Generating Vague Geographic Information through Data Mining of Passive Web Data



Paul Brindley

School of Computer Science

University of Nottingham

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

July 2016

Abstract

Vagueness is an inherent property of geographic data. This thesis develops a geocomputational method that demonstrates that vague information has the potential to be incorporated within GIS in straightforward manner. This method applies vagueness to the elements of place: *types*, *names* and *spatial boundaries*, generating vague geographic objects by extracting and filtering the differing opinions and perceptions held within web derived data. The aim of the research is threefold: (1) to investigate an approach to automatically generate vague, probabilistic geographical information concerning place by mining differing perspectives from passive web data; (2) to assure the quality of the vague information produced and test the hypothesis that its results are indistinguishable from directly surveying public opinion; and (3) to demonstrate the value of integrating vague information into geospatial applications via examples of its use.

To achieve the first aim, the thesis develops methods to extract differing perspectives of place from web data - constructing (i) vague place type settlement classification and (ii) vague place names *and* boundaries for 'neighbourhood' level units. The methods developed are automated, suitable for generating output at a national scale and use a wide range of different source data to collect the differing opinions.

The second aim assesses the quality of the data produced, determining if output extracted from the web was representative of that obtained from asking people directly. Statistical analysis of regression models demonstrates that data were representative of that collected through asking people directly both for vague settlement classifications and vague urban locale boundaries. Importantly, the validation data, drawn from public opinion, also supported the notion that vagueness was omnipresent within geographic information concerning place.

The third aim was addressed through the use of case studies in order to demonstrate the added value of such data and subsequent integration of vague geographic objects within other socio-economic data. Critically, the incorporation of vagueness within place models not only add value to geographic data but also improve the accuracy of real-world representations within GIS.

List of Published Papers

The following list contains work from the thesis that has been published:

1. Brindley, P., Goulding, J. and Wilson, M.L. (2014) Mapping Urban Neighbourhoods from Internet Derived Data. In Proceedings of the 22nd GIS Research UK (GISRUK) Conference (Glasgow, UK, April 16-18th), pp. 355-364.

2. Brindley, P., Goulding, J. and Wilson, M.L. (2014) A data driven approach to mapping urban neighbourhoods. In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (Dallas, Texas, USA, November 4-7th), pp. 437-440.

Acknowledgements

There are a number of people without whom this thesis would not have been possible. I would like to acknowledge my supervisors: Max L. Wilson and James Goulding. Their advice and support during this lengthy process has been greatly appreciated. I am grateful for their continued support and encouragement. They have guided me down paths I would otherwise not have taken and have always been to hand at times of need. I would also like to thank Gary Priestnall and Professor Chris Jones for agreeing to be my internal and external examiners respectively.

I wish to thank the funders of this PhD (Horizon CDT and the Horizon Digital Economy Research Institute) without which I could not have even entertained the notion of undertaking my doctoral studies[1]. I am grateful to the Research Team at Ordnance Survey for kindly supplying data to help with the research and also for providing my internship opportunity that allowed me to gain valuable experience applying my proposed methods in a wider context.

Thank-you to all those people who have made contributions (large and small) and helped form and shape this thesis. This is especially true of those who have helped the validation of my methods by providing their views on vague spatial units.

Last but by no means least - I am indebted to my family. My sister, Lisa for helping distribute additional questionnaires to support the validation process. To my parents whose example has taught me to work hard for all that I aspire to achieve in life. To my daughter, Hannah, for putting up with the endless hours spent working when I suspect that she would have preferred I was on the Wii! To our cat Wispa for sleeping beside me whilst I have typed away. Finally, I should acknowledge the innumerable sacrifices made by my wife, Sara who has not only kept me fed and watered during the process but has also provided me the necessary support throughout the process.

To one and all, I thank-you...

[1] Paul was supported by the Horizon Centre for Doctoral Training at the University of Nottingham (RCUK Grant No. EP/G037574/1) and by the RCUK's Horizon Digital Economy Research Institute (RCUK Grant No. EP/G065802/1)

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. In line with the stipulations of the University of Nottingham, this thesis contains fewer than 100,000 words and 300 pages (including appendices, bibliography, footnotes, tables and equations).

Paul Brindley
July 2016

I would also like to acknowledge the copyright and source of the data contained within many of the figures. The following figures all contain OS data © Crown copyright (and database right) 2015: 3.3, 5.5-5.9, 6.1-6.3, 6.7, 6.9, 6.11-6.14, 7.1-7.4, 7.6, 7.13-7.14, 7.16-7.17, 8.1-8.7, 8.9, B.1, D.1-D.10 and H.1-H.4

Table of contents

List of figures **xi**

List of tables **xiv**

1 Introduction: Project Outline, Aims and Objectives **1**

1.1 Project outline 1
1.2 Aims and objectives 3
1.3 Structure of document 5

2 Vague Geographies **8**

2.1 Introduction 8
2.2 Defining vagueness within Geographic Information 8
2.2.1 Natural vagueness 10
2.2.2 Multiperson disagreement 11
2.2.3 Vagueness within GISc 13
2.2.3.1 Spatial and temporal vagueness 15
2.3 Modelling vagueness in Geographic Information 16
2.3.1 Fuzzy set theory 16
2.3.1.1 Geographic fuzzy vague objects 17
2.3.1.2 Critical reflection 18
2.3.2 Probabilistic approach 19
2.3.3 Fuzzy-bilistic: the differences between probabilistic and fuzzy 20
2.3.4 Alternative models for vague GI 21
2.4 Towards vague places 22
2.5 Using web-based passive crowdsourcing to model vague places 23
2.6 Summary 26

3 Vague Attributes: Settlement Classification **28**

3.1 Introduction 28
3.2 Settlement classification 29
3.2.1 Literature concerning existing settlement classifications . . . 29
3.2.2 Limitations of current settlement classifications 30
3.2.3 Vague settlement classification 32

3.2.4 Passive crowdsourcing . . . 33
3.3 Method . . . 35
3.3.1 Description of rules used . . . 36
3.3.2 Case study areas . . . 37
3.3.3 Wikipedia bias and data cleaning . . . 39
3.3.4 Validation method . . . 39
3.4 Results and discussion . . . 41
3.4.1 Language . . . 41
3.4.2 URL classification . . . 44
3.4.3 Synonymy and related concepts within settlement terms . . . 47
3.4.4 Relationship with population . . . 47
3.4.5 Validation . . . 51
3.4.6 Fuzzy-babilistic . . . 56
3.5 Discussion and settlement classification conclusions . . . 57

4 Vague Boundaries: Urban Locales - Existing work 60
4.1 Introduction . . . 60
4.2 'Hoods vs 'Burbs: towards a definition of urban locales . . . 61
4.2.1 Neighbourhoods . . . 61
4.2.2 Suburbs . . . 63
4.2.3 Urban locales . . . 64
4.3 The motivation for urban locales . . . 65
4.4 Current modelling methods . . . 70
4.4.1 Resident mapping . . . 71
4.4.2 Space syntax . . . 72
4.4.3 Web derived data . . . 73
4.4.4 Web derived data: Clustering social media data . . . 74
4.4.5 Web derived data: Cross referencing with existing gazetteers . 75
4.4.5.1 Obtaining point data . . . 75
4.4.5.2 Bounding points . . . 79
4.4.5.3 General limitations of cross referencing with existing gazetteers . . . 82
4.5 Summary . . . 84

5 Building Vague Urban Locales 85
5.1 Towards an urban locale framework . . . 85
5.2 Detailed method . . . 87
5.2.1 Data collection . . . 87
5.2.2 Misspelling identification . . . 92
5.2.3 Urban locale derivation . . . 94
5.2.3.1 Pre-filtering rules . . . 95

5.2.3.2 Urban locale surface generation 96
5.2.4 Linear features as breakline constrains for smoothing 97
5.2.5 Post-processing: Synonym identification 98
5.2.6 Effective storage and distribution 99
5.3 Parameter justification 100
5.3.1 KDE 100
5.3.2 KDE: Bandwidth 101
5.3.3 KDE: Cell size 105
5.3.4 Data point frequency cut-off 105
5.3.5 Adaptive parameters 105
5.4 Summary 108

6 Results and Validation of Vague Urban Locales 109
6.1 Results 109
6.1.1 Spatial boundary results 111
6.1.2 Place name results: Synonyms and misspellings 114
6.2 Validation 114
6.2.1 Comparison of urban locale names 118
6.2.1.1 Gazetteer disparity 119
6.2.1.2 Comparing web extraction output 121
6.2.2 Comparison of urban locale geography 125
6.2.2.1 Comparisons with areal data 126
6.2.2.2 Stakeholder views of urban locale output 130
6.2.2.3 Comparison with samples of residents' views . . . 132
6.3 Summary 142

7 Analyses of Vague Urban Locales 146
7.1 Web groupings 147
7.1.1 Web groupings - method 148
7.1.2 Web groupings - validation 149
7.1.3 Web groupings - results and discussion 150
7.2 Temporal aspects 155
7.2.1 Short-term change: Monthly 156
7.2.2 Long-term change: Two year span 158
7.3 Tying other data to urban locale units 161
7.3.1 Joining point data 161
7.3.2 Joining polygon data 162
7.4 Homogeneity 168
7.5 Comparison with underlying population geography 172
7.5.1 Comparison with all data returns 173
7.5.2 Comparison with urban locale data returns 173

7.6 Differing perspectives . . . 176
7.6.1 Differences within a specified urban locale . . . 177
7.6.2 The extent of differing perspectives . . . 178
7.6.3 Urban locale inter-relationships . . . 181
7.7 Summary . . . 184

8 Discussion and Case Study Application of Vague Urban Locales 186
8.1 Discussion . . . 186
8.1.1 Advantages of method . . . 186
8.1.1.1 Data volumes . . . 187
8.1.1.2 Use of percentage data . . . 189
8.1.1.3 Linear features as breakline constrains for smoothing 190
8.1.1.4 Other benefits . . . 193
8.1.2 Vague place names: Identification of misspellings and synonyms 193
8.1.3 Fuzzy-babilistic . . . 196
8.1.4 Limitations of method . . . 197
8.1.5 Synopsis . . . 201
8.2 Case study applications . . . 202
8.2.1 Identification of case study boundaries for research . . . 202
8.2.2 Local community groups case study . . . 204
8.2.3 Neighbourhood planning case study . . . 205
8.3 Further research . . . 208
8.3.1 Further development of the method . . . 208
8.3.2 Re-application of the work . . . 210
8.4 Summary and urban locale conclusions . . . 211

9 Conclusions: Generating Vague Places from web data 215
9.1 Introduction . . . 215
9.2 Summary of research findings . . . 215
9.3 Contributions of this research . . . 223
9.3.1 Vague attributes for settlement classification . . . 223
9.3.2 Vague place names for urban locales . . . 224
9.3.3 Vague spatial boundaries for urban locales . . . 224
9.4 Final remarks . . . 225

References 227

Appendix A Comparison of settlement classification data between web extraction method and questionnaire responses 241

Appendix B Data coverage using different KDE bandwidths 248

Appendix C Exploring domain name frequencies of urban local data 253

Appendix D Data output for all case-studies 257

Appendix E Levels of urban locale name agreement between different data-sources 262

Appendix F Web grouping: ancillary information 267

Appendix G Homogeneity testing: ancillary information 270

Appendix H Relationship with underlying population 277

Appendix I Relationship with Flickr 283

List of figures

2.1 Conceptual model of uncertainty within spatial data 9
2.2 Ambiguous images of everyday objects 12
2.3 Ambiguous images 13
2.4 Weekendness: concept of fuzzy set theory 17
2.5 Example of probabilistically defined vague boundaries 20
2.6 Concept of Place 23

3.1 Roberts' Settlement Terms and Population Thresholds 31
3.2 The Case Areas of 47 unitary authorities, districts and London Boroughs 38
3.3 Differing extents for Broomhill, Sheffield 49
3.4 Average Related Population for Selected Settlement Classification Terms 51
3.5 Robustness test: Comparing output for all websites and only those with less than 50 data returns 52
3.6 Comparison between web extraction output and questionnaire responses for selected settlements 53
3.7 Relationship between questionnaire output and web extraction data . . 56

4.1 Illustrative example of the scale of various administrative areas 68
4.2 Twelve Resident views of Hamstead Garden Suburb Boundaries . . . 72
4.3 Space syntax: Maiden Lane housing estate 73
4.4 Hoodsquare zone identification through Foursquare check-ins 74
4.5 The spatial footprint of Flickr photograph tags 77
4.6 The spatial footprint of Mid-Wales 83

