Neighbourhood Statistics) API (Office for National Statistics, 2009). This API provides an interface to get live XML feeds of different data source from the ONS (Office for National Statistics) website. The second part of this chapter describes the different data sources available in the public sector and includes a review of their structure, spatial coverage and limitations. Finally, the third part of this chapter describes the creation of two new data sources through the amalgamation of house price data and ethnicity data, and discusses a way to create new and innovative data sources (such as the names data discussed previously) by combining these available data sources. This chapter concludes with the selection of a data source to be used in creating interactive geodemographic classifications and justification of the data source selected.

5.2 LIVE XML DATA EXTRACTION WEB SERVICE

The ONS (Office of National Statistics) provides different types of data sources collected by the public sector. Census data, Land Registry data, DCSF (Department of Children, Schools and Families) Education data are examples of such data sources that ONS provides and can be downloaded (from www.neighbourhood.statistics.gov.uk). The ONS has launched an XML based API, through which users can get live XML feeds of the data. User applications can use these data for their own purposes. This API is called the NeSS (Neighbourhood Statistics) Data Exchange API. For the rest of this thesis this API will be referred as ONS NeSS API. Office of National Statistics (2009) described NeSS Data Exchange API as:

"NeSS Data Exchange API uses a combination of a bespoke NeSS XML format (named NeSS-ML) and a format developed in partnership with the Department of Communities and Local Government (CLG) called LGDX (Local Government Data Exchange). LGDX is the XML standards used by the CLG Data Interchange Hub for the exchange of information relating to performance indicators in Local Area Agreements".

This section describes the creation of a software web service to extract data from the ONS NeSS data Exchange API. This software web service is developed using Java which sends XML feeds of data to the NeSS data Exchange API and retrieves the result back from the API, and afterwards parses and inserts the XML data into the database.

This software service acts as the proof of the concept that live feeds of data from disparate data sources could be extracted and integrated for building geodemographic classifications.

5.2.1 NESS DATA EXCHANGE HIERARCHY ELEMENTS

The ONS NeSS Data Exchange works using a request/response procedure. A user application sends a request in LGDX format to the Ness Data Exchange API; in return the API sends the response, containing the requested data, back to the user application. The user application will need to parse the LGDX format to a local format in order to use the data. The ONS NeSS Data Exchange API works in the hierarchy of data elements and it is necessary that requests be made in a specific hierarchy in order to get the correct data. There are 5 NeSS hierarchy elements which are Subjects, Levels, Dataset Families, Areas, and Area Statistics. What follows is a description of each of these hierarchy elements.

SUBJECTS

Subjects in the NeSS Data Exchange API represent a data source. E.g. ‘2001 Census: Census Area Statistics’ is a subject in NeSS Data Exchange with subject id 16. All the data sources are referred as Subjects Ids when requesting data from ONS NeSS API.

LEVELS

Levels represent the geographical extent of an area. E.g. ‘Counties, Local Authorities, Output Areas etc’ are levels in the NeSS API. Within each Subject, data are stored in different levels. It is necessary that all the Levels be extracted first for a specific Subject in order to retrieve data for the required Levels.

DATASET FAMILIES

Dataset Families represent the individual variables in each Subject (Dataset). E.g. ‘Accommodation Type – Household Spaces (UV56)’ represents a dataset family in the subject ‘2001 Census: Census Area Statistics’. Data for all Levels are stored at the level of the Dataset Family in the NeSS data store.

AREAS

Areas are the geographical areas for which a user wants to extract data. E.g. ‘Newham, Warwickshire, City of London etc’ represent areas in NeSS API. Data are extracted from the API with the combination of Areas and Dataset Families.

AREA STATISTICS

Area Statistics are the type of data stored for a specific combination of ‘Area’ and ‘Dataset Family’. For example, Counts or Percentages of dataset family statistics in different areas represent Area Statistics. There are different types of statistics which the ONS NeSS API provides depending on the type of ‘Dataset Family’ (variable).

The following table shows the measurement units of variables.

Measurement Unit Code	Measurement Unit Name
	Count
%	Percentage
	Rate
	Rate per 1000
	Rate per 10000
	Rank
£	Pound Sterling
	Score
Y	Years
D	Days
H	Hours
Min	Minutes
S	Seconds
Km	Kilometres
Mil	Miles
M	Metres
M2	Metres Squared
Ha	Hectares

Table 5.2: Area Statistics Measurement Units

The following table shows the statistical units of variables.

Statistical Unit Name
Persons
Households
Businesses
Enterprise Units
Offences
Areas
Episodes
Families
Dwellings
Incidents
Cases

Users
Rooms
Lettings
Days
Meter Points
Kilowatt hours
Hereditament

Table 5.3: Area Statistics Statistical Units

5.2.2 NESS DATA EXCHANGE XML REQUEST FORMATS

ONS NeSS Data Exchange API is based on a bespoke LGDX request and responses procedure. All the requests from a client application first need to be formatted to LGDX format and sent to the API. The API sends the result back to the calling application and the resultant LGDX format is parsed to local data format and inserted into the database. This section explains the XML request formats of five hierarchy elements ‘Subject’, ’Level’, ’Dataset Family’, ‘Area’, and ‘Area Statistics’.

Each LGDX request has a standard format, which is given as:

<soap:Envelope xmlns:soap='http://schemas.xmlsoap.org/soap/envelope/' xmlns:wsu='http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-wssecurity-utility-1.0.xsd'>

<soap:Header>

<wsse:Security xmlns:wsse='http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-wssecurity-secext-1.0.xsd' xmlns='http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-wssecurity-secext-1.0.xsd' xmlns:env='http://schemas.xmlsoap.org/soap/envelope/' soap:mustUnderstand='1'>

<wsse:UsernameToken xmlns:wsse='http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-wssecurity-secext-1.0.xsd' xmlns='http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-wssecurity-secext-1.0.xsd'>

<wsse:Username>m.adnan@ucl.ac.uk</wsse:Username>

<wsse:Password Type='http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-username-token-profile-1.0#PasswordText'>adnan123</wsse:Password>

```
</wsse:UsernameToken>

</wsse:Security>

</soap:Header>

<soap:Body xmlns:ns2='http://neighbourhood.statistics.gov.uk/nde/v1-0/discoverystructs'>

</soap:Body>

</soap:Envelope>
```

This is a standard LGDX format specifying the username and password of the user accessing the data. <soap:Body> element of the LGDX request contains the requests for ‘Subject’, ’Level’, ’Dataset Family’, ‘Area’, and ‘Area Statistics’. By changing the contents of <soap:Body> with a bespoke LGDX element, a new request can be created for any specific hierarchy element. The following is a description of LGDX requests for individual hierarchy elements:

SUBJECTS

LGDX format for retrieving all ‘Subjects’ from NeSS Data Exchange API is given below:

```
<ns1:subjectsElement>
```

The response from the ONS NeSS Data Exchange API was parsed into the local data format and the following is a snapshot of the data received.

Subject ID	Subject Name
16	2001 Census: Census Area Statistics
15	2001 Census: Key Statistics
1	Access to Services

Table 5.4: Snapshot of the Subjects data

LEVELS

The LGDX format for retrieving all 'Levels' from ONS NeSS Data Exchange API is given below:

<ns2:LevelTypesElement>

This returns all the 'Levels' of the ONS NESS API. Response from the ONS NeSS Data Exchange API was parsed into the local data format and the following is a snapshot of the data received.

Level ID	Level Name
13	Local Authority
140	Middle Layer Super Output Area
15	Output Area

Table 5.5: Snapshot of the Levels data

AREAS

Retrieving 'Areas' from the ONS NeSS data Exchange API is a recursive and complex process. The LGDX format for retrieving 'Areas' from the NeSS Data Exchange API is given below:

```
<ns1:AreaChildrenElement>
  <AreaId>276707</AreaId>
</ns1: AreaChildrenElement>
```

The above LGDX format extracts the children area ids of an area id. The process of extracting all area ids of United Kingdom works by starting from the top area of United Kingdom, which has an area id of 276699. Requesting the children of this area id returns a number of area ids. Those area ids are parsed to simple text and stored into the database. The process, then, gets all the stored area ids and retrieves those of their children as well and stores them in the database. This process is repeated so that all the children of all the area ids are retrieved and stored into the database. The following table shows a snapshot of the parsed results stored in the database.

Area ID	Area Name	Level ID	Parent Area ID
410789	00MLNQ0035	15	301310
410790	00MLNQ0036	15	301310
410797	00MLNQ0043	15	301310

Table 5.6: Snapshot of the Areas data

DATASET FAMILIES

Requesting all ‘Dataset Families’ is a complex process. The LGDX format for requesting ‘Dataset Families’ from NeSS Data Exchange API is given below:

```
<ns1:DatasetFamiliesElement>
  <SubjectId>16</SubjectId>
  <AreaId>276707</AreaId>
</ns1:DatasetFamiliesElement>
```

This format requires a ‘Subject’ and an ‘Area’ to retrieve the ‘Dataset Families’. The process of extracting all ‘Dataset Families’ works by retrieving all ‘Subject’ and all ‘Areas’ and then making a combination of each ‘Subject’ and ‘Area’. This produces a large combination of ‘Subjects’ and ‘Areas’. LGDX requests are then created with each of those combinations, and ‘Dataset Families’ for those combinations are requested from the NESS Data Exchange API. The LGDX responses retrieved from the NESS Data Exchange API are parsed to simple text and stored into the database. The following table shows a snapshot of the parsed results stored in the database.

DS Family ID	Dataset Family Name	Subject ID	Area ID	From Date	To Date
125	Method of Travel to Work - Resident Population (UV39)	16	276693	2001-04-29	2001-04-29
173	Multiple Ethnic Groups (UV69)	16	276693	2001-04-29	2001-04-29
117	NS-SeC of Household Reference Person (UV33)	16	276693	2001-04-29	2001-04-29

Table 5.7: Snapshot of the Dataset Families data

AREA STATISTICS

Requesting all ‘Area Statistics’ is a complex process. The LGDX format for requesting ‘Area Statistics’ from NeSS Data Exchange API is given below:

```
<Dimension name="variablefamily" isMeasuredDimension="true">
  <Group dimension="dataset" code="75" type="all"/>
</Dimension>
<Dimension name="area" isSpatialDimension="true">
  <Item>
    <HierarchyArea>
      <AreaId>285493</AreaId>
    </HierarchyArea>
  </Item>
</Dimension>
```

This format requires a ‘Dataset Family’ (i.e. a Variable) and an ‘Area’ to retrieve the ‘Area Statistics’ (i.e. ‘Area Statistics’ of an ‘Area’ for a ‘Variable’). The process for extracting all ‘Area Statistics’ for all ‘Areas’ starts by extracting all the ‘Dataset Families’ and ‘Areas’ from the database and making LGDX requests for all the combinations of ‘Dataset Family’ and ‘Area’. This produces a large number of LGDX requests. Each LGDX for the combinations of ‘Dataset Family’ and ‘Area’ is then sent to NeSS Data Exchange and resultant ‘Area Statistics’ are parsed and stored into the database. The following table shows a snapshot of the parsed results stored in the database.

DS Family ID	Statistics Name	Type	Area ID	Counts
125	Underground, metro, light rail or tram	Count	276693	709386
125	Train	Count	276693	950023
125	Bus, minibus or coach	Count	276693	1685361

Table 5.8: Snapshot of the Area Statistics data

5.2.3 DATA FLOW DIAGRAM

Data flow for the data retrieval from the ONS NeSS Data Exchange API works in different small steps. It requires that the data for a hierarchy element has been retrieved by the calling application before the next hierarchy element is called. Thus it requires the data of previous hierarchy elements in retrieving the data of next hierarchy elements. The following figure shows the data flow diagram for the data extraction from the ONS NeSS Data Exchange API.

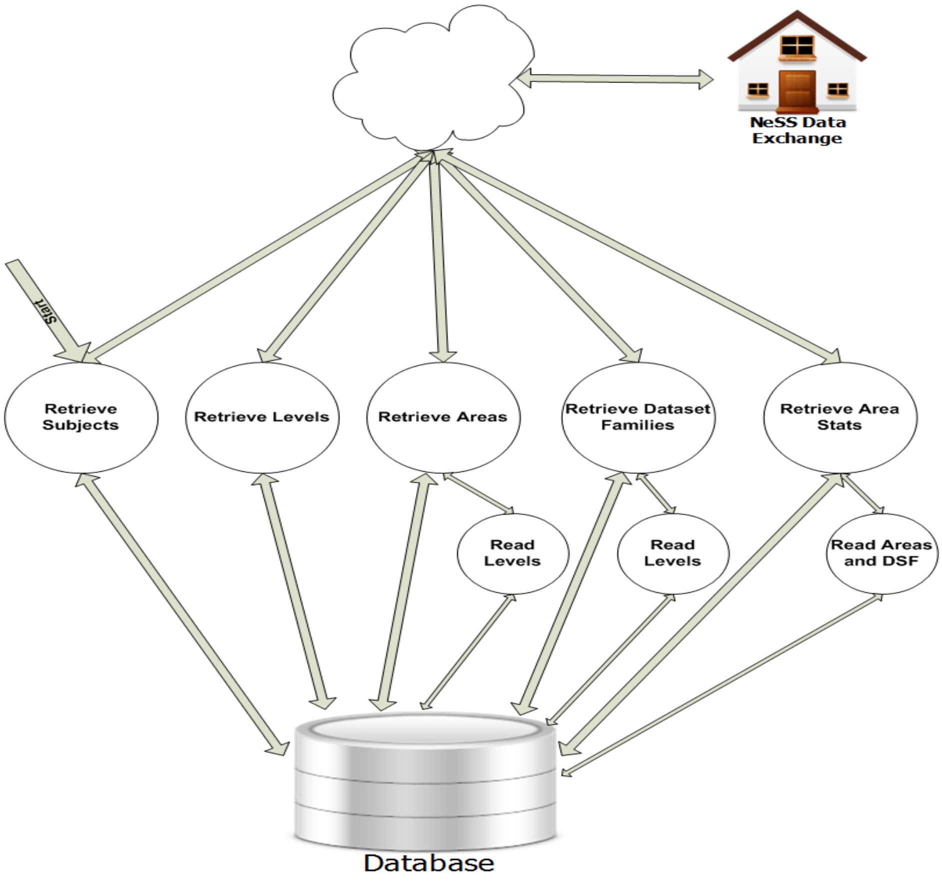


Figure 5.1: Data flow diagram

In the above data flow diagram, flow starts from left and the system starts by retrieving all 'Subjects' from the NESS data store and stores them in the database. There is another module which extracts all 'Levels' afterwards from the API and stores them to the database. Extraction of 'Areas' is based on reading 'Levels' and then using them to extract specific 'Areas' for those 'Levels'. Extraction of 'Dataset Families' is based on reading 'Levels' and then using them to extract particular 'Dataset Families' for those 'Levels'. In the same way, the final step is to read 'Areas' and 'Dataset Families' and using their combination to extract 'Area Statistics'. 'Area Statistics' combined with 'Areas' and 'Dataset Families' is the final data required. 'Area Statistics' are the final data required.

5.2.4 DATABASE STRUCTURE

Once data have been stored in database, and aggregated to appropriate geographical levels, it is important to describe the structure of the database. Because data need to be stored using some kind of data structure, a database provides that structure. A database structure represents how different entities (tables) interact with each other to make relationships. A database structure is represented in terms of a database diagram

DATABASE DIAGRAM

The database diagram of the database tables is shown below.

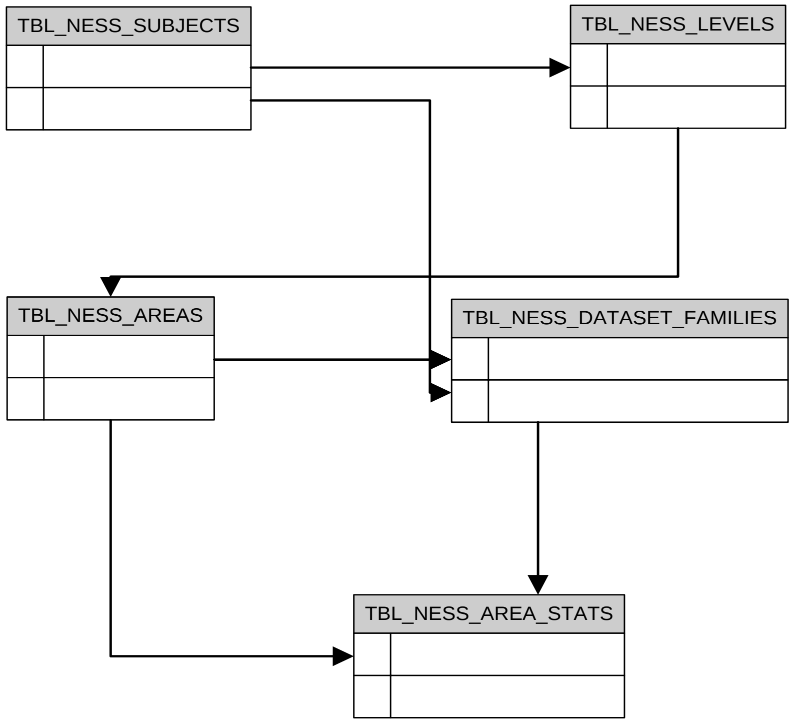


Figure 5.2: Database Diagram

TBL_NESS_SUBJECTS table stores the entire 'Subjects' (i.e. dataset names) in the database. TBL_NESS_LEVELS stores the level names and for different 'Subjects' in the NESS Data Store. TBL_NESS_AREAS stores 'Area names' for 'Levels' and subsequently TBL_NESS_DATASET_FAMILIES stores the 'Dataset Family' information for different 'Areas' and 'Subject'. Finally, TBL_NESS_AREA_STATS stores the actual area statistics of the 'Dataset Families'.

5.2.5 NESS DATA EXTRACTION SERVICE

The ONS NeSS Data Extraction service is based on a Java based application which extracts data using the NeSS API. Using LGDX based XML requests, this service sends requests to NeSS API and parses and stores the results into the database. This service has three main functionalities.

a) Retrieving Areas

b) Retrieving Dataset Families

c) Retrieving Area Statistics

RETRIEVING AREAS

Retrieving Areas works by extracting 'Levels' information from TBL_NESS_LEVELS information, and then using it to extract all the 'Area names' from the ONS NeSS API. This works in an iterative manner, where children of an 'Area' are requested from the ONS NeSS API and then the received data are used to request further children, and so on. Retrieved records are stored in TBL_NESS_AREAS.

RETRIEVING DATASET FAMILIES

Retrieving 'Dataset Families' works by extracting 'Areas' from TBL_NESS_AREAS, and then using the 'Area Names' to extract the 'Dataset Families' of any specific 'Subject'. 'Dataset Family' represents a variable in a dataset or 'Subject'. Retrieved records are stored in TBL_NESS_DATASET_FAMILIES.

RETRIEVING AREA STATISTICS

Retrieving ‘Area Statistics’ works by extracting ‘Dataset Family’ and ‘Area’ information from TBL_NESS_DATASET_FAMILIES and using it to request ‘Area Statistics’ from the ONS NeSS Data Store API. Retrieved records are saved in TBL_NESS_AREA_STATS.

Once functionality has been defined for the extraction service, it is important to illustrate the Use Case diagram of the application. This has been an important tool in showing the functional part of the application in diagrammatic form. The following section describes the Use Case Diagram of this NESS data extraction service.

USE CASE DIAGRAM

Copeland (2008) describes how a use case is a scenario that describes the use of a system by an actor to accomplish a specific goal. An actor is a user playing a role with respect to the system. Actors are generally people although other computer systems could be actors as well. A scenario is a sequence of steps that describes the interaction between an actor and the system. The use case model consists of the collection of all actors and all use cases.

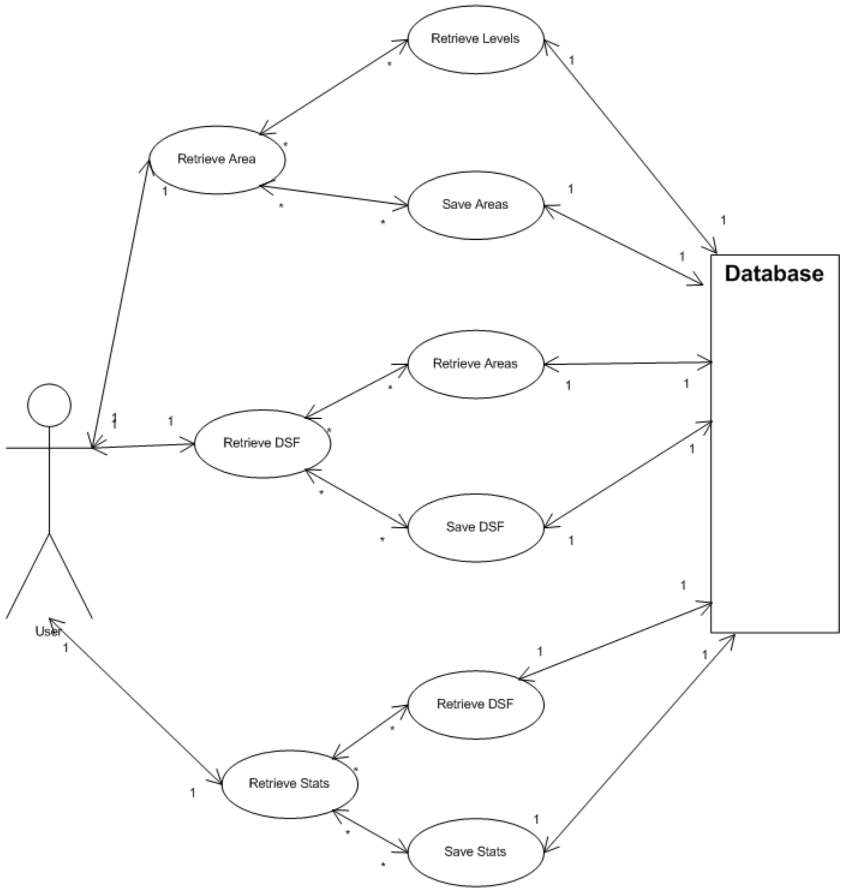
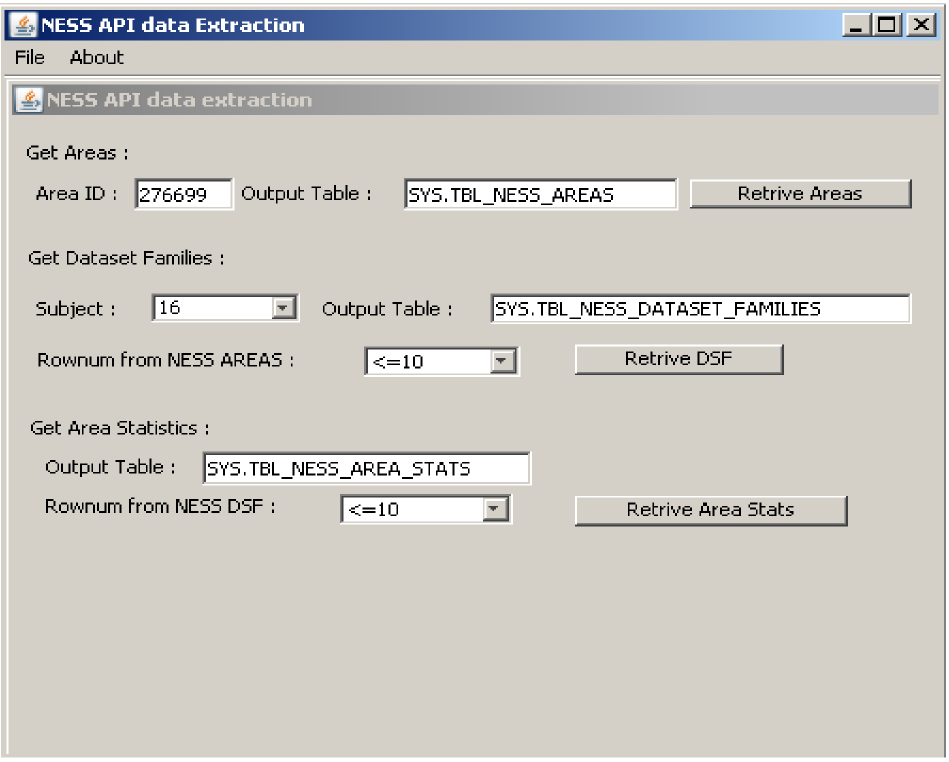


Figure 5.3: Use Case diagram of the NESS data extraction service

SCREEN SHOT OF DATA EXTRACTION SERVICE

The following is the screen shot of the ONS NeSS data extraction service. It shows the different controls that a user can exert.



Area ID : 276699 Output Table : SYS.TBL_NESS_AREAS

Subject : 16 Output Table : SYS.TBL_NESS_DATASET_FAMILIES

Rownum from NESS AREAS : <=10

Output Table : SYS.TBL_NESS_AREA_STATS

Rownum from NESS DSF : <=10

Figure 5.4: Ness data extraction service

END POINT OF THE DATA EXTRACTION SERVICE

This data extraction service is a proof of the concept that web services could be created for the extraction and integration of data from a number of data sources. This will be help in creating geodemographic information systems where the data is integrated from live data sources for building the geodemographic classification in an online environment.

5.3 REVIEW OF DIFFERENT DATA SOURCES

Selection of appropriate data sources is one of the most important parts of creating a geodemographic classification. There are different data sources available in public domain at different spatial extents, and these can be clustered using a range of different techniques. This section is aimed at reviewing the different data sources available in the public sector in accordance with their structure, spatial coverage, and limitations. This review is important from the point of view of selecting appropriate data sources for the creation of geodemographic classifications.

The data sources reviewed in this section are:

a) 2001 Census: Census Area Statistics

b) 2001 Census: Key Statistics

c) Access to Services

d) Community Wellbeing/Social Environment

e) Education, Skills, and Training

f) Health and Care

g) Housing

h) Indices of deprivation and classification

i) People and Society: Income and Lifestyle

j) People and Society: Population and Migration

k) Work Deprivation

l) Ethnicity data

The following is a description of each of these data sources.

A) 2001 CENSUS: CENSUS AREA STATISTICS

This data source is available to download free of charge from the ONS (Office of National Statistics) website. They also provide XML feeds of this data source through their NeSS API. This dataset provides detailed information on a total of 59 Census variables. This data source has remained the main and important resource for the creation of geodemographic classifications because of the wide range of socio-economic variables it covers.

Some of the variables it covers are shown in Table 5.9.

Variable Name
Accommodation Type – Household Spaces
Accommodation Type – All People
Age
Age – Workplace Population
Amenities
Approximate Social Grade
Approximate Social Grade – Workplace Population

Table 5.9: Some of the variables of the2001 Census: Census Area Statistics data set

Unless otherwise dated these data are available at the finest level, which is the OA (Output Area), and can therefore be aggregated to higher geographies.

B) 2001 CENSUS: KEY STATISTICS

This data set is available to download free of charge from the ONS (Office of National Statistics) website. They also provide XML feeds of this data set by their NeSS API. This dataset covers all main census topics, presented as counts and percentages. This dataset provides detailed information on a total of 31 Census variables. Some of the variables covered by this dataset are shown in Table 5.10.

Variable Name
Age Structure
Cars or Vans
Communal Establishment Residents
Country of Birth
Economic Activity – All People
Economic Activity – Females
Economic Activity – Males

Table 5.10: Some of the variables of the 2001 Census: key statistics data set

Unless otherwise stated these data are available at the finest level, which is the OA (Output Area), and can therefore be aggregated to higher geographies.

C) ACCESS TO SERVICES

This data set is available to download free of charge from the ONS (Office of National Statistics) website. They also provide XML feeds of this data set through their NeSS API. This dataset covers different aspects of access to specific health, education and legal facilities. This dataset has 7 variables. Table 5.11 shows the variables of this data set.

Variable Name
Cars or Vans (KS17)
Cars or Vans (UV62)
Distance Travelled to Work
Distance Travelled to Work – Workplace Population
Distance Travelled to Work – Daytime Population
Distance Travelled to Work – Resident Population
Travel to Work

Table 5.11: Variables of Access to Services data set

Unless otherwise stated these data are available at the finest level, which is the OA (Output Area), and can therefore be aggregated to higher geographies.

D) COMMUNITY WELL BEING/SOCIAL ENVIRONMENT

This dataset is available to download free of cost from ONS (Office of National Statistics) website. They also provide XML feeds of this data set by the NeSS API. This dataset contains information to support work on community involvement, social inclusion and improving overall standards, including those relating to street cleanliness. This dataset has 8 variables which are given in the following table (Table 5.12).

Variable Name
Communal Establishment Residents (KS23)
Communal Establishment Residents (UV71)
Communal Establishments
Communal Establishments – People
Health and provision of Unpaid Care
Provision of Unpaid Care

Religion (KS07)
Religion (UV15)

Table 5.12: Variables of Community Well Being/Social Environment data set

Unless otherwise stated these data are available at the finest level, which is OA (Output Area), and can therefore be aggregated to higher geographies.

E) EDUCATION, SKILLS, AND TRAINING

This data set is available to download free of charge from the ONS (Office of National Statistics) website. They also provide XML feeds of this data set by the NeSS API. This dataset includes information on educational attainment, school absence, enrolment to higher education, and number of students. This dataset has 4 variables which are given in Table 5.13.

Variable Name
Economic Activity – Full-time Students
Qualifications
Qualifications and Students
Schoolchildren and Students in Full-time Education living away from home during term time

Table 5.13: Variables of Education, Skills, and Training data set

Unless otherwise stated these data are available at the finest level, which is OA (Output Area), and can therefore be aggregated to higher geographies.

F) HEALTH AND CARE

This data source has data for health, life, expectancy, hospital episodes, healthy lifestyle behaviours and provision of unpaid care. This data set is available to download free of charge from ONS (Office of National Statistics) website. They also provide XML feeds of this data set by the NeSS API. This dataset has 9 variables which are given in the following table (Table 5.14).

Variable Name
Communal Establishment Residents (KS23)
Communal Establishment Residents (UV71)
Communal Establishments
Communal Establishments – People
General Health
Health and provision of Unpaid Care
Households with Limiting Long-term Illness and Dependent Children
Limiting Long-term Illness
Provision of Unpaid Care

Table 5.14: Variables of Health and Care data set

Unless otherwise stated these data are available at the finest level, which is OA (Output Area), and can therefore be aggregated to higher geographies.

G) HOUSING

This data set is available to download free of charge from ONS (Office of National Statistics) website. They also provide XML feeds of this data set by the NeSS API. This dataset includes data for housing demand and supply, tenure and condition, overcrowding and homelessness. This dataset has 22 variables and some of the variables are shown in Table 5.15.

Variable Name
Accommodation Type – Household Spaces
Accommodation Type – People
Amenities
Communal Establishment Residents (KS23)
Communal Establishment Residents (UV70)
Communal Establishments
Communal Establishments – People
Dwelling Stock by Council Tax Band
Dwelling

Table 5.15: Some of the variables of Housing data set

Unless otherwise stated these data are available at the finest level, which is OA (Output Area), and can therefore be aggregated to higher geographies.

H) INDICES OF DEPRIVATION AND CLASSIFICATION

This data set is available to download free of charge from the ONS (Office of National Statistics) website. They also provide XML feeds of this data set by the NeSS API. This dataset includes the Indices of Deprivation, Socio-Economic Classification and Area Classification. This dataset has 10 variables which are given in Table 5.16.

Variable Name
Approximated Social Grade
Approximated Social Grade – Workplace Population
Households by Selected Household Characteristics
National Statistics 2001 Area Classification of Output Areas
National Statistics Socio-economic Classification
National Statistics Socio-economic Classification – Workplace Population
NS-SeC of Household Reference Person
NS-SeC of Household Reference Person – People Under Pensionable Age
Residents in Households by NS-SeC of Household Reference Person Under Pensionable Age
Rural and Urban Area Classification of Output Areas

Table 5.16: Variables of Indices of Deprivation and Classification data set

Unless otherwise stated these data are available at the finest level, which is the OA (Output Area), and can therefore be aggregated to higher geographies.

I) PEOPLE AND SOCIETY: INCOME AND LIFESTYLE

This data set is available to download free of charge from ONS (Office of National Statistics) website. They also provide XML feeds of this data set by the NeSS API.

This dataset includes information on income, including direct measures and indirect indicators, as well as lifestyles of families and households. This dataset has 15 variables which are given in Table 5.17.

Variable Name
Cars or Vans (KS17)
Cars or Vans (UV62)
Dependent Children
Household Composition
Household Composition – Households
Household Composition – Households – Alternative Classification
Household Composition – People
Household Composition – People – Alternative Classification
Household Type
Living Arrangements (KS03)
Living Arrangements (UV82)
Lone Parents Households with Dependent Children
People aged 18 to 64 in Single Adult Households
Religion (KS07)
Religion (UV15)

Table 5.17: Variables of People and Society: Income and Lifestyle data set

Unless otherwise stated these data are available at the finest level, which is the OA (Output Area), and can therefore be aggregated to higher geographies.

J) PEOPLE AND SOCIETY: POPULATION AND MIGRATION

This data set is available to download free of charge from ONS (Office of National Statistics) website. They also provide XML feeds of this data set by the NeSS API. This dataset contains data for demographic change, such as births, deaths and migration. This dataset has 19 variables, some of which are given in Table 5.18.

Variable Name
Age
Age – Workplace Population
Age Structure
Country of Birth (KS05)
Country of Birth (UV08)
Ethnic Group
Marital Status

Table 5.18: Some of the variables of People and Society: Population and Migration data set

Unless otherwise stated these data are available at the finest level, which is the OA (Output Area), and can therefore be aggregated to higher geographies.

K) WORK DEPRIVATION

This data set is available to download free of cost from ONS (Office of National Statistics) website. They also provide XML feeds of this data set by their NeSS API. This dataset contains business and economic activity date, work-related benefits claimants, and participation on government training programmes. This data set has 21 variables, some of which are shown in the following table.

Variable Name
Distance Travelled to Work
Distance Travelled to Work – Workplace Population
Economic Activity
Economic Activity – All People
Hours Worked
Industry of Employment
Occupation Groups

Table 5.19: Some of the variables of Work Deprivation data set

Unless otherwise dated this data is available at the finest level, which is OA (Output Area), and can therefore be aggregated to higher geographies.

L) ELECTORAL ROLL ETHNICITY DATA

2001 census data include ethnicity information of the population. The following table 5.20 lists the ethnic groups covered by 2001 census data. The source of this table is http://www.neighbourhood.statistics.gov.uk.

Ethnic Group
White: British
White: Irish
White: Other
Mixed: White and Black Caribbean
Mixed: White and Black African
Mixed: White and Asian
Mixed: Other Mixed
Asian or Asian British
Asian or Asian British: Indian
Asian or Asian British: Pakistani
Asian or Asian British: Bangladeshi
Asian or Asian British: Other Asian
Black or Black British: Caribbean
Black or Black British: African
Black or Black British: Other Black
Chinese or Other Ethnic Group: Chinese
Chinese or Other Ethnic Group: Other Ethnic Group

Table 5.20: Ethnic groups covered by 2001 census data

However, there are other ways of finding the detailed ethnic origins of the population. Onomap (www.onomap.org) is a way of classifying people based on their cultural, ethnic, and linguistic roots. Based on Mateos (2007), it is a methodology that allows users to classify any number of names (forenames and surnames) into groups of shared ethnic and linguistic origin. Onomap comes as Java software which takes a CSV file as input and classifies names in the file according to their cultural and ethnic origins. Onomap ascribes names with an appropriate ethnicity by using its own forename and surname dictionaries.

This software can help to create new data sets which containing ethnicity information and it could be a good resource for the creation of new groups for geodemographic classifications. The ethnic origin categories of Onomap were used in 'Worldnames' website (described in chapter 4) for the ‘Ethnicity Search’ and displayed users the roots of their searched ‘surname’.

The Electoral Roll registers every individual who is eligible to vote and contains forename, surname, address, and postcode of each person. Electoral Roll data were used as an input in the creation of Onomap (Mateos 2007) and the results tagged to the Electoral Roll. The format of the final data set is shown in the following table.

Surname	Forename	Cell Code	Cell Type
Steward	Royston	CL213	Scottish
Portch	Hillary	EN110	English
Sow	Baba	AF436	Nigerian

Table 5.21: Format of Electoral roll ethnicity data

Where ‘Cell Type’ represents the ethnic group of the individual person and ‘Cell Code’ is the internal ethnicity code of an ethnic group within the Type. With the addition of ethnic group to the data set, it can be combined with other datasets to provide a good resource for geodemographic classifications. It could also be useful for performing ethnic group analysis of individuals for any geographical region of United Kingdom.

The electoral roll data is available at the postcode level, so this data set can be aggregated to any higher spatial levels (output area, lower super output area, ward etc.).

5.4 CREATION OF NEW DATA SORUCE WITH THE AMALGAMATION OF UK LAND REGISTRY DATA AND ELECTORAL ROLL ETHNICITY DATA

The UK Land Registry (http://www.landreg.gov.uk) provides data for all registered land and properties in England and Wales. Two data sources from UK Land Registry have been used here to create two new data sources. The two data sources are 'House price data' and 'Registered land data'. From the perspective of creating geodemographic classification these data sources can help to incorporate the socio-economic status of people in the form of the price they pay for houses – and detailed data on the attributes of built environments provides a useful complement to social, economic and demographic information on household and individual characteristics. This section describes the creation of two new data source, by the amalgamation of 'Land Registry House Price data' and 'Land Registry registered land data' with the 'UK's Electoral roll ethnicity data'. The resultant data sources contain modelled ethnicity data at building level and house price level for London. These data can be used alongside Census data for the creation of summary neighbourhood geodemographic indicators.

5.4.1 LAND REGISTRY HOUSE PRICE DATA

The creation of this data source involved address matching between the 'UK Land Registry house price data' and 'Electoral roll data'. Data from the Land Registry come in the format detailed in Table 5.22.

Postcode	Address	Property Type	Price
BR1 1AE	Hawksworth House, Flat 2, Tetty Way, Bromley, Greater London	Flat	196125
BR1 1BQ	Newman Court, Flat 34, North Street, Bromley, Greater London	Flat	202000
BR1 1RF	26, West Street, Bromley, Greater London	Semi	230000

Table 5.22: Format of Land registry's address data

In this Table ‘Address’ and ‘Postcode’ denote the location indicator of an individual property, and ‘Price’ represents the value of the property in GBP.

The format of the UK’s electoral roll address data is shown in the following table:

Address_1	Address_2	Address_3	Address_4	postcode
79 QUEEN ST	ABERDEEN	ABERDEENSHIRE		AB10 1AN
FLAT 1	16 NETHERKIRKGATE	ABERDEEN	ABERDEENSHIRE	AB10 1AU
2B UPPERKIRKGATE	ABERDEEN	ABERDEENSHIRE		AB10 1BA

Table 5.23: Format of Electoral roll data

‘Address_1’, ‘Address_2’, ‘Address_3’, and ‘Postcode’ represent an individual address in the UK’s Electoral RollElectoral Roll data set. The address is represented in different formats. Matching the above two data sets involved a recursive process of matching multiple addresses and then finding the best match amongst them.

Following algorithm was applied to perform address matching of the two data sets. The two datasets have been given the following abbreviations:

Land Registry house price data = LRHPD

Electoral roll data = ERD

1) For each postcode in LRHPD, select all the addresses.
2) Match the postcode of LRHPD with the postcodes of ERD and extract the addresses from ERP of the matched postcode.
3) Divide the addresses of LRHPD into tokens based on special characters. This is shown by an example below:

   If the address is ‘11A, RAVENSBOURNE ROAD, BROMLEY, GREATER LONDON’, then it will have 7 tokens as ‘11A’, ’RAVENSBOURNE’, ‘ROAD’, ‘BROMLEY’, ‘GREATER’, ‘LONDON’.

4) For the addresses of matched postcode in ERD, join the fields ‘Address_1’ and ‘Address_2’ to get a single address field. So, Address_1=’11A RAVENSBOURNE ROAD’ and Address_2=’BROMLEY’ will be combined and become Address=’11A RAVENSBOURNE ROAD BROMLEY’.

5) Divide the single address fields of ERD into tokens based on special characters. So, the single address field Address=’11A RAVENSBOURNE ROAD BROMLEY’ will have 4 tokens as ‘11A’, ‘RAVENSBOURNE’, ‘ROAD’, and ‘BROMLEY’.

6) Convert ‘APARTMENT’, ‘APPT’, ‘APT’, ‘FLAT’ from both addresses to a common token ‘FLAT’.

7) Match the address tokens of LRHPD with the address tokens of ERD for all matched records for a single postcode. Each character match was assigned a weight of 0.2 and each integer tokens match was given a weight of 0.6. The following table shows such a matching between the source address and the destination addresses. For this example source address is ’11A, RAVENSBOURNE ROAD, BROMLEY, GREATER LONDON’.

Destination Address	Total Weight
11 RAVENSBOURNE ROAD BROMLEY	0.6
11A RAVENSBOURNE ROAD BROMLEY	1.2
15 RAVENSBOURNE ROAD BROMLEY	0.6
15A RAVENSBOURNE ROAD BROMLEY	0.6
17A RAVENSBOURNE ROAD BROMLEY	0.6
19 RAVENSBOURNE ROAD BROMLEY	0.6
21 RAVENSBOURNE ROAD BROMLEY	0.6
25 RAVENSBOURNE ROAD BROMLEY	0.6
27 RAVENSBOURNE ROAD BROMLEY	0.6
FLAT 1 65 RAVENSBOURNE ROAD BROMLEY	0.6
GRD FL 31 RAVENSBOURNE ROAD BROMLEY	0.6
FLAT 1A 55 RAVENSBOURNE ROAD BROMLEY	0.6
33B RAVENSBOURNE ROAD BROMLEY	0.6
3A 55 RAVENSBOURNE ROAD BROMLEY	0.6

Table 5.24: Format of Electoral Roll data

8) The record with the highest 'Total Weight' is the final matched record for the source address.

9) Repeat the above process for all addresses of all postcodes in LRHPD.

Once the above algorithm finishes, the final step is the joining of the resultant data source with the 'Electoral Roll Ethnicity Data', discussed in the previous section, to create the resultant 'House Price Level Ethnicity Data Set'.

5.4.2 LAND REGISTRY REGISTERED LAND DATA

This section describes the creation of another data set which combines 'UK Land Registry registered land data' with 'UK Electoral Roll data'. 'UK Land Registry registered land data' contains the listing of all the buildings in London. A property may be divided into multiple flats, so this data set accommodates divisions in build structures. The format of the dataset is shown in Table 5.25.

Base	Primary Address	Secondary Address	Street	Postcode
DWELLING	5		ROYSTON ROAD	RM30SR
DWELLING	LAMPETER HOUSE	3	KINGSBRIDGE CIRCUS	RM38NH
DWELLING	PAINES BROOK COURT,14	FLAT 23	HARLESDEN ROAD	RM39JN

Table 5.25: Examples of Land Registry registered land data

'Base' identifies which type of property a particular building is. 'DWELLING' represents the properties which are residential. 'Primary Address', 'Secondary Address', and 'Street', and 'Postcode' represent aspects of the individual address of a building.