5.1 Urban locale evolution 86
5.2 Identifying urban locale terms between street and settlement content . 88
5.3 Case study settlements 91
5.4 Case study settlements: Relationship with population size 92
5.5 Pre-processing: conversion of aggregated linear features to polygons . 98
5.6 Linear features as breakline constrains for smoothing: Example of Radford, Nottingham 99
5.7 Differences between KDE and simple moving average: number of postcodes, Sheffield 101

5.8 Spreading of data beyond coherent areas: within Morpeth 103
5.9 Spread of data coverage using different KDE bandwidths 103

6.1 Example data output (percentage of data) 113
6.2 Example data output: Urban Locales in Sheffield 114
6.3 Example dataoutput: Urban Locales in and around Camden and Westminster, London 116
6.4 Example of urban locales exported to KMZ for easy distribution to non-GIS users 117
6.5 Venn diagram of Nottingham Urban Locales by datasource 119
6.6 Venn diagram of Sheffield Urban Locales by datasource 120
6.7 Comparing the geography of Highfield urban locale, Sheffield 127
6.8 Exploring the edge of an urban locale with a chair of a local community group 131
6.9 Approximate location of resident perspectives of home urban locale name within Nottingham 134
6.10 Comparison of resident views of urban locale and different sources of urban locale geography 135
6.11 Raw point data collected by Crespmoor Neighbourhood Forum 137
6.12 Comparing the geography of resident perceptions for an urban locale with web extraction output 139
6.13 Level of geographic agreement between resident perceptions for a single urban locale with a range of other data output 140
6.14 Study areas for testing diversity of opinion in selected streets 143

7.1 Example map of web categorisation: The Park, Nottingham 151
7.2 The degree of support between the three different web groupings for Radford and Broxtowe, Nottingham 152
7.3 Example map of web categorisation: Broxtowe, Nottingham 154
7.4 Urban locale data reproduction from four different monthly data runs . 158
7.5 Map of temporal change: majority function 159
7.6 Example of temporal change in Manor Park urban locale, Sheffield . . 161
7.7 Term frequency–inverse document frequency (tf–idf) of Twitter sentiment for two urban locales in Sheffield 162
7.8 Areal interpolation: area weighting and point-in-polygon techniques . 164
7.9 Areal interpolation example for Output Area E00069672 166
7.10 Example Census of Population output 167
7.11 Example of homogeneity for a single Census variable (Percentage of people aged 30-44) within Sheffield urban locales 170
7.12 Comparing variations of homogeneity with differently sized areal units 171

7.13 Residuals for the relationship between the total number of data returns and numbers of residential and non-residential properties 174
7.14 Residuals for the relationship between named urban locale data and numbers of residential and non-residential properties 176
7.15 CMYK color model 178
7.16 Cyan-Magenta-Yellow-Key (CMYK) composite for example urban locales in London 179
7.17 Geographic footprint of the differing perspectives of urban locale: the number of different locales existing and potentially overlapping . . . 181
7.18 Urban locale relationships for Beeston, Nottingham 183

8.1 Differences between absolute and percentage returns 190
8.2 Linear features as breakline constrains for smoothing: Example of Castle Marina, Nottingham 191
8.3 Cyan-Magenta-Yellow-Key (CMYK) composite to explore the impact of using linear features as breaklines: for the centre of Nottingham . . 192
8.4 Voids due to sparse data if all A-roads are considered as breaklines: Example of Hockley, Nottingham 192
8.5 Synonym identification: the case of 'Pottington business estate' and 'Pottington industrial estate' urban locales in Barnstable 195
8.6 Synonym identification: the case of 'Archway' and 'Upper Holloway' urban locales in London 195
8.7 Defining urban areas 199
8.8 Defining Finsbury Park for a case study area 203
8.9 Example of aiding Neighbourhood Planning - designation of Neighbourhood Areas 207
8.10 A framework for the use of urban locales within postal addresses in differing countries 211

List of tables

1.1 Thesis Outline 5

3.1 Open Street Map Classification Recommendations 31
3.2 'Break terms' used in the method 37
3.3 Examples of records not meeting the rule criteria 38
3.4 Examples of records with no settlement classification reference for the search query 40
3.5 Examples of web based settlement classification for selected settlements to the north of Leicester 42
3.6 The most common settlement terms extracted from the Bing API request data 43
3.7 The most commonly cited settlement words (splitting up the terms into individual words) 45
3.8 The most common domain names posting settlement classification from the web extraction data 46
3.9 Semantic Content: Relationship between Settlement Terms 48
3.10 Settlement Terms and Household Population 50
3.11 Comparison of Settlement Classification between Questionnaire Responses and Web Extraction data 55

5.1 Example spelling comparisons 94
5.2 Data coverage using differing bandwidths 104
5.3 Examples of places excluded on the basis of having fewer than 40 datapoints 106

6.1 Summary statistics for data output 110
6.2 Domain name frequencies for selected case study areas 112
6.3 Examples of output cleaning within the 'spelling' identification process 115
6.4 Examples of output from synonym identification 115
6.5 Number of Urban Locales identified by different sources 119
6.6 Comparability of web extraction output with other gazetteers with varying level of agreement between the datasources 122

6.7 Urban locales identified by web extraction output but that were not included in other gazetteers . . . 124
6.8 Urban locales not identified by web extraction output . . . 124
6.9 Comparability of urban locale output for the different gazetteers with varying level of agreement between the datasources . . . 125
6.10 Average number of cells identified by the source for the 19 comparator urban locales with agreement with geographies from other data sources 129
6.11 Number of cells not in agreement with any of the other four sources . 129
6.12 Percentage of cells 'missed' by the method where evidence from the other four sources would suggest the locale's presence . . . 129
6.13 Combining ranks from Table 6.10, Table 6.11 and Table 6.12 . . . 130
6.14 Cases where resident view matched with web extraction data but not where data were in majority agreement . . . 133
6.15 Comparing data volumes for Neighbourhood Forum resident views and web extraction output . . . 138
6.16 Comparing resident boundaries for a selected urban locale with varying other datasource methods . . . 141
6.17 Testing the level of variation within the urban locale name in selected streets, Sheffield . . . 143

7.1 Validation of web group assignment using manual checking . . . 150
7.2 Web grouping classification output . . . 150
7.3 Web grouping classification z-scores for selected urban locales in Nottingham . . . 153
7.4 R-squared output from linear grid regression for web grouping classification, selected urban locales in Nottingham . . . 155
7.5 The number of times data records occurred within the four monthly runs (%) . . . 156
7.6 Level of agreement with monthly data for selected urban locales in Sheffield . . . 157
7.7 Temporal change: change in urban locale majority . . . 160
7.8 Urban locale change, Sheffield 2012-2014 . . . 160
7.9 House Prices (June 2012-July 2013) for selected areas in Nottingham 163
7.10 Change in Census 2001 to 2011 for selected areas in Sheffield . . . 165
7.11 Size of source datasets used within testing homogeneity of units . . . 169
7.12 Regression output: relationship between all data and delivery points: Sheffield . . . 173
7.13 Regression output: relationship between named urban locale data and delivery points: Sheffield . . . 175
7.14 The fifteen highest and lowest average regression residuals when aggregated to urban locales . . . 177

7.15 Degree of differing perspectives of urban locale: within Sheffield and Nottingham 181
7.16 Degree of differing perspectives: The 15 urban locales in Nottingham with the greatest level of overlap 182
7.17 Degree of differing perspectives: The 15 urban locales in Sheffield with the greatest level of overlap 183

8.1 Comparing data volumes with differing approaches 188
8.2 Comparing approaches to perceiving space 196

Chapter 1

Introduction: Project Outline, Aims and Objectives

1.1 Project outline

Vagueness is an inherent property of geographical data (Fisher, 2000). Defined as something that is not clearly expressed or is stated in indefinite terms, vagueness is omnipresent in nearly all human phenomena. It has been described as both a blessing and a curse for Geographic Information Science (GISc). A curse, for systems are not designed to handle vagueness but at the same time providing the opportunity to devise innovative ways in order to help overcome the issue (Davies, 2009). This thesis demonstrates that vagueness can in fact be relatively simply embraced within GISc. Geographic Information Systems (GIS) build *models*, or simplifications, of the real world. The critical question is how realistic those models should be? We live in a complex world that includes vagueness - to be as representative as possible, our geographic models should, when suitable, also incorporate elements of vagueness. This thesis contributes to this debate by demonstrating the general need for vague geographic information - specifically within settlement classification and neighbourhood level places.

Vagueness in GI (Geographic Information) can transpire within either any attribute or categorisation of the entity (is this object a city or a town, a mountain or a hill, a wood or a thicket?) or the spatial footprint of an object (where is the edge of the city, mountain or woodland?). The investigation of vagueness within GI is not new - see for example work by Fisher (1999, 2000, 2003), Evans and Waters (2007), or Hall and Jones (2008) which are discussed in more detail in the following chapter. What the existing literature lacks, however, is that it does not explicitly consider an

approach to model vagueness that is automated, at a national scale and that use a wide range of different source data to collect differing opinions. This thesis addresses this problem by building on the work of Jones et al. (2008a); Pasley et al. (2008); Purves et al. (2005b, 2007) in order to extract information from the World Wide Web (or web) in order to construct vague places (both in terms of attribute categorisation and spatial boundaries). In doing so, it generates both a vague settlement classification (via probabilistic attributes or 'place types') and vague neighbourhood level spatial boundaries, without prior knowledge of the named places themselves.

Settlement classifications are a useful tool for policy purposes and statistical reporting (Bibby and Shepherd, 2004). The motivation for vague settlement classifications is that whilst existing typologies frequently offer a top down approach whereby data are categorised into distinct classes based on similar attributes, they do not usually allow an object to be in more than one category or class. Different people, however, will inevitably perceive objects in a variety of ways. Under such principles, objects are likely to be viewed in multiple classes and could be considered with a probability of belonging to the class. For example, a settlement might be considered by 60% of people to be a small town but as a large village by others (40%). A vague settlement classification will more accurately reflect how people perceive places.

The term '*urban locales*' was coined in this work to represent all neighbourhood level elements, encompassing any sub-division of urban towns or cities (such as neighbourhoods and suburbs). The motivation for vague urban locale boundaries is that, whilst such units are considered as "the building blocks of public services society" (Pickles, 2010), no set of official boundaries exist for Great Britain. This is due to the subjective nature of the data and the difficulties in collecting such information. Administrative units are not the same as the neighbourhood units that people recognise and associate with (Flemmings, 2010; Schockaert and De Cock, 2007; Twaroch et al., 2009a). Although a number of methods exist to extract neighbourhood level information from the web (such as using Internet photograph sites[1] or searches for facilities within specified neighbourhoods[2]), they are predominately reliant on the use of existing gazetteer information - which are not necessarily comprehensive (Brindley et al., 2014b) and are susceptible to issues of ambiguity. This work demonstrates the potential of generating such local boundary information through the extraction of 'neighbourhood' level data appearing in postal addresses held on the web. It is, thus, possible to map the output and produce dynamically determined regions that reflect the *real* social boundaries in which we live. Such units hold extensive utility not only for

[1] Grothe and Schaab (2009); Hall et al. (2011); Hollenstein and Purves (2010); Keßler et al. (2009a); Purves et al. (2011)

[2] Jones and Purves (2008); Pasley et al. (2007); Purves et al. (2007); Schockaert and De Cock (2007)

researchers investigating a range of social processes but also in the wider community, ranging from local web search querying to geodemographic profiling.

1.2 Aims and objectives

The principal aim of this thesis is to:

1. *investigate an approach to automatically generate vague geographical information concerning place through extracting differing perspectives from the web*.

In order to achieve this primary aim, vagueness is applied to all three of the elements of place as identified by Goodchild and Hill (2008): place types, place names and spatial boundaries of places.

In addition to the creation of vague information for the three place elements, this will be complemented by further aims to:

2. *assure the quality of the vague information produced*, and
3. *demonstrate the value of the vague information with examples of its use*.

In order to achieve these aims; the following research objectives have been identified:

i to investigate the concept of vagueness within geographic information concerning place,

ii to explore how a vague settlement classification could be formulated from data held on the web,

iii to construct vague **place types** from web data,

iv to quality assure the vague place type data with a mixture of statistical techniques and surveys and with an emphasis on investigating the extent to which the vague settlement classification generated from web data are representative of those obtained through interviewing people,

v to examine the extent of vagueness within **place names** extracted from web data,

vi to investigate how bounded urban locales can be formulated,

vii to construct vague **spatial boundaries** of urban locales using data mined from the web,

viii to quality assure the vague urban locale data with a mixture of statistical techniques and surveys and with an emphasis on investigating the extent to which the vague geographic information generated from web data are representative of those obtained through interviewing people,

ix to show evidence of the added value of the vague data when linked to other socio-economical data from sources such as the Census of Population, and

x to evaluate the success of the work and examine potential uses for such information and discuss possibilities for further work.

The research concerning the vague spatial boundaries of urban locales proved particularly fruitful. In order to capitalise on this, a number of additional questions were formulated to either further develop and expand the method (see the first three bullet points below) or to demonstrate the value of the research by addressing specific questions (the last three of the bullet points below):

- **Web groupings:** *To what extent is it possible to define urban locale areas for different user groups?* Addressing such a question is of value in exploring if varying user groups might generate different urban locale geographies. Furthermore, greater confidence in the output of data may be gained if the geographies of urban locales from different user groups are in consensus.
- **Temporal aspects:** *To what extent is it possible to explore temporal change in urban locales?* Given the dynamic world in which we live, the degree of change in urban locale boundaries through time is of general interest.
- **Tying other data to urban locale units:** *How feasible is it to tie other data (either point or area measures such as Census data) to the newly formed geographies of urban locale?* Aggregating other socio-economic data to urban locales may add value to both source and input data - creating output for the areas people connect and associate with in everyday life.
- **Homogeneity:** *How homogeneous are the urban locales in terms of socio-economic characteristics?* One measure of defining 'neighbourhood' is through the aggregation of data to form homogeneous areas. In this light, exploring the degree of homogeneity/heterogeneity of urban locales is of interest.
- **Comparison with underlying population geography:** *How well does the frequency of urban locale data compare to the underlying population? Are there*

unexpected areas with low/high data returns? Such a question may reveal interesting insights about the urban locale data and also aid validation of the output.

- **Differing perspectives:** *To what extent do the urban locale spaces overlap or nest, creating differing perspectives? Can the hierarchical nature of space be examined?* The degree of overlapping of urban locales is critical in establishing the extent to which locale boundaries should be *discrete* or represented as *vague* entities. Furthermore, it may provide insight into the complex hierarchy and relationships between places within our urban areas.

1.3 Structure of document

In order to achieve the research objectives set out in the previous section, this thesis is organised into a number of chapters as outlined in Table 1.1. Chapter 2 discusses the literature generally concerning vagueness within geographic information. Chapter 3 relates specifically to place type (or attribute) vagueness and explores this in relation to settlement classification. As such it includes both a literature review and study of vague settlement classification. Chapter 4 turns attention to the literature concerning vague geographic boundaries and names of urban locales. Whilst, Chapters 5 to 8 consist respectively of methods, results, analysis and finally discussion and application of vague boundaries and place names. Chapter 9 concludes the findings of this research regarding vagueness in all three place elements. Each chapter (excluding this introduction) relates to one or more of the stated research objectives as shown in Table 1.1.

Table 1.1 Thesis Outline

No.	Chapter name	Aim	Objective
2	Vague Geographies	1	i
3	Vague Attributes: Settlement Classification	1,2 & 3	i, ii, iii & iv
4	Vague Boundaries: Urban Locales - Existing work	1	i & vi
5	Building Vague Urban Locales	1	v, vi, & viii
6	Results and Validation of Vague Urban Locales	2	vii & viii
7	Analyses of Vague Urban Locales	1 & 3	viii & ix
8	Discussion and Case Study Application of Vague Urban Locales	1 & 3	v, ix & x
9	Conclusions: Generating Vague Places from web data	1 & 3	x

Chapter 2 (Vague Geographies) introduces the general concept of vagueness and links this specifically to geographic vagueness of place. Specifically, the roles of natural vagueness and vagueness arising from multi-person disagreement are considered. Vagueness is contemplated within all three elements of place - as defined by Goodchild

and Hill (2008): place types, place names and spatial boundaries of places. Finally, the role of the web as a potential source for gathering different perspectives of geographic place information through multi-person disagreement is discussed.

Chapter 3 (Vague Attributes: Settlement Classification) explores generating a vague set of place types (settlement classification) using data mined from the web. Data are validated from 281 questionnaires completed by Nottinghamshire residents who categorised twenty different places.

Chapter 4 (Vague Boundaries: Urban Locales - Existing work) turns attention to vague spatial boundaries. It introduces the main motivations for defining suburb and neighbourhood boundaries and the current literature on the subject. It develops the concept of an 'urban locale' (any named part of an urban area) and identifies and assesses existing methods of urban locale identification.

Chapter 5 (Building Vague Urban Locales) describes the method for generating vague, probabilistic spatial boundaries for urban locales using data extracted from the web. As part of the process, the method for identifying vague place names from the urban locales is also discussed.

Chapter 6 (Results and Validation of Vague Urban Locales) illustrates example results for the bounded urban locales and also validates the data using a number of statistical tests and qualitative investigations. Due to the vague and subjective nature of urban locale boundaries and difficulties in data collection - there is rarely a universal ground truth to directly compare against. In order to overcome this, a range of validation comparisons are made with a variety of other data and information. Initially, urban locale *names* are compared with existing gazetteer sources (such as Geonames and Yahoo! GeoPlanet), with detailed examination of the disparities. Secondly, in those few cases where data allows, boundary comparisons can be made between differing sources (such as Wikimapia, Quattroshapes and Ordnance Survey's[3] Multi Resolution Data Programme (MRDP)). Also, the opinions of boundaries from a small number of selected stakeholders were sought (consisting of local planning officers and the chair of a local community group). Finally, residents' opinions were collected. These perhaps represent the closest one can get to potential "ground truth", and were broken into three different elements. Firstly, resident opinions were gathered across the city of Nottingham in order to compare against the other urban locale source data. Secondly, the geography of an individual locale was established from nearly 400 resident responses collected by a Neighbourhood Forum and compared with the other urban locale datasets. Finally, residents in areas of Sheffield (where the web extraction method identified a number of different urban locales co-existing within the same

[3]Britain's mapping agency - http://www.ordnancesurvey.co.uk

space) were interviewed in order to compare web extraction data and actual resident perceived urban locales.

Chapter 7 (Analyses of Vague Urban Locales) discusses additional questions concerning the identification of the vague urban locale boundaries. Three additional questions develop and expand the method. The first extension forms separate perspectives of space depending on the source of the original data. Thus, it is possible to distinguish between a real estate view of the world, compared to that constructed from business information. The second, investigates the extent to which it is possible to explore temporal change in urban locale boundaries by comparing data runs for the city of Sheffield in 2012 and 2014. The third, demonstrates how other data (such as Population Census) can be tied to the vague urban locale information. The last three additional questions demonstrate the value of the vague information with examples of its use. These include assessing the level the homogeneity of socio-economic characteristics within the urban locales; comparing the data with the underlying population geography; and exploring the extent to which the urban locales overlap and nest to form different perspectives of the same space.

Chapter 8 (Discussion and Case Study Application of Vague Urban Locales) consists of a discussion of the overall method (including merits and limitations), demonstration of the value of the work using a number of different applications (for researchers, local community groups and for neighbourhood planning) and finally an identification of areas of related further research.

Chapter 9 (Conclusions: Generating Vague Places from web data) provides a conclusion to the study by summarising the findings of the research and reviews how well each of the aims of this study have been fulfilled.

Chapter 2

Vague Geographies

2.1 Introduction

We live in a complex world, full of uncertainty and vagueness. In order to model this vagueness within a GIS (Geographic Information System) a number of issues need addressing. Firstly, the concept of vagueness needs discussing and *defining*, including a dialogue concerning the phenomena vagueness is applied to - in this case Geographic Information (GI). Secondly, some *method* of modelling the vagueness will be required. A number of approaches including fuzzy set theory and probabilistic methods will be discussed. Thirdly, the method will need to be *applied* to a specific case study. For the purposes of this work, vagueness will be considered within the concept of place (specifically place names, place types or categorises and the geographic boundaries of place). Finally, some dataset or *datasource* is needed for the model. In this regard, the use of the web for mining differing perspectives and opinions of vague places will be considered and will specifically seek to collect information from a range of different datasources.

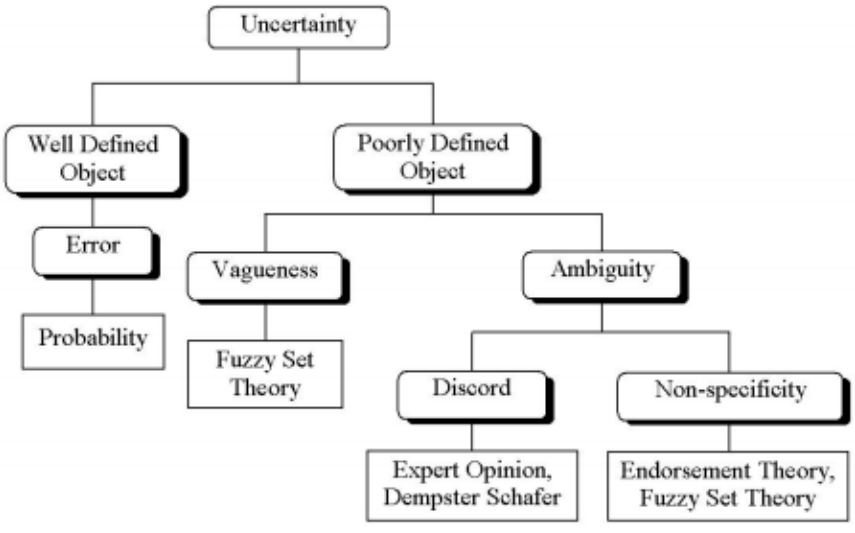
2.2 Defining vagueness within Geographic Information

Vagueness is the lack of precise definition and occurs when something is not clearly expressed or is stated in indefinite terms. The investigation of vagueness within GI is not a new phenomena - see for example work by Fisher (1999, 2000, 2003), Evans and Waters (2007), or Hall and Jones (2008) which are discussed subsequently. Previous work, however, fails to explicitly consider an approach to model vague places that is automated, national in scale and character (in that it can consider all similar geographic

features over a national coverage, for example all villages within the UK) and that draws data from a range of different source data in order to collect varying opinions. Before further exploring vagueness within GI, however, it is pertinent to briefly discuss the concept of vagueness more generally.

Vagueness is omnipresent in language. Concepts used in everyday conversation are frequently vague, often to avoid unnecessary complexity and qualifying precision. How 'tall' is 'tall', how 'near' is 'near', when is 'fresh' 'fresh' or 'local' 'local'? Without such vague statements, however, we would be forever qualifying ourselves which would be extremely time consuming having to rely continually on measurable characteristics. The concept of vagueness can be extended from common language to that of technical terminology and to physical objects (Sainsbury, 1989; Williamson, 1994). Fisher (1999) argues that *vagueness* is one of the three types of uncertainty - as shown in Figure 2.1. The others being *error* (the deviation of a value from truth - Roberts (1997)) and *ambiguity* (the possibility of interpreting 'something' in multiple, different ways, for example 'bank' could refer to a financial institution or the side of the river, as well as other less well known meanings). Figure 2.1 will be discussed in more detail in subsequent sections. Fisher et al. (2007) argues that a great deal of vagueness is "not a property of the world itself, but rather of our human process of cognition and our habit of naming the world and of classifying the phenomena we encounter" (p.312).

Fig. 2.1 Conceptual model of uncertainty within spatial data



Source: Fisher et al. (2006)

Hall and Jones (2008) distinguish between the following forms of vagueness within GISc:

- Multivariate classification - where vagueness is due to multiple non-independent classification criteria. For example, soil types where multiple variables are combined into a single category based on the mix and best fit to a set of 'text-book' classes (Evans and Waters, 2007). Vagueness arises because the underlying data are diffuse whilst the output classification appears more uniform;
- Precision vagueness (also termed averaging or scale dependent vagueness) is due to representational and scale factors where spatial extents represent the average of some varying boundary. For example, a single line to represent the distinction between land and sea with the constantly changing tide; or a seasonal variation in the term "seaside" (Evans and Waters, 2007);
- Multiperson disagreement (or contested vagueness) - is vagueness due to the differing interpretations by various people; and
- Natural (or conceptual) vagueness - is where the underlying concept is ill-defined, due to the sorites paradox as discussed below.

The focus of this work concerns the last two types of vagueness: natural vagueness and multiperson disagreement. Both will now be further discussed in more detail.

2.2.1 Natural vagueness

Vagueness has been philosophically debated since the time of the ancient Greeks. The Sorites Paradox (or 'paradox of the heap') appears to have been formulated as a philosophical paradox by Eubulides of Miletus (Fisher, 2000) and queries the notion of precise identity (Miller, 2006). It can be applied equally to nearly all human constructs including geographic objects. The paradox posed the question that if we agree that one, or two, or three grains of sand cannot be called a 'heap', at what point will one additional grain of sand create a heap? As such, a heap of sand can, thus, be seen as possessing vague or indefinite properties.