The following abbreviations are used for the purpose of the address matching algorithm:

Land Registry building data = LRBD

Electoral roll data = ERD

1) For each postcode in LRBD, select all the addresses.
2) Match the postcode of LRBD with the postcodes of ERD and extract the addresses from ERP of the matched postcode.
3) Join the ‘Secondary Address’, ‘Primary Address’, and ‘Street’ of LRBD into a single address. For example, if Primary Address=’PAINES BROOK COURT, 14’, Secondary Address=’FLAT 23’, and Street=’HARLESDEN ROAD’, then the final single address will become ‘FLAT 23 PAINES BROOK COURT, 14 HARLESDEN ROAD’.
4) Divide the single address of LRBD into tokens based on special characters. This is shown by an example below:

   If the address is ‘FLAT 23 PAINES BROOK COURT, 14 HARLESDEN ROAD’, then it will have 8 tokens as ‘FLAT’, ’23’, ‘PAINES’, ‘BROOK’, ‘COURT’, ‘14’, ‘HARLESDEN’, and ‘ROAD’.
5) For the addresses of the matched postcode in the ERD, the fields ‘Address_1’ and ‘Address_2’ are joined to get a single address field. So, Address_1=’11A RAVENSBOURNE ROAD’ and Address_2=’BROMLEY’ are combined and become Address=’11A RAVENSBOURNE ROAD BROMLEY’.
6) Divide the single address fields of ERD into tokens based on special characters. For the single address field Address=’11A RAVENSBOURNE ROAD BROMLEY’ have 4 tokens as ‘11A’, ‘RAVENSBOURNE’, ‘ROAD’, and ‘BROMLEY’.
7) Convert ‘APARTMENT’, ‘APPT’, ‘APT’, ‘FLAT’ from both addresses to a common token ‘FLAT’.
8) Match the address tokens of LRBD with the address tokens of ERD for all matched records for a single postcode. Each character match is assigned a weight of 0.2 and each integer tokens match requires a weight of 0.6. The following table shows such a matching between the source address and the

destination addresses. For this example source address is ’FLAT 15 RUSSELL WILSON COURT, 150 LITTLE ASTON ROAD’.

Destination Address	Total Weight
FLAT 1 RUSSELL WILSON COURT	0.8
FLAT 4 RUSSELL WILSON COURT	0.8
FLAT 5 RUSSELL WILSON COURT	0.8
FLAT 8 RUSSELL WILSON COURT	0.8
9 RUSSELL WILSON COURT 150 CHURCH ROAD	1.4
FLAT 14 RUSSELL WILSON COURT	0.8
RUSSELL WILSON COURT FLAT 15 150 CHURCH ROAD	2.2
FLAT 11 RUSSELL WILSON COURT	0.8

Table 5.26: Format of Electoral Roll data

9) The record with the highest ‘Total Weight’ is the final matched record for the source address.

10) Repeat the above process for all addresses of all postcodes in LRBD.

Once the above algorithm finishes, the final step is the joining of the resultant data source with the ‘Electoral Roll Ethnicity Data’, discussed in the previous section, to create the resultant ‘Building level ethnicity data set’.

5.5 DATASOURCES FOR CREATING GEODEMOGRAPHIC CLASSIFICATIONS

(Vickers & Rees, 2007) created the 2001 Output Area Classification by selecting 41 variables from the 2001 Census of Population. These variables were chosen from a number of domains covering demographic, household composition, housing, employment, and socio-economic characteristics.

These variables could be used for the creation of geodemographic classifications because they are the indicators of socio-economic characteristics of the population.

All of these variables hold data at output area level. The following table 5.26 shows the 41 variables and their respective domains and sub-domains. These variables will be incorporated in the 'GeodemCreator' software (described in Chapters 7 and 8) as default variables for the creation of open geodemographic classifications.

Variable (Sub-domain)	Domain
V1: Age 0-4 **(Age)** V2: Age 5-14 **(Age)** V3: Age 25-44 **(Age)** V4: Age 45-64 **(Age)** V5: Age 65+ **(Age)** V6: Indian, Pakistani or Bangladeshi **(Ethnicity)** V7: Black African, Black Caribbean or Other Black **(Ethnicity)** V8: Born Outside the UK **(Country of Birth)** V9: Population Density **(Population)**	**Demographic**
V10: Divorced **(Living Arrangements)** V11: Single person household (not pensioner) **(Living Arrangements)** V12: Single pensioner household **(Living Arrangements)** V13: Lone Parent household **(Living Arrangements)** V14: Two adults no children **(Living Arrangements)** V15: Households with non-dependent children **(Living Arrangements)**	**Household Composition**
V16: Rent (Public) **(Tenure)** V17: Rent (Private) **(Tenure)** V18: Terraced Housing **(Type and Size)** V19: Detached Housing **(Type and Size)** V20: All Flats **(Type and Size)** V21: No central heating **(Quality/Crowding)** V22: Rooms per household **(Quality/Crowding)** V23: People per room **(Quality/Crowding)**	**Housing**
V24: HE Qualification **(Education)** V25: Routine/Semi-Routine Occupation **(Socio-Economic class)** V26: 2+ Car household **(Ownership/Commuting)** V27: Public Transport to work **(Ownership/Commuting)** V28: Work from home **(Ownership/Commuting)** V29: Limiting Long Term Illness (SIR) **(Health and Care)**	**Socio-Economic**

V30: Provide unpaid care **(Health and Care)**
V31: Students (full-time) **(Employment)**
V32: Unemployed **(Employment)**
V33: Working part-time **(Employment)**
V34: Economically inactive looking after family **(Employment)**
V35: Agriculture/Finishing employment **(Industry Sector)**
V36: Mining/Quarrting/Construction employment **(Industry Sector)**
V37: Manufacturing employment **(Industry Sector)**
V38: Hotel & Catering employment **(Industry Sector)**
V39: Health and Social work employment **(Industry Sector)**
V40: Financial intermediation employment **(Industry Sector)**
V41: Wholesale/retail employment **(Industry Sector)**

Table 5.26: Census variables and their domains for creating geodemographic classifications

Although Census area statistics remain the main and important data source for creating geodemographic classifications, other data sources can also contribute important additional aspects to the classifications. The 'Approximated Social Grade' variable from the 'Indices of Deprivation and Classification' data source (discussed in section 5.3) provides an indicator of the employment social grade of people in a specific area, and could be used for creating geodemographic classifications.

New variables of ethnicity could be created from the data source 'electoral roll ethnicity data' described in the second part (5.3) of this chapter. The new variables, representing ethnicity, could be combined with the other socio-economic variables for the creation of innovative geodemographic classifications. Chapter no. 8 will discuss the creation of a geodemographic classifications using socio-economic data (41 census variables outlined in Table 5.26) and electoral roll ethnicity data (described in section 5.3).

The 'House Price Level Ethnicity data' for London, created in the section 5.4 of this chapter, could also play an important role to bring ethnicity dimensions of economic status at the finest postcode level geography.

Taken together, (a) the 41 variables (outlined in Table 5.26) from 2001 census data, (b) Indices of Deprivation data, (c) Electoral Roll Ethnicity data, and (d) the House Price Level Ethnicity data sources can each contribute towards the creation of geodemographic classifications.

5.6 CONCLUSION

This chapter has broadened the emphasis upon consistency and visualisation of names data of the preceding two chapters through a focus on the different data sources that are readily available for the creation of new and innovative geodemographic classifications. In the first section, this chapter explained the creation of a software service to extract live XML feeds of data from the ONS (Office for National Statistics) NeSS (Neighbourhood Statistics) API. This software service provides a proof of the concept that software services could be created for the extraction and integration of data from different data sources for the creation of innovative and interactive geodemographic classifications. This software service could be really useful in an environment where geodemographic classifications are to be created in an online environment with integration of data from a number of live data sources. The second part of this chapter described different data sources available in public sector through a review of their structure and spatial coverage. The third part of this chapter described the creation of two new data sources with the amalgamation of house price data and ethnicity data, and discussed a way to create new and innovative data sources by combining the available data sources. This data source could be a good indicator of socio-economic status of people in the form of the price they pay for houses. The final section of this chapter set out the data sources which can be used for the creation of geodemographic classifications. It proposed that a combination of Census Area Statistics alongside some other data sources can be used to represent salient aspects of socio-economic status. However, the use of life style surveys alongside Census Area Statistics in the creation of classifications also needs to be investigated for the creation of innovative and more responsive classifications.

After the data sources have been selected for the creation of geodemographic classifications, the final level of classifications is created by using cluster analysis. The next chapter discusses cluster analysis in detail with a review of a number of clustering algorithms.

6 CHAPTER 6: CLUSTER ANALYSIS FOR THE CREATION OF GEODEMOGRAPHIC CLASSIFICATIONS

6.1 INTRODUCTION

The previous chapter has focused on different data sources that are readily available for the creation of new and innovative geodemographic classifications. A number of data sources available in the public sector were discussed through a review of their structure and spatial coverage. The last part of the previous chapter emphasised on the creation of two new data sources with the amalgamation of house price data and ethnicity data, and discussed a way to create new and innovative data sources by the combination of the available data sources.

After the selection of the data sets and their corresponding variables, the finest level of a geodemographic classification is created by cluster analysis. This chapter discusses cluster analysis in detail, and provides a detailed overview of different clustering techniques and algorithms. *K*-means clustering algorithm remains the core algorithm for the computation of geodemographic classifications, thus section 6.2 of this chapter is dedicated to a detailed discussion of *k*-means clustering. Because the standard implementation of *k*-means requires multiple runs of individual instances to create a robust classification (Singleton & Longley, 2009), it is desirable to implement short cuts of *k*-means. Some improvements to the *k*-means clustering algorithm have been implemented by Reynolds *et al* (2004) as "*k*-means ++". This chapter also provides a detailed overview of some of the alternate clustering algorithms. Hierarchical clustering, partitioning around medoids (PAM), and genetic clustering algorithms are discussed in section 6.3. Section 6.4 provides a comparison of the *k*-means clustering algorithm with Clara (Clustering Large Applications) and genetic clustering algorithms.

However, from the perspective of creating geodemographic classifications in an online environment, users require response times within seconds, or minutes at most. None of the algorithms yet created provide sufficiently efficient ways to create efficient geodemographic classifications within an acceptable time range for an online system. Therefore, section 6.5 of this chapter is dedicated to the work towards the computational improvements of *k*-means clustering algorithm. These include implementing *k*-means with PCA (principal component analysis) and a parallel implementation of *k*-means. Parallel and cloud computing is becoming very popular these days. With the innovation of Amazon Cloud (Amazon, 2010) and iCLOUD (Apple, 2011), it has become possible to run computationally intensive algorithms on the cloud of computers. This is beneficial, because the speed of an algorithm increases if run on multiple computers rather than a single computer.

However, parallelism of algorithms is not limited to grids or clouds of computers. A more radical approach is the use of NVidia graphics cards for the parallelism of computationally expensive algorithms. NVidia graphics cards have multiple GPUs (Graphical Processing Units) in them. These GPUs are the shorter form of a CPU (Central Processing Units) and they facilitate rendering of images. These GPUs can also be used for parallel implementation of computationally intensive algorithms as, for example, when adapting a message-driven parallel application to GPU accelerated clusters (Philips et al, 2008). Accelerated large graph algorithms (Harish & Narayanan, 2007) and biomedical image analysis (Hartley et al, 2008) have used Computer Unified Device Architecture (CUDA: a parallel computing architecture developed by Nvidia) in order to run algorithms in parallel.

Computer Unified Device Architecture (CUDA) has been used for the implementation of parallel version of *k*-means (parallel *k*-means) on NVidia graphics cards. The parallel *k*-means clustering algorithm is described in the section 6.5.2 and a comparison is provided between the parallel *k*-means and standard *k*-means clustering algorithms. The parallel *k*-means clustering algorithm can be applied to the computation of geodemographic classifications online.

6.2 CLUSTER ANALYSIS

Cluster analysis groups similar items into categories that share common characteristics. In the context of creating geodemographic classifications, cluster analysis is performed by processing a matrix where each row represents an area and each column represents the socio-economic attribute of that area. An example matrix is shown below:

Areas	V1	V2	V3	V4	V5	V6	……
Area1							
Area2							
Area3							
Area4							
Area5							
Area6							
…….							

Table 6.1: Layout of the data for creating a geodemographic classification

An example of cluster analysis is shown in the figure 6.1. A cluster analysis seeks to group areas into homogeneous groups based on their socio-economic attributes. So after a cluster analysis has been performed, items sharing similar constellations of characteristics are assigned to a single group. Cluster analysis is performed by using clustering algorithms. Some noteworthy clustering techniques/algorithms are hierarchical clustering (Kaufman and Rousseeuw, 1990), partitioning around medoids (PAM: Kaufman and Rousseeuw, 1990), *k*-means (Kaufman and Rousseeuw, 1990) and consensus clustering (Cheshire, Adnan et al., 2011). The idealised outcome of clustering is shown in Figure 6.1.

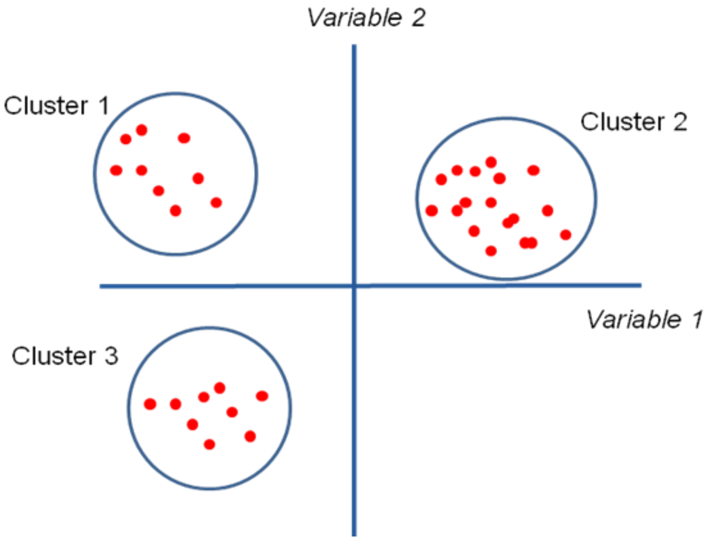


Figure 6.1: A hypothetical example of cluster analysis performed on two variables with three resultant clusters of values

The result of this illustrative cluster analysis is the assignment of a cluster number to each area as shown in Table 6.2. These numbers are the individual classes or groups of a geodemographic classification. These numbers are then typically assigned a distinctive colour, for displaying the results in the map.

Areas	Cluster
Area1	1
Area2	1
Area3	2
Area4	1
Area5	3
Area6	3
…….	

Table 6.2: An example result of the cluster analysis

6.3 CLUSTERING ALGORITHMS

There are a number of clustering algorithms available to perform cluster analysis. The techniques that they deploy vary, for example in the ways that initial seeds are assigned, or the particular distance measures used for performing cluster analysis. Each clustering algorithm has its own merits and drawbacks. This section provides a thorough discussion of different clustering algorithms with a view of creating tools for real-time geodemographics classification. The aim is to identify a clustering algorithm which can produce results in real-time, and can also maintain the integrity of the results produce. The clustering techniques and algorithms discussed in this section are:

a) Hierarchical Clustering

b) *K*-means Clustering

c) Partitioning Around Medoids (PAM)

d) Genetic Clustering

6.3.1 HIERARCHICAL CLUSTERING

K-means, PAM, Clara, and Genetic clustering all result in a simple or 'flat' partition, rather than any hierarchical ordering of clusters. By contrast, hierarchical clustering seeks to build a hierarchy of clusters that can be depicted using a "tree" or "dendogram". There are two approaches to hierarchical clustering. We can go "from the bottom up", grouping small clusters into larger ones, or "from the top down", dividing big clusters into smaller ones. These are called **agglomerative** and **divisive** clustering, respectively.

The structured results of hierarchical clustering are more informative than the unstructured set of clusters returned by flat clustering. Flat clustering algorithms require pre-specification of the number of clusters required; however, hierarchical clustering does not require this, and the analyst can decide to terminate the clustering in the light of the structure revealed by the data.

Figure 6.3 explains the "from the top down" approach of hierarchical clustering. This example is the creation of surname regions (Cheshire, 2011), and shows clearly how the hierarchy can be terminated at any convenient level of recursion.

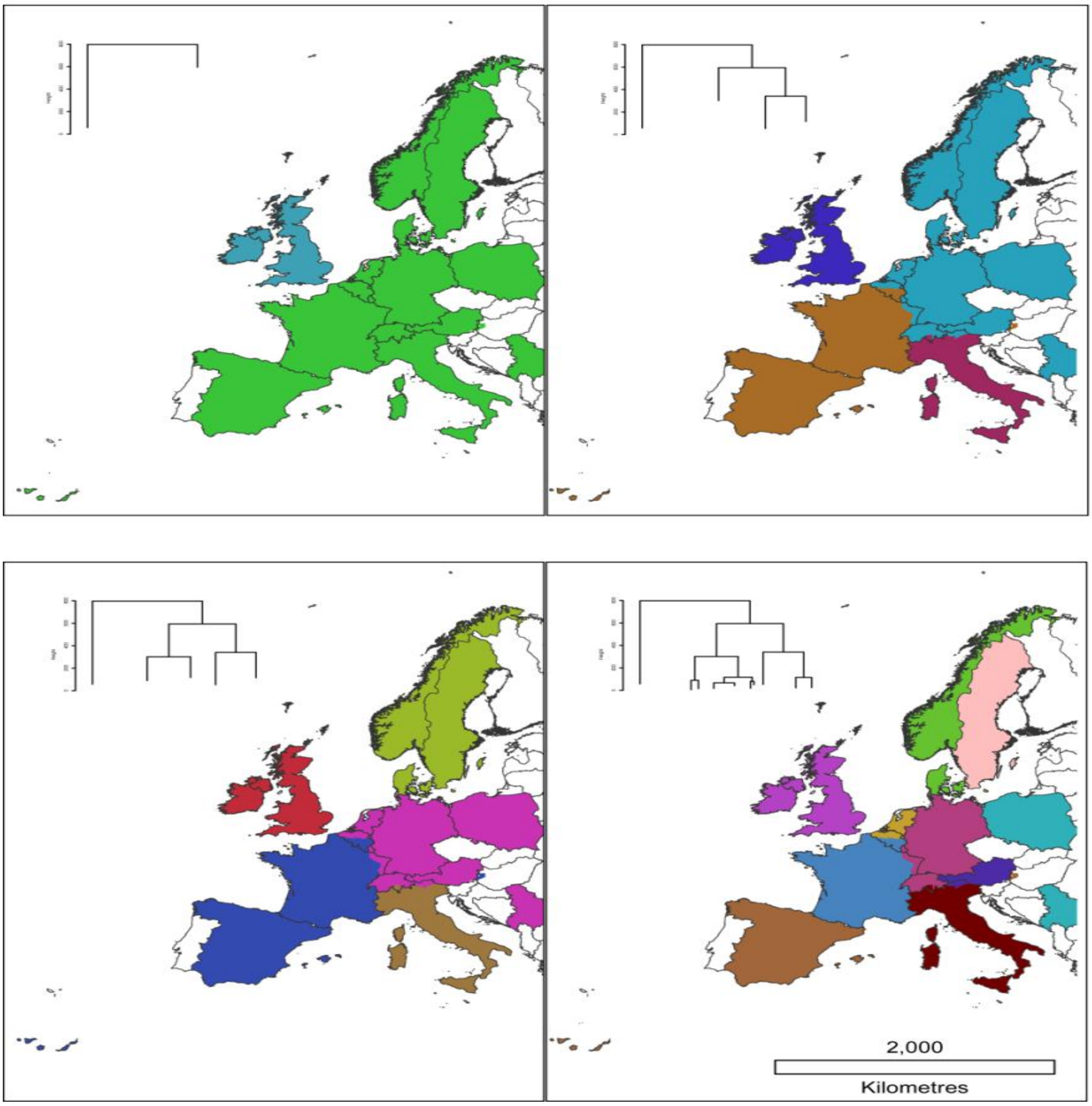


Figure 6.2: Creating surname regions by using hierarchical clustering (source: Cheshire, 2011)

In the context of creating geodemographic classifications, hierarchical clustering is not really useful. It has never been used in the creation of geodemographic classifications, because the desired outcome of a geodemographic classification is distinctive individual groups/classes and not a hierarchy of different classes.

6.3.2 K-MEANS CLUSTERING

K-means is a very popular clustering algorithm. It has been used for different applications, and the finest level in most geodemographic classifications is created using the *k*-means algorithm which seeks to find the set of cluster centroids that minimises expression (6.1) below.

$$V = \sum_{j=1}^{k} \sum_{i=1}^{k} (x_j - \mu_i)^2 \tag{6.1}$$

where *k* is the number of clusters, and μ_i is the mean centroid of all the points x_i in cluster *i.*

The *k*-means algorithm begins by randomly allocating a set of *k* seeds within the data matrix and proceeds by allocating each data point to its nearest seed in multidimensional space. A cluster centroid is then calculated for each cluster, and a new partitioning of the data points is made around the new set of centroids. The centroids are then recalculated for the new clusters of points, and the algorithm repeats these steps until a convergence criterion is met (usually when switching of data points no longer takes place between the clusters). Singleton and Longley (2009) have illustrated how the resulting classifications are sensitive to the placement of the initial seeds, with consequences for the performance of the cluster model. They suggest that, in order to optimise a classification, a model is run multiple times in order to suggest an optimal convergence solution based upon Equation (6.1).

To ascertain the degree of instability of the *k*-means clustering algorithm, the algorithm was run on the 41 variables of the 2001 Census of Population (described in the section 5.5) Output Area dataset for London. *K*-means was run on the dataset for 100 iterations. The hardware used for this purpose was an "Intel® Xeon® CPU 5150 @ 2.66 GHz, 3.00 GB of RAM". The results were mapped for different iterations. Figures 6.3 through 6.5 show three results from the 100 runs. These results illustrate the instability of *k*-means clustering. Different iterations give different results, so from the perspective of creating geodemographic classifications it becomes necessary to run *k*-means multiple times, and retrieve the best result.

The best result is described as the *k*-means run having minimum 'within sum of squares distance' between the data points. Singleton and Longley (2009) suggest running *k*-means 10,000 times on a dataset for creating a geodemographic classification.

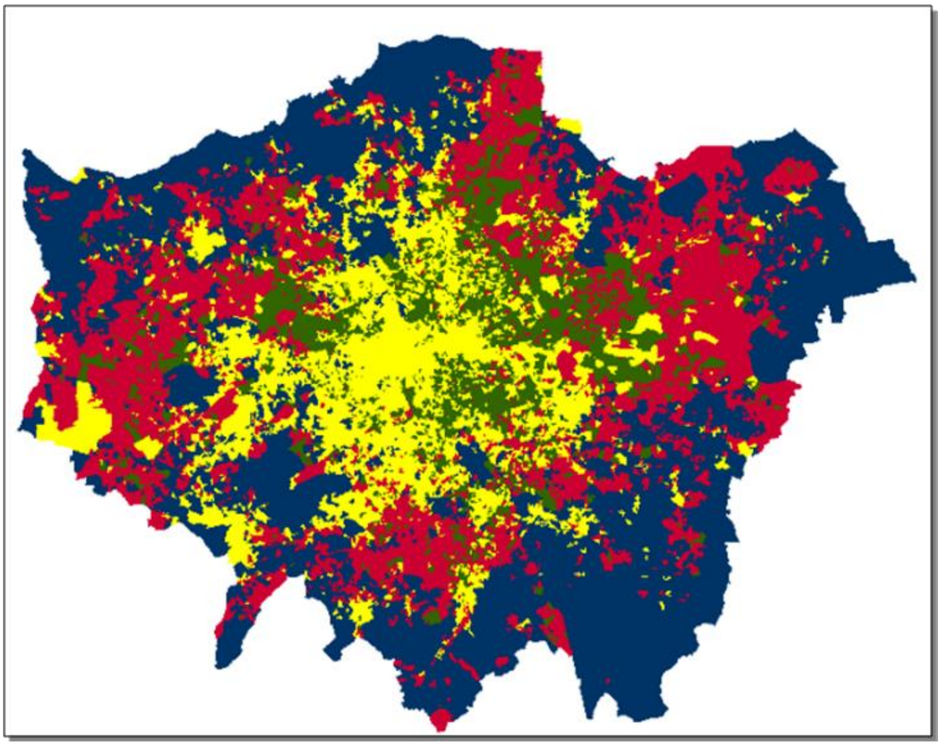


Figure 6.3: Applying k-means to an OAC dataset for London

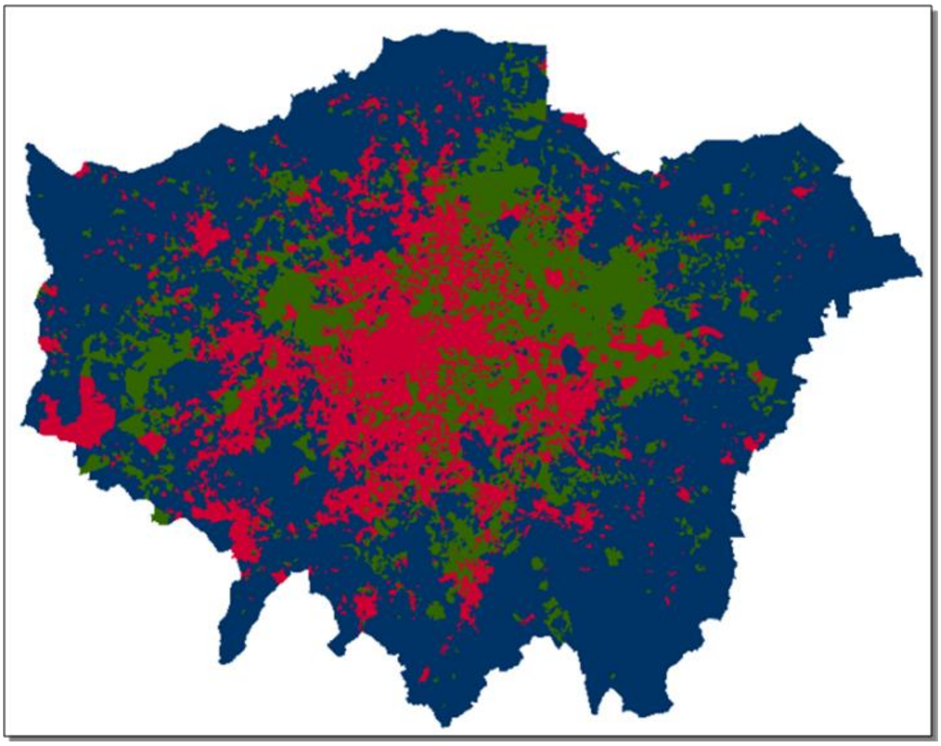


Figure 6.4: Applying k-means to an OAC dataset for London

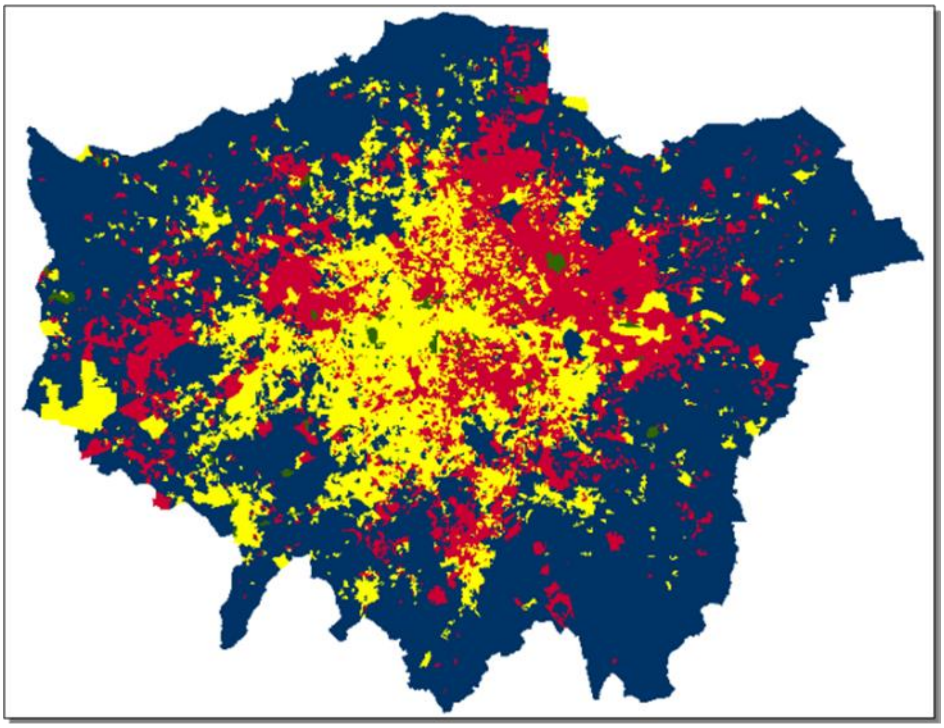


Figure 6.5: Applying k-means to an OAC dataset for London

This process is computationally very inefficient, firstly because each *k*-means takes time to compute, and secondly because the results from only a single cluster analysis is saved. For classifications created offline this inefficiency is acceptable, yet this is not the case where users are seeking an interactive solution. An online tool thus requires results to be returned more quickly. For example, if the input UK data for the Office for National Statistics Output Area Classification (OAC: *n* = 41 variables) are clustered on a high specification computer (Intel® Xeon® CPU 5150 @ 2.66 GHz, 3.00 GB of RAM) where *k* =52, then the processing time for this to converge is 4.23 seconds. Thus for this classification to run 10,000 times, a user can expect to wait for 42,300 seconds (11.75 hours) to obtain results.

Considerable work has been undertaken to improve the efficiency of *k*-means. For example, Reynolds *et al* (2004) describe a new way of choosing the initial seeds, describing the algorithm as "*k*-means++". This method selects initial centres based on the density of data points and improves the overall processing time because data points converge around seeds more quickly. A more radical method of improving classification efficiency is to supplement *k*-means with modern classification procedures. Two such techniques are partitioning around medoids

(PAM: Kaufman and Rousseeuw, 1990) and clustering using genetic algorithms (GA: Painho and Bação, 2000; Maulika and Bandyopadhyay, 2000).

6.3.3 PAM (PARTITIONING AROUND MEDOIDS) CLUSTERING

The PAM procedure is represented in Equation 6.2. This algorithm attempts to assign points from within a multidimensional data matrix into clusters based in their "nearness" to a series of randomly selected "representer" points. Unlike k-means, representer points are actual data points from within the data matrix, rather than any random seed point within the Euclidean space. In PAM "nearness" is calculated using a pre-computed dissimilarity matrix across all variables and data points within a data set. This offers improved efficiency over *k*-means because it reduces the processing of on-the-fly distance calculations, and additionally is less sensitive to outliers – because positioning the centroid utilises a median rather than the mean in the optimisation procedures. Effectively, this minimises

$$V = \sum_{j=1}^{k} \sum_{i=1}^{k} \left| x_j - \mu_i \right| \tag{6.2}$$

where the variables are defined as in Equation (6.1).

The pre-computation of a distance matrix is memory intensive and PAM struggles when applied to large data sets. Thus, Kaufman and Rousseeuw (1990) developed a sampling algorithm called clustering large applications (Clara). Clara draws multiple samples of the dataset, applies PAM to each sample, and returns its best clustering as output. Because Clara applies PAM on samples rather than on the whole dataset, it can cope with larger data volumes.

6.3.4 GENETIC CLUSTERING ALGORITHMS

Brunsdon (2006) and Fernández et al (2005) demonstrated that genetic algorithms (GA) can be combined with PAM to offer further improvements in classification efficiency by supplementing the initial random representor point selection with a genetic algorithm which generates multiple possible sets of representor points. Genetic algorithms run repeated analyses and works through a breeding procedure which preserves the characteristics of the best data points for the subsequent generation. The subsequent generation is created by mutation (change of a random position of the “chromosome” i.e. change of a data point in a cluster) and crossover (change of slices of “chromosomes” between parents i.e. change of slices of data points). After a number of generations, the genetic algorithm converges on the optimal solution of the clustering problem. For this thesis, an R-based genetic algorithm was used varying values of number of generations to devise a clustering solution, retaining a number of data items from one generation to another, and invoking chance mutation.

6.4 COMPARISON OF K-MEANS, CLARA, AND GENETIC CLUSTERING ALGORITHMS

This section develops a comparison between *k*-means, Clara, and genetic algorithm (GA). Three metrics are compared: computation efficiency (time), classification optimization efficiency using average silhouette width, and computation efficiency using different variable standardization techniques. The aim of this analysis is to examine which type of classification procedure would be most appropriate for computing real time geodemographic segmentations online. To compare *k*-means, Clara, and GA the input data for the National Statistics Output Area Classification (Vickers and Rees, 2007) aggregated at three geographical levels Output Area (OA), Lower Super Output Area (LSOA), and Ward for the UK was used. The input data file was a matrix of 223,060 rows and 41 columns (representing the 41 socio-economic variables). An Intel® Xeon® CPU 5150 @ 2.66 GHz with 3.00 GB of RAM was used for these comparisons.

6.4.1 MEASURING CLUSTERING EFFICIENCY BASED ON COMPUTATIONAL TIME

In order to measure computing time for each of the algorithms *k*-means, Clara, and GA were run for 1-100 cluster solutions at all of the three different geographic levels of the OAC dataset for the UK. Later, the CPU clock time (in seconds) for each algorithm was compared. Figures 6.6 - 6.8 show the relationship between CPU clock time (in seconds) and the number of clusters (1-100) by using *k*-means, Clara, and GA for the three different geographical levels. These algorithms were run a single time for each value of *k*.

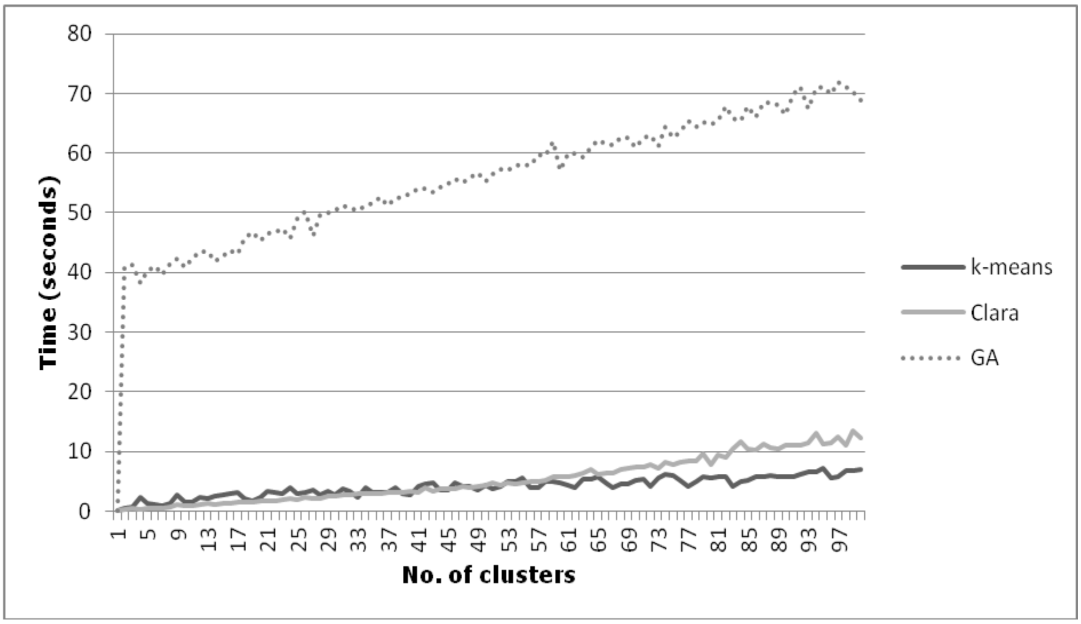


Figure 6.6: Comparing computational efficiency in an Output Area (OA) level dataset covering the UK

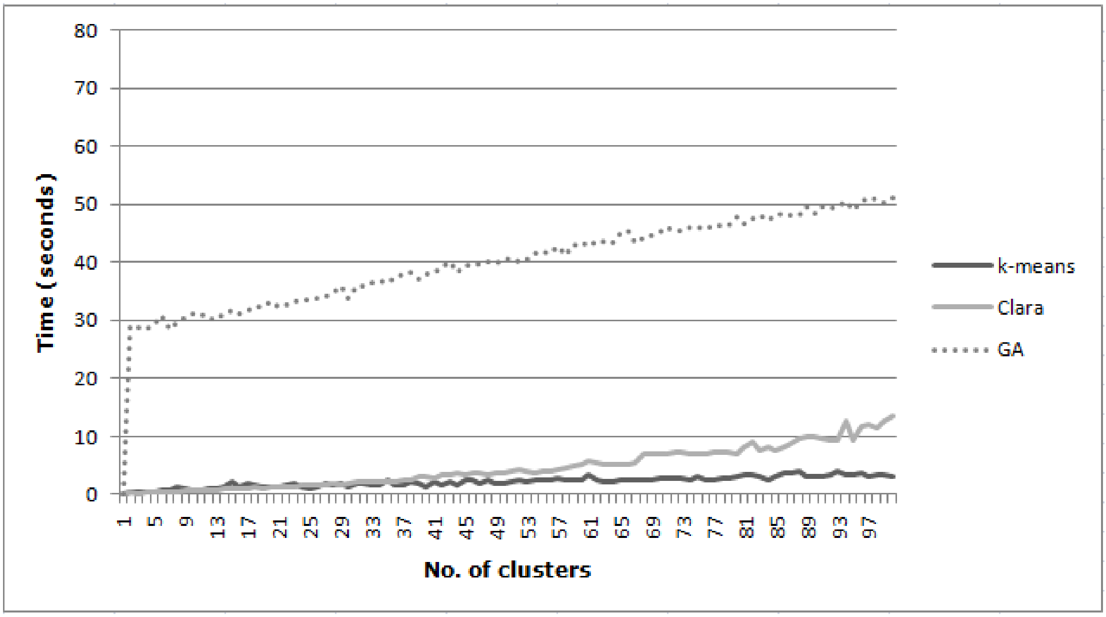


Figure 6.7: Comparing computational efficiency in a Lower Super Output Area (LSOA) level data set covering the UK

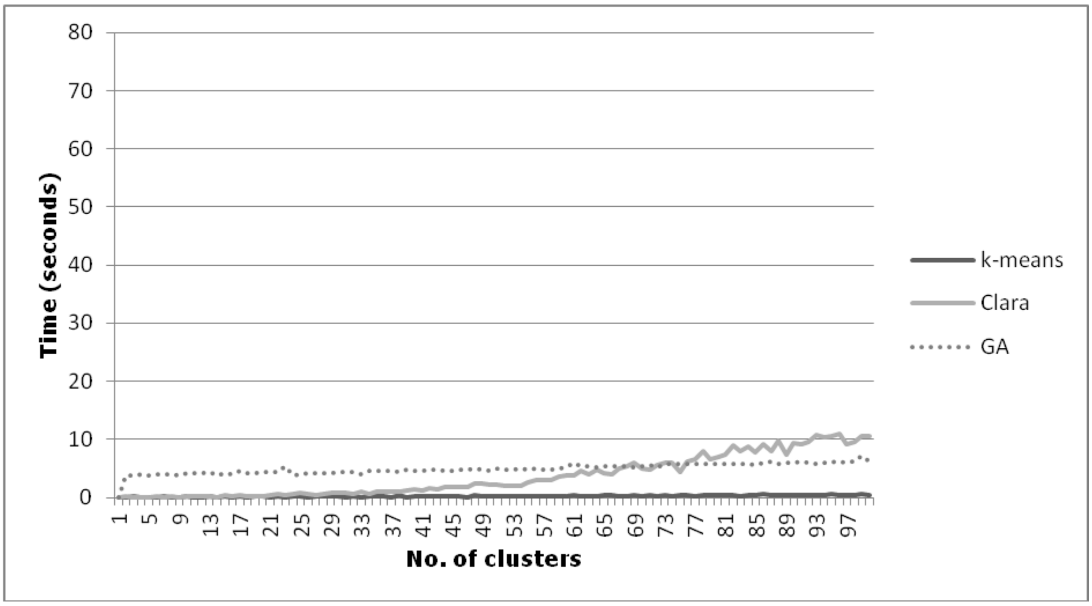


Figure 6.8: Comparing computational efficiency in a Ward level data set covering the UK

The figures show that for large number of data points (OA and LSOA level) Clara runs faster than *k*-means and GA given the number of clusters is a small number (<55 for OA level and <30 for LSOA). However, when the number of clusters is a large number, *k*-means runs faster than Clara. For small number of data points (Ward), *k*-means runs faster than Clara and GA. As noted earlier, the *k* means algorithm performance is sensitive to the selection of initial seed points, and as such the actual computation time for *k* means may have to be multiplied numerous times if a global optimal solution is sought.

6.4.2 MEASURING CLUSTERING EFFICIENCY BASED ON AVERAGE SILHOUETTE WIDTH

Kaufman and Rouseeuw (1990) introduced silhouette width as a plot showing which data points lie within a cluster and which ones are between clusters. Width of a cluster can be considered as a measure of a good or bad clustering outcome, and enables multiple algorithms to be compared. A large silhouette width indicates a good clustering solution and a small silhouette width indicates an average or bad clustering solution. An average silhouette width is the average of all the silhouette width of different clusters in a clustering problem. Reynolds and Richards et al (2006) demonstrated how average silhouette width (Kaufman and Rouseeuw, 1990) could be implemented as a method of comparing clustering efficiency. They present the following equation for silhouette width:

$$S(k) = \frac{x(k) - y(k)}{\max\left(x(k), y(k)\right)} \tag{6.3}$$

where $y(k)$ is the average distance of k from all other objects in the cluster C_k. For each $C \neq C_k$ average distance of k from the object C is given by $d(k, C)$. $x(k)$ is the smallest result after computing $d(k, C)$ for all clusters $C \neq C_k$. The mean of $S(k)$ for all objects k, is said to be the “average silhouette width” of that cluster solution. $S(k)$ ranges between 1 for a good clustering solution and -1 which would be a bad clustering solution (Reynolds and Richards et al, 2006).