As an example, imagine a single solitary tree. On its own we can safely say that it is not a 'forest'. By adding a second tree the answer remains that it cannot be a forest. If a third tree were added - is this a forest? Clearly not, and yet, if there were n trees but no forest, then adding another tree to make n+1 trees still would not make a forest. The addition of a single tree can never turn the collection of trees into a forest and thus, even if there were thousands of trees, under this line of argument, they could not be called a forest. A threshold could be set whereby when n exceeded the threshold (for example 250 trees) then the trees could be deemed a forest. This would mean, however, that the removal of one tree (say to 249 trees) would mean that the area could

no longer be called a forest. Who would keep a constant count of the trees? When would a sapling be deemed a tree? How close would the trees have to be to remain contiguous? The complexity grows with no clear answer.

The debate concerning the existence of abstract objects is well discussed within the nominalism branch of philosophy (for example see Wittgenstein (1965)). Rehearsing such ideas is outside the scope of this research which is more concerned with a pragmatic approach based on the application within a GIS. The principal point of the argument within this body of work is that many words and concepts (including geographic features) are poorly defined and whilst they may be frequently used in common language to describe objects, they are not precise, and never can be. They are as such inherently vague in meaning and application (Fisher, 2000). Many objects or phenomena are Sorites susceptible: whether entities are tall, bald, new, old, thin, big, small, hot, cold, near, far, local, soon, high crime, high pollution, how many houses does a village require or when does a pond become a lake? The list is exhaustive.

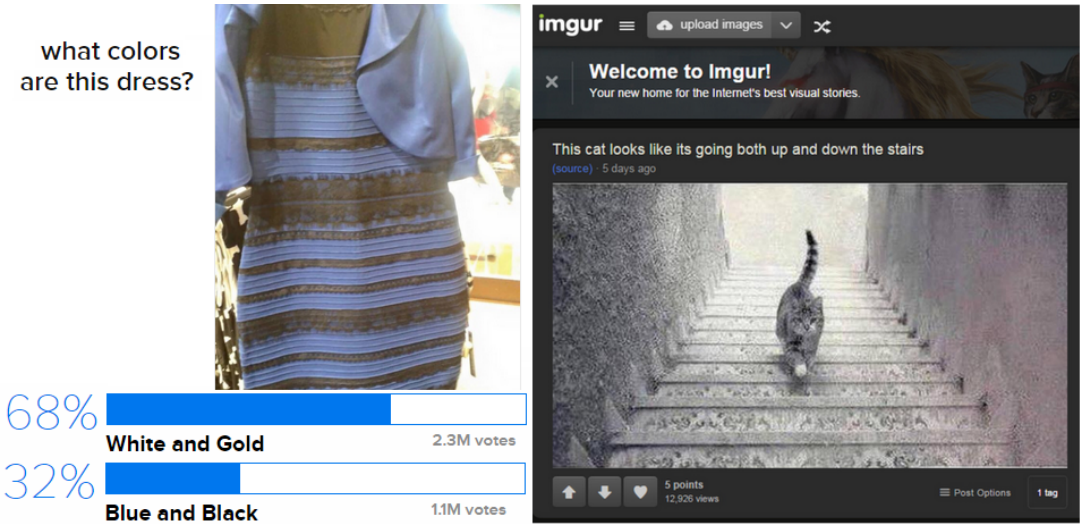
Of course many vague terms are relative. 1000m might be considered *near* in the context of comparing the distance of London to Paris but not when talking about the distance to the nearest Post Box. Although relatives may help, they do not eliminate vagueness. In line with the Sorites Paradox, would 2000m be considered near when discussing London to Paris distance? What about 3000m, or 3001m and so on... Such delimitations require a specified threshold of some measurable quanta or the opinion of an individual or expert (Fisher, 1999). Such thresholds, however, may be both contested and dependant on context and individual perspectives. If an object is hard to define, so too will be any threshold values used to categorise the object (Fisher, 2000). Although vagueness can aid communication it is challenging to incorporate when translated to formal systems such as computers. Statements such as: "the man is tall", are difficult for computers to translate into definite rules. On the other hand, "the man's height is 175cm", does not explicitly state whether the man is tall or not.

2.2.2 Multiperson disagreement

Evidence of multiperson disagreement can be found in popular trending 'optical illusions' found on the web and shown in Figure 2.2. In such cases there might actually be a right or wrong answer but that may remain illusive without further information. For example the dress in Figure 2.2a is reportedly actually blue and black (interestingly only one third of 3.4 million voters guessed 'correctly' in the associated poll), whilst due to the raised position of the cat's tail in Figure 2.2b, the cat is actually going downstairs. The critical point is that in the complex world that we live, it should be quite normal that objects can be interpreted differently. The situation is further

exacerbated when naturally vague objects are considered, where opinions may be even more diverse and there is no right answer.

Fig. 2.2 Ambiguous images of everyday objects



a) 'What Colors Are This Dress?[1]' b) Going both up and down the stairs[2]

All of the examples in Figure 2.3 represent images where the same individual object may be perceived in multiple ways. Examples (a) and (b) within Figure 2.3 contain ambiguous images (also known as reversible or bi-stable) that can form two separate entities within a single image: a duck or rabbit in the first case and a vase or two silhouette faces in the second. These optical illusions exploit graphical similarities between the different entities within the picture. The third example represents a hybrid image developed at MIT in which the image can be interpreted in one of two different ways depending on the viewing distance. It is created from low-frequency components of a photograph of Marilyn Monroe (top inset) and high-frequency components of a photograph of Albert Einstein (bottom inset). Figure 2.3 demonstrates that we should not find it so hard to conceive the fact that, in our complex world, objects can be interpreted differently. Issues of vagueness might not always reduce to a right and wrong interpretation but merely different perspectives - all with an equal level of 'correctness'.

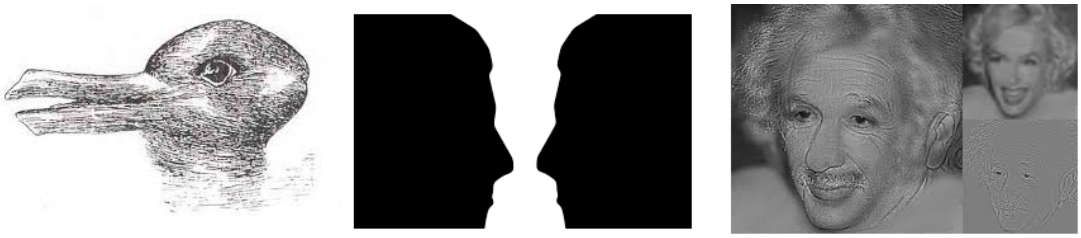
A further example of varying interpretations of like-concepts is through *synonyms*, where different words are used to describe the same or similar concept. Take the idea of 'beautiful' which could also be described using a host of other words including: pretty, lovely, attractive, gorgeous, dazzling, ravishing, graceful, elegant, fine, exquisite, aesthetic, stunning, glorious, heavenly, radiant, glowing, blooming, sparkling and so on. Furthermore, concepts and words might have different meanings in differing

[1] http://www.buzzfeed.com/catesish/help-am-i-going-insane-its-definitely-blue#.vh73LnrBv and http://en.wikipedia.org/wiki/The_dress_(viral_phenomenon

[2] http://imgur.com/gallery/I56Uepu

places. For example bread 'rolls' may be called cobs in the Midlands and Northern England but baps, buns and rolls elsewhere. There are often slight differences between interpretation that humans may be at relative ease disambiguating between. It is, however, rather more problematic for any automation of the process to interpret the slight differences in meaning within synonyms. What is the difference between a stream, brook, creek or beck? Whatever definitions are formed - it is unlikely that everyone would completely agree on the distinctions. Do nicknames and aliases for places such as Stannz or Stannzville for St. Ann's in Nottingham refer to exactly the same object or are they in some manner a subset (with either a slightly different geography or embodying differing meaning)? Furthermore, the different interpretations are likely to be dependant on the application and context rather than any underlying universal truth.

Fig. 2.3 Ambiguous images



a) Joseph Jastrow's duck rabbit[3] b) Edgar Rubin's Rubin vase[4] c) Einstein/Monroe hybrid image[5]

2.2.3 Vagueness within GISc

Vagueness can occur within GISc relations/operations (such as near, north of and so on) and to objects (where does this mountain start/end?). There is considerable existing literature covering both these elements. Work concerning GISc relations/operations includes a number of studies that have explored the use of particular spatial relations and operations through empirical experiments with human subjects (Fisher and Orf, 1991; Montello et al., 2014; Worboys, 2001); different datasource precision (Du et al., 2013); or using the Bing Internet index[6] for co-references of places from GeoNames in the form: 'A near B' (Derungs and Purves, 2014). Investigation of general spatial relations and language usually contain elements of vague properties (for example in the work of Wallgrün et al. (2014) and Stock et al. (2013)). Whilst such work is of intrinsic merit and worth - it is not the concern of this research - which instead focuses on vagueness in *geographic objects*.

[3] http://en.wikipedia.org/wiki/Joseph_Jastrow
[4] http://en.wikipedia.org/wiki/Rubin_vase
[5] http://en.wikipedia.org/wiki/Hybrid_image
[6] in the en-us market using theMicrosoft n-gram API

Within GISc, 'objects' are seen as the basis for describing geographical phenomena, however, they are frequently conceived as uncontroversial or neutral, with a single function (Ahlqvist et al., 2005) and with definite, fixed boundaries that can be represented in a GIS (Fisher and Unwin, 2005). This thesis will employ the definition of objects within Bibby (2005) as simply something that can be expressed in speech or writing as a noun (phrase), in other words something that might be imagined by a sentient subject. As such, objects will embody multiple meanings in different contexts (Bibby, 2005) and can rarely be represented within GIS with perfect certainty. Vagueness is ubiquitous and is inherent in geographical data (Bennett, 2001; Fisher, 2000) and although the importance of vague objects within GISc is widely recognised it is frequently viewed as a "problematical issue" (Foody, 2003).

It is the standpoint of this work that it is the *language*, and its consensus, that binds together particular bundles of matter to create meaningful objects. Ahlqvist et al. (2005) argue that unpacking objects and examining the processes that lead to their construction reveals that many objects, if not all, are vague (in both semantic and spatial definition), ambiguous and imprecise. In a computational sense, objects are instances of classes (or categories) but as these categories do not exist in the world, they exist only in the mind and are therefore subject to cross-cultural differences in interpretation and implementation (Ahlqvist et al., 2005). In essence, "one man's meat is another man's poison" or "one man's trash is another man's treasure" - we all interpret things differently, through multiperson disagreement.

It is simple to see how people can make different interpretations of the same object given the complexity of the world we live in. GIS frequently carve up space into neatly ordered bundles of spatial objects but such modelling represents generalisations of the real world. The same spatial object can constitute multiple things and be viewed by different individuals accordingly. This might be in attribute data: the A52 between Nottingham and Derby is known as both 'Derby Road' *and* 'Brian Clough Way'; or in function: a river might be considered an area for fishing by one user but as a place to go punting for a different user. Frequently, differences in view will involve temporal change. Consider the 'ship of Theseus'[7] whereby the ship was constantly being refitted and old planks were replaced by new ones (see Varzi (1998)). Given complete change in the components or parts of the ship, the question arises, at what point is the vessel considered to be a different object to the starting entity? Similarly, does the existence of a 'park' persist when the users change from children playing during the day to drug dealers frequenting the same space at night? Is a shop still a 'shop' when it is closed and you cannot purchase goods from it? Objects may be coextensive, so a wood may also be a nature reserve. Alternatively aggregations of individual objects may overlap to form different higher order entities that might not be coextensive, such as a town and

[7] also known as 'Theseus' paradox'

a regeneration area. There is a need to embrace differing interpretations of the same object rather than attempt to reduce them down to a single viewpoint. Giunchiglia et al. (2012) argues that the formation of specified standards do not represent effective interoperability solutions. For example, if a specific term becomes a standard for describing an object (such as 'harbour') – the alternative terms (for example 'seaport') can then become marginalised and not included in subsequent analysis, even if they share the same concept.

Vagueness is an inherent property of geographical data (Fisher, 2000) and can transpire within either the spatial footprint of an object (where is the edge of the city, mountain or woodland?) or any categorisation of the entity (is this object a city or a town, a mountain or a hill, a wood or a thicket?). Lüscher et al. (2009) use the example of ponds - although they can be defined as water bodies smaller than a lake, in reality a continuum exists between ponds and lakes. Frequently, grey areas exist where either or both label(s) might be applied.

Instead of always considering vagueness and uncertainty as a problem, however, it is increasingly considered as an asset (Devillers et al., 2010), demonstrating the world in a more realistic manner. The stance of this work is that vagueness is not an problem as such, rather it is an inherent way in which we perceive and classify the world. As such, we do not need to 'solve' vagueness but we should be developing methods to incorporate the concept within the geographic objects that we model.

2.2.3.1 Spatial and temporal vagueness

Vagueness (specifically the concerns here of natural vagueness and multiperson disagreement) can manifest itself within both spatial and temporal objects. We live in an ever changing complex world where the role of change within vagueness should be considered. For example, take the level of consensus regarding the question if Pluto is a "planet" or not. There would have high consensus within the general public that Pluto was a planet up to 2006 when it was officially 'downgraded' to a "dwarf planet". Now, it is less clear how many people would carry this official view or if they would still consider Pluto a planet in its own right. Planet status can continually change because it is, in part, dependant on the vague definition of the celestial body clearing other objects in close proximity[8]. Whilst the temporal vagueness of objects is important, it is not the main concern of this work - which rests with *geographic vagueness*, as discussed below. This is not to disregard temporal effects, but merely to emphasise the focus on the spatial element which will also include the combined spatial-temporal domain when applicable. For example, many vague processes will

[8] http://en.wikipedia.org/wiki/Definition_of_planet

exhibit spatiotemporal change - to name just a small section, high crime areas or areas of high air pollution, even definitions of what we consider spatially 'close' changes with advances in travel.

In summary, GISc cannot present a truly representational view of the world, human categorisation of objects creates (natural) vagueness and those vague entities can be interpreted differently through multiperson disagreement. GISc can, however, provide a formal framework for the reconciliation of different perspectives (caused by vagueness and ambiguity). As indicated within Figure 2.1, one method to incorporate vagueness and different viewpoints is through the use of fuzzy set theory.

2.3 Modelling vagueness in Geographic Information

Montello et al. (2003) state that examples of vague concepts vary in the degree to which they are members of a category (in which case they are fuzzy) *or* the chance that they are members of a category (they are probabilistic). Both approaches will now be discussed in turn and briefly evaluated.

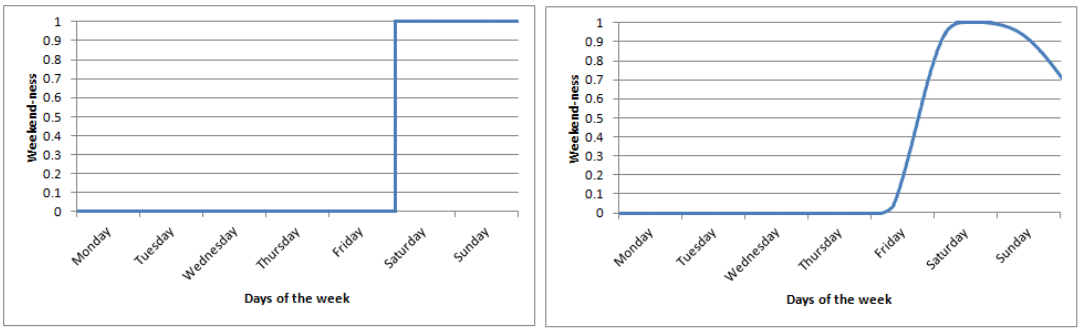
2.3.1 Fuzzy set theory

The differing views or perspectives of vague objects fits well with the notion of fuzzy set theory which attempts to capture the different degrees of membership to a class or object. In line with Figure 2.1, Fisher et al. (2006) argues that vagueness can be treated with fuzzy set theory. Fuzzy set theory was introduced by Zadeh (1965) as an alternative to Boolean (also known as Cantor or classical) sets. Membership of an object in a Boolean set is absolute, it either belongs (and is assigned a membership value of 1) or it does not belong (membership value of 0). In contrast, a fuzzy set is defined by memberships values consisting of real numbers that continuously vary between 0.0 and 1.0. All intervening values define an intermediate degree of belonging to the set. For example, a membership value of 0.25 reflects a smaller degree of belonging to the set than a membership value of 0.5 and thus, in this example, the object described is less like the central concept of the set.

In fuzzy set theory, the truth of any statement becomes a matter of degree. For example, consider 'the weekend'. Using Boolean sets, Saturday and Sunday would be assigned values of 1 and all other days would be 0. Many, however, would consider that the weekend starts as work finishes on Friday. Some may even consider that Sunday is less like the weekend than Saturday. In Figure 2.4a (the Boolean model) at midnight on

Friday, just as the second hand sweeps past 12, the weekend-ness membership value jumps from 0 to 1, whilst in contrast, the Fuzzy membership values in Figure 2.4b define the degree to which any given instant belongs in the weekend rather than an entire day.

Fig. 2.4 Weekendness: concept of fuzzy set theory



a) Boolean logic b) Fuzzy set theory

Fuzzy memberships are commonly identified by one of three methods (Fisher, 2003):

i the Similarity Relation model involves a pattern recognition approach commonly using the Fuzzy c Means algorithm or fuzzy neural networks;

ii the Semantic Import model, uses formulae to derive fuzzy membership as specified by the user or another expert;

iii the Experimental model empirically derives membership by experiment, usually with human subjects.

2.3.1.1 Geographic fuzzy vague objects

Fuzzy set theory can and has been readily applied to geographic objects. There has been a considerable interest in the fuzzy modelling of land cover classification through remote sensing especially concerning the uncertainty of sub-pixel variations (Zhang and Foody, 1998). The use of fuzzy sets within soil modelling has been well documented (Burrough, 1989). Fuzzy set theory has been applied to landscape features such as morphometric classes in the Lake District (Fisher et al., 2004), mountain peaks in Scotland (Fisher et al., 2007) and ephemeral lakes in Australia (Humayun and Schwering, 2013). Fisher (1991) applied fuzzy sets to viewshed analysis. Whilst a substantial number of applications have been concerned with the natural environment, there are also examples of the use of fuzzy set theory within human geography. Waters and Evans (2003) asked users to apply a digital spray-can over maps in order to identify fuzzy high crime areas in Leeds (UK) and Montello et al. (2003) constructed a fuzzy

representation of downtown Santa Barbara (USA) by asking a number of residents (n=36) to draw their perceived boundaries with differing levels of certainty. Coombes (2004) highlights the importance of fuzzy membership between the concepts of urban and rural areas, where it would be unlikely to identify a clear 'definitive answer' with a single criterion.

2.3.1.2 Critical reflection

As shown within Figure 2.1, Fisher et al. (2006) argue that vague objects must be poorly defined and well defined objects cannot be vague. This body of work, however, argues that this alone is not satisfactory and must be taken in light of the appropriate geographic scale of measurement and application. After all, it is well recognised that it is impossible to measure anything with complete accuracy and certainty. Such *problem of definition* is well discussed by Taylor (1997) who demonstrates that even when a carpenter uses a tape measure to assess the height of a door - there is always some remaining uncertainty due to the precision of measurement (for example is the 211.3cm measurement really 211.3000cm or 211.3001cm, the doorway may not be a uniform height but vary slightly due to local variations such as paint, dirt, chips and so on). Even the well defined objects proposed by Fisher (2000) (such as census geographies and land ownership) are inherently impossible to spatially define with complete certainty, for even a line must contain an element of width to it. Such an effect was noted in the following statement by Black (1937):

> *"It is a paradox, whose important familiarity fails to diminish, that the most highly developed and useful scientific theories are ostensibly expressed in terms of objects never encountered in experience. The line traced by a draftsman, no matter how accurate, is seen beneath the microscope as a kind of corrugated trench, far removed from the ideal line of pure geometry.... While the mathematician constructs a theory in terms of 'perfect' objects, the experimental scientist observes objects of which the properties demanded by theory are and can, in the very nature of measurement, be only approximately true."* (p427)

It is, thus, impossible to split and define any unit with complete precision and accuracy. Such issues are rife within GIS when mapping objects and everything comes down to appropriate geographic scales. The surface of a table is only smooth when viewed through the human eye and not a magnifying lens which would show it as a rough surface. There are, however, geographic features whose variation operate at scales more aligned with human interpretation. For example, the vagueness of where a mountain starts, where the city centre might be, or the extent of an area of high air pollution, all operate at geographic scales that we are comfortable with and can more

easily be interpreted as vague. When objectifying entities within GISc we must to a degree forget the fact that things can never be 'perfectly' measured and rest upon the very human assumption that the measurement need only be suitable for the intended need and purpose. The appropriateness of the scale will depend on the intended purpose and will be open to interpretation in fitting with this work - in itself vaguely defined. Hence, the question of how well defined the objects are, can only ever be taken within the context of the appropriate scale required of the entities themselves.

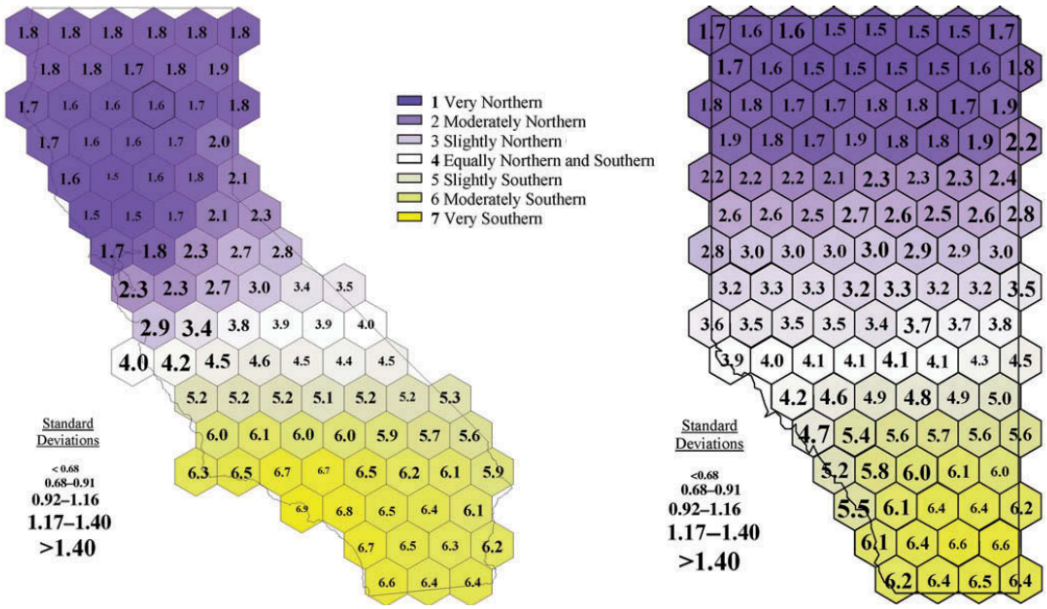
2.3.2 Probabilistic approach

Probability is a standard technique within current school level mathematics and can be found described in any introductory statistics book (for example Rogerson (2014), Field (2013) or Pallant (2013)). Membership values range from 0 to 1 and may be assigned based on a variety of ways, ranging from subjective beliefs to the use of relative frequencies based on past events (Rogerson, 2014). If we know that a certain outcome will definitely occur - a value of 1 will be assigned, similarly if an outcome is impossible it will be assigned as 0. A toss of a coin however has two possibilities (heads or tails) and has a 50:50 chance of occurring and would be assigned a value of: 0.5 Heads and 0.5 Tails. In the context of categorisation of spatial data it is possible to apply probabilistic membership functions to the class of object and its spatial boundaries. For example, one might consider the probability that a specified space could be categorised as a certain feature, for example a mountain. If 80 out of 100 people describe the object as a mountain and only 20 state that it was a hill, probability would assign a membership value to the object of 0.8 - mountain and 0.2 - hill. Thus, if someone else were to be asked to classify the same object, there would be a 0.8 chance (or 80 %) of them calling it a mountain. Similarly, spatial boundaries can be probabilistically modelled by dividing up the geographic space and determining the likelihood of occurrence for each specified unit of space.

The use of probabilistic methods are widespread both within mathematics and beyond. Just one example is that from Information theory which explores the construction of probability distributions on the basis of partial knowledge (see Jaynes (1957)). Such work is beyond the focus of this research which instead lies with probability within GISc. Probabilistic methods are commonly used within GISc, especially in modelling uncertainty (Humayun and Schwering, 2012; Worboys and Duckham, 2004) but they have also been applied to a wide variety of applications ranging from: positional accuracy (Heuvelink and Brown, 2008); place name ambiguity (Li et al., 2008); disease mapping, such as cancer risk by Goovaerts (2008); land cover mapping (de Bruin, 2000); crime mapping, such as geographic profiling (Wilson and Filbert,

2008); forest fires (Davis and Miller, 2004); landslides and weather forecasting (Ames and Anselmo, 2008) to name just a small selection. As an example, Montello et al. (2014) applied a ranking test in order to establish probabilistic boundaries between 'Northern' and 'Southern' regions in both Alberta and California. He did so by dividing the geographic space into hexagon cells and asking just over a hundred participants to score each cell on a scale of 1 to 7 (1 for very Northern; 4 for equally Northern / Southern; and 7 for very Southern). The scores were then averaged for each cell in order to define the vague and uncertain boundaries between the regions, as shown in Figure 2.5.