In order to measure relative efficiency of the algorithm optimisation procedures, the average silhouette widths were calculated for *k*-means, Clara, and GA for 1-100 cluster solutions on the three different levels of geographies for UK. Figures 6.9 - 6.11 show the relationship between average silhouette width and cluster frequency for *k*-means, Clara, and GA on the three levels of geographic data for UK. These algorithms were each run a single time for each value of *k.*

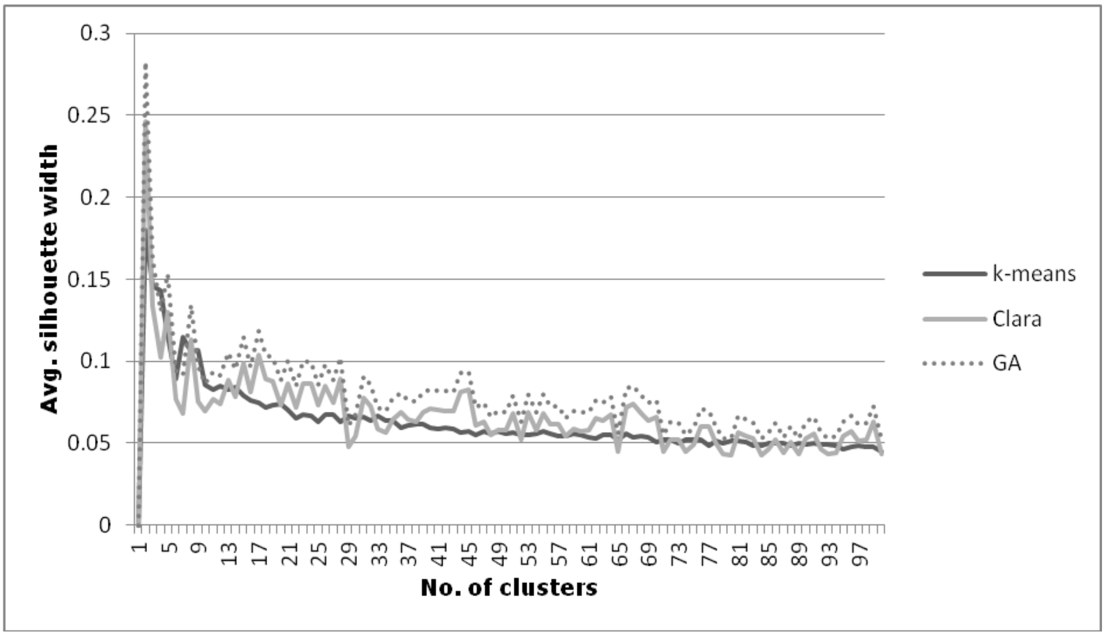


Figure 6.9: Comparing classification optimization efficiency in an Output Area (OA) level data set covering the UK

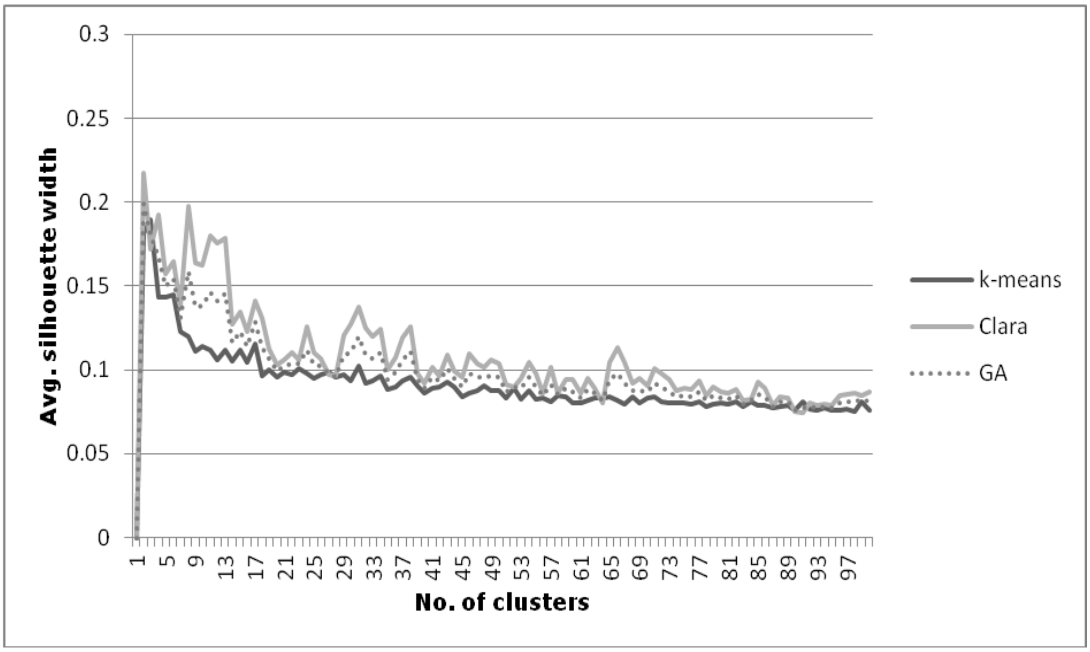


Figure 6.10: Comparing classification optimization efficiency in a Lower Super Output Area (LSOA) level data set for the UK

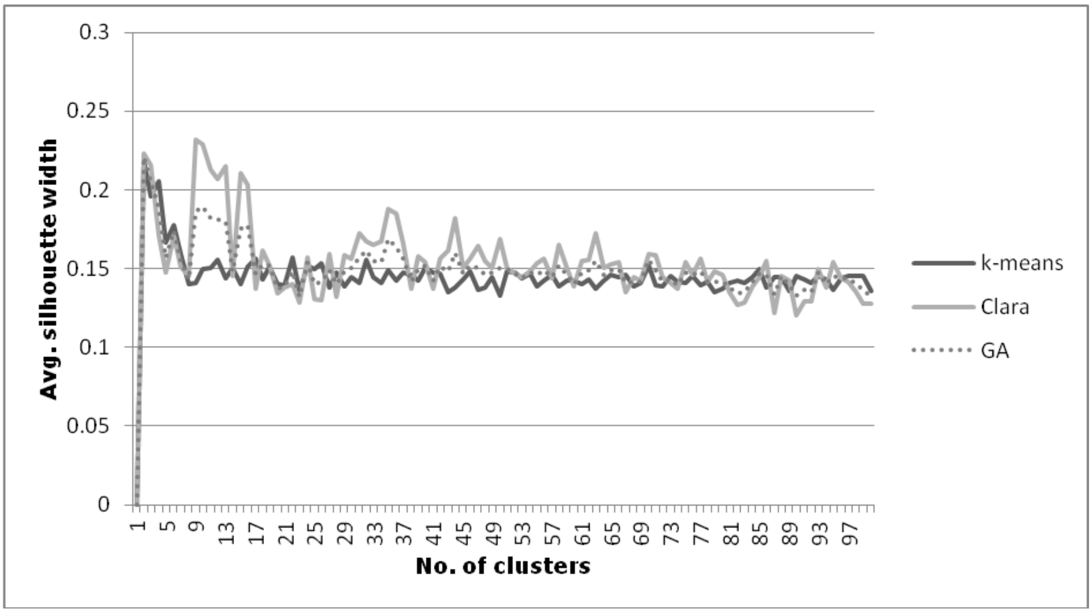


Figure 6.11: Comparing classification optimization efficiency in a Ward level data set for the UK

These charts show that for large numbers of data points, GA works better than *k*-means and Clara. However, for small numbers of data points, Clara still gives better results than *k*-means and GA.

6.4.3 MEASURING CLUSTERING EFFICIENCY BASED ON VARIABLE STANDARDISATION TECHNIQUES

An important aspect of every geodemographic classification is the standardization technique used. Here we measure the efficiency of clustering algorithms by using three different variable standardization techniques: z-scores, range standardization, and principal components analysis.

z-scores provide the most widely used method of variable standardization. If x_i be the value of a variable for area *i* and x_{mean} is the average value of the variable across all *n* areas, then the z-score is defined as:

$$Z_i = \frac{x_i - x_{mean}}{\sigma_x} \tag{6.4}$$

Z-scores are widely used in variable standardisation, but are vulnerable to outliers in the dataset. The *k*-means clustering algorithm that is usually used for creating geodemographic classifications is very sensitive to outliers. Because of multiple runs, the process requires a lot of computational time to achieve the final optimal solution. However, in an online environment where computational time is critical, z-scores may add to the computational burden. Vickers & Rees (2007) uses an alternative variable standardisation technique called the range standardisation method. This method standardises the values of each variable within the 0 and 1 interval. For a variable (x_i) the range standardisation index for area *i* is given by:

$$R_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{6.5}$$

Voas & Williamson (2001) uses principal components analysis as a variable standardisation technique. This is a powerful tool for analysing structures in data, although it is rather easy to analyse a data set with smaller dimensions than one with large dimensions. The components produced by principal components analysis represent weighted combinations of the original variables. The first component accounts for the largest component of the variance, and subsequent components account for successively smaller amounts of the remaining variation. Range standardisation and principal components analysis have each had their critics in the literature because, respectively, range standardisation compresses the data in the range of 0-1, while principal components analysis arguably places emphasis on the part of the dataset which accounts for maximum variance. This is why either of these variable standardisation techniques may omit some interesting patterns in the dataset.

This section develops a comparison of the computational run time of the three clustering algorithms against the number of clusters in a solution.

Z-SCORES

In order to measure the computation time efficiency of the algorithms using *z*-scores as a variable standardisation technique, *k*-means, Clara, and GA were run for 1-100 cluster solutions again on the three different geographic levels for the UK, and then CPU clock time (in seconds) was compared for each algorithm to converge on a specified frequency of clusters (see Figures 6.12 – 6.14). As with the previous analysis, the algorithms were run only a single time for each value of *k*.

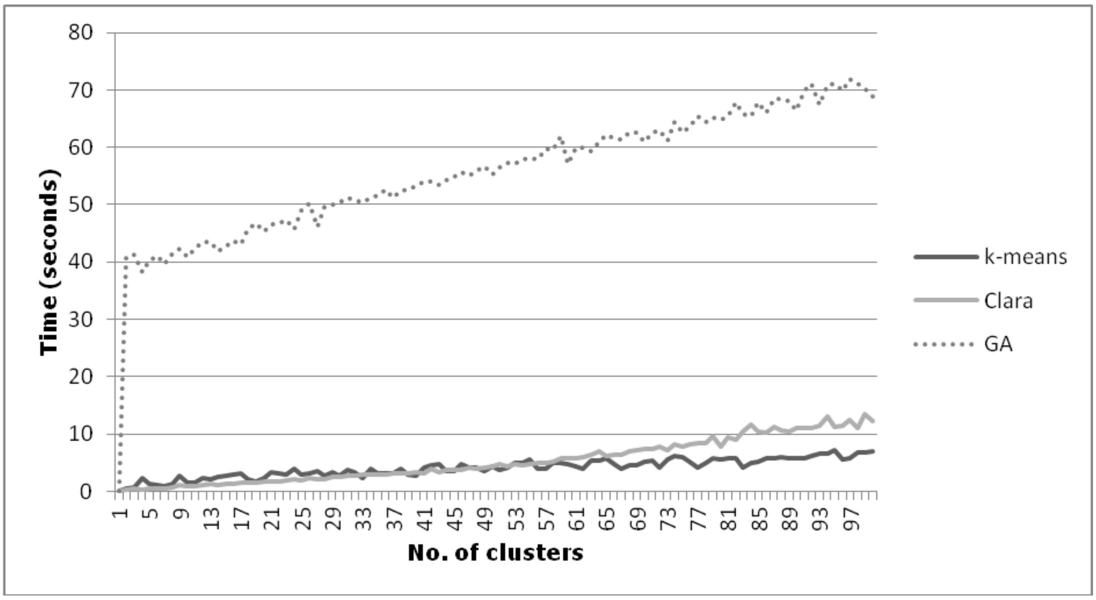


Figure 6.12: Comparing computational efficiency using z-scores as the standardisation technique for a UK Output Area (OA) level data set

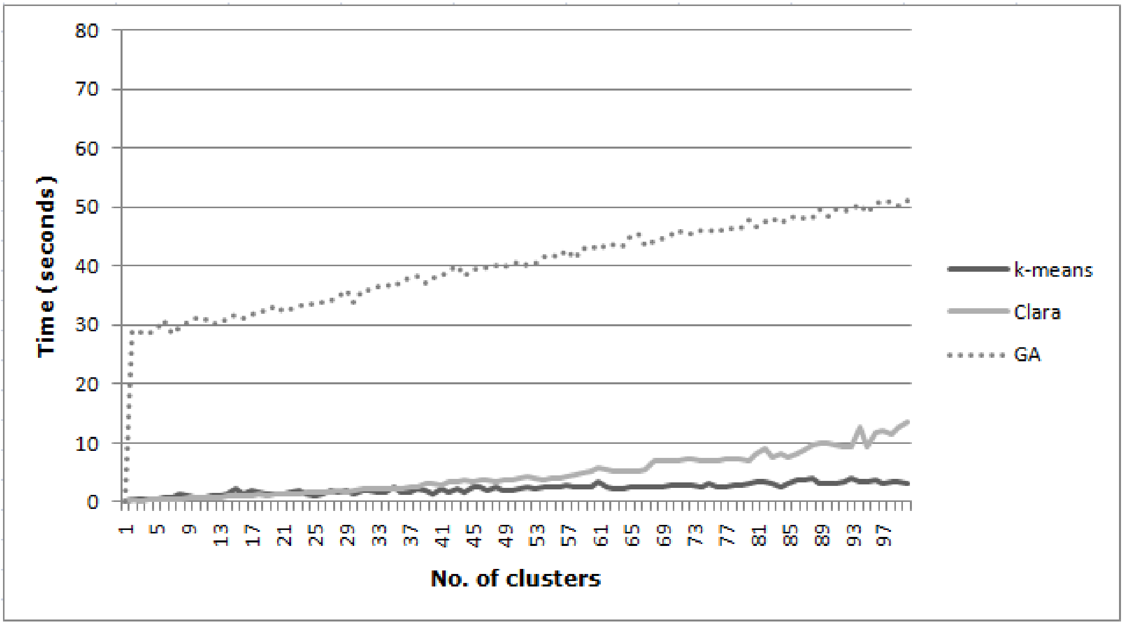


Figure 6.13: Comparing computational efficiency using z-scores as the standardisation technique for a UK Lower Super Output Area (LSOA) level data set

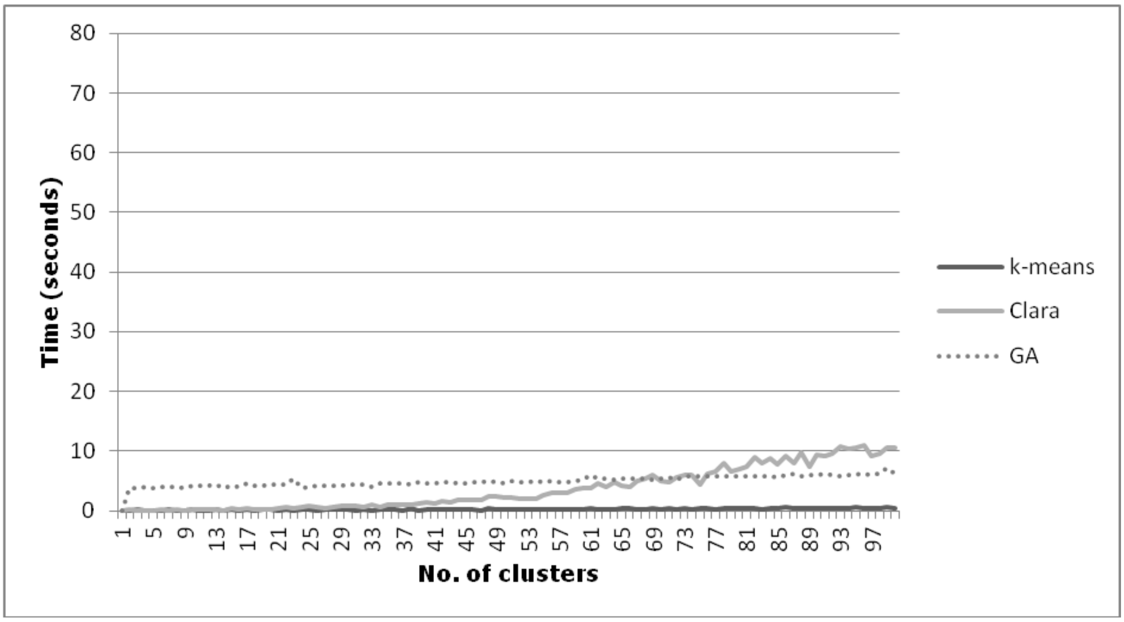


Figure 6.14: Comparing computational efficiency using z-scores as the standardisation technique for a UK Ward level data set

These charts indicate that for a data set with high dimensionality (i.e. the OA and LSOA levels) Clara runs faster than *k*-means and GA when the number of clusters is small (<55 for OA level and <30 for LSOA). However, when the number of clusters is large, *k*-means runs faster than Clara. For small numbers of data points (Ward level), *k*-means runs faster than Clara and GA.

RANGE STANDARDISATION

A repeat of the analysis was conducted with z-scores being supplemented for range standardisation (Figure 6.15 to Figure 6.27).

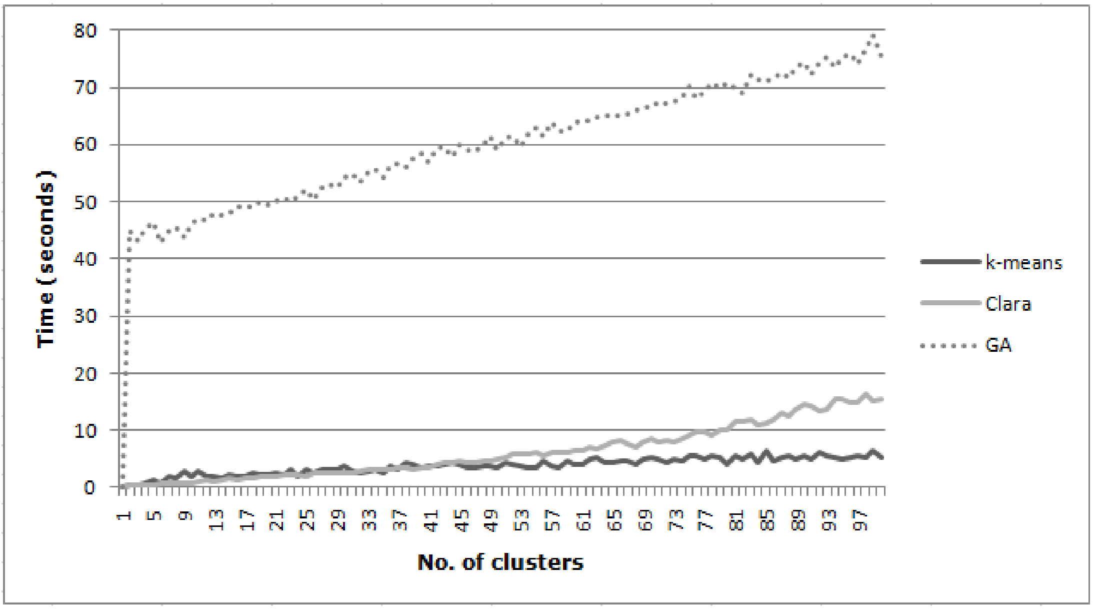


Figure 6.15: Comparing computational efficiency using range standardisation as the standardisation technique for the UK Output Area (OA) level data set

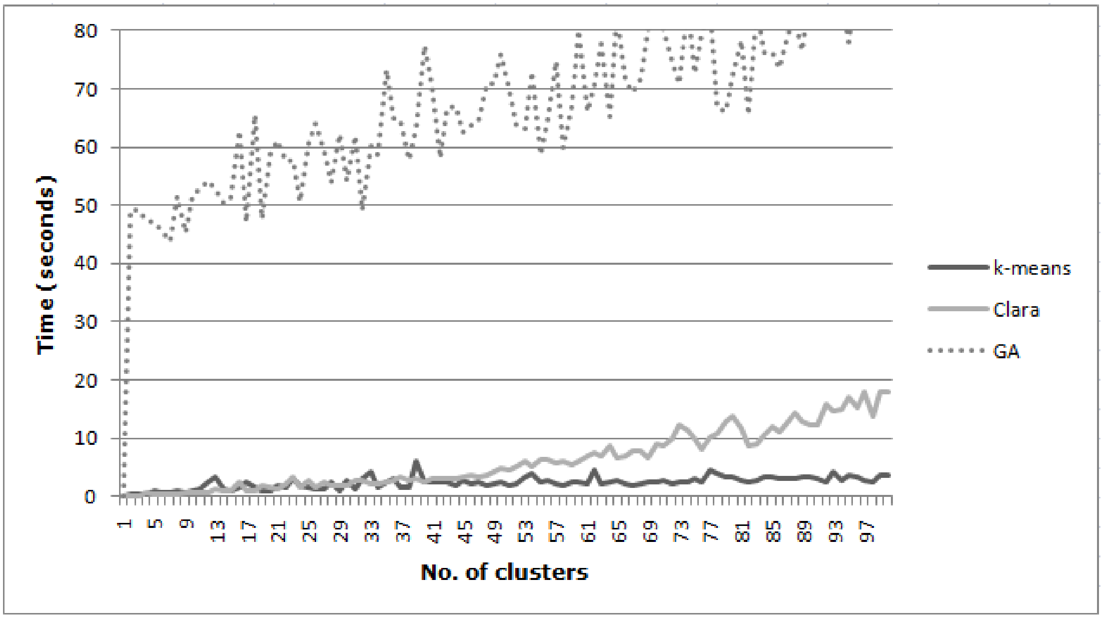


Figure 6.16: Comparing computational efficiency using range standardisation as the standardisation technique for the UK Lower Super Output Area (OA) level data set

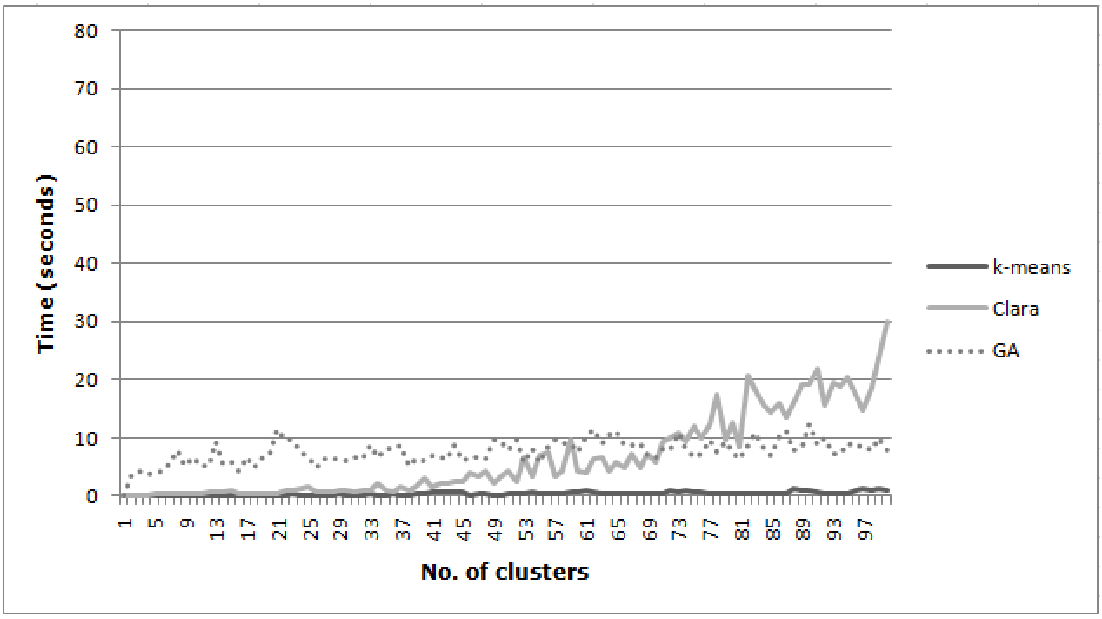


Figure 6.17: Comparing computational efficiency using range standardisation as the standardisation technique for the UK Ward level data set

The results are similar *to z*-scores and broadly show that again for large numbers of data points (OA and LSOA level) Clara runs faster than *k*-means and GA when the number of clusters is small (<44 for OA level and <40 for LSOA). Again, when the number of clusters is increased, *k*-means runs faster than Clara. For smaller datasets with lower dimensionality (e.g. Ward level), *k*-means runs faster than both Clara and GA.

PRINCIPAL COMPONENT ANALYSIS

A repeat of the analysis was conducted with principal component analysis (PCA) being supplemented for z-scores.

For the OA level data, principal components analysis was applied on the dataset and 21 principal components were identified. These accounted for 90.69% of the variance in the dataset. The three clustering algorithms were run on these components and results are shown in the Figure 6.18.

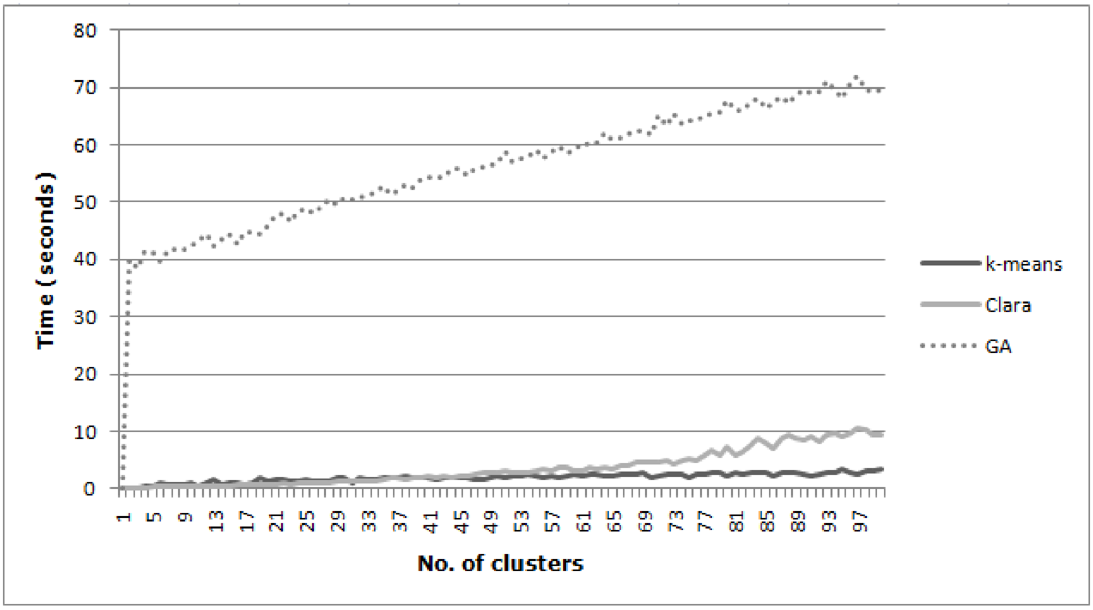


Figure 6.18: Comparing computational efficiency using PCA as the standardisation technique for the UK Output Area (OA) level data set

For LSOA level data, 16 principal components were created that accounted for 90.8% of variance within the dataset. The three clustering algorithms were then computed on these components, the results of which are shown in Figure 6.19.

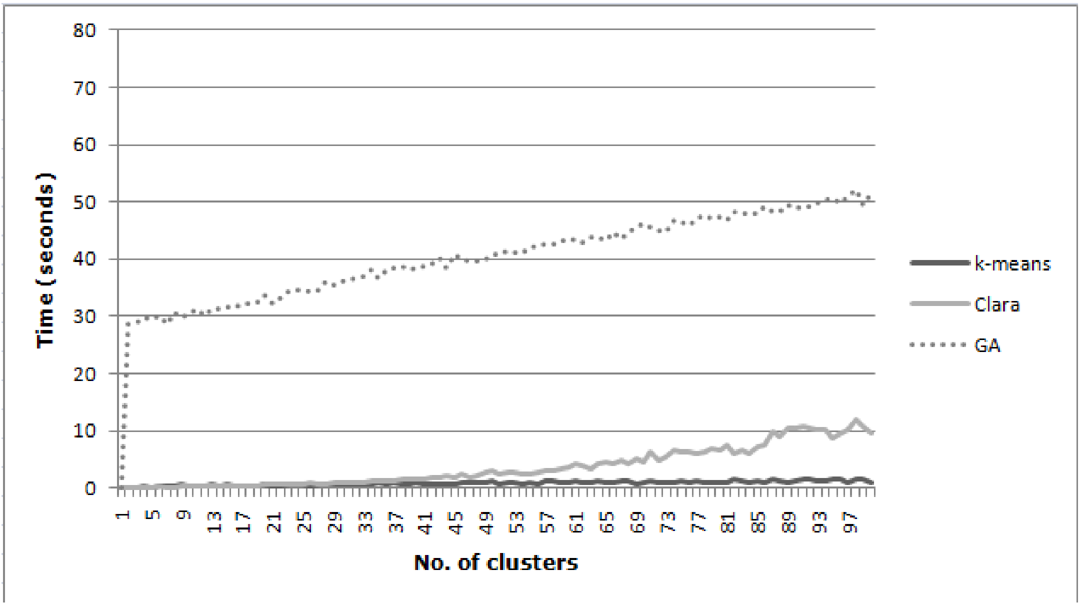


Figure 6.19: Comparing computational efficiency using PCA as the standardisation technique for the UK Lower Super Output Area (OA) level data set

For the Ward level data, principal components analysis was used to identify 10 principal components that accounted for 91.11% of the variance in the dataset. The three clustering algorithms were again run on these principal components, with the results graphed in Figure 6.20.

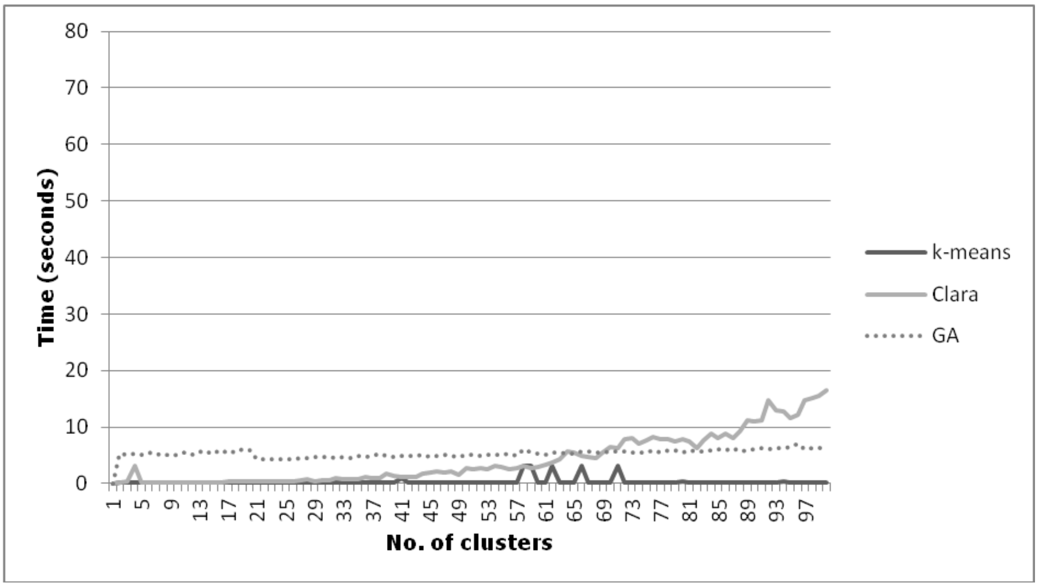


Figure 6.20: Comparing computational efficiency using PCA as the standardisation technique for the UK Ward level data set

Another similar result is returned by these analyses, again showing that for large numbers of data points (OA and LSOA level) Clara runs faster than *k*-means and GA when the number of clusters is small (<45 for OA level and <30 for LSOA). However, when there are a larger number of clusters, *k*-means runs faster than Clara. For small numbers of data points (Ward level), *k*-means runs faster than Clara and GA.

SUMMARY OF THE COMPARISON BETWEEN K-MEANS, CLARA, AND GENETIC CLUSTERING

Based on the results of the comparison, it can be concluded that for the better partitioning of the data (as measured by average silhouette width) GA works better for larger data sets, but Clara is more appropriate for smaller data sets. For speed of computation, Clara works better than GA and *k*-means for larger datasets where the number of clusters is a small. For larger data sets where a greater number of clusters is required, *k*-means runs faster than Clara and GA. When examining how these different algorithms are affected by a choice of different standardisation procedures there was little difference in computational times between methods. From the perspective of creating geodemographic classifications online, these algorithms and standardisation procedures could be chosen on the fly by the classification information system, based upon dataset size and user inputs.

6.5 *K*-MEANS COMPUTATIONAL IMPROVEMENTS

Previous section gave the comparison of three different clustering algorithms using different measures. This section is dedicated to the *k*-means computational improvements from the view of creating geodemographic classifications in an online environment. As speed of computation is the most important factor in an online environment, this section is based on making *k*-means clustering algorithm work faster.

6.5.1 APPLYING *K*-MEANS ON PRINCIPAL COMPONENTS OF A DATASET

K-means has an interesting relationship with PCA (Principal Component Analysis). (Ding & He, 2004) identified that PCA projects to the subspace where the global solution of K-means clustering lie, and thus facilitate K-means clustering to find near-optimal solution.

To test the relationship, *k*-means was run on the 41 Output Area Classification (Vickers & Rees, 2007) variables and 26 Principal Components (accounted for 90% variance in the dataset). The results were mapped for London. Figures (6.20 – 6.21) show that the results are similar when *k*-means was run on a dataset and its principal components.

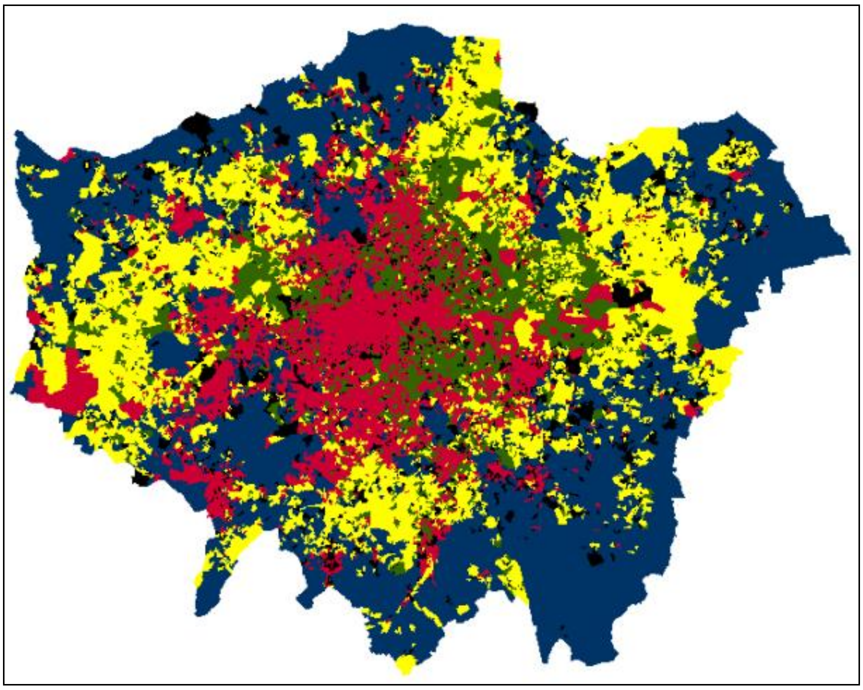


Figure 6.21: ***K*-means applied on 41 OAC variables**

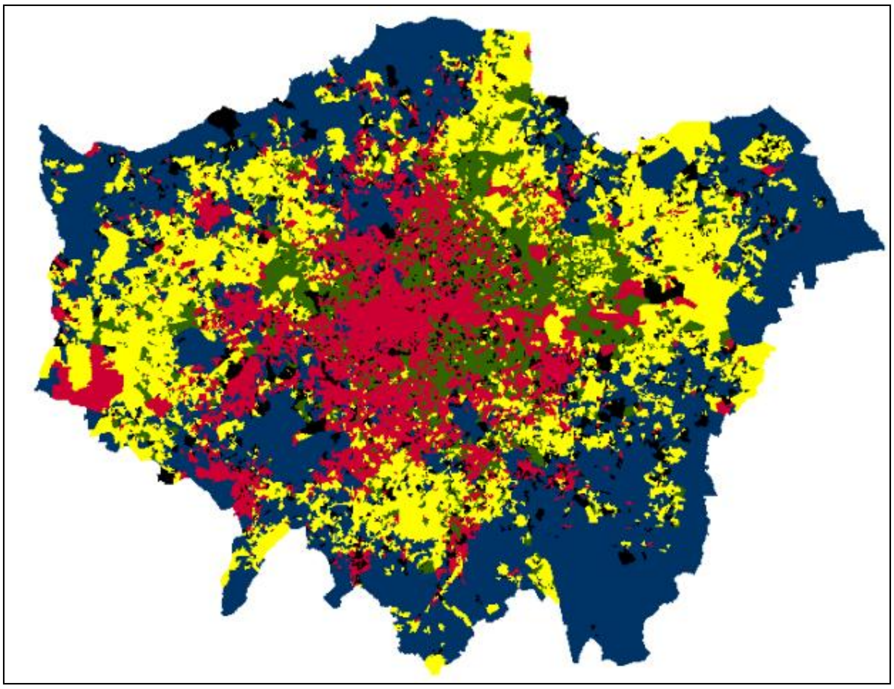


Figure 6.22: ***K*-means applied on 26 principal components**

However, the results are same but applying *k*-means on principal components of a dataset takes less time than applying *k*-means on the whole dataset. Thus PCA could be applied as a standardisation technique, if an enhanced computational performance is needed from *k*-means clustering algorithm.

6.5.2 PARALLEL IMPLEMENTATION OF *K*-MEANS

This section describes a parallel implementation of *k*-means using NVIDIA's Computer Unified Device Architecture (CUDA)[i]. CUDA allows different processes to run in parallel on the Graphical Processing Units (GPUs) of NVIDIA's graphics cards.

CUDA is a general-purpose parallel computing architecture that uses the GPUs of NVIDIA graphics cards to solve complex computational problems. A typical CUDA enabled NVIDIA graphics card has a number of GPUs and a set of memory capable of storing a reasonably large amount of data. For example, "GeForce 8400M GT" graphics card has 16 GPUs and 512MB of internal memory. CUDA requires that the computational problem to be programmed in the C language for parallel processing.

In order to improve the overall efficiency of *k*-means the remainder of this section reviews its implementation under CUDA. The *k*-means algorithm is described in detail in section 6.3 of this chapter.

Some parallel implementations of *k*-means exist with CUDA. For example, Takizawa and Kobayashi (2006) proposed a parallel *k*-means solution for solving "image texture size problem". Hall and Hart (2004) also proposed another parallel solution for solving the problem of "limited instance counts and dimensionality" for complex shapes. However, these implementations only work in specified environments and there are no global parallel *k*-means solutions that are suitable for creating geodemographic classifications.

The proposed parallel *k*-means algorithm via CUDA works as follows:

Total number of runs is specified by *N*.

a) Central Processing Unit (CPU) prepares the data points and counts the number of GPUs available on the NVIDIA graphics card. Afterwards the CPU uploads the data points and code instructing one *k*-means run to each GPU.

b) GPU performs *k*-means clustering on the data points by minimizing expression (1). When an optimal solution is achieved, GPU returns the result to CPU and claims the next *k*-means run from CPU if there are any.

c) CPU stores the results returned by GPUs in a local data structure contained in Random Access Memory (RAM). CPU keeps on delegating requests to GPUs until number of runs are less than *N*.

d) If number of runs is equal to *N*, CPU compares the “within sum of squares distance” optimisation criteria of all the runs.

e) The optimal solution is the one that has minimum “within sum of squares distance”.

6.5.3 COMPARING *K*-MEANS AND PARALLEL *K*-MEANS

This section demonstrates the comparison of *K*-means and Parallel *K*-means clustering algorithms. The hardware used for this evaluation comprised an “Intel Core2 Duo 2.10GHz” CPU, 4GB RAM, and “GeForce 8600M GS” NVIDIA graphics card. The graphics card has 16 GPUs and 512 MB of RAM. For comparison the input data for the National Statistics Output Area Classification (Vickers & Rees, 2007) was used. The comparison was done at three spatial levels i.e. Output Area (OA), Lower Super Output Area (LSOA), and Ward.

Figures 6.22 - 6.24 shows the results of running *K*-means and Parallel *K*-means on National Statistics Output Area Classification (Vickers & Rees, 2007) data for three different spatial levels.

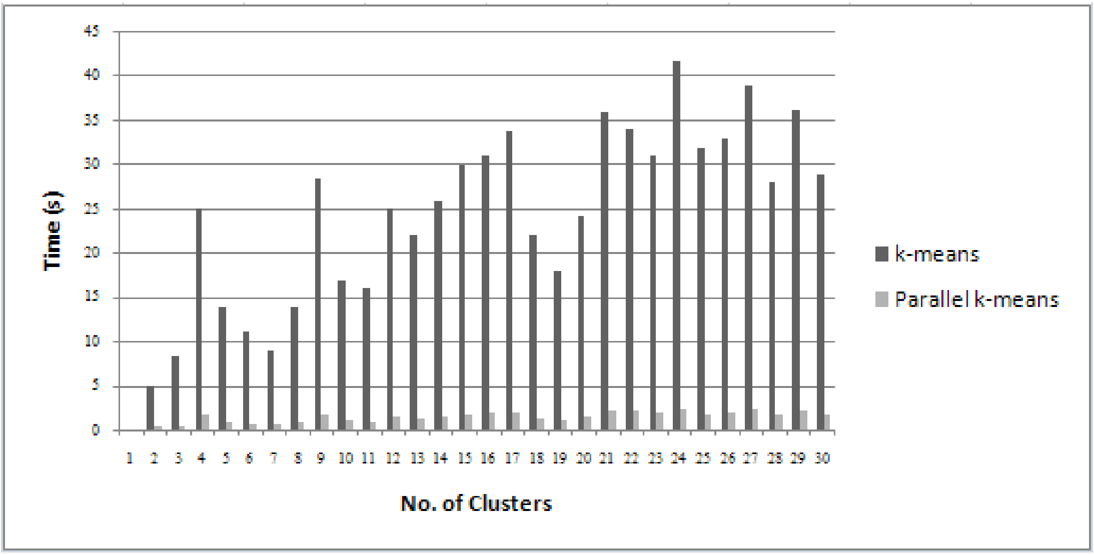


Figure 6.23: Output Area (OA) level results for the two clustering algorithms

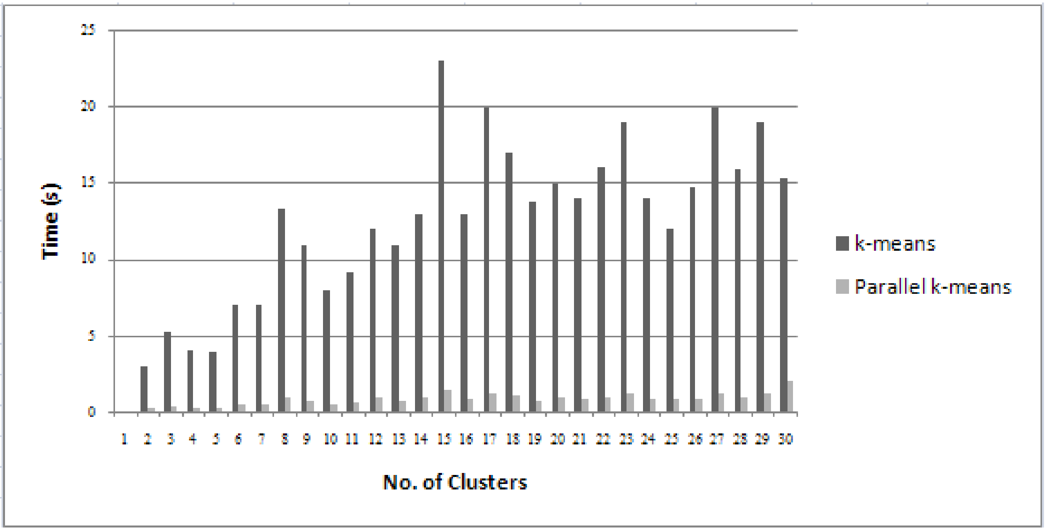


Figure 6.24: Lower Super Output Area (LSOA) level results for the two clustering algorithms

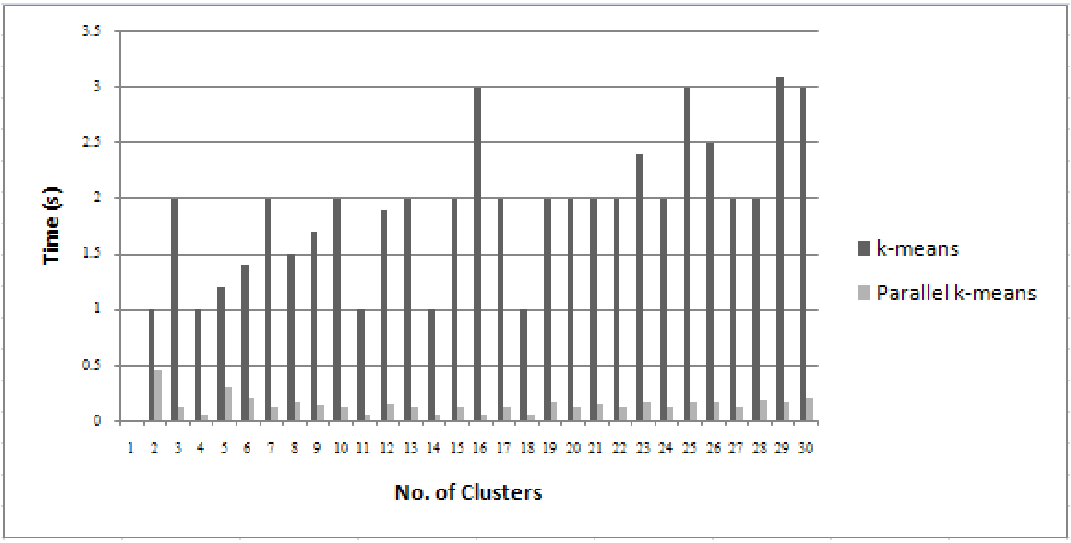


Figure 6.25: Ward level results for the two clustering algorithms

These results indicate that the parallel implementation of *k*-means out performs *k*-means considerably in terms of computational time across all different spatial levels featured in the evaluation. For different spatial levels, parallel *k*-means runs 12 times faster than the standard *k*-means cluster algorithm. This makes parallel *K*-means the best choice for its usages in an online environment.