Fig. 2.5 Example of probabilistically defined vague boundaries



a) California (n=80) b) Alberta (n=24)

Source: Montello et al. (2014)

2.3.3 Fuzzy-bilistic: the differences between probabilistic and fuzzy

Taking the words of Bruno de Finetti (1974) neither probability or fuzzy set theory exist but are a "subjective description of a person's uncertainty". There is a great deal of similarity between probability and fuzzy set theory. Principally, they are both measured on a scale of 0 to 1. Probability attempts to make predictions about future events from a state of partial knowledge. Each event will have a boolean existence - it will either occur or not occur. In contrast, fuzzy sets are based on vague definitions, not randomness. They are about degrees of truth - about fuzziness or partial truths, where outcomes themselves can be fuzzy or mixed in nature. Probability theory says nothing about how to reason concerning objects and events that are not entirely true or false. Fisher (2000) argues that probability is concerned with identifying whether

or not an object belongs to a Boolean set. As such, he states that probability deals with problems related to lack of data, whilst in contrast fuzziness deals with lack of definition. Clearly, by definition, there is overlap for those cases that lack both definition and data - demonstrating that there is not necessarily a simple duality or choice between the methods but that overlap exists.

Support for such an argument is given by Montello et al. (2003) who defined the downtown area of Santa Barbara in America by asking residents' opinion. They concluded that the output could equally be fuzzy or probabilistic in nature with much of the difference depending on precisely what is required and the question posed. They stated that given a membership value (z) for an area (A) it could be interpreted as:

1 a probability, $0 \leq z \leq 1$, under two different approaches:

   - in the strictly frequentist sense, as the probability that a randomly chosen person would assign the point to area *A*, or
   - subjectively as a measure of confidence in assigning the point to area *A*.

2 as a measure of the membership of z ($0 \leq z \leq 1$) in the fuzzy set *A*.

2.3.4 Alternative models for vague GI

There are other models that can represent vague GI. For example, *rough set* theory uses a three valued logic of: (true; false; maybe) to assign membership values (Bittner and Stell, 2003; Worboys, 1998), as opposed to the classic two value or boolean logic (true, false). The *egg-yoke* method is similar in that it defines a vague region analogous to an egg; whereby the 'yoke' represents a minimal extent and the 'white' corresponds an intermediate area and reflects a maximum extent (Cohn and Gotts, 1996). *Supervaluation* (as described within Kulik (2001)) assumes that there is not one but a number of different interpretations. If all interpretations are in agreement of existence, then the membership is definitely true (supertrue and creating the area termed the positive extension); in contrast if there is complete agreement of non-existence then the membership will be definitely false (superfalse - with negative extension). Between these two extremes lies an intermediate area (called the penumbra) with differing degrees of truth. Similarly, the concept of *broad boundaries* defines objects with lower and upper approximations of extent (Clementini and Di Felice, 1996, 2001).

What all these different theories have in common is that they establish 'middle cases' between the boolean logic of 'true = 1' and 'false = 0', creating intermediate areas which contain differing degrees of membership. Analogous to the earlier discussion

concerning the problem of definition, the open question is the degree (or granularity / scale) to which the intermediate area is divided. Such questions, however, cannot be addressed without consideration of the context in which data are to be applied.

The approach adopted here is in line with the standpoint of Montello et al. (2003) where the outcome could be described as either probabilistic *or* fuzzy in nature depending upon the specific perspective taken. The standpoint will be re-assessed when further information concerning the specifics of the data collected are available (see appropriate sections within Chapters 3 and 8). The critical point remains is that the vagueness of geographic objects may represent degrees of membership within the *attribute data* of the object (including the categorisations of objects) but also reflect *degrees of belonging to the entity* itself and translate a degree of vagueness in its *geographic position or boundaries*.

2.4 Towards vague places

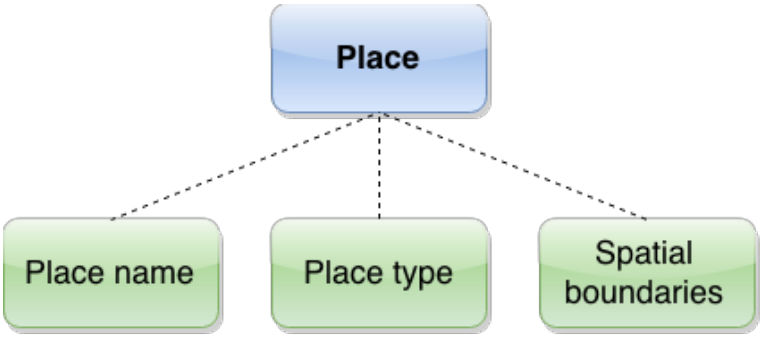
> *"If two different authors use the words 'red', 'hard', or 'disappointed', no one doubts that they mean approximately the same thing... But in the case of words such as 'place' or 'space', whose relationship with psychological experience is less direct, there exists a far-reaching uncertainty of interpretation".* Albert (Einstein, 1954), Foreword to Concepts of Space

If vagueness is to be modelled using a suitable approach - the method must be applied to a specific context. The interests of this work are geographic in nature and specifically rest with the notion of place. 'Place' is central to the concerns of geographers (Cresswell, 2009) and is a fundamental concept of everyday life and reflects the way people perceive, understand and interact with their environment (Tuan, 1977). Cresswell (2009) argues that place is a meaningful site that combines location, locale and sense of place. Where Cresswell defines 'Location' as the geographic space specified through a set of coordinates (the 'where' of place); 'Locale' as the material setting for social relations – the way a place looks (including the buildings, streets, parks, and other visible aspects of a place); and 'Sense of place' as the more ill-defined meanings associated with the feelings and emotions a place evokes.

The concept of place, however, is at best a challenge for GISc (Davies et al., 2009). Places can be as small as a hearth, chair or bed, or as large as the earth (Tuan, 1975). In line with the work discussed previously concerning objects - Tuan (1977) argues that a place comes into existence when a location is identified or given a *name*; thus, separating it from the undefined geographical space around it. To name a place is

to give it explicit recognition (Tuan, 1975). It is this aspect of place that this body of work is interested in - concerning named places or toponymy, rather than the wider sense of place as defined by Cresswell (2009). In this light, place can thus be defined as a geographic location that has been identified and referenced as a social construct (Goodchild and Hill, 2008). Places can, at a minimum, be represented by the tuple N,F,T as shown in Figure 2.6, where N is at least one name, F is at least one representation of geographic location (ranging from a single representative point to a set of boundary coordinates), and T is at least one category or class (such as a system of feature classification) (Goodchild and Hill, 2008). From this standpoint, additional information can be added, for example Jones et al. (2001) also included elements such as a date stamp, variant name spellings and topologic relations.

Fig. 2.6 Concept of Place



Source: Goodchild and Hill (2008)

Each of the elements within Figure 2.6 can have a vagueness associated with them. Place names may be vague due to multiperson disagreement with potentially ambiguous names which might also contain aliases (or synonyms); place types, in the specific form of settlement classification may consist of multivariate classification vagueness as well as multiperson disagreement and natural vagueness (when does a village become a town?); spatial boundaries may also contain multiperson disagreement and natural vagueness (if I take a further step, do I leave the place I was in?). Place names are considered throughout the rest of this work but specifically within Chapters 5 and 8. Vague place types are explored in Chapter 3 whilst spatial extension/boundaries are investigated in Chapters 4-8.

2.5 Using web-based passive crowdsourcing to model vague places

The remaining principal question of this chapter concerns how to obtain place information to estimate the vague membership values. Goodchild (2008) poses the question

that if an object cannot be defined - how is it possible to measure it? Existing literature concerning physical geographic space frequently estimate vague extents on the basis of the underlying measured physical quanta (for example elevation, see Fisher (1991); Fisher et al. (2007) or Humayun and Schwering (2013) for example). In contrast, vagueness in more social phenomena is traditionally represented by the variations in output when asking a sample of peoples' opinions (such as in Montello et al. (2014, 2003); Waters and Evans (2003) or Rosser and Morley (2010)). In such cases, however, it is rare to be able to generate output for any extensive geographic space and research is usually limited to small case studies and not suitable for implementation at a national scale. For such reasons, there is a growing body of research that seeks to harvest data from the web.

The extraction of useful knowledge from datasets has become increasingly significant, especially as a result of increasing web data (Foody, 2003). The web has developed into an extensive knowledge base (Liu et al., 2014) and on an average day we write over 10,000 new Wikipedia articles[9], create 1.3 million blog posts on Wordpress[10], generate more than 500 million tweets[11], and share over 4.75 billion new pieces of content on Facebook[12]. The sheer size of the data involved is enormous and is increasing rapidly. The Indexed Web contains over 4.54 billion pages (February, 2015)[13] and in 2014 Google processed 40,000 search requests per second, equating to 1.2 trillion per year (Keen, 2015). The figures behind the data explosion are colossal.

Given that the majority of information contains some geospatial reference (Hahmann and Burghardt, 2013), the web is a promising source of *geographic* content (McCurley, 2001) and search engines have become the main method of identifying and locating information on the web (Bar-Ilan, 2004). The benefit of extracting data from the web is clear, with its potentially large data volumes that are relatively straight-forward and expeditious to collect. There remains, however, the possibility of data bias and poor information within web documents (Liu et al., 2014). The web has been utilised as a knowledge source in a variety of contexts such as within question-answering (Clarke et al., 2001; Kwok et al., 2001; Radev et al., 2002) and studying language usage (Kilgarriff and Grefenstette, 2003).

The notion of mining the web for geospatial features is not new. Such an approach requires two pieces of information to be collected from the web: the place name of interest and some measure (commonly co-location entities) to georeference or ground the object of interest. 'Spatially-aware Information Retrieval on the Internet' (SPIRIT

[9] http://stats.wikimedia.org/EN/TablesArticlesNewPerDay.htm
[10] https://wordpress.com/activity
[11] https://about.twitter.com/company
[12] https://www.facebook.com/permalink.php?id=567087986693671&story_fbid=690010757734726
[13] http://www.worldwidewebsize.com

- Jones et al. (2008a); Pasley et al. (2008); Purves et al. (2005b, 2007)) used search engine queries to retrieve web documents that contained a user specified vague target place name and the names of associated co-location places. Given that the geographic units of interest were generally large scale (for example regions such as the 'West Midlands' and the 'Cotswolds') a wide range of co-location entities could be used to georeference the object. Features such as cities, towns, villages and streets could all represent points within the larger target feature and thus provide evidence for the object's spatial extent. The work of Jones et al. (2008a) used the search term: "~hotels [target region]"[14] to find the co-location places which could then be used to extract spatial boundaries of the large vague geographic objects using the highest ranked 100 returns from web search engines.

The concern of this work is not with *all* information held on the web, but specifically the different perspectives of vague geographic objects that might be found. The focus is certainly not with administrative views that usually form well defined objects. The scope is, however, wider than (Web 2.0) user generated content, as geographic data can be found embedded in a wide range of web-pages and documents (within contact details for example). For these reasons, the term 'passive crowdsourcing' will be used here to describe the process of automatically retrieving web-page content containing views and opinions of vague geographic objects that 'the crowd' at some point made publicly available through the web. The term 'passive crowdsourcing' is more usually associated specifically with the mining of user generated content from Web 2.0 social media such as Twitter but also relates to any user data, such as including GPS traces from phones or in car navigation systems (Batty et al., 2010; Dobson et al., 2011). In a wider sense, however, we are all *users* of the web and *generating* data or *content* as we post and create web-pages and it is in this wider sense that the term will be used within this work (such an approach has also been implemented by Wallgrün et al. (2014)). In this context crowdsourcing is defined as asking the general public to help contribute to a shared goal (Ridge, 2014) but the process is passive - there are no active calls or queries sent to the public and there is no active input required on the part of users, as data are collected through web scraping. Users are, thus, unaware of the 'shared' goal.

One key issue with crowdsourced data, especially when extracted passively, is that it is not authoritative and, thus, may not be "trusted" (Dobson et al., 2011). For such reasoning, data adopted here will need to be thoroughly validated and only incorporated when a significant number of similar opinions reinforce the particular standpoint. In order to obtain a broad mass of differing viewpoints the use of web search engines are proposed to passively crowdsource data from a range of sources, rather than target specific sites which may provide, in some manner, biased perspectives. The extraction

[14] where ~ is a synonym operator which extends the query for synonyms of hotels such as "inn" and "accommodation"

of information from 'snippets' (containing web-page title, URL and a short textual description) from web search engines has previously been used by Purves et al. (2005b) and Arampatzis et al. (2006) to extract place name information from the web. The specific nature of the search queries is dependant on the precise requirements of the desired data and will be discussed in subsequent chapters (Chapters 3 and 5). The different opinions and perspectives harvested will then be used to form the membership sets for the vague interpretation of the objects. In essence, it is using the 'Experimental model' (Fisher (2003) - as previously described in Section 2.3.1) to form vague membership by harvesting specific content people have posted to the web.