6.6 COMPUTATION OF GEODEMOGRAPHIC CLASSIFICATIONS IN AN ONLINE ENVIRONMENT

As explained in section 6.5.1 of this chapter, *k*-means runs faster on principal components than the whole dataset while delivering similar results. Hence, principal components analysis can be used as the data standardisation technique for computing online geodemographics. Parallel *k*-means shows greater computational efficiency when compared with the standard *k*-means clustering algorithm. Thus, computation of an online geodemographic classification can be achieved by combining PCA as the standardisation technique and Parallel *k*-means as the clustering algorithm.

There are hardware implications for running parallel *k*-means clustering algorithms, as it needs a powerful NVidia graphics card installed on the machine.

However, with the decrease in computation prices in recent years this implication is acceptable. Also, the hardware implications are only for the server computer in an online environment, as all the computations are performed on a main computer in the online environment.

6.7 CONCLUSION:

This chapter has set out the procedures of cluster analysis and a range of different clustering algorithms in detail. *K*-means remains the core clustering algorithm for computing geodemographic classifications. This chapter provides a thorough discussion of *k*-means clustering algorithm and provides a thorough comparison of *k*-means using Clara and with Genetic Clustering. Different metrics were used to test the clustering algorithms and the evidence suggests that different algorithms perform better in different situations. For a geodemographic information system, it is envisaged that different algorithms and standardisation procedures could be chosen on the fly by the geodemographic information system (based upon dataset size and user inputs)

Improvements in *k*-means are proposed in Sections 6.5.1 (specifically, use of *k*-means with principal components analysis) and 6.5.2 (parallel *K*-means) of this chapter. Parallel *k*-means offers greater computational efficiency when compared to *k*-means clustering. We can conclude that for an online geodemographic information system principal component analysis (PCA) can be used as the data standardisation technique of choice, and that parallel *k*-means clustering is suitable for the computation geodemographic classifications.

The next two chapters described the development of a pilot desktop software utility for building geodemographic classifications.

7 CHAPTER 7: FUNCTIONAL AND TECHNICAL APPLICATION SPECIFICATION FOR THE CREATION OF GEODEMOGRAPHIC CLASSIFICATIONS

7.1 INTRODUCTION

Previous chapters have set out the motivation for building real-time geodemographics classification, using methods, tools and data that are open and comprehendible by users. These principles apply to the variables and the weights used, techniques used to standardise variables, and the clustering algorithms used for building a geodemographic classification. Chapter 6 set out the different techniques of cluster analysis available for building a geodemographic classification and set out a number of important enhancements to the *k*-means clustering algorithms. These enhancements in *k*-means are important for building geodemographic classifications that can be specified and estimated quickly and efficiently (possibly in online environments).

This chapter describes the development of a pilot desktop software utility for building geodemographic classifications. This software is a proof of concept for building real-time geodemographics classifications, and might be enhanced in future research to operate in an online environment. The software has been named 'GeodemCreator'. It is a freeware desktop software utility, and any user can use it to build their own geodemographic classifications.

From a software engineering perspective, functional and technical application specifications should be defined before the actual coding of the software starts. Functional specification of software describes the overall functionality in terms of contextual information and flow charts. Software has different components, and the functional part should explain and outline the individual components of the software in the form of flow charts, interaction diagrams, and contextual information.

A technical application specification specifies:

- the intended audience who will use the software product
- the operating environment in which software will run (i.e. operating systems, and any required supporting software)

- the software scope
- the technology used to develop the software product
- the software design (state transition diagrams, class diagram, use case diagram)
- the overall structure of the software product
-

The functional and technical application specifications are defined in Sections 7.2 and 7.3 of this chapter.

7.2 FUNCTIONAL APPLICATION SPECIFICATION

A functional application specification explains the overall functionality of the software in terms of contextual information and flow charts. 'GeodemCreator' is designed to work in two modes of operations, which are:

- Basic mode
- Advanced mode

Before describing these modes in sections 7.2.2 and 7.2.3, it is important to describe the socio-economic data provided in the 'GeodemCreator' by default.

7.2.1 SOCIO-ECONOMIC DATA PROVIDED IN THE 'GEODEMCREATOR' BY DEFAULT

Vickers & Rees (2007) created the 2001 Output Area Classification by selecting (but not weighting) 41 variables from the 2001 Census of Population. These variables were chosen from a number of domains covering demographic, household composition, housing, employment, and socio-economic characteristics.

The same variables and their domains are provided in the 'GeodemCreator' as default variables. These variables are included at the output area level; it means 'GeodemCreator' will enable users to create output area classifications by the default variables provided in the software. The following table 7.1 shows the 41 variables and their respective domains.

Variables	Domains
V1: Age 0-4 V2: Age 5-14 V3: Age 25-44 V4: Age 45-64 V5: Age 65+ V6: Indian, Pakistani or Bangladeshi V7: Black African, Black Caribbean or Other Black V8: Born Outside the UK V9: Population Density	**Demographic**
V10: Divorced V11: Single person household (not pensioner) V12: Single pensioner household V13: Lone Parent household V14: Two adults no children V15: Households with non-dependent children	**Household Composition**
V16: Rent (Public) V17: Rent (Private) V18: Terraced Housing V19: Detached Housing V20: All Flats V21: No central heating V22: Rooms per household V23: People per room	**Housing**
V24: HE Qualification V25: Routine/Semi-Routine Occupation V26: 2+ Car household V27: Public Transport to work V28: Work from home V29: Limiting Long Term Illness (SIR) V30: Provide unpaid care **V31: Students (full-time)**	**Socio-Economic** **The variables in bold are for the 'Employment'**

V32: Unemployed **V33: Working part-time** **V34: Economically inactive looking after family** V35: Agriculture/Finishing employment V36: Mining/Quarrting/Construction employment V37: Manufacturing employment V38: Hotel & Catering employment V39: Health and Social work employment V40: Financial intermediation employment V41: Wholesale/retail employment	**domain.**

Table 7.1: Census variables and their domains for creating geodemographic classifications

7.2.2 BASIC MODE FUNCTIONALITY

In the 'basic' mode, the software will offer a limited range of functions to users. This mode is intended for users who are not expert at creating geodemographic classifications and who want the software to create a classification for them without the need to specify variables, their weightings, the number of geodemographic classes (groups), or the choice of clustering algorithm.

In the 'basic' mode of operation, the software allows users to create their geodemographic classifications in the following domains.

Domains
Demographic classification
Household composition classification
Housing classification
Socio-economic classification
Employment classification
An overall classification based on all the above public sector domains

Table 7.2: Domains for the 'Basic' mode

The individual variables for each domain are outlined in the section 7.2.1. The software allows the Output Area level classification for the above mentioned domains to be created for the following geographical areas in the United Kingdom.

Geographical Areas
United Kingdom
East Midlands
East of England
London
North East
North West
Scotland
South East
South West
Wales
West Midlands
Yorkshire and Humber
Northern Ireland

Table 7.3: Geographical Areas for the classifications

FLOW CHART OF THE USER'S AND SOFTWARE'S OPERATIONS

In the 'basic' mode, a user specifies the 'Classification Type' and 'Geographical area' for building a classification. The software, then selects a number of pre-specified variables and their weightings, builds the classification, and returns the output to the user.

Figure 7.1 shows the flowchart of these operations.

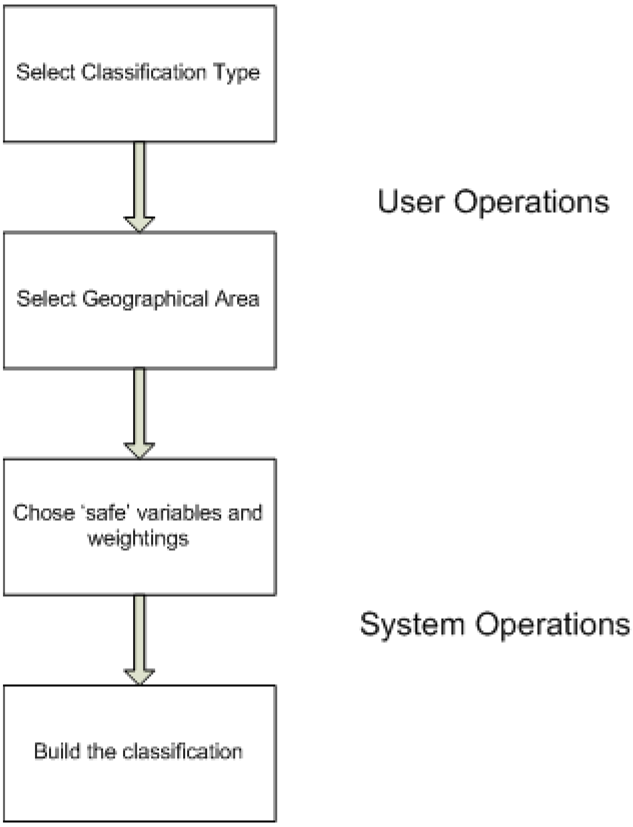


Figure 7.1: Flowchart of 'basic' mode operations

In this mode of operation, users do not have the ability to specify any variables, their weightings, and the number of output geodemographic groups (clusters). Instead, the software selects a number of pre-specified variables and their weightings for creating a geodemographic classification. This is intended to enable inexperienced users to create their own bespoke geodemographics classifications, with emphasis upon the general area of application that they are working in. The menu of variables is proposed with public sector resource allocation decisions in particular focus. The features of variable selection and the specification of the output geodemographic groups (clusters) are available in the advanced mode of the software, where a user can perform more advanced operation while creating a classification.

7.2.3 ADVANCED MODE FUNCTIONALITY

In the 'advanced' mode, users have greater control of creating a geodemographic classification. This mode is for more expert users who know about the procedures and methods for creating a geodemographics classification. Users can specify the variables, the geographical level used to create the classification for, and the number of geodemographic groups. Additionally, users can upload their own data to the software, and use their own variables (in addition to the default variables in the software) in order to create a classification. This gives users control over what data go into the classification, how the variables are weighted, and how the software builds the classification.

In addition to providing their own data sources, users can avail themselves of any of the default data sources provided in the software. Table 7.3 shows a list of the domains available by default in the software. The source of these domains is the output area classification data (Vickers & Rees, 2007). The spatial coverage for these domains is output area.

Default Domains
Demographic
Household composition
Housing
Socio-economic
Employment

Table 7.4: Domains for the 'Advanced' mode

The individual variables for each domain are outlined in the section 7.2.1. In the 'advanced' mode, users can create their Output Area classification for the following geographical areas in the United Kingdom.

Classification type
United Kingdom
East Midlands
East of England
London
North East
North West
Scotland
South East
South West
Wales
West Midlands
Yorkshire and Humber
Northern Ireland

Table 7.5: Geographical Areas for the classifications

FLOW CHART OF THE USER'S INTERACTION WITH THE SOFTWARE

In 'advanced' mode, the flow of operations is of course less straightforward than in the 'basic' mode. This is to allow users independence on the creation of their bespoke classifications.

Figure 7.2 shows the flow of operations for the user’s interaction with the software.

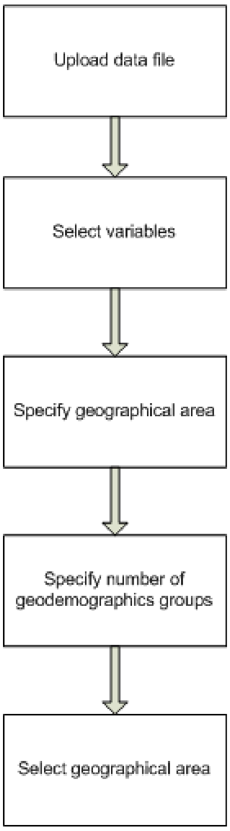

Upload data file
Select variables
Specify geographical area
Specify number of geodemographics groups
Select geographical area

Figure 7.2: Flowchart of user interaction with the software (Advanced mode)

FLOW CHART OF THE SOFTWARE OPERATIONS (ADVANCED MODE)

After a user has specified the inputs, the software is required to perform certain actions in order to create the classification. The software will merge the user's data with the default data (Demographic, Household composition, Housing, Socio-economic, and Employment). The software will show particular information messages to the users regarding the standardisation of variables to a particular format. This is designed to enable users to input data in the right format, which is very important in building a correct geodemographic classification.

Vickers & Rees (2007) identified that highly correlated variables are not useful in creating a geodemographic classification because they introduce data redundancy. An example is the obvious perfect inverse correlation between "percentage male population" and "percentage female population" for any geographical area. The software provides the functionality to test the correlation of the variables uploaded by the user in order to see if any are highly correlated.

Finally, based on the user's input regarding correlated variables, the software will build the geodemographic classification. Figure 7.3 shows the flow chart diagram of the software's interaction with the user.

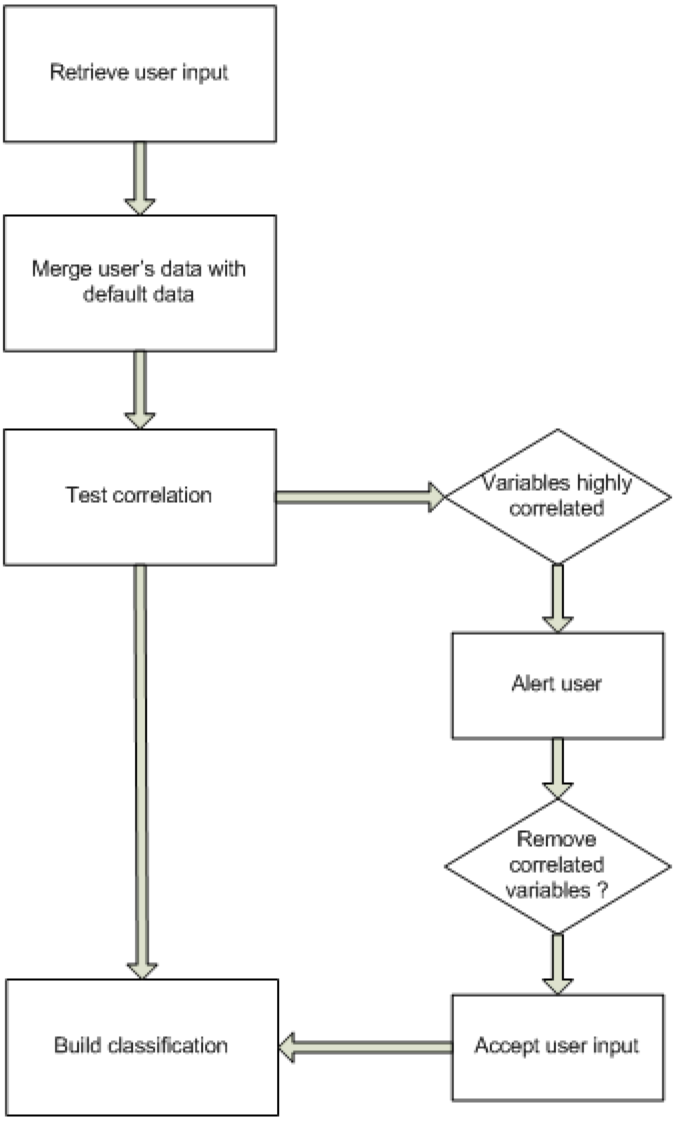

Retrieve user input
Merge user's data with default data
Test correlation
Variables highly correlated
Alert user
Remove correlated variables ?
Accept user input
Build classification

Figure 7.3: Flow chart of the software's interaction with the user (Advanced mode)

In the basic operation, a user does not have the control to specify any variables. However, in advanced level mode of operation a user has more control on how the classification is created.

7.2.4 INCLUSION OF METADATA

Metadata are really important from the perspective of software usability. This section describes the metadata that the software provides to alert users when performing certain operations.

The software will display metadata regarding the following topics.

- Inclusion of highly correlated variables in the classification
- Variable standardisation
- Constancy of variables across a geographical area

INCLUSION OF HIGHLY CORRELATED VARIABLES

Vickers & Rees (2007) describe how the inclusion of pairs of variables with strong correlations within a data set is undesirable for cluster analysis, because they introduce data redundancy. In such circumstances one should exclude one of the variables from the pair, when creating a classification.

Summary information for highly correlated variables will be shown to the users when they upload their own data points to the software. The software will test the correlation of every variable against all the other variables selected, and presents diagnostics to assist in the decision to include or exclude variables based upon the calculated correlation levels.

VARIABLE STANDARDISATION

Vickers & Rees (2007) used range standardisation in the creation of the Output Area Classifications (OAC) and describe it as the preferred method of standardization. A number of data standardisation techniques were discussed in Chapter 6. For the purpose of the 'GeodemCreator' software, all of the default data included in the software are pre-standardised by range standardisation.

The metadata calculations for variable standardisation are presented to advanced user when they upload any of their own data points to the software. The software presents message to users advising them that the data have to be range standardised in order to compute correct geodemographic classifications.

CONSTANCY OF VARIABLES ACROSS A GEOGRAPHICAL AREA

Vickers & Rees (2007) set out the importance of the constancy of variable across a geographical area. An example is the religion question in the 2001 Census data. This question was asked in all countries of the UK (England, Wales, Scotland, and Northern Ireland). In Northern Ireland, the results were reported by splitting the data into different Christian categories, while all the other religions were aggregated into a single 'other' category. In England and Wales, all types of Christians were reported as a single variable and other religions were reported separately. This inconsistency can be a problem in the creation of a geodemographics classification.

The metadata for the consistency of data classes and values will be displayed to the 'Advance' user if they use their own data. Information messages will be shown to the users explaining them that data should be consistent over a geographical region in order to computer a correct geodemographic classification. This will make sure that users use data in the right format which is consistent for a particular geographical area, and will help in creating correct geodemographic grouping.

7.3 TECHNICAL APPLICATION SPECIFICATION

A technical application specification is important for the development of any software product. The functional application specification defines the contextual information and flow charts for the software: however, it is the technical application specification that defines:

- the intended audience
- the operating environment
- the software scope
- the technology

- the software design
- the overall structure of the software product

This section gives a detailed description of each of the above points.

7.3.1 INTENDED AUDIENCE

The intended audience for this software can be any user who, subject to basic or more advanced proficiencies, wants to create their own geodemographic classification. A user could be an inexperienced, in which case the 'Basic' mode of the software makes available a number of pre-selected variables and weightings. Alternatively, a user could be an expert user who wants to create a classification by using their own data. Such users can employ the 'Advanced' mode of the software, where users have more control of how the classification is created. This software is intended to be a free software utility that any user can use to create their own local area bespoke geodemographic classification.

7.3.2 SOFTWARE SCOPE

Software scope defines the overall functionality of the software in the form of contextual information. The development of any software product has time and money associated with it, and the definition of the scope helps in identifying the cost, working hours, and time required for the successful completion of the software.

The GeodemCreator provides the following functionalities to the user for building a geodemographic classification.

1. The software enables the user to build geodemographic classifications in the 'basic' and 'advanced' modes.

2. In the 'basic' mode, users do not have the control to specify variable selection and weightings. Instead, a classification can be built under this mode as follows:

    a. Users select a geodemographics domain (Demographics, Household composition, Housing, Socio-economic, Employment).

    b. Users specify the geographical area for which they want the classification for (UK, London, East of England etc).

    c. Software produces the classification according to the selected domain and geographical area.

    d. The output will be a CSV file with a cluster number assigned to an output area. In addition, the software creates the cluster profiles in the form of radial graphs.

3. In the advanced mode, users have the ability to specify variables and their weightings. In this mode, a classification can be built as follows:

    a. Users upload a data file containing socio-economic variables for a particular geographical area.

    b. Users can opt to use their own variables only, or select a number of variables provided by software in various domains (Demographics, Household composition, Housing, Socio-economic, Employment). Alternatively, they can mix and match their variables with the variables provided by the software. In addition, users can specify the geographical area to build the classification for and the number of output geodemographic classes.

    c. The software builds the classification according to selected variables and their weightings.

    d. The output will be a CSV file with a cluster number assigned to an output area. In addition, the software creates the cluster profiles in the form of radial graphs.

7.3.3 OPERATING ENVIRONMENT

For the development of any software, it is always useful to define the operating environment in which the software will run. The operating environment defines the operating system and hardware required. Because 'GeodemCreator' is developed using Java (http://www.sun.java.com) and R (http://www.r-project.org), it is a platform independent software and can run on any operating system i.e. Windows, Linux, Unix, Mac OS.

The software requires Java and R installed on the machine. A user needs to install 'rjava' library in 'R'. A complete step by step instructions on the installation of 'GeodemCreator' are given in the Appendix 7 and are also included in the CD (with the file name How_To_Install.txt) which accompanies this thesis.

The hardware requirement for this software can be any machine having 2GHz of processor and a minimum 3 Giga Bytes memory.

7.3.4 TECHNOLOGY

This section explains the use of technology for the development of the software. There are different technologies and standards available, and choosing the right one can make the software easy to develop. The use of the right development technology also enables the software developer to alter the software in future, if required.

The choice of selecting a particular software development technology is based on the programming paradigm it uses. Object oriented programming is popular these days for software end web application development. The object oriented programming paradigm uses objects and their interactions to design and develop software or web applications. The use of object orientation in software development has increased in the recent years since applications are well structured, scalable to new changes, and offer better operating performance on different platforms and operation systems. Examples of object-orientated development frameworks include .NET (c#) and Java. From the view of creating software for building geodemographic classifications, object oriented programming presents a better choice for development and copes with the new challenges involving performance and scalability of applications in real world environments.

The Object Oriented Process model starts in an evolutionary spiral. It is iterative in nature. Thus in each iteration, an output is generated in the form of software. 'GeodemCreator' was developed in two iterations. The output of the first iteration was a working version of the software with 'Basic' mode functionality. The output of the second iteration was a working version of the software with the 'Advanced' mode functionality.

Because both .NET (c#) and Java use an object oriented paradigm. Both are good for software development. However, c# is not a cross platform language, as it is integrated with a Microsoft product (Microsoft Visual Studio). It means that applications developed in c# can only run on Microsoft Windows platforms, which is a major reason for rejecting c# as the software development technology. Java is a cross platform language, and it can run on any operating system (Linux, Unix, Microsoft Windows, Apple Operating System). This provides a rationale to select Java as the software development technology for developing this software.

Thus, Java software will interact with 'R' to cluster the data. R is a free software environment for statistical computing and graphics. It is cross platform software, and can run on Unit, Windows, Apple Operating System. Java can perform different functions and will interact with 'R' via JRI (Java to R Interface). User documentation (http://www.rforge.net/JRI/) confirms that the JRI makes it possible to run R inside Java applications as a single thread. It loads the R dynamic library into Java and provides a Java API to R functionality. This enables the Java application to use R packages for performing cluster analysis, as R packages perform cluster analysis better than Java application.

7.3.5 SOFTWARE DESIGN

Software design is an important part of the software engineering process, and before any coding of the software starts a software design should be completed. Reeves (2005) described how most current software development processes try to segregate the different phases of software development into separate pigeon-holes. The top level design must be completed and frozen before any code is written.

This section explains the design of the software for building geodemographic classifications in the form of:

- State-Transition diagram
- Class diagram
- Use case diagram

STATE-TRANSITION DIAGRAM

Copeland (2008) describes how state-transition diagrams describe all of the states that an object can have, the events under which an object changes state (the transitions), the conditions which must be fulfilled before the transition will occur (guards), and the activities undertaken during the life on an object (actions). State transition diagrams are useful to describe the behaviour of objects (i.e. entities that make up the system). Behaviour of all objects in the system describes the overall functioning of any software or web application.

As far as the scope of the GeodemCreator is concerned, there are two modes of operation i.e. 'Basic' and 'Advanced'. What follows are the state transition diagrams of each of the modes.

The following figure shows the state transition diagram for the ‘Basic Mode’.

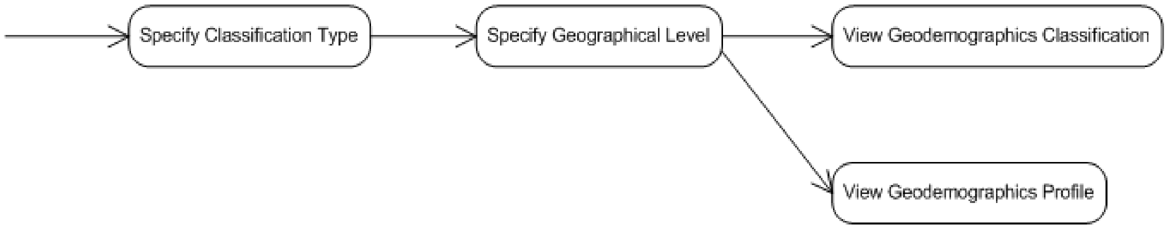

Specify Classification Type
Specify Geographical Level
View Geodemographics Classification
View Geodemographics Profile

Figure 7.4: State Transition Diagram for 'Basic' Mode

Thus under 'basic' mode, users can select the type of geodemographics classification to create, and then specify the geographical area they want to build the classification for. The software creates the final classification by performing cluster analysis, and user can view the final geodemographic classification produced. A user can also see the geodemographic profiles in the form of radial graphs.

The following figure shows the state transition diagram for the ‘Advanced' mode.

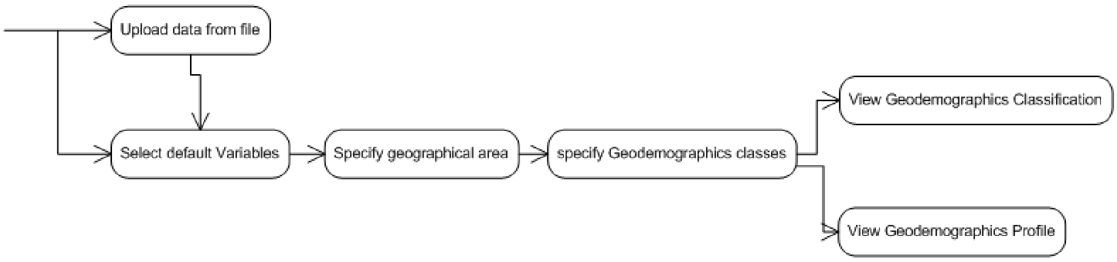

Upload data from file
Select default Variables
Specify geographical area
specify Geodemographics classes
View Geodemographics Classification
View Geodemographics Profile

Figure 7.5: State Transition Diagram for 'Advanced' Mode

In the 'Advanced' mode, users can upload their own variables and select variables from a number of the default variables provided by the software. Users can selected the geographical area they want to product the classification for, and finally specify the number of output geodemographic groups. The software will create the classification by performing cluster analysis on the selected variables and the number of cluster solutions. Finally, the users can view the geodemographics classification produced and their geodemographics profiles in the form of radial charts.

CLASS DIAGRAM

Copeland (2008) describes how a class diagram shows the classes that make up a system and the static relationships between them. Classes are defined in terms of their name, attributes (or data), and behaviours (or methods). The static relationships are association, aggregation, and inheritance. An association between two classes defines a one-to-one or one-to-many or many-to-many link. An aggregation means that a class contains another class's object. And inheritance defines that a child class inherits some of the attributes and methods of its parent class.

A class defines an individual entity of the system, and a class diagram shows all the entities and their relationships. Thus, a class diagram shows the overall picture of the software in graphical form.

Figure 7.6 shows the class diagram of the 'GeodemCreator' software.

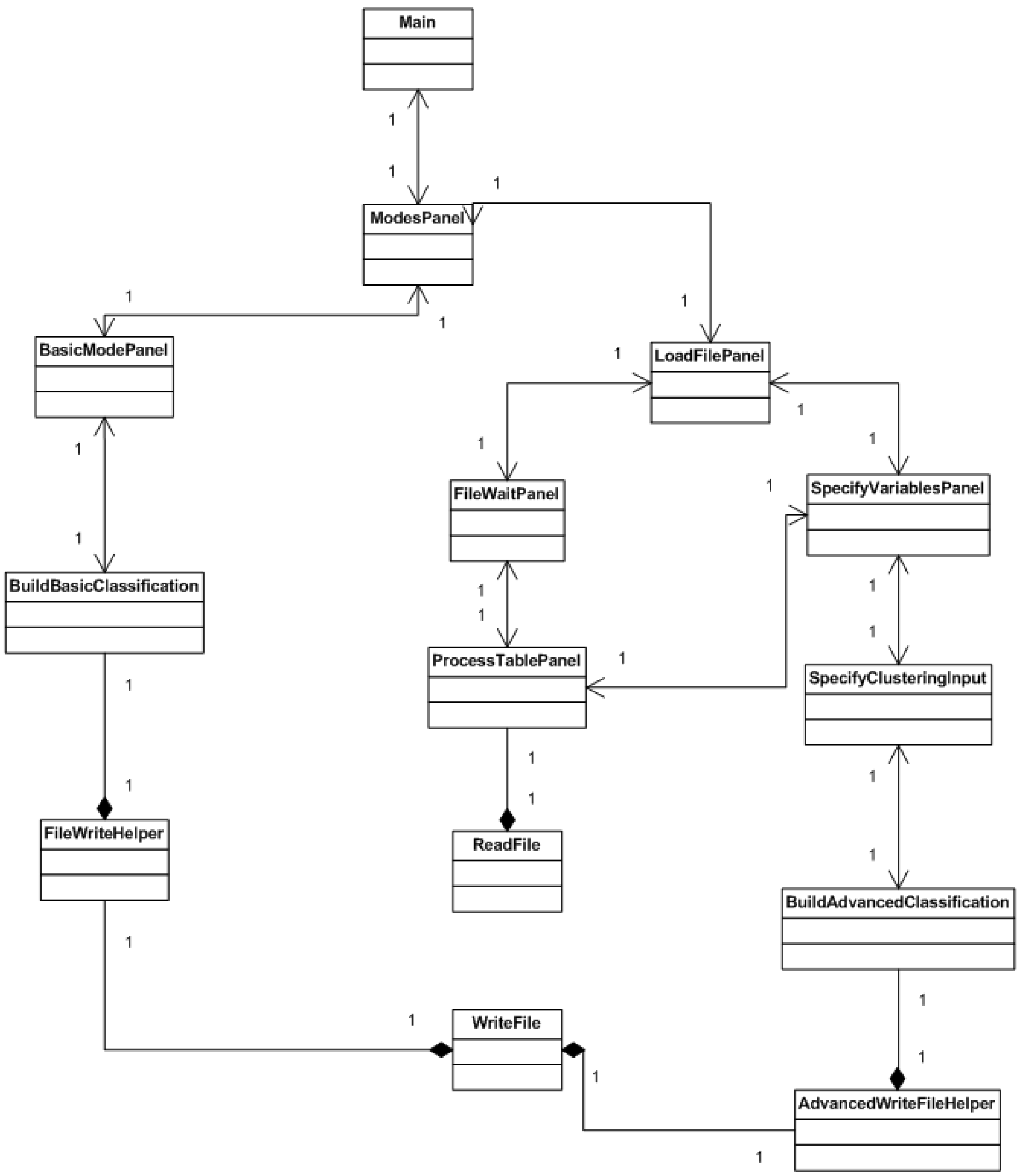


Figure 7.6: Class diagram of the 'GeodemBuilder' software

In the class diagram, diamond represents an aggregation (containment) of a class in another, while an arrow shows the association (linkage) of a class with another. The 'BuildBasicClassification' and 'BuildAdvancedClassification' classes build the geodemographic classifications in 'Basic' and 'Advanced' modes respectively. The 'FileWriteHelper' and 'AdvancedWriteFileHelper' classes write the radial charts. The

'ProcessTablePanel' class reads the user input file and shows it as a table to the user, and 'SpecifyVariablesPanel' class is a panel for the users to specify their variables in the 'Advanced' mode.

USE CASE DIAGRAM

Copeland (2008) described that a use case is a scenario that describes the use of a system by an actor to accomplish a specific goal. An actor is a user playing a role with respect to the system. Actors are generally people, although other computer systems could be actors as well. A scenario is a sequence of steps that describes the interaction between an actor and the system. The use case model consists of the collection of all actors and all use cases.

Use cases describe what functionality a user needs in the system, and they can help in identifying components of the system and their relationships with each other. In addition, use cases can also help in designing test cases. Test cases are used to test software once it is ready to use. So, use cases are really helpful in the development of software of high quality and standard.

The following diagram (Figure 7.7) shows the use case diagram of the GeodemCreator Software:

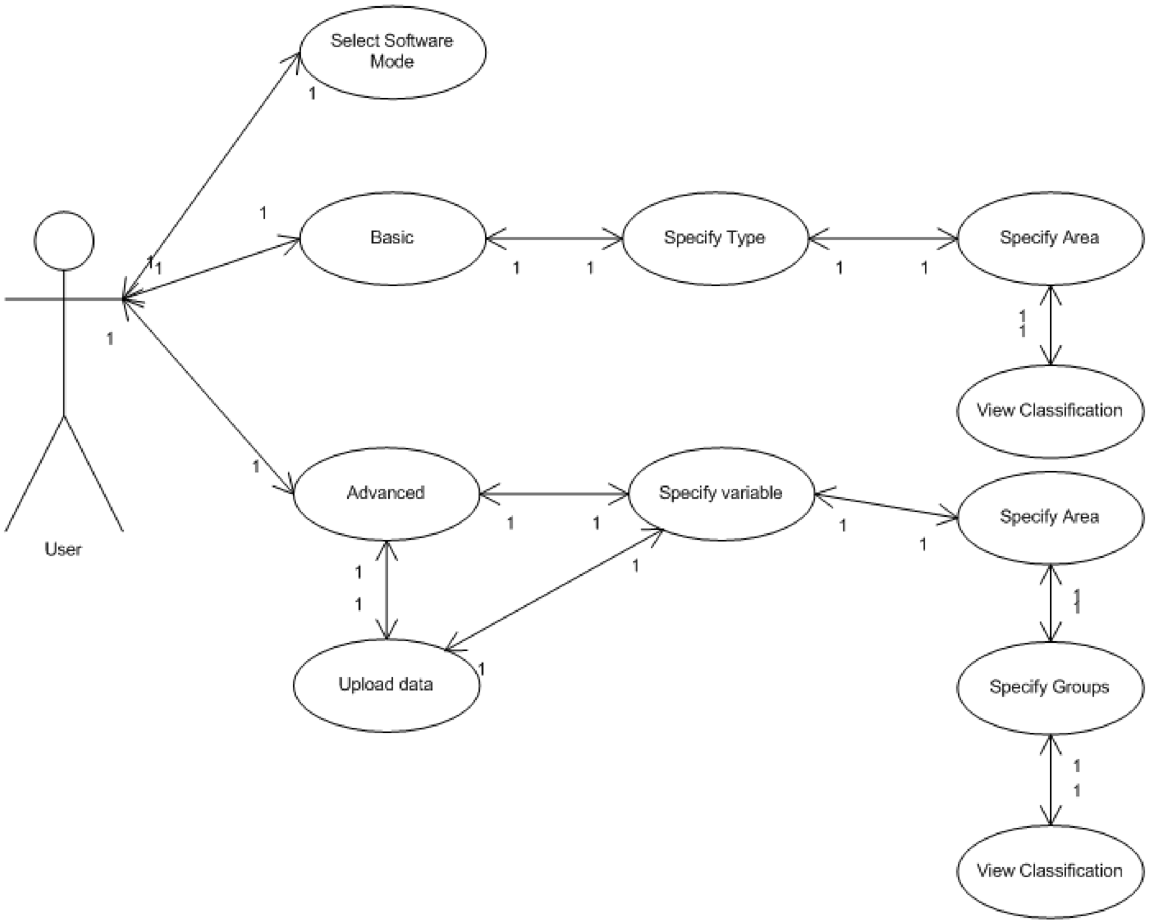


Figure 7.7: Use Case diagram of the 'GeodemBuilder' software

Looking at the use case diagram, we can see that a user can select either the 'Basic' or 'Advanced' mode of the software. Once the 'Basic' mode has been selected, a user specifies the type of geodemographic classification and the geographical area to build the classification for. The software creates the classification, and user visualises the classification. In addition, a user also visualises the geodemographic profiles produced by the software. In the 'Advanced' mode, a user can upload a data file or select variables from the dataset already included in the software or mix and match both. A user can select the variables, specifies the geographical area to build the classification for, and specifies the number of output geodemographics classes. And finally, the user visualises the classification.

7.3.6 OVERALL STRUCTURE OF THE 'GEODEMCREATOR' SOFTWARE

Based on state transition diagrams, class diagram, and use case scenario, we can draw an over structure of the GeodemCreator software. Figure 7.8 shows the overall structure of the software. This figure shows the integration of all the components of the software and demonstrates the functionality of the 'Basic' and 'Advanced' modes.

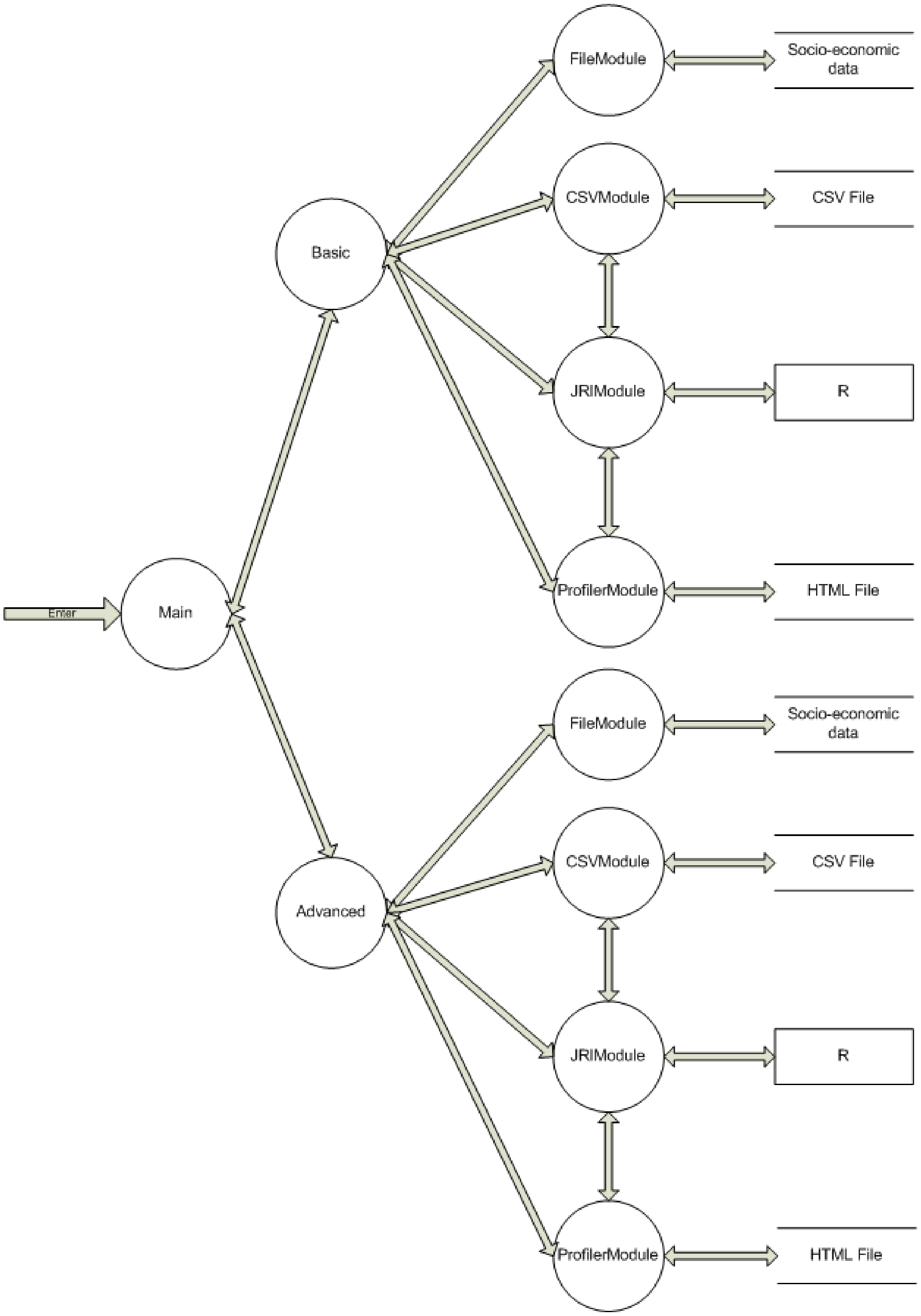


Figure 7.8: Overall structure of the 'GeodemBuilder' software

7.4 CONCLUSION

Previous chapters explained the motivation for creating open bespoke real-time geodemographics classifications, in which the methods and tools to build geodemographic classifications are open to users and general public scrutiny. In pursuing this vision, this chapter has explained the proof of concept for the development of such a software product in the form of a desktop software utility 'GeodemCreator'. This chapter has set out the architecture of the software product in the form of its functional and technical application specifications.

The functional application specification of the software defined the use of two modes of operation i.e. 'Basic' and 'Advanced'. The 'Basic' mode is for the inexperienced users, who do not have much knowledge of selecting variables and their weightings for building a classification. These users can select the geodemographic classification type and the area for which the classification is intended, and the software will automatically select a number of variables according to the selected domain. An 'Advanced' mode is for experienced users, who have greater knowledge of how a geodemographic classification is created. Under the 'Advanced' mode, users can use their own variables in the classification. They can do that by uploading a data file in the software. Users can select between a range of different variables and specify the number of output geodemographic groups (clusters). This gives users control of how a classification is created and helps them to create different bespoke geodemographic classifications.

The second section of this chapter explained the technical application specification in terms of software engineering processes. The intended audience and operating environments were clearly defined in order to define the users who will use the software. Operating system and hardware requirements were clearly defined as well. Because of the use of Java technology for software development, the software is platform independent and can run on any operating system. This was set out in detail in Section 7.4. Software scope defines the overall functionality included in the software. The final part of this section defined the software design. State transition diagrams, Class diagrams, and Use case diagrams were used to show the structure and processes of the software.

The next chapter demonstrates the deployment of the GeodemCreator using case studies by building a number of geodemographic classifications. It also presents some screen shots of the software, as well as a user guide for building a bespoke geodemographic classification.

8 CHAPTER 8: A PILOT GEODEMOGRAPHIC DECISION SUPPORT SYSTEM

8.1 INTRODUCTION

Previous chapters have emphasised the merits of real-time geodemographic classifications, in which the data and methods of building the classifications are open. Moreover, if classifications have to be created in an interactive environment, they must be produced quickly and efficiently. Chapter 7 provided the starting point for the work towards the development of a software utility which acts as a proof of concept that a software utility could be developed for creating interactive, multi source, real-time geodemographic classifications. Chapter 7 described the functional and technical application specification of 'GeodemCreator'. This free software utility allows users to build classifications in 'Basic' or 'Advanced' mode. The preceding chapter also discussed the design of the software in detail, including a detailed description of possible user interactions.

This chapter describes the use of 'GeodemCreator' as a general purpose geodemographic decision support tool. The software can be used for building national or region-specific bespoke classifications. The first section of this chapter describes how, in practice, a geodemographic classification may be created using 'GeodemCreator'. This section describes the procedure using a sequence of screen shots of the software and textual descriptions – and the reader can follow these procedures using the CD software that accompanies the hard copy of this thesis. The second part of the chapter is dedicated to three different case studies that describe the creation of three geodemographic classifications:

- An output area classification of Greater London: This case study describes the creation of an output area classification of Greater London by using 'GeodemCreator', and proves the concept of its use to create a bespoke local area classification.

- An output area classification of the United Kingdom: This case study describes the creation of an output area classification of the entire United Kingdom by using 'GeodemCreator'. This case study is the re-creation of the Output Area Classification produced by Vickers & Rees (2007). This case study gives a proof of the concept of building national level geodemographic classifications with 'GeodemCreator'.

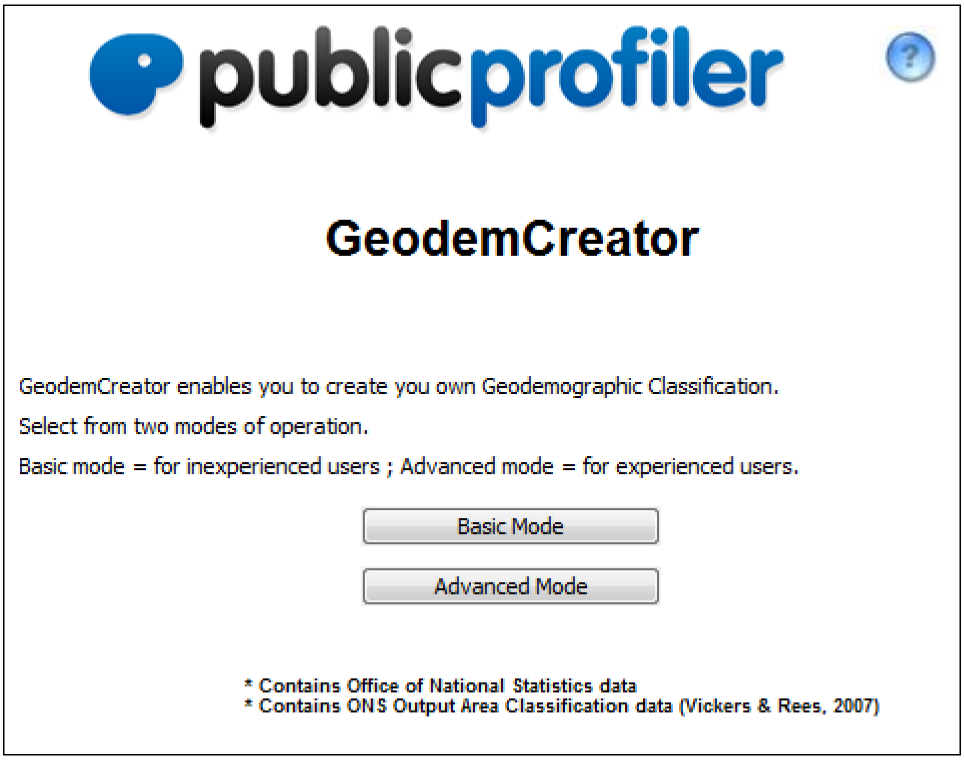
- An output area level ‘socio-economic and ethnic’ classification of Greater London: This case study describes the creation of an output area level ‘socio-economic and ethnic’ classification of Greater London by using ‘GeaodemCreatoer’. The case study gives a proof of the concept of building local area classifications using census data and other ancillary data sources. Data from ‘Worldnames’ database created in chapter no. 3 was used for the purpose of creating this classification.

8.2 GEODEMCREATOR

The functional and the technical application specification of 'GeodemCreator' were discussed in detail in the Chapter 7. This Section is dedicated to showing the working version of 'GeodemCreator' under different scenarios.

The software requires Java and R installed on the machine. A user needs to install 'rjava' library in 'R'. A complete step by step instructions on the installation of 'GeodemCreator' are given in the Appendix 7 and are also included in the CD (with the file name How_To_Install.txt) which accompanies this thesis.

Figure 8.1 shows a screen shot of 'GeodemCreator' when it is first opened. A user can either select the 'Basic' or 'Advanced' mode of operations. These modes were explained in the previous chapter.



GeodemCreator enables you to create you own Geodemographic Classification.

Select from two modes of operation.

Basic mode = for inexperienced users ; Advanced mode = for experienced users.

* Contains Office of National Statistics data
* Contains ONS Output Area Classification data (Vickers & Rees, 2007)

Figure 8.1: Using 'GeodemCreator' for building a classification

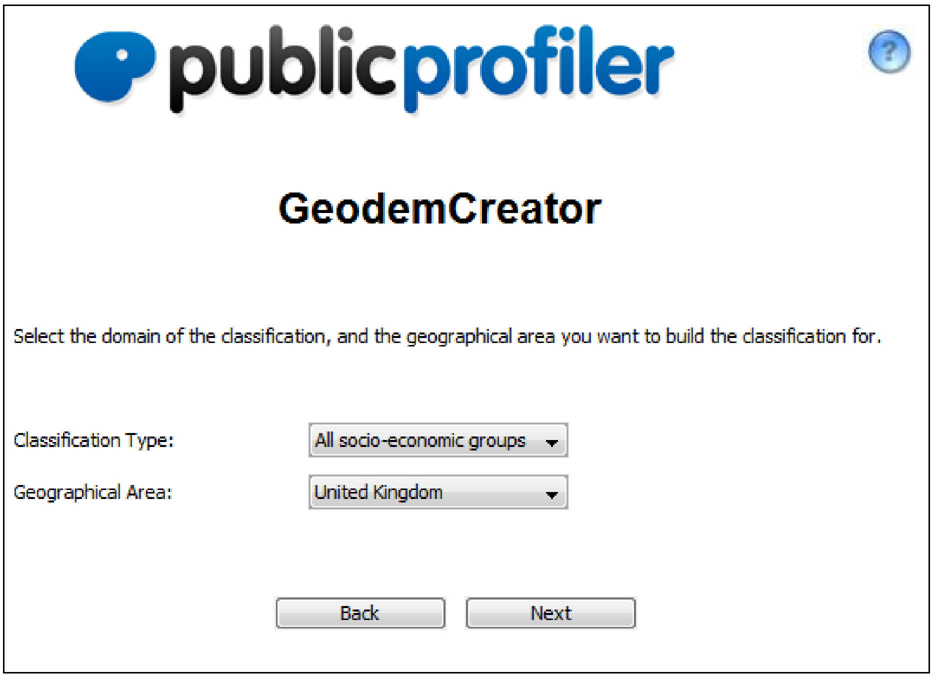
This section shows the screen shots of 'GeodemCreator' in three different scenarios.

- Building a classification in the 'Basic' mode
- Building a classification in the 'Advanced' mode with system data
- Building a classification in the 'Advanced' mode with user's data

8.2.1 BUILDING A CLASSIFCIATION IN THE 'BASIC' MODE

The 'basic' mode is intended for inexperienced users who do not have much knowledge of building geodemographic classifications. This mode offers a restricted functionality to users i.e. they do not have the choice to select variables or to define the number of geodemographic groups. In this mode, the software automatically picks up the variables and the optimum number of cluster solutions.

The following figure 8.2 shows the 'Basic' mode of 'GeodemCreator'. A user needs to select a classification type and a geographical area to build a classification in this mode. The classification types and geographical areas are explained in detail in the previous chapter (Chapter no. 7).



GeodemCreator

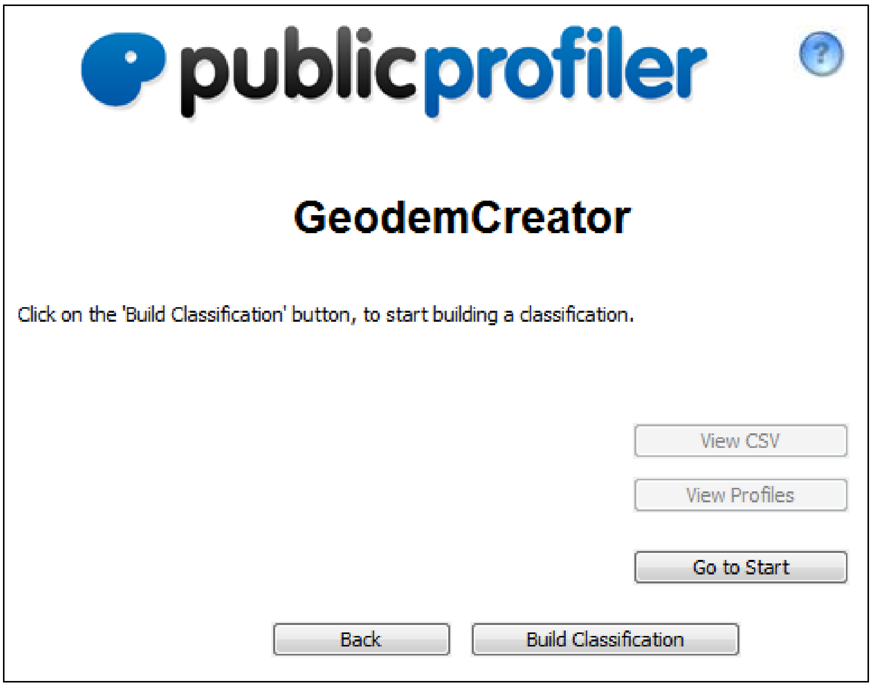
Select the domain of the classification, and the geographical area you want to build the classification for.

Classification Type: All socio-economic groups

Geographical Area: United Kingdom

Figure 8.2: Building a classification in 'Basic' mode

Once the user has selected the classification type and the geographical area to build the classification, the software shows the screen shown in Figure 8.3. A user needs to click the 'Build Classification' button to build the geodemographic classification.

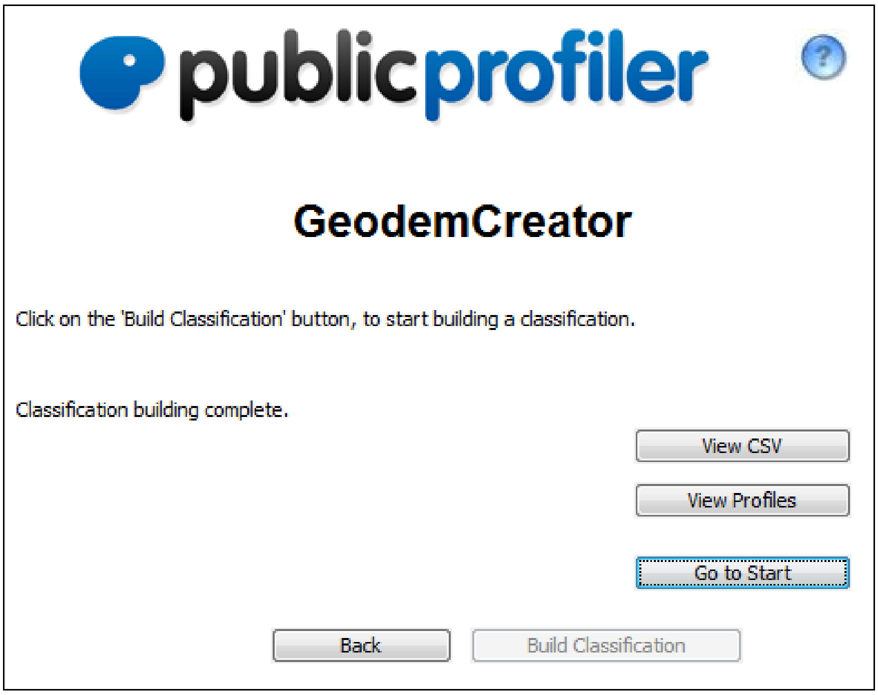


GeodemCreator

Click on the 'Build Classification' button, to start building a classification.

Figure 8.3: Building a classification in 'Basic' mode

Based on user inputs, the software builds the classification and the screen in Figure 8.4 is shown to the user once the classification building process is finished.



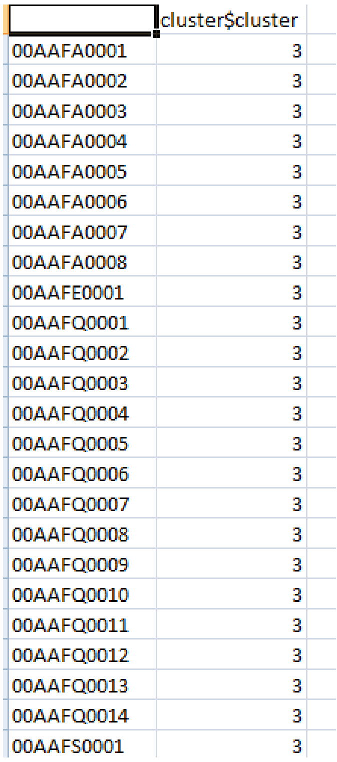
GeodemCreator

Click on the 'Build Classification' button, to start building a classification.

Classification building complete.

Figure 8.4: Building a classification in 'Basic' mode

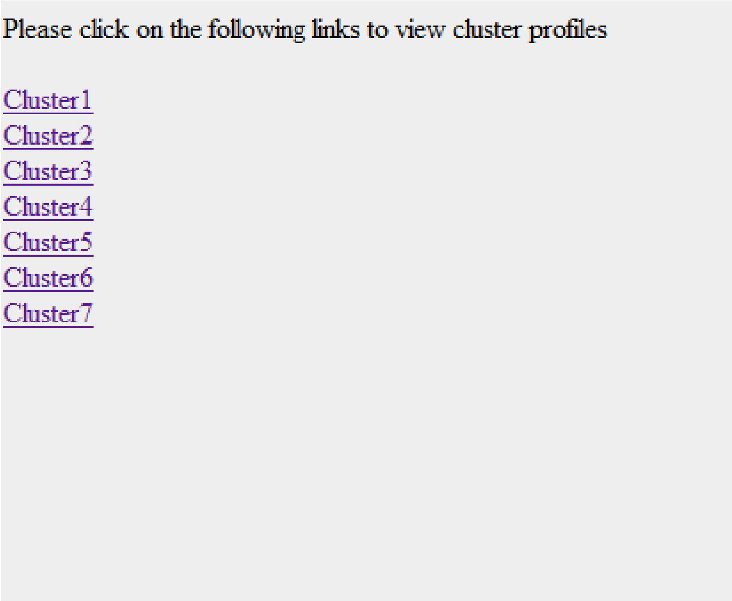
The software creates a number of results as the final geodemographic classification. A user can view the CSV file containing the final geodemographic classification. The CSV file contains a cluster number assigned to each output area. Figure 8.5 shows one of the output CSV files where a unique cluster number is assigned to a particular output area. This cluster number is used to colour the output area for visualisation purposes. Making a visualisation from the output CSV is described in the second part of this chapter (Case Studies).



	cluster$cluster
00AAFA0001	3
00AAFA0002	3
00AAFA0003	3
00AAFA0004	3
00AAFA0005	3
00AAFA0006	3
00AAFA0007	3
00AAFA0008	3
00AAFE0001	3
00AAFQ0001	3
00AAFQ0002	3
00AAFQ0003	3
00AAFQ0004	3
00AAFQ0005	3
00AAFQ0006	3
00AAFQ0007	3
00AAFQ0008	3
00AAFQ0009	3
00AAFQ0010	3
00AAFQ0011	3
00AAFQ0012	3
00AAFQ0013	3
00AAFQ0014	3
00AAFS0001	3

Figure 8.5: Output CSV file

Vickers & Rees (2007) and Petersen et al. (2010) created radar charts that are associated with each of the individual clusters. These charts plot the cluster centres for each individual variable used in the classification, in the form of a radial plot. These radar charts are of assistance when naming the individual clusters. When a user clicks the 'View Profiles' button on the output screen (Figure 8.6), an HTML files opens with a number of links for individual cluster profiles in the form of radar charts.



Please click on the following links to view cluster profiles

Cluster1
Cluster2
Cluster3
Cluster4
Cluster5
Cluster6
Cluster7

Figure 8.6: Output Cluster profiles

By clicking on the individual links, users can view the individual radar charts of a particular cluster solution. Figure 8.7 shows the radar chart of a cluster solution with the names of the individual variables.

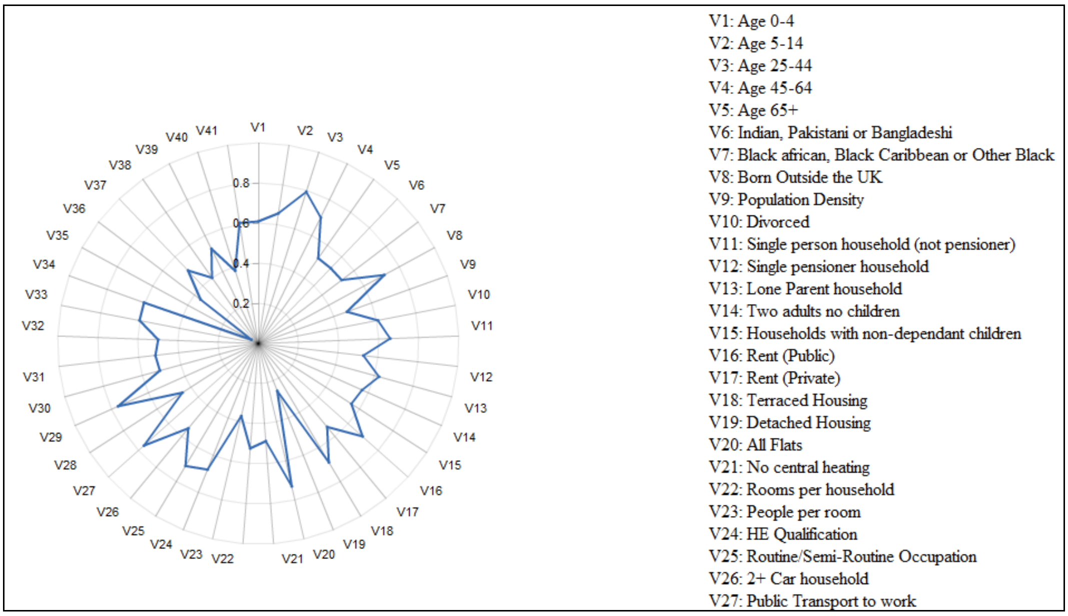
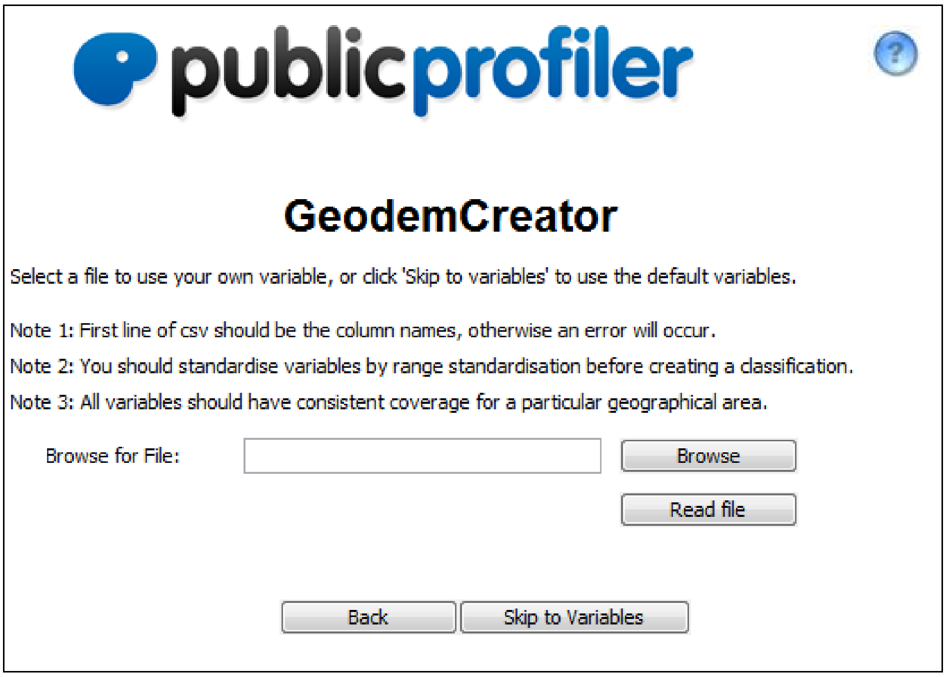


Figure 8.7: A radar chart of running the software 'All Socio-economic groups' domain

8.2.2 BUILDING A CLASSIFCIATION IN THE 'ADVANCED' MODE WITH SYSTEM DATA

As described in the previous chapter, the 'Advanced' mode is intended for experienced users who have prior knowledge of building geodemographic classifications. Users need to specify the variables and the number of cluster solutions to build a classification in this mode.

Figure 8.8 shows the 'Advanced' mode of 'GeodemCreator'. A user can use an input file to build a classification or just press the 'Skip to Variables' button to use the variables provided by the software by default (as explained in the previous chapter). Users can also see considerable information about the standardisation of the variables (if users want to use their own data) and constancy of the data over a geographical area.



GeodemCreator

Select a file to use your own variable, or click 'Skip to variables' to use the default variables.

Note 1: First line of csv should be the column names, otherwise an error will occur.

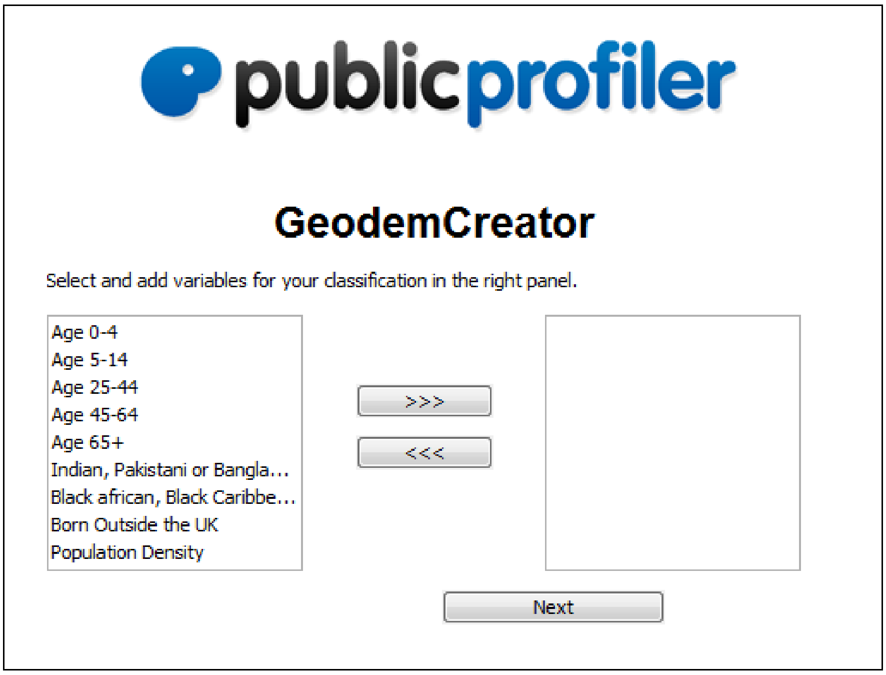
Note 2: You should standardise variables by range standardisation before creating a classification.

Note 3: All variables should have consistent coverage for a particular geographical area.

Browse for File:

Figure 8.8: Building a classification in 'Advanced' mode

Once the user has selected 'Skip to Variables', the following screen (Figure 8.9) gives the option of selecting specific variables for the classification. The user can scroll down through the list of available variables or read in additional ones. Users can simply move the variables from the left panel to the right panel. The variables moved to the right panel will be used in building the geodemographic classification.



GeodemCreator

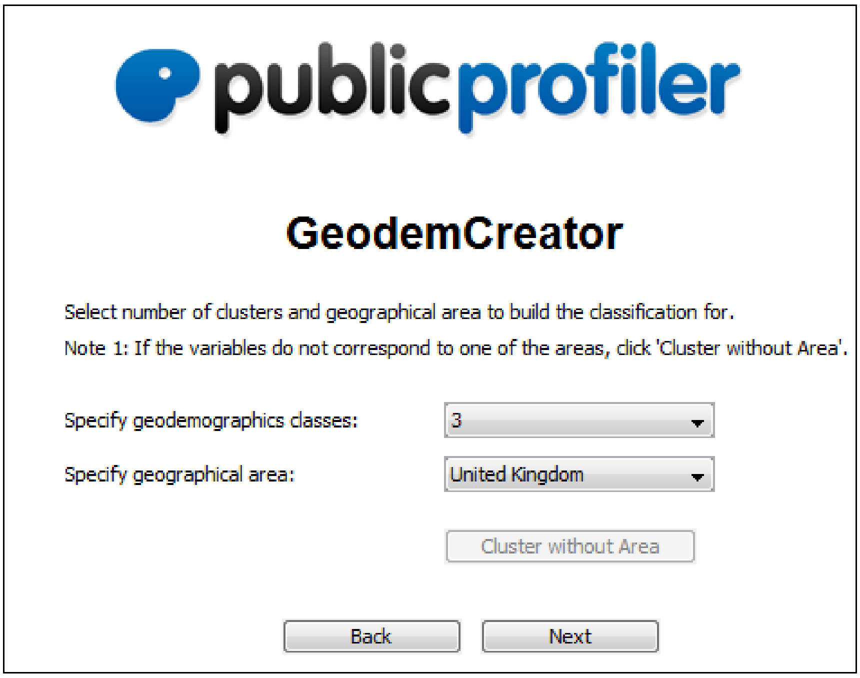
Select and add variables for your classification in the right panel.

Age 0-4
Age 5-14
Age 25-44
Age 45-64
Age 65+
Indian, Pakistani or Bangla...
Black african, Black Caribbe...
Born Outside the UK
Population Density

Figure 8.9: Building a classification in 'Advanced' mode (Variable Selection)

Once the variables have been specified, the user needs to specify the number of cluster solutions and the geographical area they want to produce the classification for. This selection procedure is shown in Figure 8.10. The geographical areas for the 'Advanced' mode are as explained in the previous chapter.

Thus this enables users to either build a national level classification for the United Kingdom or a local level classification based upon one of twelve constituent regions in the country. The 'cluster without Area' button is disabled if users are not uploading their data to the software. This button should be used if users are using their own data and do not want to join it with any of the default spatial levels of the software. This functionality enables the users to build classifications at other spatial levels than output area.



GeodemCreator

Select number of clusters and geographical area to build the classification for.

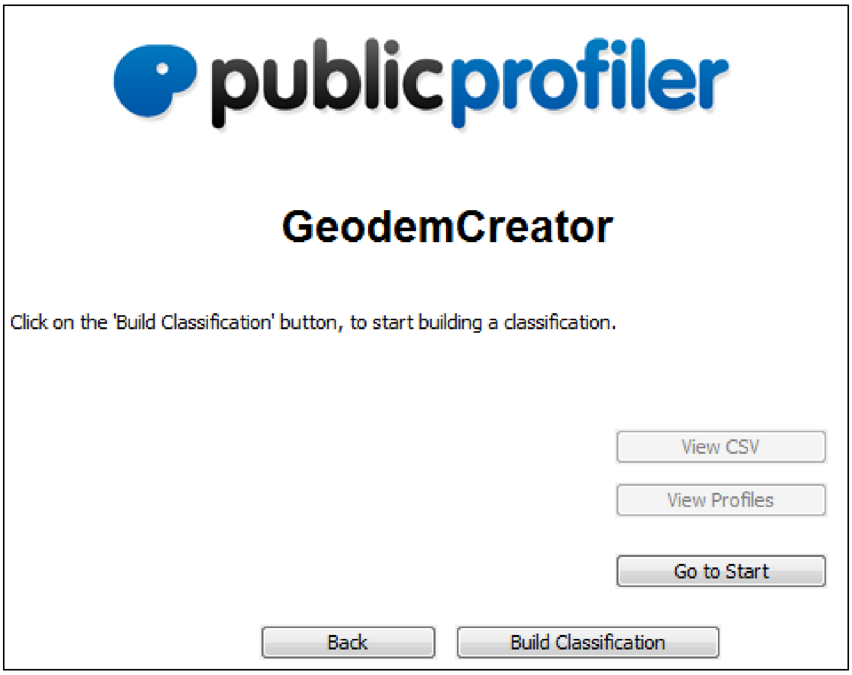
Note 1: If the variables do not correspond to one of the areas, click 'Cluster without Area'.

Specify geodemographics classes: 3

Specify geographical area: United Kingdom

Figure 8.10: Building a classification in 'Advanced' mode (Cluster number specification)

Once users have specified everything, the software shows the next screen (Figure 8.11). Users need to press the 'Build Classification' button to start building their geodemographic classification. Finally, users can press the 'View CSV' and 'View Profiles' buttons to view the outputs.



GeodemCreator

Click on the 'Build Classification' button, to start building a classification.

Figure 8.11: Building a classification in 'Advanced' mode (Results)

Once the software finishes building the classification, a user can view the output CSV file and cluster profiles in the form of radar charts. Figure 8.12 shows the radar chart of one of the cluster solutions. Of course, only the variables selected by the user in building the geodemographic classification can be viewed.

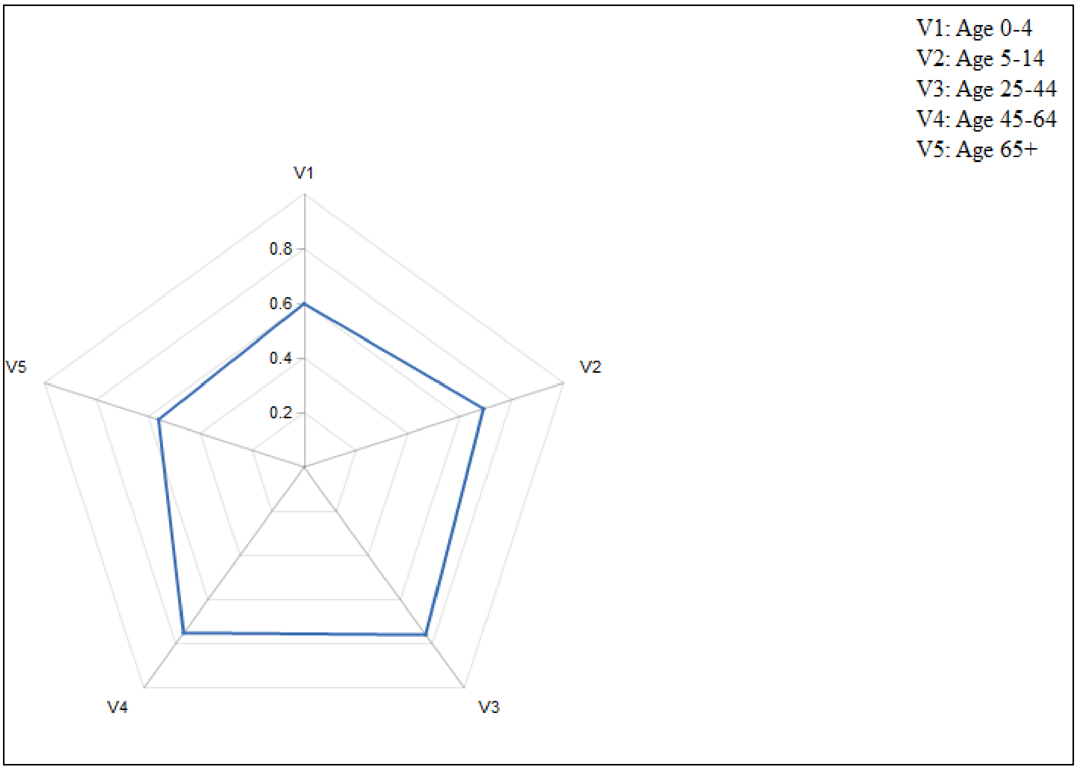
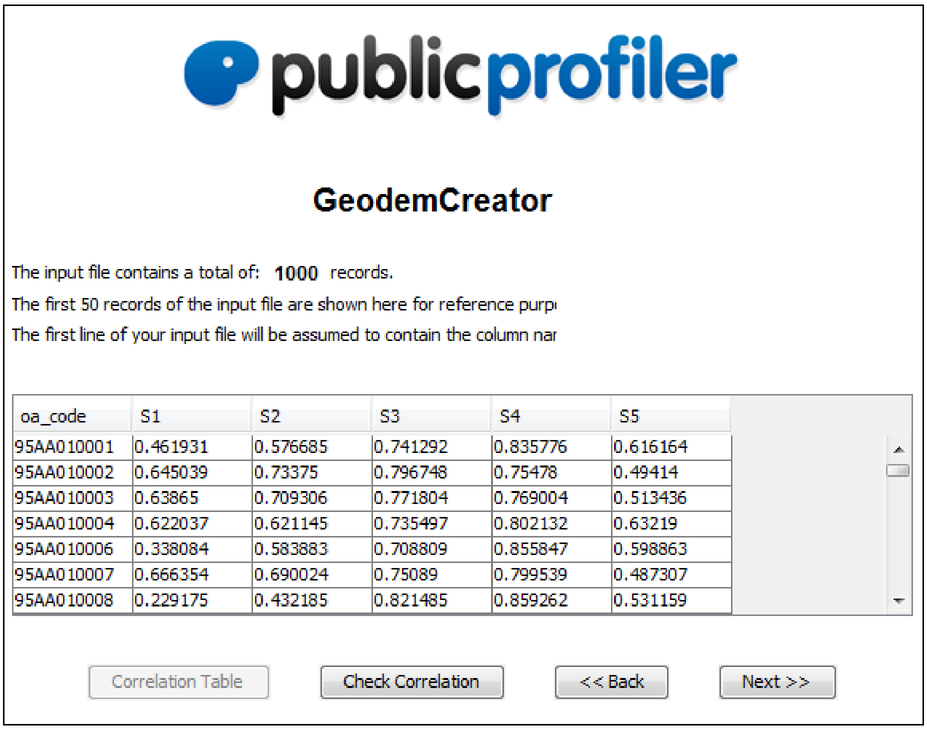


Figure 8.12: Building a classification in 'Advanced' mode (output radial chart)

8.2.3 BUILDING A CLASSIFICATION IN THE 'ADVANCED' MODE WITH ANCILLARY DATA

Section 8.2.2 explains how the system can be used to build a classification in 'Advanced' mode. However, this classification uses only the 2001 OAC Census data provided by default in the software. A user may want to use their own variables to supplement or even wholly replace these variables when building a classification. This can be done by uploading a file of variables to the software.

Figure 8.13 shows the screen after the user has uploaded the data to the software. This screen provides information for the data that have been uploaded, and is useful for having a quick summary of the data file used. User can use the 'Check Correlation' button to test the correlation of the variables.



GeodemCreator

The input file contains a total of: **1000** records.

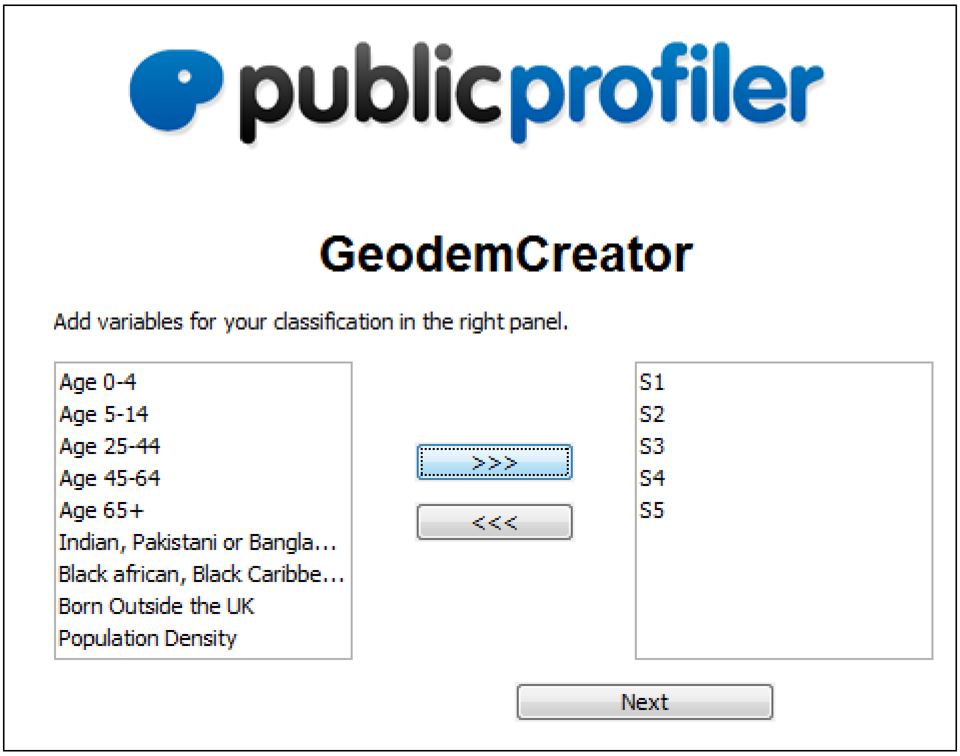
The first 50 records of the input file are shown here for reference purp

The first line of your input file will be assumed to contain the column nar

oa_code	S1	S2	S3	S4	S5
95AA010001	0.461931	0.576685	0.741292	0.835776	0.616164
95AA010002	0.645039	0.73375	0.796748	0.75478	0.49414
95AA010003	0.63865	0.709306	0.771804	0.769004	0.513436
95AA010004	0.622037	0.621145	0.735497	0.802132	0.63219
95AA010006	0.338084	0.583883	0.708809	0.855847	0.598863
95AA010007	0.666354	0.690024	0.75089	0.799539	0.487307
95AA010008	0.229175	0.432185	0.821485	0.859262	0.531159

Figure 8.13: Uploading user’s own data to build a classification

Once users have uploaded their data files to the software, the variable selection screen will contain the user's variables and the census variables provided by default in the software. The user can then build the geodemographic classification by mixing and matching the variables. This is shown in Figure 8.14.



GeodemCreator

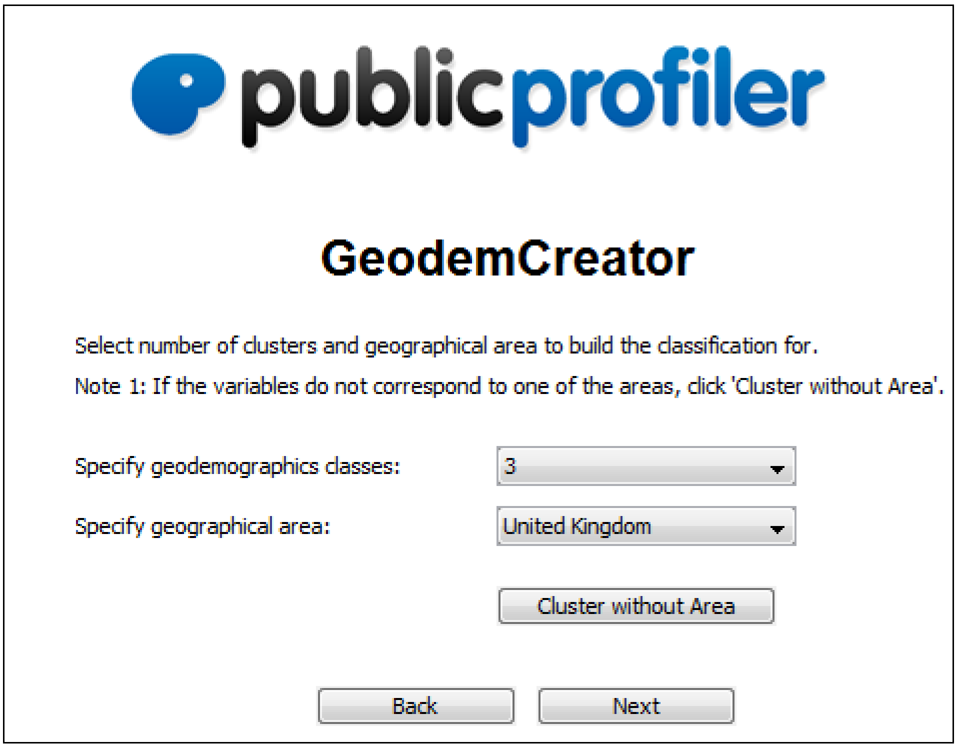
Add variables for your classification in the right panel.

Age 0-4
Age 5-14
Age 25-44
Age 45-64
Age 65+
Indian, Pakistani or Bangla…
Black african, Black Caribbe…
Born Outside the UK
Population Density

S1
S2
S3
S4
S5

Figure 8.14: Using user's own data to build a classification

If the uploaded data file corresponds to one of the default geographical areas (these areas are broken down by output areas), the user can select the geographical area from the drop down menu as shown in Figure 8.15. Alternatively, the user can press the button 'Cluster without Area', and the system will produce the classification independent of the default geographical areas, thus enabling users to create zone-free classifications.