2.6 Summary

Existing research highlights the importance of incorporating vagueness within geographic information. The uncertainty of vagueness and ambiguity manifest themselves within the differing perspectives that people derive when interpreting geographic objects. This condition of multiperson disagreement is exacerbated when concerning naturally vague objects where there is no right or wrong - merely differing views or opinions. The approach taken within this body of research will explore vague notions of places. This will be achieved through passive crowdsourcing of the web for differing perspectives on place both in terms of place type / settlement classification (Chapter 3) and 'neighbourhood' level spatial boundaries (Chapters 4-8). Such an approach builds on the work of Jones et al. (2008a) who used the form: "~hotels [target region]" to extract spatial boundaries of large vague geographic objects (such as the Cotswolds and Mid Wales) using the highest ranked 100 returns from web search engines. Any method developed will aim to extract robust perspectives using a range of different source data (rather than concentrating on a single source - such as hotels/accommodation) and will collect large data volumes (rather than limiting data to a specified number of the highest ranked returns). Critically, this work not only seeks to identify vague *boundaries* but also to the other elements of place - vague *place attributes* and *place names*.

Chapter 3

Vague Attributes: Settlement Classification

3.1 Introduction

The previous chapter argued the need to define vague rather than discrete objects and that within geographic places such vagueness could transpire in the place name, spatial extent (or boundaries) and the place type (or class of object). This chapter will investigate the last of these possibilities by demonstrating the generation of vague attributes of higher order place types - settlement classification. Due to different interpretations of vague objects like settlements, people can classify entities in a range of varying categorisations (town, village and so forth). Thus, issues of natural vagueness and multiperson disagreement can combine to create vague settlement classifications where the same object can be categorised into more than one output grouping (for example a small town might also be considered a large village by others). This chapter firstly discusses existing literature concerning settlement classifications before proposing and implementing a methodology that draws on the high data volumes that people upload to the web by using a search engine to passively mine the web for data concerning settlement classification. The output is in the form of a vague classification where a settlement might be considered as a 'city' by half of the sources but a 'large town' by the other half. Validation of the vague settlement classification method will be through the process of asking people how they would classify a number of different settlements and comparing the two sets of information. Critically, the chapter seeks to explore the extent to which settlement classifications generated from web based content are representative of those derived from interviewing people directly.

3.2 Settlement classification

Classifications or 'typologies' are ways of grouping objects that are similar in some capacity - marking them out as distinct from other objects, thus, making sense of an otherwise extremely complicated picture. Settlement classifications specify higher order place types (such as towns and villages) and are a useful tool for policy purposes and statistical reporting (Bibby and Shepherd, 2004). Aggregation of other data to settlement categorisations can be used in order to search for systematic patterns that otherwise might not be identifiable within the data.

There is a need to be able to identify places with similarities and differences because policy cannot be built on bespoke solutions for every individual place (Lupton et al., 2011). Lupton et al. (2011) state that understanding how socio-economic conditions and the performance of public sector organisations vary from place to place is a routine element within public policy. Such practises have been used for measures of service provision (Smailes, 1944); new housing and house prices (Bibby and Brindley, 2006), travel to work distances and new commercial development patterns (Bibby, 2009). They can, thus, aid the understanding of variations in socio-economic conditions and public sector organisation performance which is important for designing, implementing and monitoring policy (Lupton et al., 2011; Pateman, 2004).

Settlement classifications can also provide more detailed information relating to the type of settlement within place name gazetteers. Many gazetteers contain very limited information related to the place category – for example Geonames[1] uses the term 'populated place' for over 90% of all settlements in Great Britain. Classifications have been used within place identification in order to distinguish between settlements with the same name without reference to further locational information (such as the district), for example the *hamlet* of Newcastle (in Monmouthshire) and the *city* of Newcastle. This can aid issues of ambiguity during the matching of different settlement data (Hahmann and Burghardt, 2010; Purves et al., 2005b).

3.2.1 Literature concerning existing settlement classifications

Settlement is best represented as a vague continuous form from plots and dwellings to large cities, and yet, hierarchies require such a continuum to be carved up in some manner, predominately through some complex modelling exercise. Pateman (2004) lists the most common practises of division based on population size, population concentration and remoteness (measured as physical distance, travel time, or population

[1] http://www.geonames.org

density over large spatial scales). Roberts (1996) stresses the importance of size and form within settlement classification. Numerous settlement classifications exist (Bibby and Brindley, 2013; Bibby and Shepherd, 2004; Frey, 2004; NISRA, 2005; Roberts, 1996) along with those used within online mapping resources (including for example Open Street Map (OSM)[2]; DBpedia[3]; Yahoo! Geoplanet Place types[4]), based predominantly on measures of population (size), function and context.

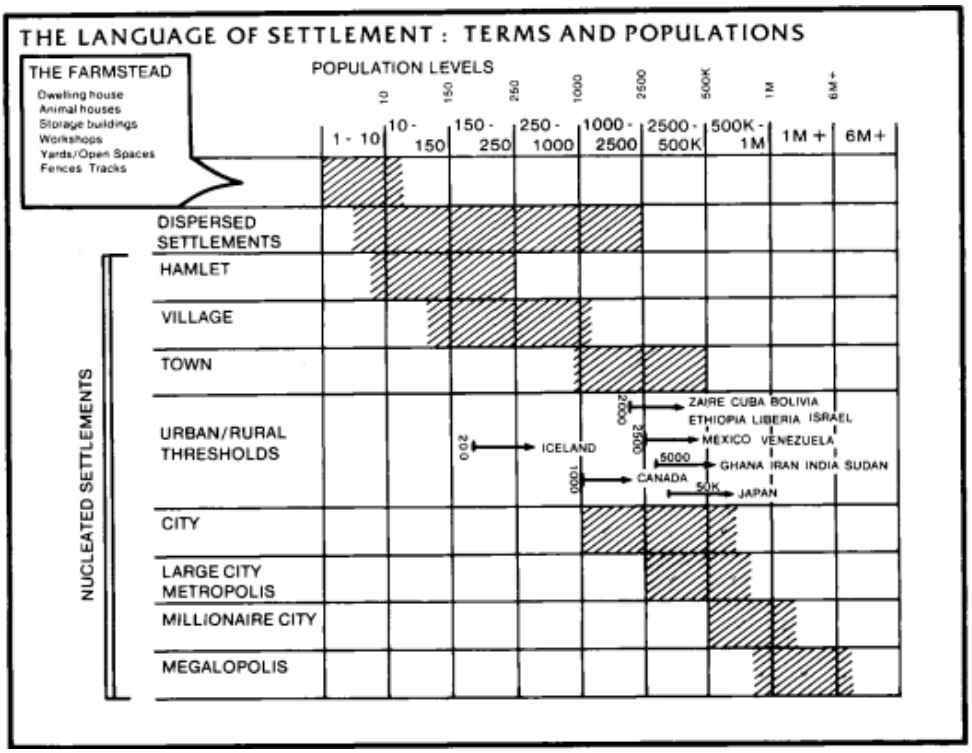
The settlement morphology of Bibby and Shepherd (2004) and updated by Bibby and Brindley (2013) underpin England and Wales' official rural-urban definitions of the Office for National Statistics (ONS)[5]. A series of 'household' density measures were used to differentiate different settlement contexts into nine categories: Urban (10k), Fringe, Peri-urban, Town, Village, Village envelope, Hamlet, Isolated farm and Other (ie scattered dwellings). Similarly, the work by the Northern Ireland Statistics and Research Agency (NISRA, 2005) also based their classification purely in terms of population. Frey (2004) created six community types ranging from low density areas to major cities within his American based study using population and density measures. Figure 3.1 outlines the terms and population categories used by Roberts (1996) in his conceptual settlement categorisation. Importantly, Roberts (1996) allows the populations of his categories to overlap - demonstrating the uncertainty within such distinctions. The Open Street Map classification recommendations can be seen in Table 3.1 - demonstrating the perceived complexity of sub-parts within settlements, which can be defined as suburbs, neighbourhoods or quarters. Critically however, these studies are all based on the author/creators' opinions and do not take consideration of a wider audiences' view and may not necessarily incorporate the category terminology used by the general public. Furthermore, they employ discrete, mutually exclusive categorisations whereby a place may only fall within one of the settlement classification classes. Hence, and counter to the arguments of vague geographies discussed within the previous chapter, in such a scheme it is impossible for a place to be conceived as being both a 'large village' *and* 'small town', depending on the viewpoint taken. These limitations will now be discussed in more detail.

3.2.2 Limitations of current settlement classifications

With the exception of the work by Bibby and Shepherd (2004) and Bibby and Brindley (2013) (which are based on contextual population densities for hectare cells), classifications based solely in terms of population usually suffer from the limitation that they

[2]http://wiki.openstreetmap.org/wiki/Key:place
[3]http://dbpedia.org/ontology/Settlement
[4]https://developer.yahoo.com/geo/geoplanet/guide/concepts.html
[5]https://www.gov.uk/government/collections/rural-urban-definition

Fig. 3.1 Roberts' Settlement Terms and Population Thresholds



Source: Roberts (1996)

Table 3.1 Open Street Map Classification Recommendations

Term	Size	Can have parts:
City	>100,000 population	Suburb, quarter, neighbourhood.
Town	>10,000 population and local facilities	Suburb, quarter, neighbourhood.
Village	<10,000 population	Neighbourhood.
Hamlet	100 inhabitants	
Isolated dwelling	Only a few households	
Farm	Named individual farm	
Suburb	Distinct section of urban settlement with its own identity, eg: -Annexed towns/villages which were formerly independent -Municipalities within a city or next to a large town -Historic districts of settlement -Industrial districts or recreational areas within settlements	Quarter, neighbourhood.
Quarter	Larger than a neighbourhood but smaller than a suburb.	Neighbourhood.
Neighbourhood	Named part of a village, town or city. Smaller than a suburb and quarter.	

Source: http://wiki.openstreetmap.org/wiki/Key:place

are dependant upon the prior bounding of the settlement objects (in other words they must first have defined the settlements' extents). Knowing precisely where a settlement

starts and ends is no trivial matter (Tayyebi et al., 2014) - a fact complicated by near constant temporal change. Terminology of category names and concepts within existing settlement classifications are usually subjectively selected by the author(s) and may not be representative of the language used by people more generally to describe place. Using common terms should enlighten which concepts are most generally significant and aid understanding of the terminology used. Coombes and Raybould (2001) stress that any measure should be transparent and interpretable – being understood by a lay person. A shared conceptualisation of the classification could be assisted by employing terminology that people use and comprehend. Vernacular geography is concerned with the geography and names of places that people commonly utilise. The recent television advert from the newly launched property website - onthemarket.com, uses a number of categories of property to promote its service including: 'OnTheCoast', 'OnTheHill', 'OnTheClose' and 'OnTheTopFloor'. It is, however, impossible to search their properties by such descriptions. Imagine the power of being able to search for 'village' properties, or more specifically properties within 'picturesque villages' - using the language that people utilise to describe places. Hollenstein and Purves (2010) argue that incorporating vernacular geography is an important aspect of any information system which attempts to replicate users' cognitive models of space, in contrast to the more commonly utilised traditional administrative geographies.

Ontologies (including classifications) should be created depending on their intended purpose or use (Gruber, 1993; Janowicz and Keßler, 2008) and yet it is sometimes difficult for the producer to anticipate who all the exact users of the data might be. Furthermore, the producer and the user may not share the same common understanding of the classes (Wadsworth et al., 2006). After all, different groups perceive place with different meanings, uses and values (Holloway and Hubbard, 2001) and there is never one single sense of place which everyone shares (Massey, 1991).

One of the main limitations of existing settlement classification work is that they are constricted by forcing objects into a single class rather than allowing them to be members of multiple classes (for example village *and* town), facilitating differences in opinion. The following section discusses the notion of vague settlement classifications, that might be able to capture the way different people think of settlements rather than forming discrete categorisations.

3.2.3 Vague settlement classification

Whilst the bricks and mortar of settlements may be considered to exist in essence, the collective aggregation as named objects are not natural entities but merely convenient modes of expression (Vining, 1955). As argued in the previous chapter, it is generally

acknowledged that such human constructions will exhibit vague concepts and objects with overlapping meanings (Fisher et al., 2007). There are two alternatives – one may proceed "avoiding splitting continua into different classes, as one can always debate the difference between a 'village' and a 'town', or a 'river' and a 'creek'" (Nagel, 2005, p. 4) or, as this chapter argues, it is desirable to embrace the fact of vagueness and attempt to incorporate it by using such methods as vague classification.