GeodemCreator

Select number of clusters and geographical area to build the classification for.

Note 1: If the variables do not correspond to one of the areas, click 'Cluster without Area'.

Specify geodemographics classes: 3

Specify geographical area: United Kingdom

Figure 8.15: Using user supplied data to build a classification (Cluster number specification)

This functionality provides considerable flexibility to users in building a geodemographic classification since it allows them to build a classification independent of the default geographical areas provided in the software. This makes it possible to build local area classifications at different spatial levels (i.e. Output Areas, Lower Super Output Areas, and Wards etc). The software clusters the data for the areas for which the ancillary data values are provided, and enables users to create zone-free classifications independent of any geographical area aggregations.

8.3 CASE STUDIES

This section describes three case studies of building three different geodemographic classifications using 'GeodemCreator' software. These classifications are:

- An Output Area classification of Greater London
- An Output Area Classification of the United Kingdom
- An Output Area level 'Socio-economic and Ethnic' classification of Greater London

This will provide proof of concept of building local area and national classifications by using 'GeodemCreator' software. The next sub-section 8.3.1 describes the variables used for creating these classifications and the next sub-sections 8.3.2, 8.3.3, and 8.3.4 contain a description and the outputs of the classifications.

8.3.1 VARIABLES USED FOR BUILDING THE TWO CLASSIFICATIONS

The first two classifications (output area classification of Greater London, output area classification of the United Kingdom) are built from the 41 census variables provided in the 'GeodemCreator' as default variables. These variables are outlined in the previous chapter (section 7.2.1). The variables are categorised in different domains i.e. demographic, household composition, housing, employment, and socio-economic characteristics.

The third classification (an Output Area level 'Socio-economic and Ethnic' classification of Greater London) was built by using 53 variables. These include 41 Census and 12 Ethnicity Data variables. The 41 census variables and their domains are outlined in the previous chapter (section 7.2.1).

The 'Worldnames' database created in Chapter 3 was used for the purpose of creating the 12 ethnicity variables. Counts of the 'Forename' and 'Surname' pairs were extracted for each output area of the Greater London. Later, Onomap (Mateos, 2010) was used to code ‘forename’ and ‘surname’ pairs to their corresponding 12 ethnic groups. This gave the counts of each ethnic group in every output area of the Greater London. These 12 ethnic groups were converted to variables and are outlined in the following table (Table 8.1).

Geographical Areas
V42: ‘European’ ethnic group
V43: ‘East Asian & Pacific’ ethnic group
V44: ‘Muslim’ ethnic group
V45: ‘Greek’ ethnic group
V46: ‘English’ ethnic group
V47: ‘Nordic’ ethnic group
V48: ‘African’ ethnic group
V49: ‘Japanese’ ethnic group
V50: ‘Hispanic’ ethnic group
V51: ‘Celtic’ ethnic group
V52: ‘Jewish’ ethnic group
V53: ‘South Asian’ ethnic group

Table 8.1: Ethnicity variables for creating the geodemographic classification

8.3.2 AN OUTPUT AREA CLASSIFICATION OF GREATER LONDON

This case study describes the creation of an output area classification of Greater London using 'GeodemCreator'. This classification can be created by running the software in 'Basic' mode and creating a classification by selecting 'All socio-economic groups' and 'London' from the domain and geographical area options. This classification specifies how a regional variant of a well known geodemographic classification can be created using 'GeodemCreator'.

VARIABLES USED TO PRODUCE THE CLASSIFICATION

The variables used for this classification are all the 41 variables outlined in the section 7.2.1 and originally used by Vickers and Rees (2007). This means that this classification will cover all the domains described in the previous section. 'GeodemCreator' automatically chooses the 41 variables when a user selects 'All scoio-economic groups' and 'London' from the domain and geographical area options.

IDENTIFYING THE NUMBER OF CLUSTERS

Vickers & Rees (2007), citing the reduced average distance to the cluster centre when the number of clusters was incremented, picked seven clusters when building the OAC geodemographic classification for United Kingdom. Bearing this point in mind, Figure 8.16 plots the change in 'within sum of squares' by incrementing 'number of clusters', obtained by running *k*-means on the Output Area dataset for London.

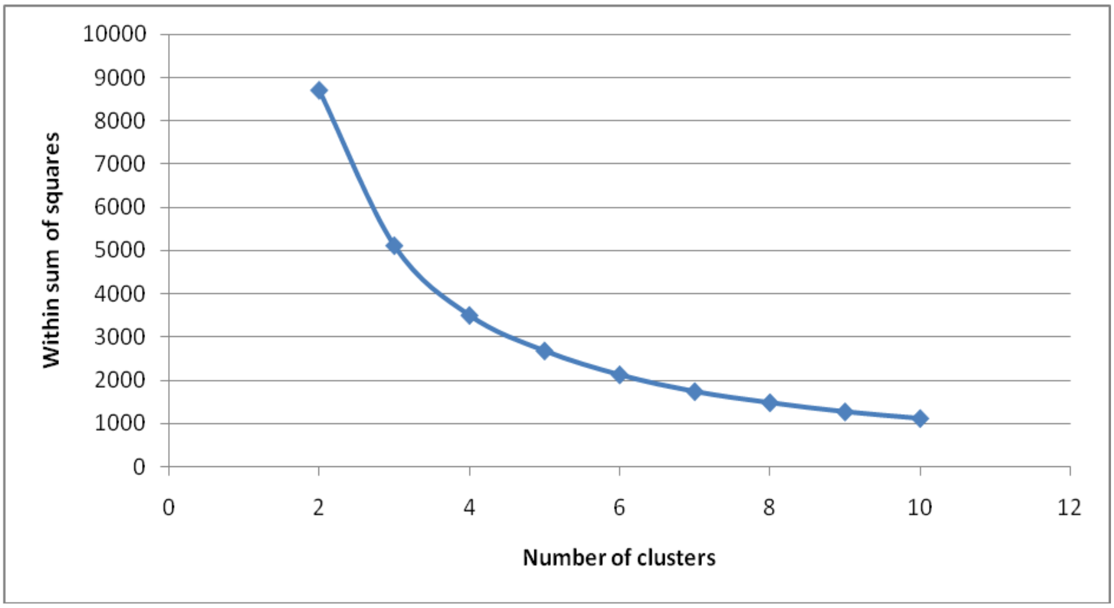


Figure 8.16: Within sum of squares by number of clusters (Case study no. 1)

K=7 could be taken as a desirable number of clusters. Reductions in the 'within sum of squares' figure are not great from k=7 to k=10. 'GeodemCreator' defaults to seven clusters when the user chooses 'All socio-economic groups' and 'London' from the domain and geographical area options.

VISUALISATION OF THE OUTPUT AREA CLASSIFICATION

'GeodemCreator' uses k=7 for building a classification when a user is in 'Basic' mode and selects 'All socio-economic groups' and 'London' from the domain and geographical area options.

These options and default were used and 'GeodemCreator' produced the desired geodemographic classification. The resultant CSV file was joined with the Greater London Output Area shape file, resulting in the map shown in Figure 8.17. R was used to produce the map and the generalised source code for this is provided in Appendix 6 section.

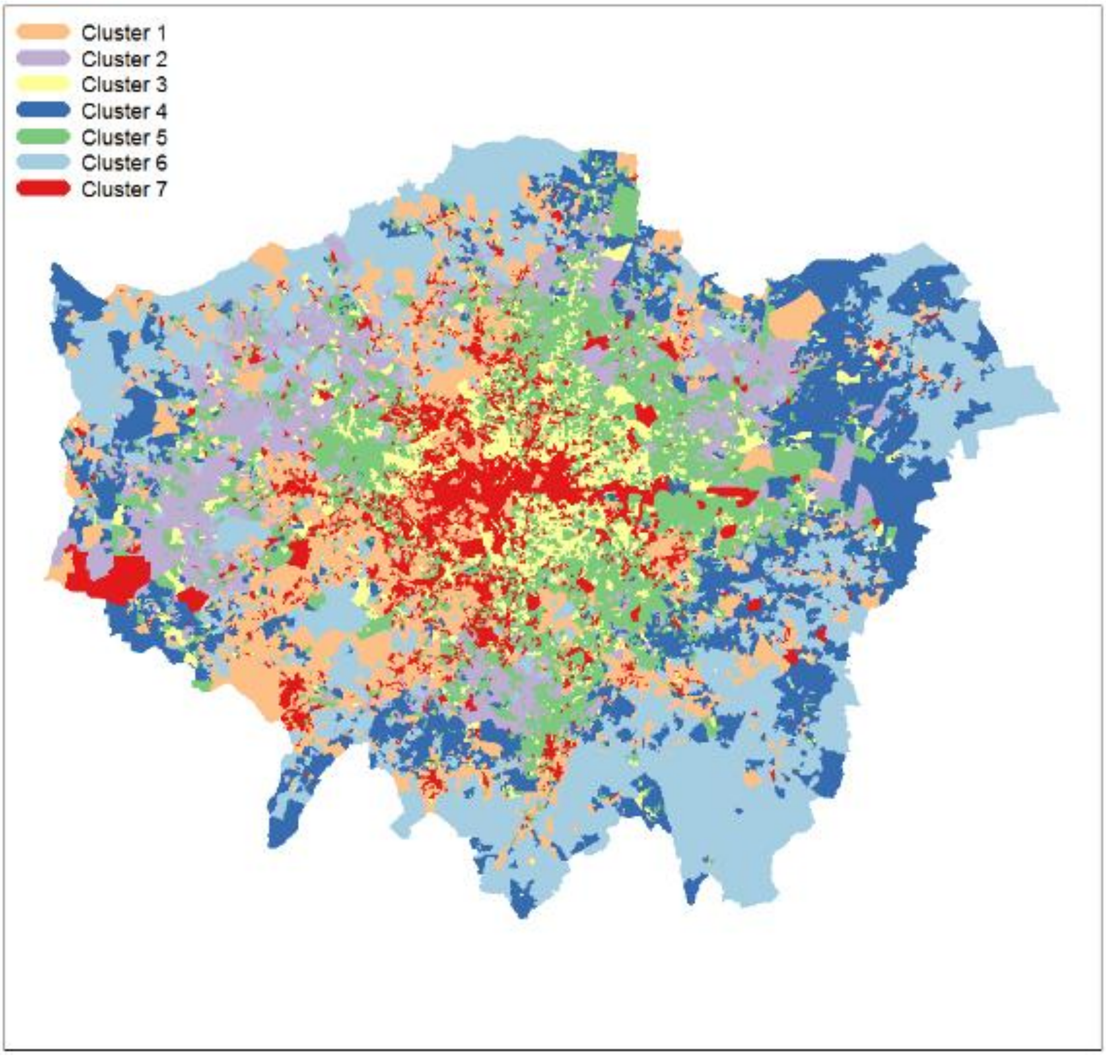


Figure 8.17: Output Area Classification of London (Case study no. 1)

'GeodemCreator' also plots radar charts for individual clusters. These radar charts are really important in defining and naming individual clusters, thus making it possible to support decision making using the clusters produced. The description of each of the clusters is given below, based on their radar charts. Petersen et al. (2010) was consulted to name the individual clusters.

CLUSTER NO. 1

Figures 8.18 radar chart of cluster no. 1. All the variables v1-v41 are as described in Section 7.2.1.

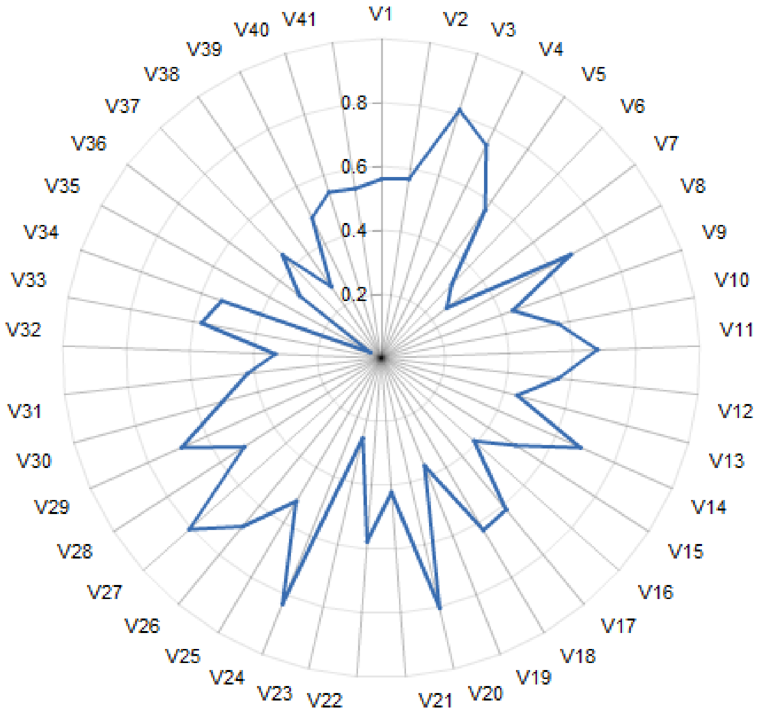


Figure 8.18: Radar chart of cluster no. 1 (Case study no. 1)

The variables having high values in this cluster are:

- V3: Age 25-44

- V4: Age 45-64
- V8: Born outside UK
- V11: Single person household (not pensioner)
- V14: Two adults no children
- V20: All Flats
- V23: People per room
- V27: Public transport to work

Based on the high values of the variables, this cluster could be named '**Areas of City Commuters**'.

CLUSTER NO. 2

Figures 8.19 radar chart of cluster no. 2. All the variables v1-v41 are as described in Section 7.2.1.

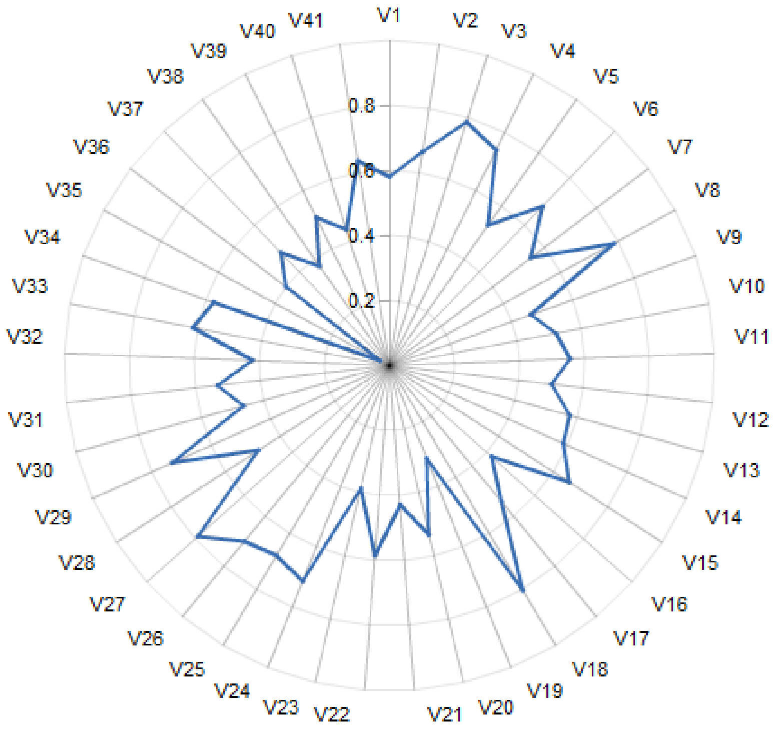


Figure 8.19: Radar chart of cluster no. 2 (Case study no. 1)

The variables having high values in this cluster are:

- V3: Age 25-44
- V4: Age 45-64
- V6: Indian, Pakistani or Bangladeshi
- V8: Born outside UK
- V18: Terraced Housing
- V23: People per rooms
- V27: Public transport to work
- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**Areas of Asian Population**'.

CLUSTER NO. 3

Figures 8.20 radar chart of cluster no. 3. All the variables v1-v41 are as described in Section 7.2.1.

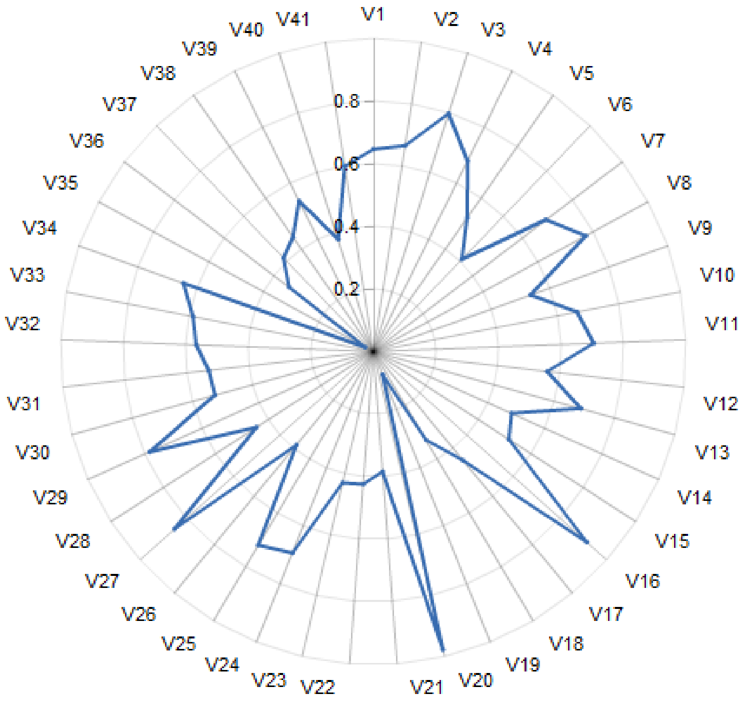


Figure 8.20: Radar chart of cluster no. 3 (Case study no. 1)

The variables having high values in this cluster are:

- V3: Age 25-44
- V8: Born outside UK
- V11: Single Person household (not pensioner)
- V13: Lone Parent household
- V16: Rent (Public)
- V20: All Flats
- V27: Public Transport to work
- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**Areas of Council Flats**'.

CLUSTER NO. 4

Figures 8.21 radar chart of cluster no. 4. All the variables v1-v41 are as described in Section 7.2.1.

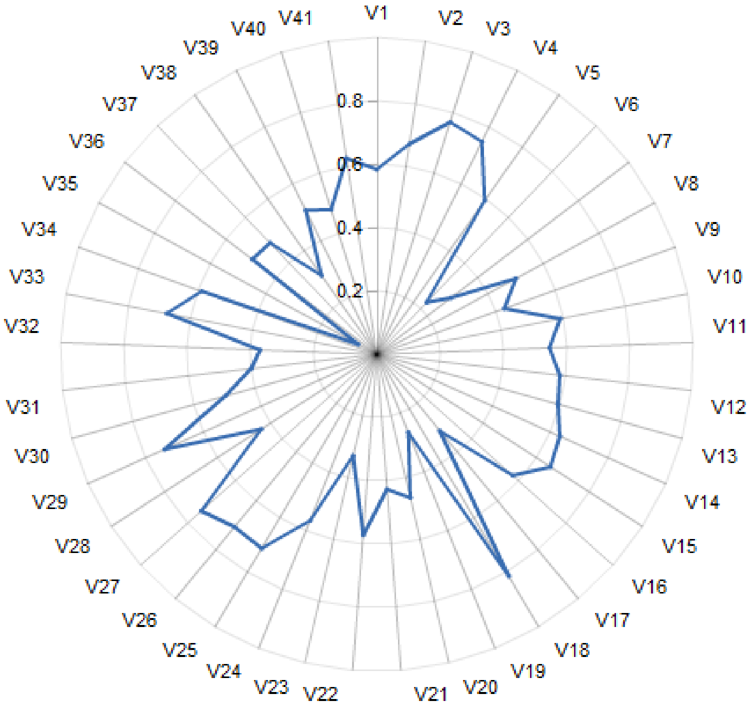


Figure 8.21: Radar chart of cluster no. 4 (Case study no. 1)

The variables having high values in this cluster are:

- V3: Age 25-44
- V4: Age 45-64
- V18: Terraced Housing
- V25: Routine/Semi-Routine Occupation
- V27: Public Transport to work
- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**Areas of Blue Collar Communities**'.

CLUSTER NO. 5

Figures 8.22 radar chart of cluster no. 5. All the variables v1-v41 are as described in Section 7.2.1.

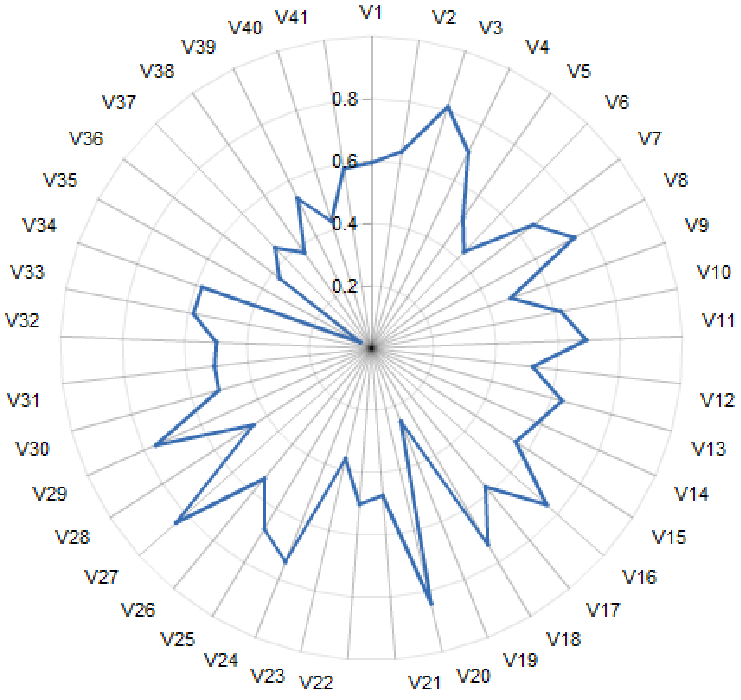


Figure 8.22: Radar chart of cluster no. 5 (Case study no. 1)

The variables having high values in this cluster are:

- V3: Age 25-44
- V8: Born outside UK
- V11: Single pensioner household
- V16: Rent (Public)
- V20: All Flats
- V24: HE Qualification
- V27: Public Transport to work
- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**Areas of Terraced Housing**'.

CLUSTER NO. 6

Figures 8.23 radar chart of cluster no. 6. All the variables v1-v41 are as described in Section 7.2.1.

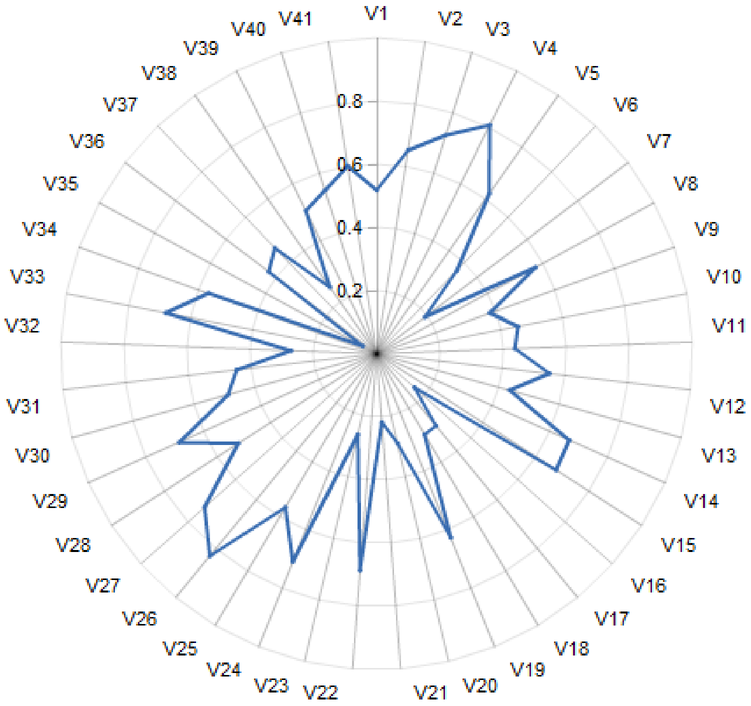


Figure 8.23: Radar chat of cluster no. 6 (Case study no. 1)

The variables having high values in this cluster are:

- V4: Age 45-64
- V14: Two adults no children
- V15: Households with non-dependent children
- V22: Rooms per household
- V26: 2+ Car household

- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**London Suburbs**'.

CLUSTER NO. 7

Figures 8.24 radar chart of cluster no. 7. All the variables v1-v41 are as described in Section 7.2.1.

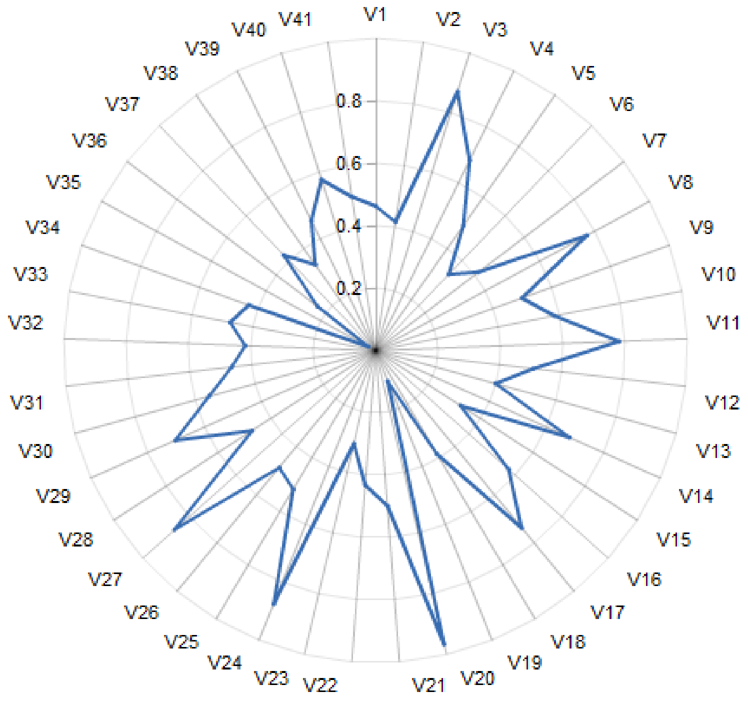


Figure 8.24: Radar chart of cluster no. 7 (Case study no. 1)

The variables having high values in this cluster are:

- V3: Age 25-44
- V8: Born Outside the UK
- V11: Single person household (not pensioner)

- V17: Rent (Private)
- V20: All Flats
- V24: HE Qualification
- V27: Public Transport to work

Based on the high values of the variables, this cluster could be named 'Central **Areas**'.

END POINT OF THE OUTPUT AREA CLASSIFICATION OF LONDON

This section has explained the use of 'GeodemCreator' for producing an output area classification for London. This is the example of a local area geodemographic classification that can be produced with the data and methods available. These local area classifications can be useful in describing interesting patterns of the population instead of just deriving inferences from a national level classification.

8.3.3 AN OUTPUT AREA CLASSIFICATION OF THE UNITED KINGDOM

This case study is a re-creation of the output area classification created by Vickers & Rees (2007). This acts as a proof of concept that national level classifications can be created by using 'GeodemCreator' software. 'GeodemCreator' builds the classification very quickly (within minutes), and makes it possible to produce a classification without the need to know the variables selected, the weighting procedures used, and the clustering algorithm applied.

The national level output area classification of the United Kingdom can be created by running the software in 'Basic' mode and creating the classification by selecting 'All socio-economic groups' and 'United Kingdom' from the domain and geographical area options.

VARIABLES USED TO PRODUCE THE CLASSIFICATION

The variables used for this classification are all of the 41 variables described in the section 7.2.1: however, in this section, output area values used for the United Kingdom in the cluster analysis. In this case study the classification will cover all the domains for all the Output Areas throughout the United Kingdom. By default, 'GeodemCreator' automatically chooses the 41 variables when a user selects 'All scoio-economic groups' and 'United Kingdom' from the domain and geographical area options.

IDENTIFYING THE NUMBER OF CLUSTERS

Section 8.3.2 discussed how to choose an optimal number of clusters. Using the same technique, Figure 8.25 shows the plot of the 'within sum of squares' by 'number of clusters' by running *k*-means on the Output Area dataset for the entire United Kingdom.

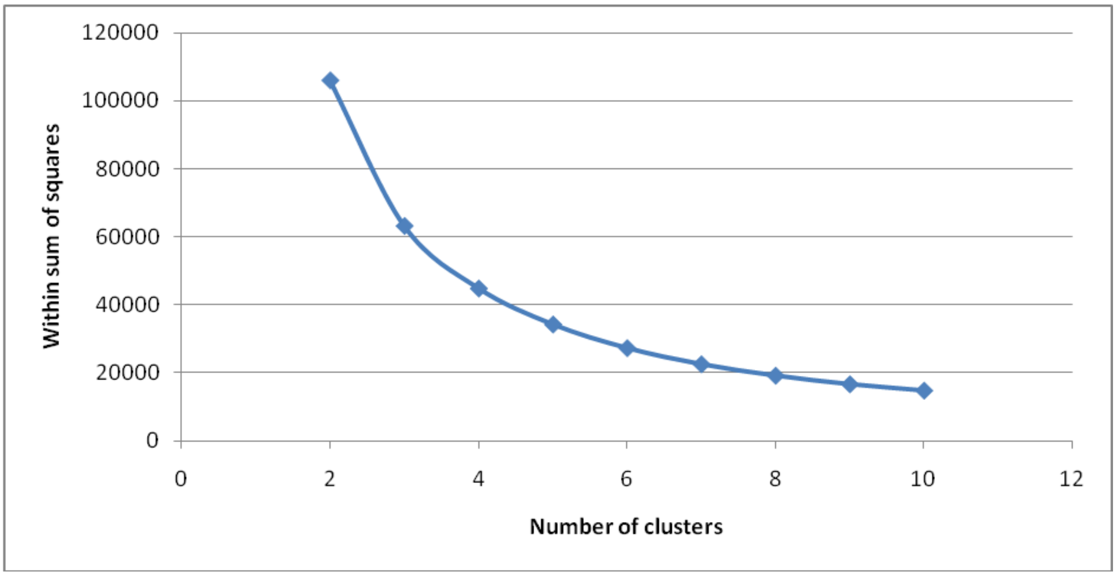


Figure 8.25: Within sum of squares by number of clusters (Case study no. 2)

k=7 can be used as the optimal solution because the within sum of squares remains almost constant after k=7 cluster solutions.

'GeodemCreator' automatically uses seven as the desired number of clusters when a user chooses 'All Socio-Economic Groups' and 'United Kingdom' from the domain and the geographical area options.

VISUALISATION OF THE OUTPUT AREA CLASSIFICATION

The appropriate options were selected from 'GeodemCreator' and the output area classification was produced for the United Kingdom. The output of the classification (i.e. CSV file) was joined with different shape files and a number of maps were produced. R was used to join the CSV file with the shape file: the generalised source code for this is given in Appendix 6 section.

Figures 8.26 - 8.29 show the resultant maps. The figures show the Output Area Classification of England and Wales, London, Wales, and Northern Ireland respectively. A number of maps were produced in order to see the classification clearly for different areas.

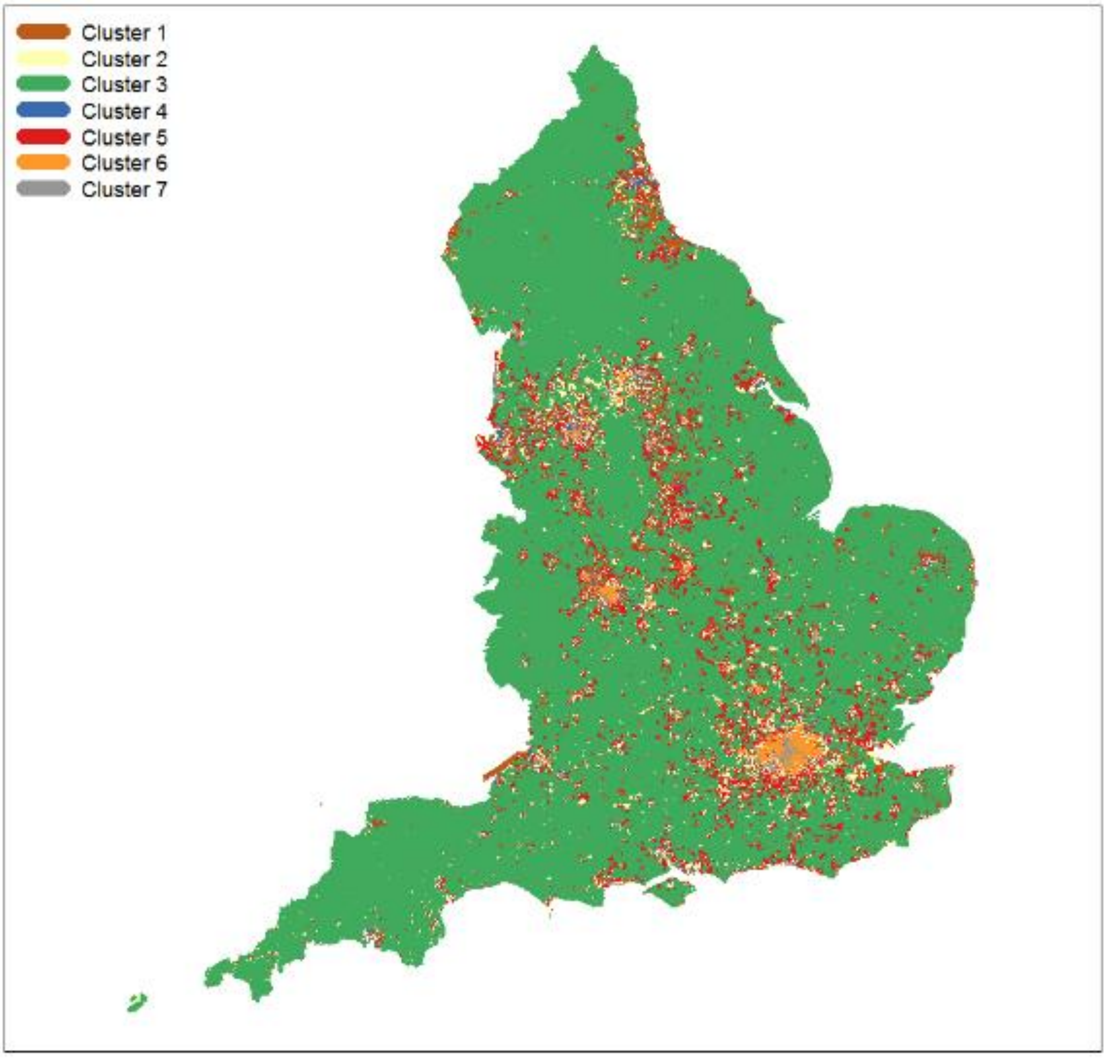


Figure 8.26: Output Area Classification of England and Wales (Case study no. 2)

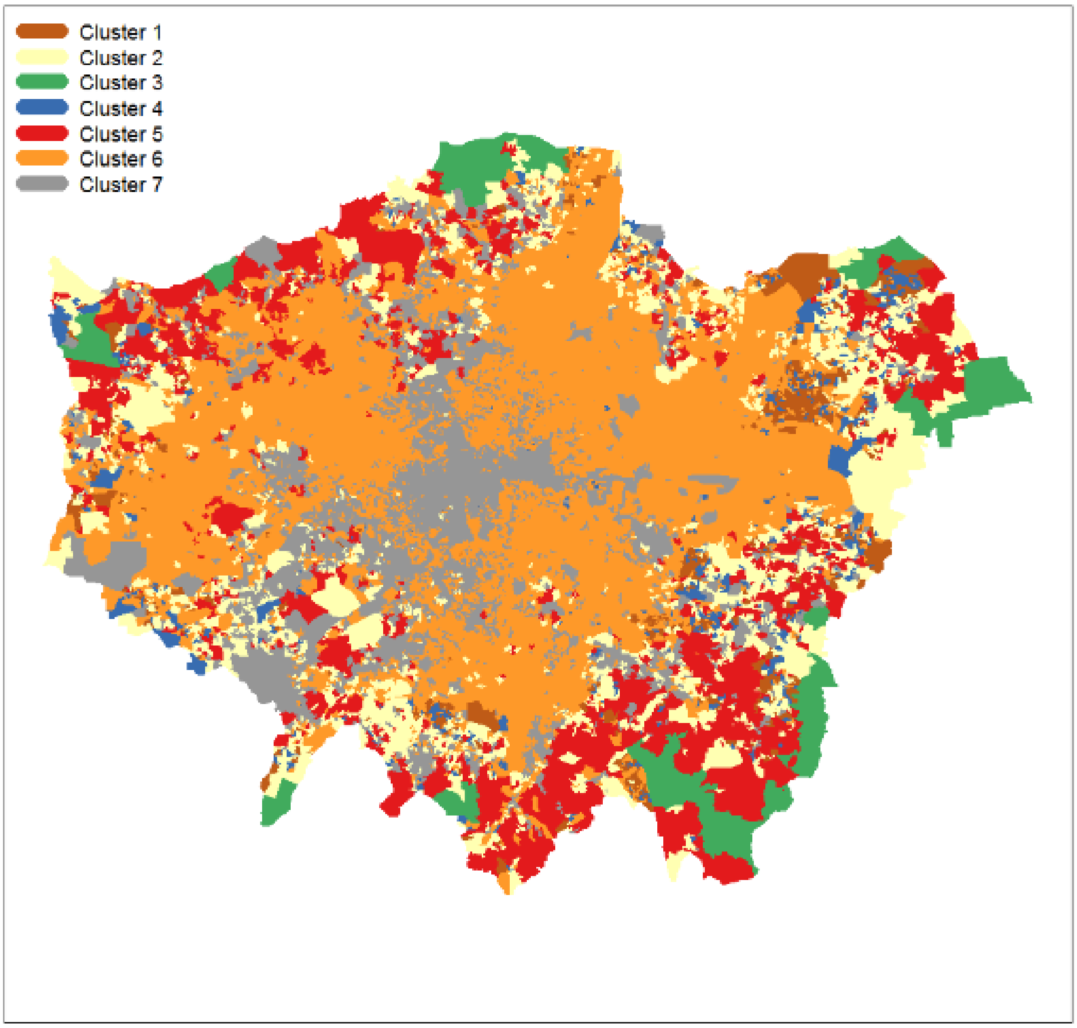


Figure 8.27: Output Area Classification of London (Case study no. 2)

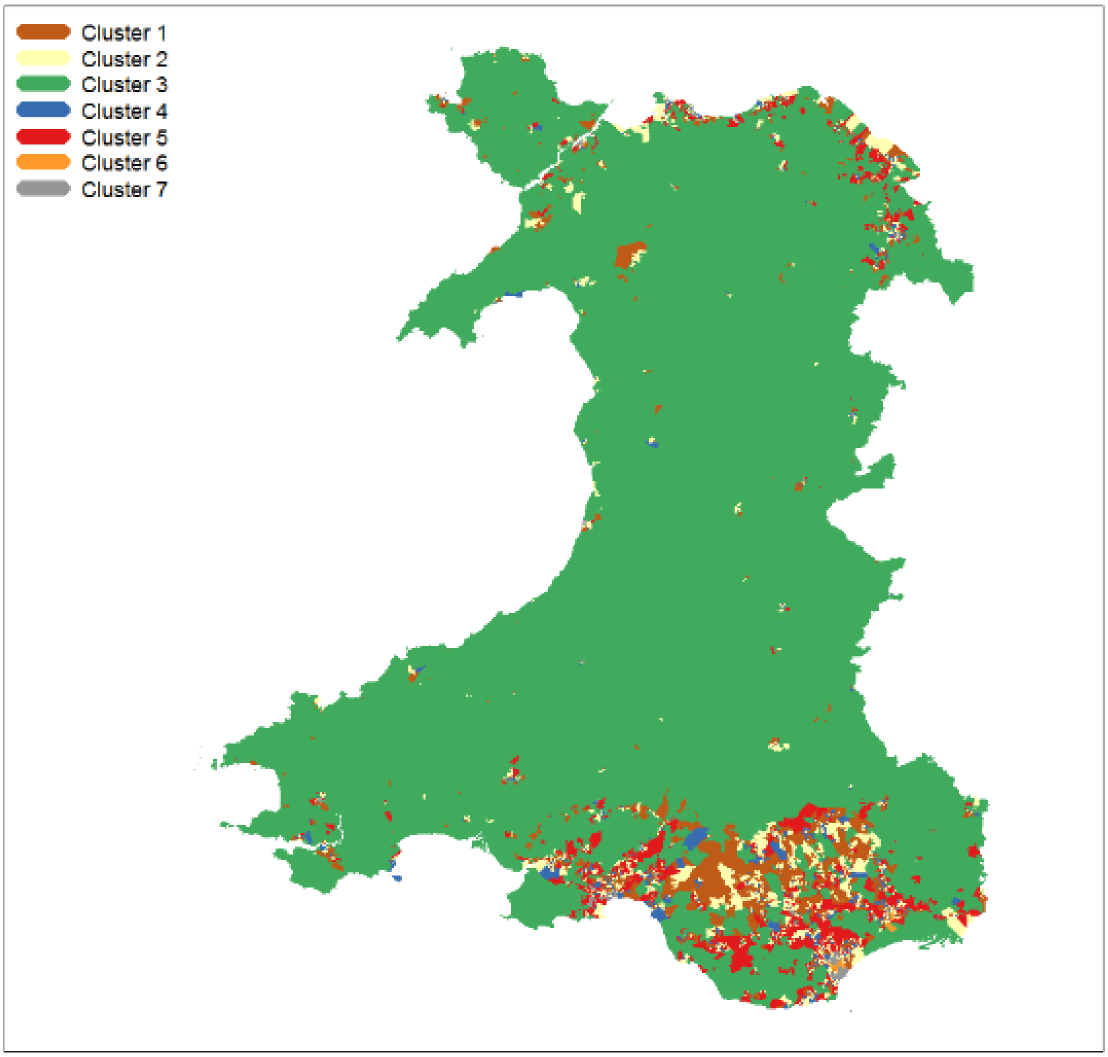


Figure 8.28: Output Area Classification of Wales (Case study no. 2)

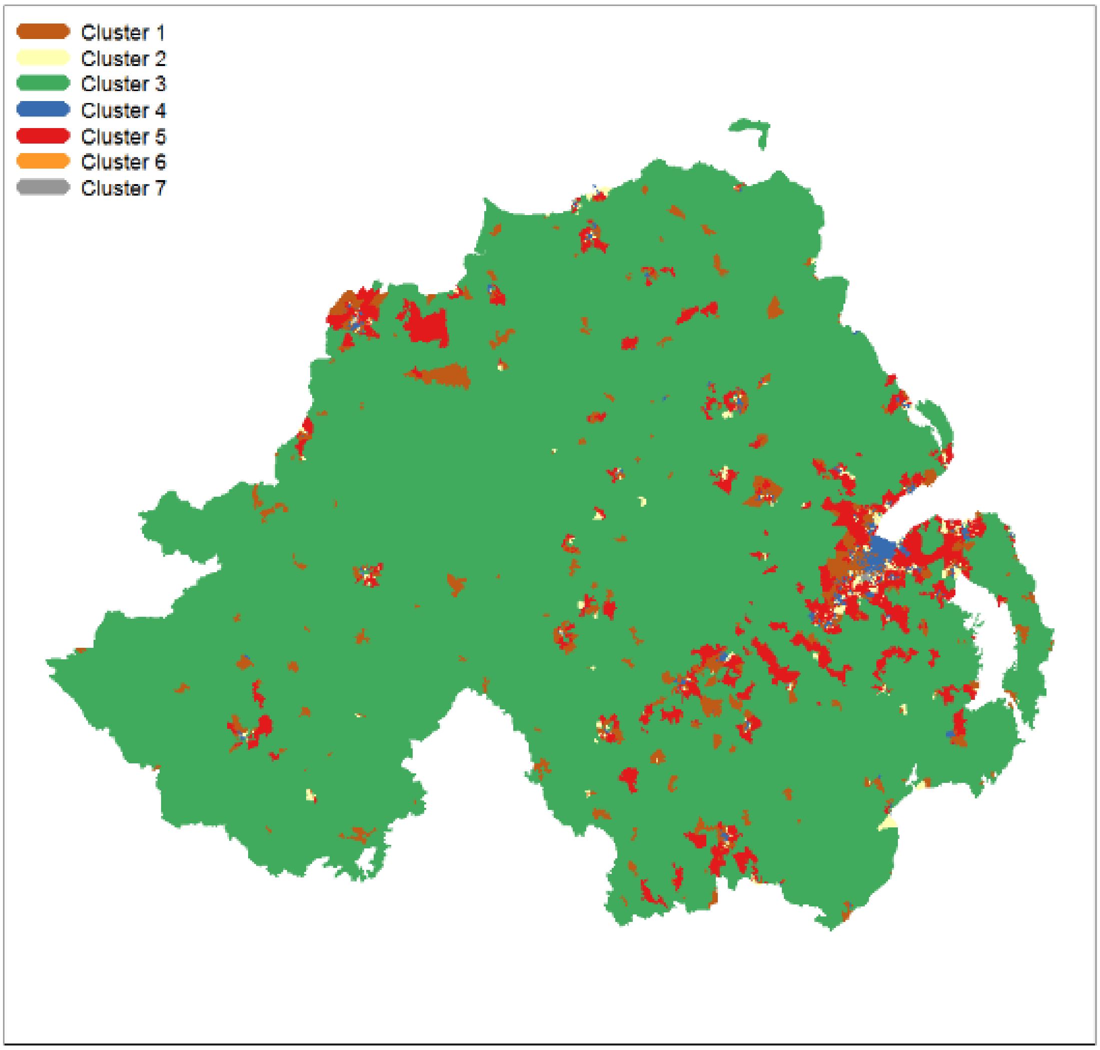


Figure 8.29: Output Area Classification of Northern Ireland (Case study no. 2)

'GeodemCreator' also plots radar charts for individual clusters. These radar charts are very useful when defining and naming individual clusters, and thus making decisions from the clusters produced. The description of each of the clusters is given below the respective radar charts. Vickers & Rees (2007) was used to name the individual clusters.

CLUSTER NO. 1

Figure 8.30 radar chart of cluster no. 1. All the variables v1-v41 are as described in Section 7.2.1.

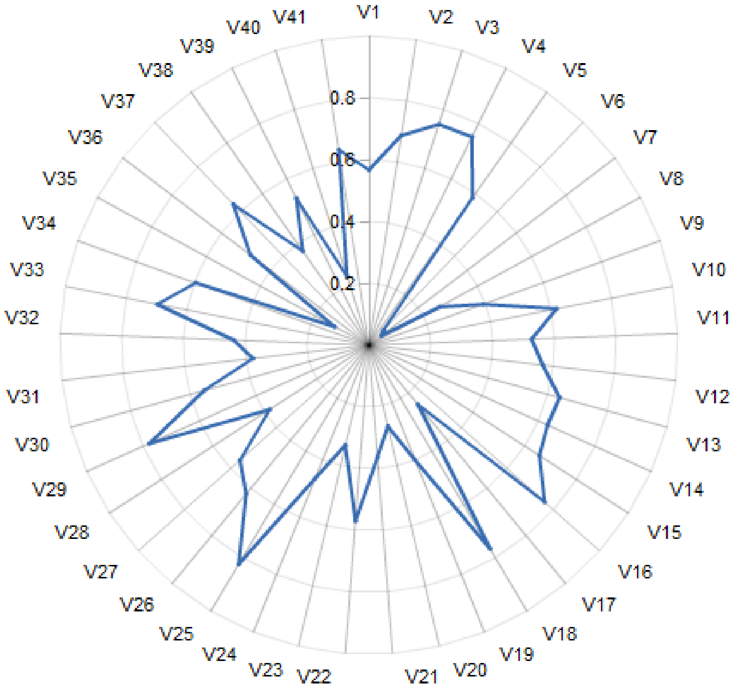


Figure 8.30: Radar chart of cluster no. 1 (Case study no. 2)

The variables having high values in this cluster are:

- V3: Age 25-44
- V4: Age 45-64
- V16: Rent (Public)
- V18: Terraced Housing
- V25: Routine/Semi-Routine Occupation
- V29: Limiting Long Term Illness (SIR)

- V33: Working part-time

Based on the high values of the variables, this cluster could be named '**Blue Collar Communities**'.

CLUSTER NO. 2

Figure 8.31 radar chart of cluster no. 2. All the variables v1-v41 are as described in Section 7.2.1.

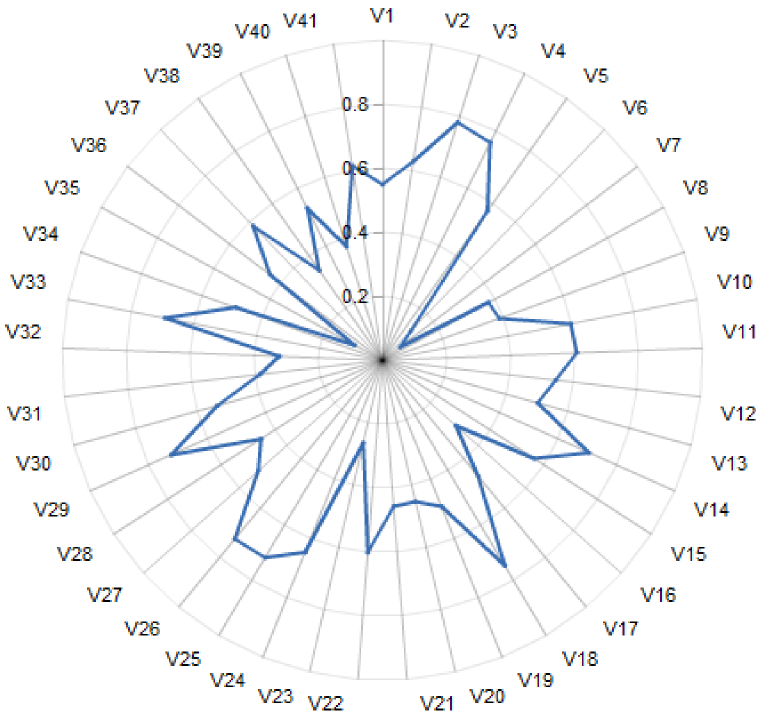


Figure 8.31: Radar chart of cluster no. 2 (Case study no. 2)

The variables having high values in this cluster are:

- V3: Age 25-44
- V4: Age 45-64
- V14: Two adults no children

- V18: Terraced Housing
- V25: Routine/Semi-Routine Occupation
- V26: 2+ Car houshold
- V29: Limiting Long Term Illness (SIR)
- V33: Working part-time

Based on the high values of the variables, this cluster could be named '**Typical Traits**'.

CLUSTER NO. 3

Figure 8.32 radar chart of cluster no. 3. All the variables v1-v41 are as described in Section 7.2.1.

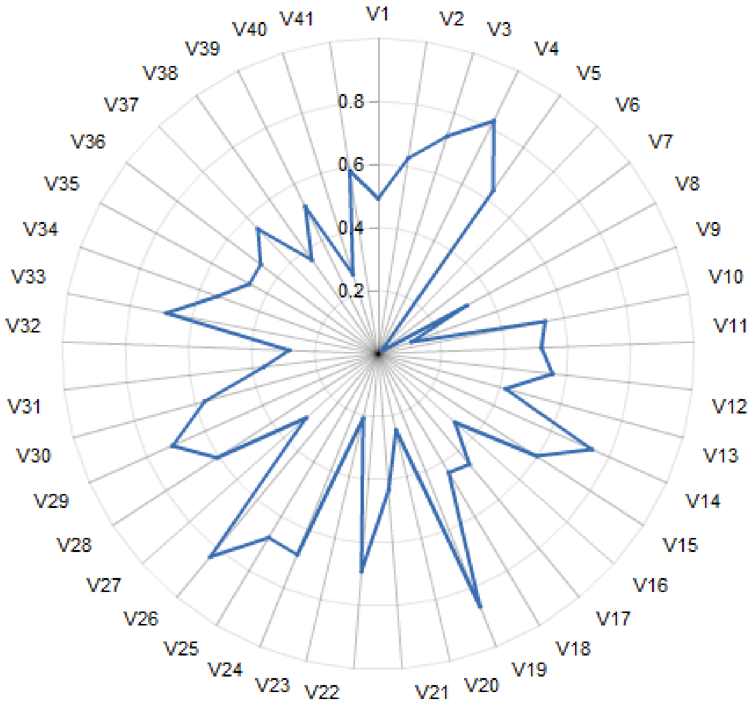


Figure 8.32: Radar chart of cluster no. 3 (Case study no. 2)

The variables having high values in this cluster are:

- V4: Age 45-64
- V14: Two adults no children
- V19: Detached Housing
- V22: Rooms per household
- V24: HE Qualification
- V25: Routine/Semi-Routine Occupation
- V26: 2+ Car houshold
- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**Countryside**'.

CLUSTER NO. 4

Figure 8.33 radar chart of cluster no. 4. All the variables v1-v41 are as described in Section 7.2.1.

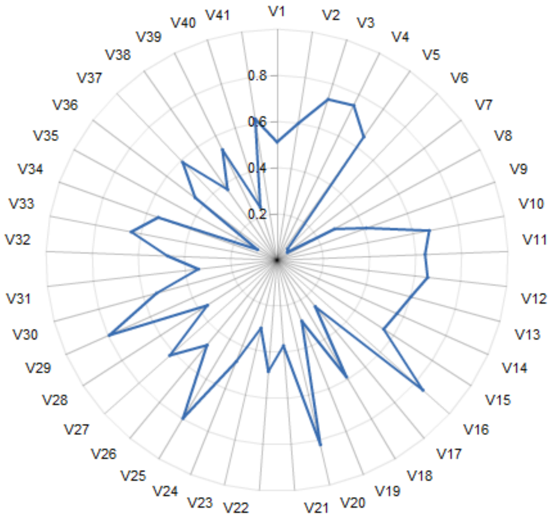


Figure 8.33: Radar chart of cluster no. 4 (Case study no. 2)

The variables having high values in this cluster are:

- V3: Age 25-44
- V4: Age 45-64
- V16: Rent (Public)
- V20: All Flats
- V25: Routine/Semi-Routine Occupation
- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**Constrained by Circumstances**'.

CLUSTER NO. 5

Figures 8.34 radar chart of cluster no. 5. All the variables v1-v41 are as described in Section 7.2.1.

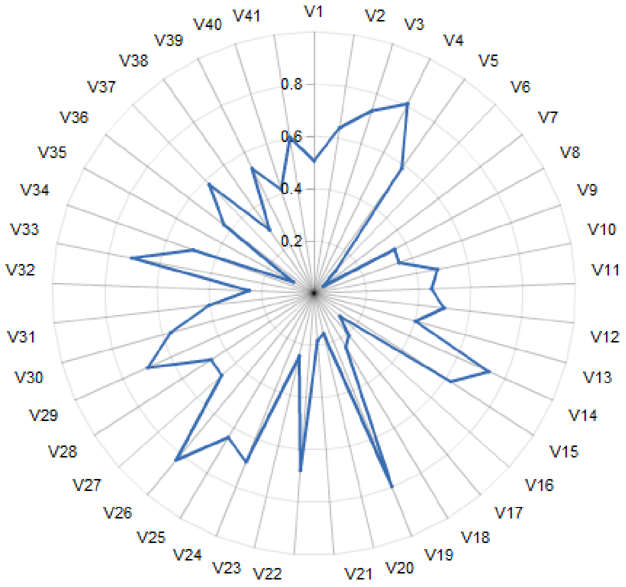


Figure 8.34: Radar chart of cluster no. 5 (Case study no. 2)

The variables having high values in this cluster are:

- V4: Age 45-64
- V14: Two adults no children
- V19: Detached Housing
- V22: Rooms per household
- V24: HE Qualification
- V26: 2+ Car household

Based on the high values of the variables, this cluster could be named '**Prospering Suburbs**'.

CLUSTER NO. 6

Figures 8.35 radar chart of cluster no. 6. All the variables v1-v41 are as described in Section 7.2.1.

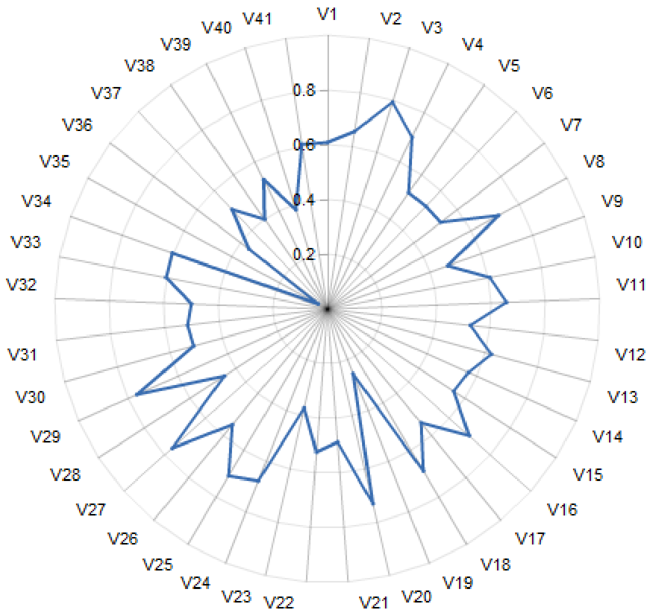


Figure 8.35: Radar chart of cluster no. 6 (Case study no. 2)

The variables having high values in this cluster are:

- V3: Age 25-44
- V8: Born Outside the UK
- V11: Single person household (not pensioner)
- V16: Rent (Public)
- V20: All Flats
- V24: HE Qualification
- V25: Routine/Semi-Routine Occupation
- V27: Public Transport to work
- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**Multicultural**'.

CLUSTER NO. 7

Figure 8.36 radar chart of cluster no. 7. All the variables v1-v41 are as described in Section 7.2.1.

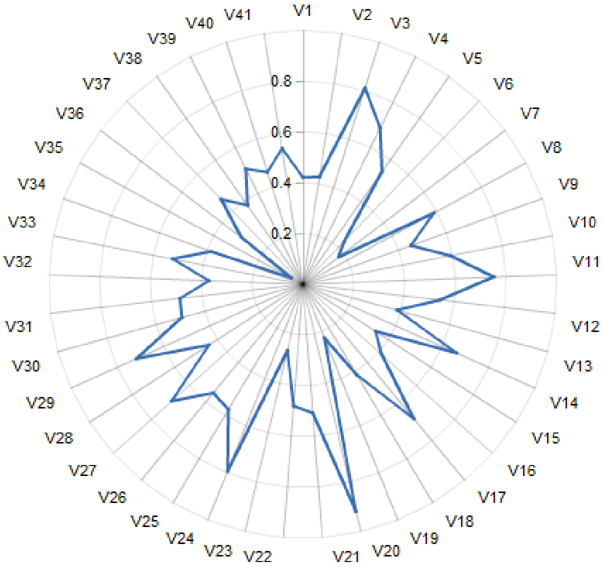


Figure 8.36: Radar chart of cluster no. 7 (Case study no. 2)

The variables having high values in this cluster are:

- V3: Age 25-44
- V11: Single person household (not pensioner)
- V14: Two adults no children
- V17: Rent (Private)
- V20: All Flats
- V24: HE Qualification
- V27: Public Transport to work
- V29: Limiting Long Term Illness (SIR)

Based on the high values of the variables, this cluster could be named '**City Living**'.

END POINT OF THE OUTPUT AREA CLASSIFICATION OF THE UNITED KINGDOM

This case study demonstrates the use of 'GeodemCreator' to create a national level output area classification of the United Kingdom. This is essentially the re-creation of the classification created by Vickers & Rees (2007), and the labels that they use to describe the original classification are cited here. The results of this classification were compared with the classification created by Vickers & Rees (2007) and they were consistent in both the cases. The results establish the consistency of the classifications created within 'GeodemCreator'.

Users can create this national level output area classification in the 'Basic' mode of the software, without the need to specify any variables, weighting, or the number of desired clusters in the solution. This gives a flexible option to the users in public sector applications to create and use geodemographic classifications in the 'Basic' mode without much knowledge of the underpinning methodologies. In addition, the running time of producing this classification is small (within minutes) as compared to the time it takes for producing a classification manually, using different software components (i.e. SPSS, Microsoft Excel etc).

8.3.4 AN OUTPUT AREA LEVEL 'SOCIO-ECONOMIC AND ETHNIC' CLASSIFICATION OF THE GREATER LONDON

In order to demonstrate the utility of the 'GeodemCreator' software for reading ancillary sources alongside Census of Population data, an Output Area Classification of the Greater London was created by supplementing the 41 census variables (socio-economic data) used in Section 8.3.2. This section explains the creation of an Output Area level 'Socio-Economic and Ethnic' Classification of Greater London.

This classification was created using the 'Advanced' mode of 'GeodemCreator'. The Ethnicity data, comprised 12 variables derived from the names data sources set out in Chapters 3 and 4. 12 ethnicity variables were imported alongside the 41 variables used in Section 8.3.2 and the combined total of 53 variables were selected for the classification. 'London' and '7' were selected as the geographical area and number of clusters, respectively.

This classification gives a proof of the concept that a classification can be created using the amalgamation of different data sources (in this case it is the amalgamation of census data and an ancillary ethnicity data source).

VARIABLES USED TO PRODUCE THE CLASSIFICATION

53 variables were used to create this classification. These include 41 Census variables and 12 Ethnicity Data variables. The 41 census variables and their domains are outlined in the previous chapter (section 7.2.1). The 12 ethnicity variables are outlined in the sub-section 8.3.1 of this chapter and were derived from the 'Worldnames' database developed in Chapters 3.

IDENTIFYING THE NUMBER OF CLUSTERS

Section 8.3.2 discussed how to choose an optimal number of clusters. Using the same technique, Figure 8.37 shows the plot of the 'within sum of squares' according to 'number of clusters' by running *k*-means on the 53 variables (socio-economic and ethnicity data) for Greater London.

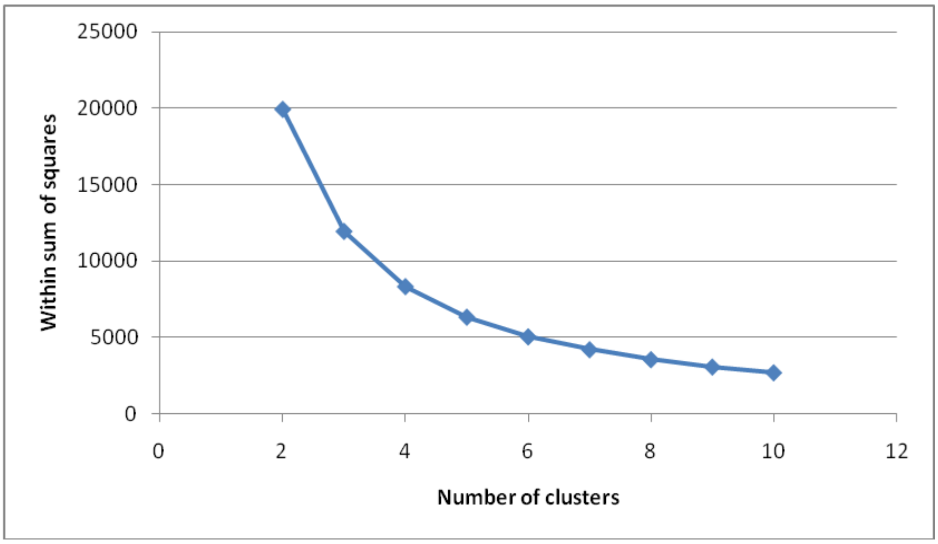


Figure 8.37: Within sum of squares by number of clusters (Case study no. 3)

k=7 can be used as the optimal solution because the within sum of squares remains almost constant after k=7 cluster solutions.

VISUALISATION OF THE OUTPUT AREA CLASSIFICATION

The appropriate options were selected from 'GeodemCreator' and the output area classification was produced for Greater London using Socio-Economic and Ethnicity data. The output of the classification (i.e. CSV file) was joined with the shape files and a map was produced. R was used to join the CSV file with the shape file: the generalised source code for this is given in Appendix 6. Figure 8.38 shows the resultant map.

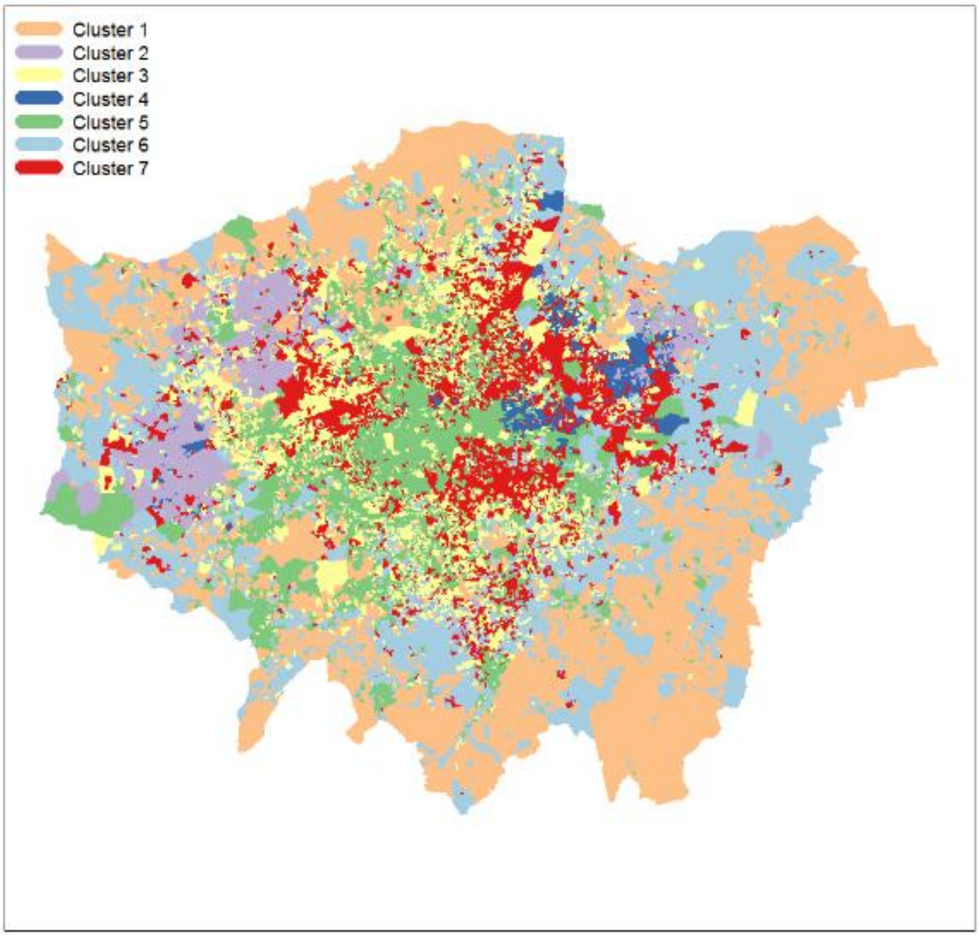


Figure 8.38: A socio-economic and ethnicity classification of Greater London (Case study no. 3)

'GeodemCreator' also plots radar charts for individual clusters. These radar charts are very useful when defining and naming individual clusters, and thus making decisions from the clusters produced. The description of each of the clusters is given beneath respective radar charts.

CLUSTER NO. 1

Figure 8.39 shows the radar chart of cluster no. 1. The variables V1 - V41 are described in the section 7.2.1, and variables V42 – V53 are described in the section 8.3.1.

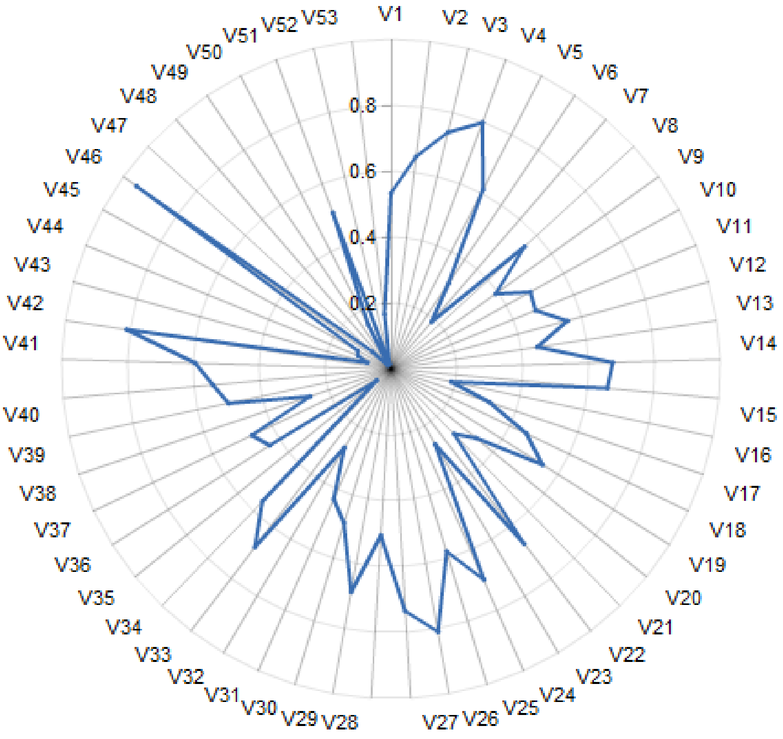


Figure 8.39: Radar chart of cluster no. 1 (Case study no. 3)

The variables having high values in this cluster are:

- V4: Age 45-64
- V22: Rooms per household
- V24: HE Qualification
- V26: 2+ Car household
- V33: Working part-time
- V42: European ethnic group
- V46: English ethnic group

Based on the high values of the variables, this cluster could be named '**English and European ethnic groups living in suburban areas**'.

CLUSTER NO. 2

Figure 8.40 shows the radar chart of cluster no. 2. The variables V1 - V41 are described in section 7.2.1, and variables V42 – V53 are described in section 8.3.1.

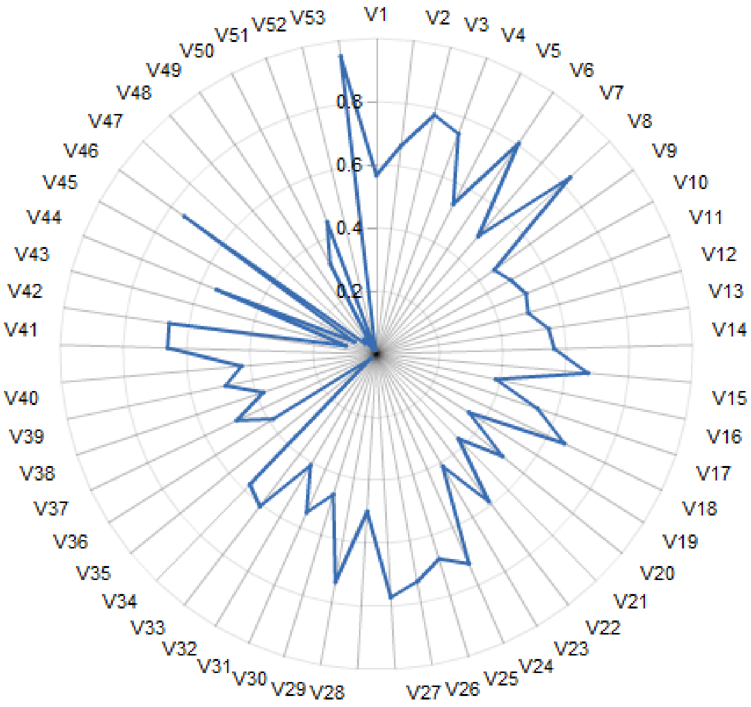


Figure 8.40: Radar chart of cluster no. 2 (Case study no. 3)

The variables having high values in this cluster are:

- V3: Age 25-44
- V4: Age 45-64
- V6: Indian, Pakistani or Bangladeshi
- V8: Born outside the UK
- V24: HE Qualification
- V26: 2+ Cars
- V27: Public Transport to work
- V29: Limited long term illness (SIR)

- V53: South Asian ethnic group

Based on the high values of the variables, this cluster could be named ''**Well off and educated Asian Families** '.

CLUSTER NO. 3

Figure 8.41 shows the radar chart of cluster no. 3. The variables V1 - V41 are described in section 7.2.1, and variables V42 – V53 are described in section 8.3.1.

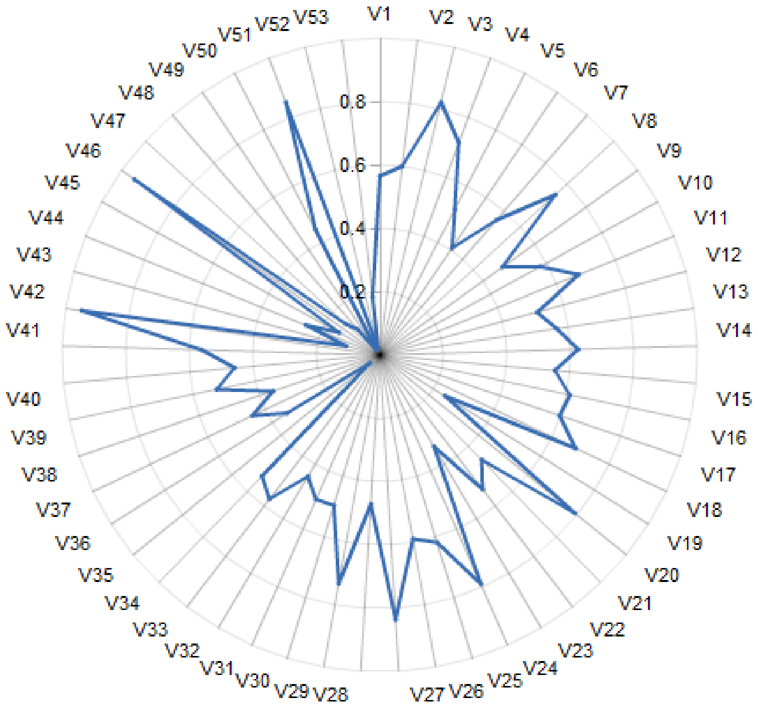


Figure 8.41: Radar chart of cluster no. 3 (Case study no. 3)

The variables having high values in this cluster are:

- V3: Age 25-44
- V18: Terraced Housing
- V20: All Flats
- V24: HE Qualification
- V27: Public transport to work

- V42: European ethnic group
- V46: English ethnic group
- V51: Celtic ethnic group

Based on the high values of the variables, this cluster could be named '**English, European, and Celtic fringe city commuters** '.

CLUSTER NO. 4

Figure 8.42 shows the radar chart of cluster no. 4. The variables V1 - V41 are described in section 7.2.1, and variables V42 – V53 are described in section 8.3.1.

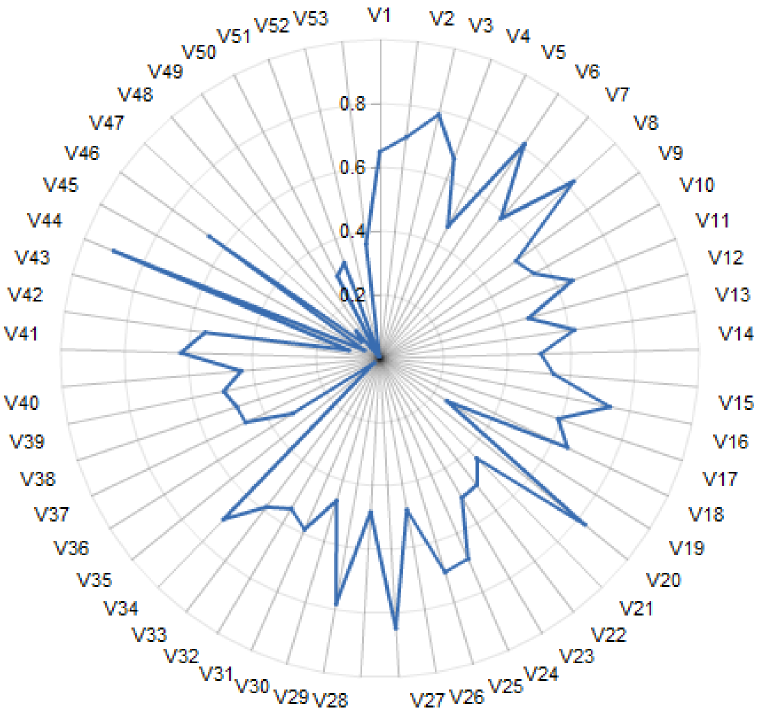


Figure 8.42: Radar chart of cluster no. 4 (Case study no. 3)

The variables having high values in this cluster are:

- V3: Age 25-44
- V6: Indian, Pakistani or Bangladeshi

- V8: Born outside the UK
- V16: Rent (Public)
- V20: All flats
- V27: Public transport to work
- V29: Limited long term illness (SIR)
- V34: Economically inactive looking after family
- V44: Muslim ethnic group

Based on the high values of the variables, this cluster could be named '**Poor Asian Families**'.

CLUSTER NO. 5

Figure 8.43 shows the radar chart of cluster no. 5. The variables V1 - V41 are described in section 7.2.1, and variables V42 – V53 are described in section 8.3.1.

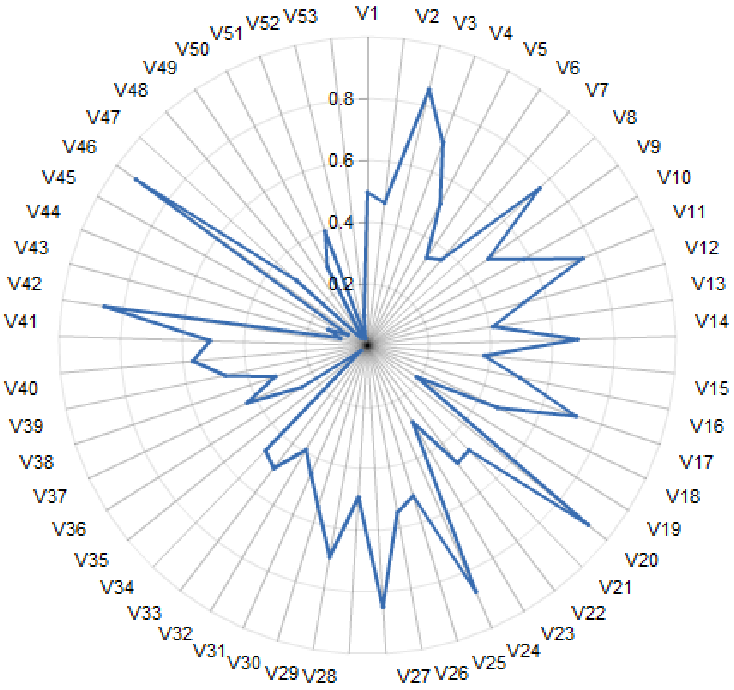


Figure 8.43: Radar chart of cluster no. 5 (Case study no. 3)

The variables having high values in this cluster are:

- V3: Age 25-44
- V8: Born outside the UK
- V11: Single person household (not pensioner)
- V14: Two adults no children
- V17: Rent (Private)
- V20: All flats
- V24: HE Qualification
- V27: Public transport to work
- V42: European ethnic group
- V46: English ethnic group

Based on the high values of the variables, this cluster could be named '**Childless European city dwellers** '.

CLUSTER NO. 6

Figure 8.44 shows the radar chart of cluster no. 6. The variables V1 - V41 are described in section 7.2.1, and variables V42 – V53 are described in section 8.3.1.

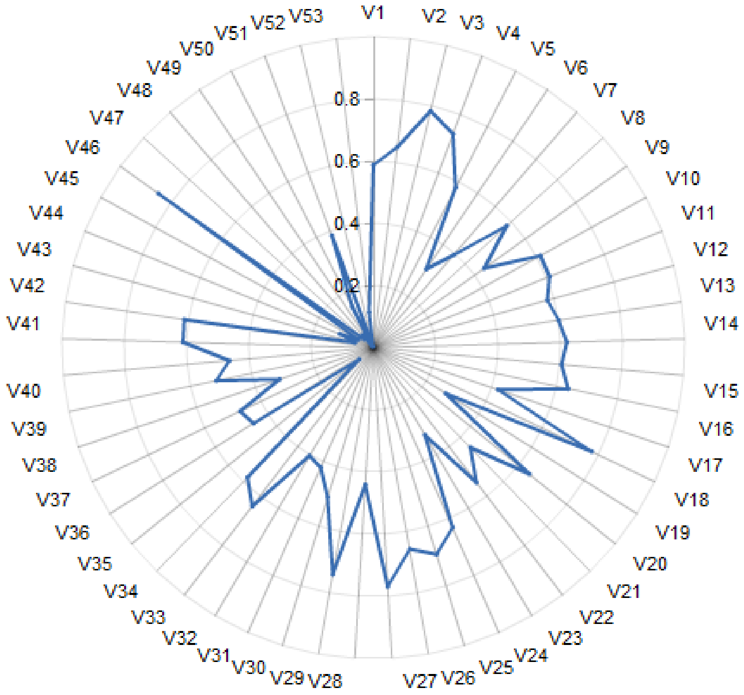


Figure 8.44: Radar chart of cluster no. 6 (Case study no. 3)

The variables having high values in this cluster are:

- V3: Age 25-44
- V4: Age 45-64
- V18: Terraced housing
- V25: Routine/Semi-Routine Occupations
- V27: Public transport to work
- V46: English ethnic group

Based on the high values of the variables, this cluster could be named **'Native Blue Collar communities**'.

CLUSTER NO. 7

Figure 8.45 shows the radar chart of cluster no. 7. The variables V1 - V41 are described in section 7.2.1, and variables V42 – V53 are described in section 8.3.1.

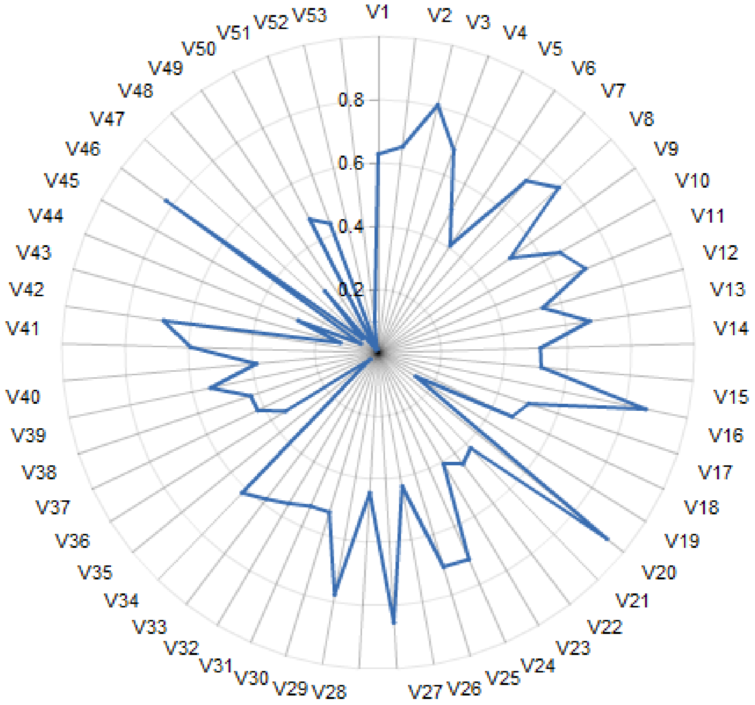


Figure 8.45: Radar chart of cluster no. 7 (Case study no. 3)

The variables having high values in this cluster are:

- V3: Age 25-44
- V8: Born outside UK
- V16: Rent (Public)
- V20: All flats
- V27: Public transport to work
- V42: European ethnic group
- V46: English ethnic group

Based on the high values of the variables, this cluster could be named '**English and Europeans living in council properties**'.

THE OUTPUT AREA LEVEL 'SOCIO-ECONOMIC AND ETHNIC' CLASSIFICATIONS OF GREATER LONDON

This case study demonstrated the use of 'GeodemCreator' to create a bespoke local area classification by the combination of socio-economic data (41 census variables) and an ancillary data source (12 variables of ethnicity data). This data source was created by using the ‘Worldnames’ database set out in Chapter 3 and visualised in Chapter 4.

The classification was created using the 'Advanced' mode of 'GeodemCreator'. The Ethnicity data, containing 12 variables, were imported to the software and all 53 variables (12 Ethnicity Variables; 41 Socio-Economic Variables provided in the software by default) were selected for the classification. Later, 'London' and '7' were selected as the geographical area and number of clusters, respectively.

This classification gives a proof of the concept that a classification can be created using the amalgamation of different data sources.

8.4 CONCLUSION

Previous chapters have set out the background to the requirement for creating real-time and open geodemographics, where the classification is created quickly and efficiently and the data and methods used to build the classification are open to the public and available for public scrutiny. A software utility was required as proof of the concept that could be used for building geodemographic classifications. Chapter 7 was the starting point of describing such a software utility, and it described the functional and technical application specification of 'GeodemCreator'. The 'GeodemCreator' was developed and this chapter described the functioning of the 'GeodemCreator' as a geodemographic decision support tool. Users can create their geodemographic classifications, draw inferences, and make decisions based on the classifications produced. While this requires knowledge and understanding of how geodemographic classifications work, both inexperienced and experienced users can use this software in one of its two different modes of operations. This gives flexibility for the end users base, where on the one hand 'GeodemCreator' could be used for creating classifications for particular projects without any decision making process involved or, on the other hand, it could be used by decision makers for the analysis of their particular geographical areas.

The first part of this chapter described the use of 'GeodemCreator' in the form of screen shots. The operation of the ‘Basic’ and ‘Advanced’ modes of operation was described with the help of a series of screen shots.

The second part of the chapter was built around case studies. Three case studies were presented. The first case study described the creation of an Output Area Classification for London, and described the steps involved in building the classification by using 'GeodemCreator'. This case study provided proof of the concept that bespoke local area classifications can be created by using 'GeodemCreator'. The cluster solutions obtained from the classifications were assigned ‘pen portrait’ names based on the characteristics of the data (i.e. characteristics of the population), as revealed by the radial plots produced by the software. Such local area classifications are of great potential utility in certain domains and could prove better than the general national level classification because of their restricted study areas. Because 'GeodemCreator' also allows users to cluster their own data in 'Advanced' mode, this opens up a new way of building small scale bespoke classifications using a free software utility.

The second case study retraced the steps of Vickers and Rees (2007) and described the creation of an Output Area Classification for the entire United Kingdom, by using various default options of 'GeodemCreator'. This case study provided proof that national level classifications can be created by using 'GeodemoCreator' that are consistent with the published literature.

The third case study described the creation of an Output Area level ‘Socio-Economic and Ethnic’ classification for London, and described the steps involved in building the classification by using 'GeodemCreator'. This case study provided proof of the concept that bespoke local area classifications can be created by the amalgamation of different data sources. The cluster solutions obtained from the classifications were assigned ‘pen portrait’ names based on the characteristics of the data (i.e. characteristics of the population), as revealed by the radial plots produced by the software. It is quite interesting to see how ethnicity data effects the classification produced. This opens up new horizons of using ‘GeodemCreator’ on different datasets for the creation of small scale geodemographi classifications.

Thus 'GeodemCreator' is a tool for creating classifications where the data and methods used are open to the users – unlike the classifications created by commercial companies. The open nature of the geodemographic classifications produced will make them available to public scrutiny, thus making decision making process transparent as well as more efficient.

'GeodemoCreator' could be enhanced in two ways. First, by providing more public data sources in the 'GeodemCreator': this would present users with more options when building bespoke classifications, and drawing inferences from the classifications. Second, this work can be extended to an online web based service for building geodemographic classifications online. This would enable a large number of users to have simultaneous access to the tool and they would be able to create the classifications over the web with the need to download and install 'GeodemCreator' and 'R'.

9 CHAPTER 9: CONCLUSIONS

Geodemographic modelling of neighbourhood conditions presents a successful applied problem-solving approach to understanding socio-spatial differentiation at the neighbourhood scale. For this reason, it is widely used in the public as well as the private sectors to predict the consumption of products, services or resources (Singleton et al., 2010)

9.1 REFLECTIONS

The thesis began with a statement of the requirement for real-time and open geodemographic classifications, in which the data and methods used to build the classifications are open to users. After setting out the contribution of novel data sources to this agenda, and ways of visualising them at a full range of scales from the sub regional to the international, the development work culminated with the development and implementation of a pilot software utility for creating national and local area geodemographic classifications. During this journey, the PhD has investigated the data sources and the methodology of building geodemographic classifications in detail.

The thesis journey started with the introduction of geodemographic classifications in general, and a review and interpretation of their rich history. Geodemographics have become an important industry now with many commercial companies creating niche, as well as general purpose, classifications. The PhD reviewed the current practices of commercial companies, alongside the available public domain alternatives. In addition, international geodemographic classifications were discussed and an introduction of the data and methods used to create the classifications was set out.

This journey then led to the investigation of the data sources which could be used to build new and innovative geodemographic classifications. A major effort in this regard was the creation of a novel data source of 'names' from telephone directories and electoral registers. Chapter 3 discussed the creation of this large database of names. These innovative and new data sources could be really useful in creating new types of geodemographic classifications. The geographical origins of 'names' and their distribution can be combined with other socio-economic data for the creation of innovative geodemographic classifications. These could also be of utility in the context of international geodemographics, as the geographical distribution of 'surnames' varies between regions, countries and continents.

While the creation of new data sources is important, the visualisation and diagnostic checking of such data is important too. This motivation led to the investigation of different ways of visualising the 'names' data over the web, and a website was created for the visualisation for the 'surname' data. This website ('Worldnames') was launched and provided very convincing testimony to public interest in aspects of geodemographics, with over a million users during the first 10 days following the launch. This website is a proof of concept of serving data over the web in the context of geodemographic classifications. From the point of view of creating online geodemographic information systems, the 'Worldnames' website also acts as a proof of concept for spatial data query and serving large amounts of data on-the-fly.

Current classifications are created using static data source i.e. census data, lifestyle surveys etc. A thorough investigation of the data sources in the public domain was undertaken in Chapter 5, where different data sources available in the public domain were investigated with a view to their potential inclusion in geodemographic classifications. As the trend nowadays is moving towards live feeds of data from APIs e.g. The ONS NeSS Data Exchange API (Office for National Statistics, 2009) and London Data Store (London Data Store, 2010. Keeping this in mind, a web service was developed to extract live feeds of data from the ONS (Office for National Statistics) NeSS API. This web service demonstrates the feasibility of assembling and (where appropriate) aggregating different live data sources into a single resource for building geodemographic classifications on the web.

During the journey, the investigation of the methodologies used to building geodemographic classifications was really important. Cluster analysis is the procedure of choice for creating the groups of a geodemographic classification. *K*-means is the traditional clustering algorithm for creating geodemographic classifications. A thorough investigation of *k*-means was undertaken, including comparison of its variant algorithms. *k*-means is an instable clustering algorithm and needs to run multiple times (10,000 times (Singleton & Longley, 2009)) for building a geodemographic classification. Thus other algorithms were also investigated. PAM (Partitioning Around Medoids) and Genetic Clustering were investigated and compared with *k*-means at different spatial levels. This comparison was important in order to come up with an algorithm that runs quicker if classifications are to be built in an online environment. Computational enhancements to *k*-means clustering were also investigated. This included using *k*-means with PCA (Principal Components Analysis), and a parallel implementation of *k*-means on Nvidia's graphics cards using CUDA (Computer Unified Device Architecture). This algorithm was called 'parallel *k*-means' and could be used in building geodemographic classifications in an online and web based environment.

Once the data and methods had been investigated from the point of view of creating real-time geodemographic classifications, a proof of the concept software utility was developed. The aim of this software utility was to create geodemographic classifications on desktop machines, and in future an online version of the software utility could be developed. The functional and technical application specifications were discussed from a software engineering perspective. These specifications are important for the development of an efficient software product.

The research subsequently led to the development of the desktop software utility 'GeodemCreator'. This is a free software utility for building geodemographic classifications. The software allows both inexperienced and experienced users to create geodemographic classifications and could be used as a geodemographic decision support tool. Users can create their geodemographic classifications, draw inferences, and make decisions arising from the classifications produced. This provides flexibility for the end user, whereby the 'GeodemCreator' could be used to create acceptable classifications for particular projects without any sophisticated expertise or, on the other hand, could be used by more sophisticated decision makers for the classification of particular geographical areas, using particular variables and weighting schemes.

In the future, 'GeodemoCreator' could be enhanced as an online web based service for building geodemographic classifications online. This could include the integration of live data sources from different locations into the tool and the live computation of geodemographic classifications. This will enable users to create geodemographic classifications via web based services by accessing the tool online. This lies beyond the scope of this PhD, and forms the part of the evolving research agenda.

9.2 ADVANTAGES AND LIMITATIONS OF 'GEODEMCREATOR'

The thesis focused on the need to create open and real-time geodemographic information systems. A number of avenues relating to data sources and methodologies of creating geodemographic classifications were investigated and described. 'GeodemCreator' software developed as a software utility which could be used for creating national and local area geodemographic classifications. 'GeodemCreator' has a number of advantages and limitations.

To date, there is no specific software product that could be used for creating geodemographic classifications. Researchers have deployed different statistical packages for analysing data and creating their classifications. Relevant software includes Microsoft Excel, SPSS, and R. However, there is no single resource that could create the geodemographic classifications according to precise user requirements. As such, 'GeodemCreator' has the advantage of being the first software utility that could be used for creating different types of geodemographic classifications. A second point is that hitherto, the creation of geodemographic classifications has itself been regarded as the task of a closed community of experts, with end users expected to uncritically accept the final classifications provided to them. 'GeodemCreator' empowers both inexperienced and experienced users to build their own bespoke geodemographic classifications. Inexperienced users do not need to know much about how the classifications are produced and by opting to use the software in 'Basic' mode can nevertheless create classifications that are safe to use. In 'Basic' mode the software automatically selects tried and tested variables, weightings, and numbers of clusters. Experienced users can choose the software to work in 'Advanced' mode, where they can specify the variables and the number of cluster solutions for a geodemographic classification. This saves a lot of time that would be required to write the source code in R or SPSS, or to perform manual statistical calculations using Microsoft Excel.

In addition to these benefits, 'GeodemCreator' nevertheless has a number of limitations. At the time of writing this thesis, 'GeodemCreator' provides only a limited number of cluster solutions as the output (i.e. k=7 as the maximum cluster solutions). This is because of the memory handling procedures in Java programs when clustering a large data source. However, this issue could be solved in the future by optimizing the data structures used in the software. 'GeodemCreator' also does not presently operationalise the concept of integration of data from a number of live XML data sources. Instead it works on static data sources which are either the part of the software (41 census variables) or are downloaded by the users onto their machines. Another limitation is that 'GeodemCreator' is a desktop utility and needs 'R' to be installed on the machine before building a geodemographic classification. It is hoped that, in the future, an online version of 'GeodemCreator' will be developed and it will remove any dependency upon particular software installation on users' machines.

Further thorough testing of 'GeodemCreator' is also needed, by running the software on different data sources of various sizes and geographical scales.

9.3 FUTURE RESEARCH

Two main avenues of future research have been envisaged at this stage. One is related to the improvements in the use of data sources, and the other is design and development of an online version of 'GeodemCreator' for creating bespoke geodemographic classifications in an online and web based environment.

9.3.1. DATA SOURCES IMPROVEMENTS

Amongst the data sources improvements, there are two major enhancements needed.

First is the use of some new and innovative data sources for the creation of new geodemographic classifications representing different aspects of the population. Data from 'Worldnames' database created in Chapter 3 and the 'house price ethnicity' data source created in Chapter 5 are the examples of the two data sources which could be used for the creation of bespoke geodemographic classifications. These data sources, combined with the socio-economic data, could bring in some interesting updates relevant to population dynamics. The 'house price ethnicity' data source was created only for London: however, the same technique could be used for creating a 'house price ethnicity' data source for the United Kingdom. Thus this data source could bring in the notion of house prices (or wealth) in terms of ethnic origins of the population in particular areas. The 'Worldnames' database could have a still bigger impact in the creation of geodemographic classifications. Previous research has shown that 'names' distributions follow a geographical pattern. Thus names data, combined with the socio-economic data, could develop new possibilities of creating geodemographic classifications by 'surname regions'. (Cheshire, 2011) demonstrates how 'surname regions' can be created across Europe by using the 'worldnames' database.

The second possible enhancement, in terms of data sources, is the design and development of web services for the possible integration of data from different live and large databases. Different data sources have been moving towards providing live feeds of the data via XML web services. Users need to connect the APIs and they can archive the data to their data files and databases. Chapter 5 demonstrated the creation of a web service for retrieving live XML feeds of data from the ONS (Office for National Statistics) NeSS API. This was the proof of the concept of retrieving data from a live XML data source. However, this web service could be enhanced in order to retrieve data from multiple data sources. A detailed study of geographical scales and coverage of data will be needed for their possible integration into a large single data source. Thus, the web service will need to retrieve data from a number of data sources, and will need to resolve the issues of geographical coverage and scale variations, but will result in the integration of all of these data sources into a single resource. This web service will be a major enhancement in the work towards creating geodemographic classifications online by using different live data sources.

9.3.2. AN ONLINE VERSION OF 'GEODEMCREATOR'

The second avenue of future research is the design and development of an online version of 'GeodemCreator' for creating bespoke geodemographic classifications in a web based environment. 'GeodemCreator' was the proof of the concept that a software utility could be designed and developed that could be used for creating national and local area geodemographic classifications. However, 'GeodemoCreator' presently relies completely on using static data sources i.e. 41 variables of census data are provided as the default data source and users can upload their own data via a CSV file. However, as mentioned in Section 9.3.1, future research will focus on the design and development of web services for the possible integration of data sources from different live and large databases. This will likely be followed by the design and development of online tools for creating geodemographic classifications in a web based environment. This will enable users to use web based tools in order to compute their geodemographic classifications without the need to install 'GeodemCreator' and 'R' on their machines. This could result in a significant increase in the potential user base as web based tools are becoming popular because of their accessibility to a wide range of users.

This will bring a major issue of computing geodemographic classifications really quickly and efficiently, as users in online environments are looking at a response time which is in seconds or (at most) within a couple of minutes. Thus the integration of data from live data sources and the computation of geodemographic classifications by cluster analysis will have to be really fast. *K*-means has remained the traditional algorithm for creating geodemographic classifications. However, Chapter 6 investigated some alternative clustering algorithms for their possible use in an online environment. A parallel implementation of *k*-means (Parallel *k*-means) was also designed and compared with the traditional *k*-means clustering algorithm. Parallel *k*-means could be a possible alternative to the established *k*-means clustering algorithm in an online environment as it computes geodemographic classifications much faster than the traditional *k*-means algorithm. This makes it a candidate estimator for an online version of 'GeodemCreator'.

At the present time, 'GeodemoCreator' does not allow the automatic generation of a map visualisation. Instead users have to use a generalised script written in 'R' (provided in Appendix 6 section). The future research also aims at producing the output of 'GeodemCreator' as an automatic generation of map visualisation.

Figure 9.1 shows a possible architecture for an online version of 'GeodemCreator'.

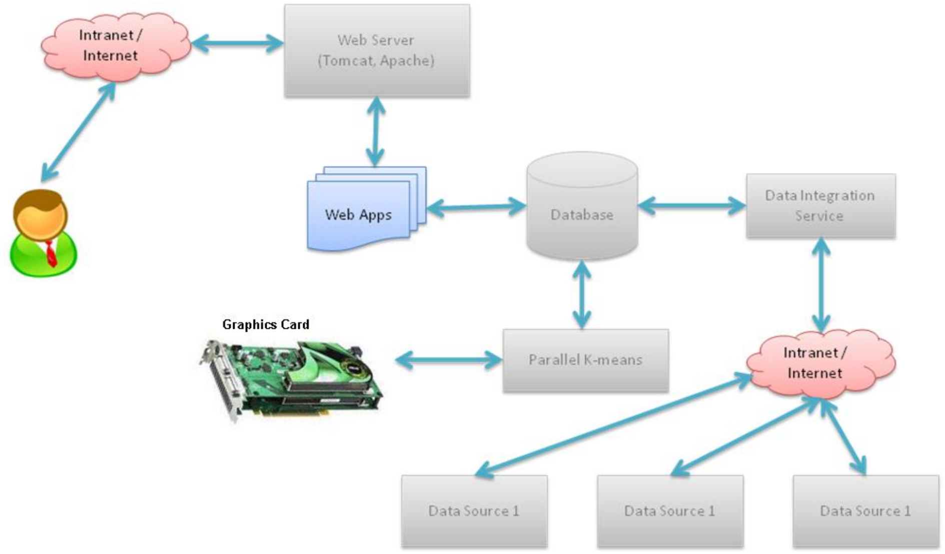


Figure 9.1: A possible architecture for an online version of 'GeodemCreator'

The 'Data Integration Service' would be used for archiving data from online data source into a database. Users would interact with the online version of 'GeodemCreator' on the web. The user would specify the variables for creating a geodemographic classification from a menu of live data sources. The user would also specify the number of cluster solutions for the final geodemographic classification and would ask the system to build the classification. Based on user input, an online version of 'GeodemCreator' would integrate the data from all the variables selected and normalize the data. Once the data are prepared, parallel *k*-means would be used for computing the geodemographic classification. As the classification is built, so the user would be able to view the classification online. The system would also show certain information messages regarding the variables used and their spatial coverage in order to give information to users of the methods and procedures used to build the classification. This would make this kind of system open and readily intelligible to the users. Moreover, users would be able to download the classifications build, and verify the results by applying their own statistical techniques, making the classifications available for any kind of public scrutiny.

9.4 CONCLUDING STATEMENT

Geodemographics are small area classifications that provide summary indicators of the social, economic and demographic characteristics of neighbourhoods. Geodemographics work on the concept that place and population are linked with each other. In their current form geodemographic classifications are created using static data sources and rather little user knowledge of the data and methods used. There is a very important requirement to create open geodemographic classifications where the data and the methods used to build the classifications are understood by the users and open to wider public scrutiny. One important avenue of research is to use new and innovative data sources to create classifications and allow users far greater control in creating the classifications using a desktop or an online tool.

This Ph.D. has been an effort to contribute to these challenging and wide-ranging requirements. These have included developing new data sources of 'names' and 'house price ethnicity' data, and developing new ways of visualising the former. Very considerable effort has been expended in investigating and evaluating the methodologies used to build geodemographic classifications. An enhancement of the traditional *k*-means clustering algorithm was required and this was achieved by the design and development of a parallel version of *k*-means algorithm (parallel *k*-means). Later, a software utility 'GeodemCreator' was designed and developed for creating national and local area geodemographic classifications. The software utility was tested by the creation of three geodemographic classifications, with evaluation presented in the form of case studies in the previous chapter.

This chapter has presented an overview of the PhD journey and suggested possible future research directions. An online data integration web service and an online tool for computing geodemographic classifications are envisaged as possible future research outcomes. It is hoped that wider efforts towards open geodemographics will benefit from the software tool 'GeodemCreator'. In the future, it is also possible that an online version of 'GeodemCreator' will be created as part of a wider effort to assist decision makers in understanding the patterns of consumption of services and goods in a full range of geographical settings.

REFERENCES

ACS. (2009). "Annual Schools Census."   Retrieved 18th June, 2009, from http://www.teachernet.gov.uk/management/atoz/a/annualschoolscensus/.

Adnan, M., Longley, P.A., Singleton, A.D., Brunsdon, C. (2010). "Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases". *Transactions in GIS*, 14 (3), 283-297.

Amazon (2010). "Amazon EC2 ". Retrieved 5th July, 2011, from http://aws.amazon.com/ec2/.

Aoidh, E. M., M. Bertolotto, Wilson, D.C. (2008). "Understanding geospatial interests by visualising map interaction behaviour". Information Visualisation. 7 (3-4), 275-286.

Apple (2011). "iCloud". Retrieved 5th July, 2011, from http://www.apple.com/uk/icloud/what-is.html.

Ashby, D. I., Longley, P. A. (2005). "Geocomputation, Geodemographics and Resource Allocation for Local Policing". *Transactions in GIS,* 9 (1): 53-72.

Batey, P., Brown, P. (1995). "From Human Ecology to Customer Targeting: the Evolution of Geodemographics". In Longley, P. and Clarke, G. eds *GIS for Business and Service Planning*. GeoInformation International, Cambridge, 77-103.

Birking, M., Clarke, G. (1988). "SYNTHESIS: A SYNTHetic Spatial Information System for urban modelling and spatial planning". Environment and Planning A*, 20*, 1645-1671.

Birkin. M., Clarke, G., Clarke, M. (2002). Retail Geography and Intelligent Network Planning. Wiley, Chichester.

Birkin, M., Clarke, G. (1998). "GIS, Geodemographics, and Spatial Modelling in the U.K. Financial Services Industry". *Journal of Housing Research*, 9 (1), 87-111.

Birkin, M., Clarke, G., Clarke, M., Wilson, A. (1996). Intelligent GIS: location decisions and strategic planning. GeoInformation International, Cambridge.

Brassington, F., Pettitt, S. (2006). Principles of Marketing. Pearson Education Limited, England.

Brunsdon, C. (2006). "A Cluster based approach to the zoning problem using and extended genetic algorithm". Proceedings of the GIS Research UK 14th Annual Conference, The University of Nottingham, United Kingdom, 5-7th April 2006.

Burnard, T. (2000). "Slave Naming Patterns: Onomastics and the Taxonomy of Race in Eighteenth-Century Jamaica" *Journal of Interdisciplinary History*, 31 (3), 325-346.

Burrows, R., Gane, N. (2006). "Geodemographics, Software and Class", *Sociology*, 40 (5), 793-812.

Busselle, J. (2008). "Raster Images versus Vector Images." Retrieved 15th January, 2009, from http://www.signindustry.com/computers/articles/2004-11-30-DASvector_v_raster.php3.

Caschera, M. C., F. Ferri, et al. (2008). "Interactive information visualisation in a conference location". *GeoVisualization of Dynamics, Movement and Change*. Workshop at the AGILE 2008 Conference (May 5, 2008, Girona, Spain)

Chappel, M. (2008). "Entity-Relationship Diagram." Retrieved 15th January, 2009, from http://databases.about.com/cs/specificproducts/g/er.htm.

Cheshire, J., Adnan, M., Gale, C. (2011). "The Use of Consensus Clustering in Geodemographics". Proceedings of the GIS Research UK 19th Annual Conference. 27th – 29th April, 2011. Portsmouth, University of Portsmouth.

Cheshire, J., Mateos, M., Longley, P.A. (2011). " Delineating Europe's Cultural Regions: Population Structure and Surname Clustering ", *Human Biology*, In Press.

Colantonio, S. E., G. W. Lasker, et al. (2004). "Use of Surname Models in Human Population Biology: A Review of Recent Developments", *Human Biology*, 75 (6), 785-807.

Copeland, L. (2008). "Testing UML Models." Retrieved 15th January, 2009, from http://www.stickyminds.com/sitewide.asp?Function=edetail&ObjectType=ART&ObjectId=6232.

Dantzker, M. L. and D. Freeberg (2003). "An Exploratory Examination of Pre-Employment Psychological Testing of Police Officer Candidates with a Hispanic Surname", *Journal of Police and Criminal Psychology*, 18 (1), 38-44.

Debenham, J., G. Clarke, Stillwell, J. (2001). "Deriving supply-side variables to extend geodemographic classification", *Cybergeo : European Journal of Geography*, 12th European Colloquium on Quantitative and Theoretical Geography.

Degioanni, A., A. Lisa, Zei, G., Darlu, P. (1996). "Italian surnames and Italian migration to France 1891- 1940", *Population*, 51 (6), 1153-1180.

Department of Communities and Local Government (2011). "Index Of Multiple Deprivation"   Retrieved 15th June, 2011, from http://www.communities.gov.uk/documents/statistics/pdf/1870718.pdf.

Ding, C., He, X. (2004). "K-means clustering via Principal Component Analysis". Italian surnames and Italian migration to France 1891- 1940". Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.

Experian. (2011). "Mosaic UK - unique consumer classification based on in-depth demographic data"   Retrieved 17th May, 2011, from http://www.experian.co.uk/business-strategies/mosaic-uk-2009.html.

Farr, M., Evans, A. (2005). "Identifying `Unknown Diabetics' using Geodemographics and Social Marketing". *Interactive Marketing* 7: 47-58.

Fernández, V., García Martínez, R., González, R., Rodriguez, L. (2005). "Genetic Algorithms Applied to Clustering", Buenos Aires, School of Engineering, University of Buenos Aires.

Gibin, M., Singleton, A., Milton,R., Mateos, P., Longley, P. (2008) An Exploratory Cartographic Visualisation of London through the Google Maps API. *Applied Spatial Analysis and Policy*. 1(2), 85-97.

Hall, J.D., Hart, J.C. (2004). "GPU acceleration of iterative clustering". In: ACM Workshop on General-Purpose Computing on Graphics Processors, p C-6

Harris, R., Sleight, P., Webber, R. (2005). Geodemographics, GIS and Neighbourhood Targeting. Wiley, London.

Harish, P., Narayanan, P., (2007). " Accelerating Large Graph Algorithms on the GPU Using CUDA". HiPC'07 Proceedings of the 14th international conference on High performance computing. pp. 197–208.

Hartley, T., Catalyurek, U., and Ruiz, A. (2008), Biomedical Image Analysis on a Cooperative Cluster of GPUs and Multicores. *Proceedings of the 22nd annual international conference on Supercomputing*, 15–25.

HESA. (2009). "Overview." Retrieved 18th June, 2009, from http: //www.hesa.ac.uk/index.php/content/view/4/54/.

Kaufman , L., Rousseeuw, P.J. (1990). Finding Groups in Data. New York: John Wiley & Sons, Inc.

Lagerberg, D., Magnusson, M., Sundelin, C. (2005). "Surname as a marker of ethnicity. A study from child health services shows that immigrant respective swedish families seem to be isolated in different ways". *Lakartidningen,* 102(30-31), 2145-8.

Lasker, G. W. (1985). Surnames and Genetic Structure, Cambridge University Press.

Lauderdale, D. S., Kestenbaum, B. (2000). "Asian American ethnic identification by surname". Population Research and Policy Review, 19, 283-300.

LCS. (2009). "Learning and skills council."   Retrieved 18th June, 2009, from http://www.lsc.gov.uk/aboutus/.

Leino, A., Mannila, H., Pitkanen, R. L. (2003). "Rule Discovery and Probabilistic Modeling for Onomastic Data", PKDD 2003, LNAI 2838, pp. 291–302.

Lewison, G., Kundra, R. (2008). "The internal migrantion of indian scientists, 1981-2003, from an analysis of surnames", SCIENTOMETRICS, 75 (1), 21-35.

London Data Store (2010). "London Data Store"   Retrieved 15th May, 2011 from http://data.london.gov.uk.

Longley, P. A. (2005). "Geographical Information Systems: a Renaissance of Geodemographics for Public Service Delivery". *Progress in Human Geography*, 29, 57-63.

Longley, P. A., Singleton, A. D. (2009). "Linking Social Deprivation and Digital Exclusion in England". *Urban Studies*, 46 (7), 1275-1298.

Longley, P. A., Singleton, A.D. (2009). "Classification through consultation: public views of the geography of the E-Society". *International Journal of Geographic Information Science*, 23 (6), 737-763.

Longley, P., Webber, R., Lloyd, D. (2007). "The Quantitative Analysis of Family Names: Historic Migration and the Present Day Neighbourhood Structure of Middlesbrough, United Kingdom", *Annals of the Association of American Geographers*, 97 (1), 31-48.

Longley, P. A., Goodchild, M. F., Maguire, D. J., Rhind, D. W. (2011). Geographic Information Systems & Science, Third Edition, Wiley.

Lucchetti, E., M. Tasso, Pizzetti, S., Caravello, G. U. (2005). "Surname distributions and linguistic-cultural identities in the Alto Adige-Siidtirol Area." *International journal of Anthropology*, 20 (3-4), 225-245.

Maulika, U., Bandyopadhyay, S. (2000). "Genetic algorithm-based clustering technique". *Pattern Recognition*, 33(9), 1455-1465.

Marr, T.R. (1904). Housing Conditions in Manchester and Salford. Manchester, Manchester University Press.

Mateos, P. (2007). An ontology of ethnicity based upon personal names: with implications for neighbourhood profiling. Geography. London, Univeristy College London.

Mateos, P., R. Webber, Longley, P. A. (2007). "The Clustural, Ethnic and Linguistic Classification of Population and Neighbourhoods using Personal Names". Centre for Advanced Spatial Analysis (UCL): London.

Office for National Statistics (2009). "Ness Data Exchange." Retrieved 3rd June, 2009, from http://www.neighbourhood.statistics.gov.uk/HTMLDocs/downloads/NeSS-Data-Exchange-Technical-Implementation-Guide-v1.0.pdf.

Oracle (2005). "Technical Comparison of Oracle Database 10g vs. SQL Server 2005: Focus on Performance." (2005).

Painho, M., Bação, F. (2000). "Using Genetic Algorithms in Clustering Problems". Proceedings of Geocomputation 2000, University of Greenwich, United Kingdom, 23 - 25 August 2000.

Petersen, J., Gibin, M., Longley, P. A., Mateos, P., Atkinson, P., Ashby, D. (2010). "Geodemographics as a toold for targetting neighbourhoods in public health campaigns". *Journal of Geographical Systems*, 13(2), 173-192.

Philips, J. C., Stone, J. E., Schulten, K. (2008). "Adapting a message driven parallel application to GPU-Accelerated clusters". SC '08 Proceedings of the 2008 ACM/IEEE conference on Supercomputing.

Reynolds, A. P., Richards, G., Rayward-Smith, V. J. (2004). "The Application of *K*-Medoids and PAM to the Clustering of Rules". Lecture Notes in Computer Science. 3177/2004, 173-178.

Reynolds, A.P., Richards, G., et al. (2006). "Clustering Rules. A Comparison of Partitioning and Hierarchical Clustering Algorithms". presented at J. Math. Model. Algorithms, 2006, pp.475-504.

Reeves, J. W. (2005). "Code as desgin." Retrieved 19th January, 2009, from http://www.developerdotstar.com/mag/articles/PDF/DevDotStar_Reeves_CodeAsDesign.pdf.

Shelton, N., Birkin, M., Dorling, D. (2006). "Where not to live: a geo-demographic classification of mortality for England and Wales, 1981-2000". *Health and Place*, 12 (4).

Singleton, A. D., Longley, P. A. (2009). "Creating Open Source Geodemographics - Refining a National Classification of Census Output Areas for Applications in Higher Education". *Papers in Regional Science*, 88(3), 643-666.

Singleton, A. D., Wilson, A.G., O'brien, O. (2010). "Geodemographics and Spatial Interaction: an integrated model for higher education". *Journal of Geographical Systems*, 1435-5930, 1-19.

Singleton, A.D. (2010). "The Geodemographics of Educational Progression and their Implications for Widening Participation in Higher Education". *Environment and Planning A*. In Press.

Sleight, P. (2004). Targetting Customers-How to Use Geodemographic and Lifestyle Data in Your Business.

Stroud, D. (2007). The 50 plus market: Why the future is age-neutral when it comes to marketing. Bell & Bain, Glasgow.

Takizawa, H., Kobayashi, H. (2006). "Hierarchical parallel processing of large scale data clustering on a pc cluster with GPU co-processing". *J. Supercomput*.,36 (3), 219–234.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–40.

Transport for London (2010). "Tfl Cycle Hire API" Retrieved 15th May, 2011 from http://bike-stats.co.uk.

UCAS. (2009). "About us." Retrieved 18th June, 2009, from http: http://www.ucas.ac.uk/about_us/whoweare/.

Voracek, M. and G. Sonneck (2007). "Surname study of suicide in Austria: Differences in regional suicide rates correspond to the genetic structure of the population" *The Middle European Journal of Medicine*. 119 (11-12), 355-60.

Vickers, D.W., Rees, P.H. and Birkin, M. (2005). "Creating the National Classification of Census Output Areas: Data, Methods and Results". *Working Paper 05/2,* School of Geography, University of Leeds, Leeds

Vickers, D.W., Rees, P.H. (2007). "Creating the National Statistics 2001 Output Area Classification". *Journal of the Royal Statistical Society, Series A*, 170(2), 379-403.

Voas, D., Williamson, P. (2001). "The diversity of diversity: a critique of geodemographic classification". *Area*, 33(1), 63 – 76.

Webber, R. (1975). Liverpool Social Area Study, 1971 Data: PRAG Technical Paper 14, London.

Webber, R., Craig, J. (1978). "A Socio-Economic Classification of Local Authorities in Great Britain". HMSO, London.

Weiss, M. (2000). The Clustered World. Boston, MA: Little, Brown and Co.

Wilson, A. G. (1971). "A family of spatial interaction models and associated developments."

Yuan, Y. (2007). "Three Hundred Big Names 1".

Yuan, Y. (2007). "Three Hundred Big Names 2".

Yuan, Y. (2007). "Three Hundred Big Names 3".

APPENDIX 1: LIST OF PUBLISHED OUTPUTS FROM PHD

The following article and conference papers are a direct output of this PhD research.

(a) Referred Academic Journals

Adnan, M., Longley, P.A., Singleton, A.D., Brunsdon, C. (2010). Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases, Transactions in GIS, 14(3), 283-297.

(b) Working Papers

Adnan, M., Singleton, A.D., Longley, P.A. (2010). Developing efficient web-based GIS applications. CASA working paper 153. London: Center for Advanced Spatial Analysis.

(c) Conference Papers

Adnan, M., Singleton, A.D., Longley, P.A. (2011). “Building Geodemographics on parallel graphics processing unit architecture". Proceedings of the 11th International Conference on Geocomputation. 20th – 22nd July, 2011. London, University College London.

Cheshire, J., Adnan, M., Gale, C. (2011). “The Use of Consensus Clustering in Geodemographics”. Proceedings of the 11th International Conference on Geocomputation. 20th – 22nd July, 2011. London, University College London.

Adnan, M., Singleton, A.D., Longley, P.A. (2011). “The Specification, Estimation, and Testing of Real-time Geodemographics using Parallel Graphics Processing Unit

Architecture". Proceedings of the GIS Research UK 19th Annual Conference. 27th – 29th April, 2011. Portsmouth, University of Portsmouth.

Cheshire, J., Adnan, M., Gale, C. (2011). “The Use of Consensus Clustering in Geodemographics”. Proceedings of the GIS Research UK 19th Annual Conference. 27th – 29th April, 2011. Portsmouth, University of Portsmouth.

Adnan, M., Singleton, A.D., Longley, P.A. (2010). “Parallel K-means clustering using Graphical Processing Units for the Geocomputation of Real-time Geodemographics”. Proceedings of the GIS Research UK 18th Annual Conference. 14th – 16th April, 2010. London, University College London.

Adnan, M., Singleton, A.D., Longley, P.A. (2009). “Real-Time Geodemographics: Requirements and Challenges”. Proceedings of the RGS-IBG Annual International Conference. 14th – 16th September, 2009. Manchester, University of Manchester.

Adnan, M. (2009). “Real-Time Geodemographics: Requirements and Challenges”. Workshop on Spatio-Temporal Analysis of Network Data and Road Developments. London, University College London.

Adnan, M., Singleton, A.D., Brunsdon, C., Longley, P.A. (2009). “Moving to real time segmentation: Efficient computation of geodemographic classification”. Proceedings of the GIS Research UK 17th Annual Conference. 1st – 3rd April, 2009. Durham, University of Durham.

Adnan, M., Singleton, A.D., Longley, P.A. (2009). “Real-time Geodemographics”. Center for Advanced Spatial Analysis (CADA) S4 Conference, 9th January, 2009. London, University College London.

(d) Prizes and Awards

- Won prize for the “Best paper by a young researcher” for the paper on “Moving to real time segmentation: Efficient computation of geodemographic classification” at GISRUK 2009 conference.

- Won "Best Student of the year Award" from UCL ESRI Developer Center (CDC), 2011.

APPENDIX 2: PREFIXES

Following is the list of all the prefixes used for cleaning the names:

Mrs
Mr
Ms
Miss
Dr
AVM
Absgsg
Adm
Amb
AMN
Archbishop
Baroness
Brig
Gen
Brigadier
Bro
Capt
Cardinal
Cmdr
CMSGT
Col
Consul
Countess
Count
Cpl
CPO
CWO
Father
Gov
Hon
Lord
Lt
Master
Mlle
Mme
Mother
MSGT
PFC
Princess

Prince
Prof
Radm
Rev
Sgt
Sir
Sister
SMSGT
Speaker
Squad
Ldr
SrA
Srta
Sr
SSGT
Swami
TSGT
Vadm
Jr
MD
PhD
Ret
DC
DD
DDSPA
DMD
DO
DPM
DVM
ESP
Filho
MDPA
MFCC
ND
ODPC
OFM
OBS
OD
PA
PC
PRS
QC
THD
VMD

VP
GbR
co
BV
RETD
COMDR
SUB
BSNL
SMT
CDR
GENL
AVSM
PT
ing
ATT
VAGT
VVS
CIV
ANESTHESIOLOGY
DIST
JUDGE
RES
CHILDREN
PHONE
LINE
TELEPHONE
CPA
OFC
OPTMTRST
LTCOL
CHILDRENS
STEETMAN
PLACE
CLINICAL
PSYCHOLOGIS
PSYCHOLOGIST
DNTST
PODIAATRST
PODIATRST
CONSLTNT
CONSULTANT
GEOLOGIST
GEOLOGST
DDS

TEENAGERS
TEEN
ACCT
ATTY
AUCTNEER
EXOTIC
HOUSE
EAST
EST
NORTH
SOUTH
GEOL
INS
INVSTMTS
OIL
PRODCR
PRODUCER
PHY
RURAL
CHILDRWNS
FACSIMILE
FAX
FACSIMILIE
TN
TEENS
MOBILE
CABIN
RANCH
II
III
IV
REV
FOREMAN
USN
MGR
MANAGER
EST
PHYS
STK
SCLPTR
CAP
STDY
CARETAKER
HEADQUARTERS

HSE
CLNG
ASSN
INC
MACHINE
FACSIMLE
CPT
PUB
LODGE
APPARTEMENTS
CASTHOF
VETTERL
DIPLING
DIPLVW
DRJUR
DRMEDUNIV
JUR
MAG
DOLM
PRIM
DORFGASTHAUS
DIGROWERBUNG
DRMED
Filho
Filho>Suff
JUN
SEN
JR
SR
GASTHAUS
ZUR
POST

APPENDIX 3: SUFFIXES

Following is the list of all the suffixes used for cleaning the names:

Jr
Sr.
II
III
M.D.
Ph.D.
(Ret.)
C.P.A.
D.D.S.
DC
DD
DDSPA
DMD
DO
DPM
DR
DVM
ESP
Esq.
Filho
I
INC
ITF
IV
IX
MDPA
MFCC
MS
ND
Neto
O.D.P.C.
O.F.M.
OBS
OD
P.A.
PC
PRS
Q.C.
R.N.

Sobrinho
THD
V
VMD
VP
JUN.
SEN.
JR.
SN.

APPENDIX 4: UNICODE TO ASCII CONVERSION TABLE

Following is the tale of all the Unicode characters and their corresponding Ascii characters used for the conversion.

Unicode Characters	Ascii Characters
A	A
À	A
Â	A
Ä	A
Å	AA
Æ	AE
Ç	C
Ð	D
É	E
È	E
È	E
È	E
Ê	E
Ë	E
Í	I
Ì	I
Î	I
Ï	I
Ñ	N
Ó	O
Ò	O
Ô	O
Ö	OE
Õ	O
ß	SS
Ú	U
Ù	U
Û	U
Ü	UE
Ý	Y

APPENDIX 5: SEPARATING STANDARDIZED NAMES INTO PARTS

Following is a generalised Java source code for separating names into parts based on special characters. Comments are preceded by '//' and are in green color.

```
Vector nameParts=new Vector();
 StringBuffer sbParts=new StringBuffer();

 for(int i=0;i<name.length();i++)
  {
    char ch=name.charAt(i);

    if(ch==' &' || ch==' +' || ch=='- ' || ch=='_ ')  //All the special characters go here
       {
         nameParts.add(sbParts.toString());
         sbParts=new StringBuffer();
       }
     else
       {
          sbParts.append(ch);
       }
   }

  nameParts.add(sbParts.toString());
```

APPENDIX 6: 'R' SOURCE CODE FOR PRODUCING MAP VISUALISATION

Once 'GeodemCreator' has produced the geodemographic classification, following 'R' source code could be used for producing a map visualization. This source code is also included in the CD (which accompanies this thesis) with the file name 'Generate_Shape.R'.

Note: Comments are preceded by '#' and are in green color.

```
setwd("")  #set working directory
library(maptools) #you should have maptools library installed
ukmap<- readShapePoly("uk.shp", IDvar="LABEL") #reads in a shape file
CS<-read.csv("output.csv", sep=",", header=T) #The output of 'GeodemCreator'
colours <- rainbow(8)
brks<-c(1,2,3,4,5,6,7,8)
colours[1] <- "#FDC086" #change colour codes if needed
colours[2] <- "#BEAED4"
colours[3] <- "#FFFF99"
colours[4] <- "#386CB0"
colours[5] <- "#7FC97F"
colours[6] <- "#A6CEE3"
colours[7] <- "#E31A1C"
colours[8] <- "#E41A1C"
cluster<-as.data.frame(CS[2])
str1<-CS[rownames(cluster),]
rownames(cluster)<-str1[,1]
mapdata<-match(ukmap$LABEL, rownames(cluster))
mapdata1<-cluster[mapdata,]
mapdata1frame<-as.data.frame(mapdata1)
rownames(mapdata1frame)<-ukmap$LABEL
names(mapdata1frame)
join<-spCbind(ukmap, mapdata1frame)
png(filename="RESFILE.png", width=1000, height=1000)
plot(join, col=colours[findInterval(join$mapdata1,brks,all.inside=FALSE)],lty=0
,axes=F)
legend("topleft", ,c('Cluster 1','Cluster 2','Cluster 3', 'Cluster 4','Cluster 5',  'Cluster
6','Cluster 7'), cex=1.5, col=colours, lty=5, lwd=13, bty="n")
box()
dev.off()
```

APPENDIX 7: HOW TO INSTALL 'GEODEMCREATOR'

A user needs to follow following steps to install 'GeodemCreator'. These steps are for installing 'GeodemCreator' on a Windows operating system. The minimum requirement for running 'GeodemCreator' is 2 Giga Hertz of processor and 3 Gigabytes of memory.

This 'How to Install' guide is also given in the CD, that accompanies this thesis, in the file called 'How_To_Install.txt'.

- Install Java 1.5 or higher. Java can be downloaded free of charge from the web link http://www.java.sun.com.

- Install 'R'. 'R' can be downloaded free of charge from the web link http://www.r-project.org/.

- Once 'R' is installed, open 'R' and install 'rJava' library. This could be done by choosing Packages -> Install Packages. Chose the appropriate 'CRAN mirror' from the window and then chose 'rJava' from the packages list.

- Note down the path, where 'rJava' is installed. There will be a folder called 'jri' (Java to R Interface) in the directory where 'rJava' has installed.

- Set the environment variables.

    - In the 'jri' folder, there will be two files 'JRI.jar' and 'jri.dll'. Path for both the 'JRI.jar' and 'jri.dll' should be in the 'PATH' environmental variable of the system.
    - If the windows has a 64-bit version, the 'jri' folder will have two more folders 'i386' and 'x64'. Both of these folders contain 'jri.dll'. 'i386' has the 32-bit version and 'x64' has the 64-bit version. Depending on the type of java installed (64 bit or 32 bit), the 'PATH' environment variables should be set to one of these folders (i.e. 'i386' or 'x64').
    - A detailed instruction of installing 'rJava' and 'jri' can also be found at the link http://www.rforge.net/JRI/.

- Copy and paste folder 'GeodemCreator' provided in the CD on your computer.

- Once this is done, run the exe file 'geodem.exe' from the folder 'GeodemCreator'.

APPENDIX 8: 'GEODEMCREATOR' SOURCE CODE

All the source code of the 'GeodemCreator' software is given in the CD, that accompanies this thesis, in the folder called 'Source_Code'.

Following are the names of the Java classes given in the CD.

- Main
- ModesPanel
- BasicModePanel
- BuildBasicClassification
- WriteFileHelper
- WriteFile
- LogoPanel
- ReadFile
- LoadFilePanel
- SpecifyVariablesPanel
- SpecifyClusteringInput
- BuildAdvancedClassification
- AdvancedWriteFileHelper
- FileWaitPanel
- ProcessTablePanel

APPENDIX 9: PEER REVIEWED PUBLICATION

Adnan, M., Longley, P.A., Singleton, A.D., Brunsdon, C. (2010). Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases, *Transactions in GIS*, 14(3), 283-297.