Official definitions of settlement status may exist. 'City' status in England is currently granted by the Queen on the advice of her ministers through competition on special occasions such as coronations, royal jubilees, or the Millennium (Beckett, 2005). 'Town' status might be considered as a reflection upon the local council governance as a town rather than parish council. Other distinctions, however, are less clear and whilst official definitions might exist, the fact that a settlement has a specific official status does not necessarily mean that the residents will call it by such official terms. This was clearly evident in pilot interviews undertaken as part of this work within the 'town' of Cotgrave in Nottinghamshire, when one participant stated "Cotgrave's a village, although it's got town status but we're not happy about it".

In support of the principle of vagueness as discussed in more detail in the previous chapter, Knights (2008) states that whilst people will frequently have a clear understanding of differences between such concepts as hamlets, villages, towns and cities in their own minds, such distinctions are not fixed, and thus, different people may have dissimilar standpoints. Such a standpoint supports the concept of vagueness through multiperson disagreement. Clear distinctions amongst cities, towns and villages purely on the basis of population or function have not been found (Crumley, 1976) and yet it is precisely such data that are commonly utilised within settlement classification to create the categorisation. In essence, one person's 'hamlet' is another's 'village'. Clearly, this does not mean that every term that could possibly be used to describe settlements should be included, rather a balance is required that allows for a wider set of categorisations whilst producing generalised categorisations. Thus, construction of a classification with vague membership would support people possessing differing perspectives – none in themselves being the sole answer or necessarily wrong. Whilst this section has argued the need to collect information pertaining to different opinions concerning settlement classification, the question remains *how* to collect such data.

3.2.4 Passive crowdsourcing

As discussed within the previous chapter, there has been a growing research base that has explored the potential of harvesting geographical information from the web through passive crowdourcing. However, whilst extensive research concerning such

extraction from the web has predominately centred on the identification of place names and mapping boundaries, there has been little attention to settlement classification, especially using automated means for data collection. Existing work concerning vernacular geography (the expression of place through the use of ordinary people's language) generally includes the extraction of place name and location from: photo sharing sites such as Flickr[6] (Hall et al., 2011; Hollenstein and Purves, 2010; Keßler et al., 2009a; Purves et al., 2011), web search engines (Brindley et al., 2014a; Jones and Purves, 2008; Pasley et al., 2007; Purves et al., 2007; Schockaert and De Cock, 2007), Gumtree (Twaroch et al., 2008), business directory information (Twaroch et al., 2009a) and extraction of information from various social media such as Twitter and Foursquare (Phithakkitnukoon and Olivier, 2011; Stefanidis et al., 2013; Wilson, 2012).

The limited work concerning extracting place semantics using web sources includes the identification of individual property level building types (such as schools, art galleries for example) (Alazzawi and Abdelmoty, 2012; Goldberg et al., 2009) or where only city forms of settlement were identified (Uryupina, 2003). The work of Uryupina (2003) was able to correctly classify objects into six classes (city, region, country, island, river or mountain) with an average accuracy of 86.5% using web derived data from Altavista web searches. The city class performed particularly poorly, however, (being incorrect in 38% of cases) within the semi-supervised learning process due to their varied naming forms. Furthermore, there is recognition that the built environment is not static (Garrioch, 2006; Pumain, 2000) and so any classification is transient (Halfacree, 2004). Drawing information from the web means that the method is dynamic in nature and both quick and easy to repeat at different times.

Whilst the previous settlement classification work outlined earlier has focused on categorising locations into predetermined classes, there appears to be no work that specifically seeks to identify the classes of settlement object (hamlets, villages, towns and so on) using web derived sources. This work contributes a new approach that automates the identification of classes of settlement, using a vague classification drawn on data from passive crowdsourcing of the web. The proposed method would have the advantage that classification would not have to be based purely on population sizes of settlement but rather by how they are being described by a range of formal and user-generated sources on the web.

[6] https://www.flickr.com

3.3 Method

As suggested in the previous chapter, in order to obtain a broad mass of perspectives that were not biased towards a small specific group, data concerning settlement classification were extracted from the web using the Bing Search API[7]. The use of an established API was preferable to screen-scraping from web search engine returns as it should represent a more stable environment and was easier to comply with terms and conditions of use (the Bing API is also discussed in Chapter 5 in relation to generating vague spatial boundaries). The extraction of information from 'snippets' (containing web-page title, URL and a short textual description) from web search engines has previously been used by Purves et al. (2005b) and Arampatzis et al. (2006) to extract place name information from the web. The work here involved sending automated requests to the Bing Search API search engine based on the principle of:

"[Place name][8] is a" [Parent term][9]

(Where the terms in square brackets are variables)

The returned snippets were interpreted using a series of simple rules (discussed in more detail later) in order to identify any potential settlement classification information. Example returns might include:

1 **Shirley** *is a* **suburb** of the city of **Southampton** and provides a good variety of facilities for all ages. There are a number of local schools, and further education is ...

2 **Clifton** *is a* **council estate and village** in the city of **Nottingham**, Nottinghamshire, England founded by Sir Robert Clifton. The estate has a number of ...

The parent terms were used to reduce ambiguity in order to ensure that the item of interest (settlement) was in fact the correct object and not relating to something else. Applying the multiple levels of ambiguity developed by Wacholder et al. (1997), this uncertainty may take the form of either referent[10] or semantic[11] ambiguity. For example, a search on Shirley (an area within Southampton) might also return information

[7] https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44

[8] extracted from the Multi Resolution Data Programme (MRDP) Names database by Ordnance Survey (OS) used for internal organisational use only

[9] a higher level of geography to reduce issues of place ambiguity – where the returned data snippet may not be in reference to the place of interest but some other object sharing the same name

[10] where the same place name occurs in numerous different locations, such as Cambridge in the UK or in Massachusetts, USA

[11] where there was uncertainty over the type of object in question, such as if Rugby relates to a place (in the UK, Australia or United States of America), the sport (either Union or League), the

related to the novel Shirley by Charlotte Brontë, the person Shirley Bassey, facilities such as the Shirley High School or Shirley Town Football Club, or other places called Shirley (such as in Croydon, Solihull or America). If a parent term such as Southampton or Hampshire was present, however, this reduces the risk of such issues of ambiguity and the search was more likely to be relevant to the intended object of interest. Parent terms used included county, district name, or other nearby major urban areas. In order to determine appropriate parent terms, each settlement was linked to the County and Local Authority District / Unitary Authority / London Borough (as relevant) it was within, as well as important nearby urban centres using the 2001 Synthetic Neighbourhood Localities (for example Test Valley District, Hampshire was split between two different 'locale' areas: Southampton in the south and Andover in the north).

Numerous extractions of web information use the phrase "is a" (Arampatzis et al., 2006; Giunchiglia et al., 2012; Purves et al., 2005b). The use of the phrase was chosen in preference of over other alternatives (such as "located in the/this/a lively residential suburb of x") as it gave very good output from a relatively small number of returns and the terms of interest (classification term and settlement) were both in close proximity, and thus, easier to extract with fewer issues of ambiguity. It would be possible to undertake analysis using a number of different search phrases – however, this would significantly increase the number of requests needed. Given the exploratory nature of this work only a single term was used for simplicity. Future work could extend the research in this regard.

3.3.1 Description of rules used

Rules were used to extract the text following "is a" within the returned snippet text. In order to isolate precise and meaningful simple terms, it was necessary to know when to stop searching the text for settlement classification terminology. In order to achieve this, a number of 'break terms' were manually identified where a shift in language would signify the end point of the settlement descriptive. The full set of 'break terms' used can be found within Table 3.2 and include such examples as 'between', 'north of', 'located' and so forth. Thus, for example, from the description "Etterby is a hamlet and township containing several handsome villas, one mile N.N.W. of Carlisle" the following were identified: a search term (Etterby is a), a break term (containing) and a parent term (Carlisle). This then allowed the identification of the potential settlement phrase between the search and break terms (a hamlet and township). Also, it was necessary that the parent terms appeared within the same segment of text and not

clothing brand Rugby Ralph Lauren, an American car in production between 1923-28, or even perhaps a relatively uncommon surname

separated by snippet punctuation '...', in order to ensure that the settlement of interest and parent terms were more likely to be related.

Table 3.2 'Break terms' used in the method

situated near	that	but is
between	to the	particularly
north of	just	offering
south of	approximately	convenient
east of	within	conveniently
west of	with	some
consisting of	forming	containing
close to	lying	administered
which	overlooking	about + [number]
roughly	being	[number]+ miles/km/minutes/mins
located	at	

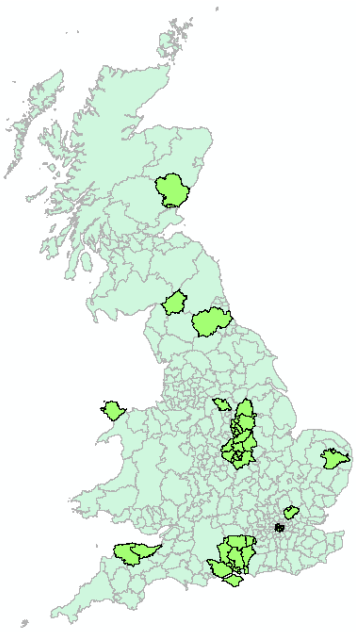
3.3.2 Case study areas

Due to the substantial number of requests required for national coverage (estimated 160,000 requests) a case study approach was adopted. Automated requests were sent to the Bing Search API for all settlements within a selected 47 unitary authorities, districts and London Boroughs. The areas were selected to provide a wide coverage in terms of different types of setting (urban, rural, geographic areas etc). The selected areas can be found in Figure 3.2 and represent over 4,200 settlements in total (approximately 10% of all settlements in Britain).

In total 14,869 requests were sent to the Bing Search API, returning 255,188 snippets. The vast majority of returns (90%), however, did not conform to the rule set applied. This was for many reasons including ambiguity between the settlement name being searched for and the 'parent term'[12] (as was the case for the first two examples within Table 3.3). Whilst a good many of these would in practice actually relate to the place of interest, the risks of including data where the place and settlement classification were not directly associated would introduce potential false positives. Other records that did not conform to the rule set included cases where the object of interest did not relate to the settlement of interest. For example, it may be a person's name (as in the last example of Table 3.3). In summary, potential false positives were filtered out through identification of data which did not meet the following requirements:

[12] due to either referent or semantic ambiguity as previously discussed

Fig. 3.2 The Case Areas of 47 unitary authorities, districts and London Boroughs



NAME	NAME continued…
Carlisle District (B)	Bassetlaw District
North Devon District	Rushcliffe District (B)
Epping Forest District	Gedling District (B)
Harlow District	Ashfield District
Basingstoke and Deane District (B)	Newark and Sherwood District
New Forest District	Broxtowe District (B)
Eastleigh District (B)	Mansfield District
East Hampshire District	West Somerset District
Winchester District (B)	Lambeth London Borough
Test Valley District	Southwark London Borough
Hart District	Wandsworth London Borough
Gosport District (B)	Hammersmith & Fulham London Borough
Fareham District (B)	Kensington & Chelsea London Borough
Havant District (B)	City of Westminster London Borough
Rushmoor District (B)	Sheffield District (B)
Hinckley and Bosworth District (B)	Angus
North West Leicestershire District	City of Leicester (B)
Melton District (B)	City of Nottingham (B)
Harborough District	City of Southampton (B)
Blaby District	County Durham
Charnwood District (B)	Dundee City
Oadby and Wigston District (B)	Isle of Wight
South Norfolk District	Sir Ynys Mon - Isle of Anglesey
Norwich District (B)	

- the place name and parent name were disjoint within the search returns (either not in the same element together, such as search return title, text and so forth, or were separated by snippet punctuation (...))
- 'break terms' were not found following the place name, which could imply that the intended place name was subject to semantic ambiguity and might not relate to the actual place specified

Table 3.3 Examples of records not meeting the rule criteria

query: **Greta Bridge** is a parent: Darlington title: Burns Cottage (Ref : MPB) in Nr Barnard Castle, Co. Durham ... text: **Greta Bridge** is a landmark hamlet, south of the A66 trans-Peninne road and only 4 miles **...** **Darlington** in Durham is 10 miles Ferryhill in Durham is 17 miles Shildon in Durham ...
query: **Weston** is a parent: Southampton title: Weston Sailing Club: Dinghy Sailing Club, **Southampton**, UK text: Welcome to **Weston** Sailing Club. **Weston** is a well established Club with a full and varied programme to suit all sailors. It is the most active dinghy club on ...
query: **Scargill** is a parent: Darlington title: An interview text: Arthur **Scargill** is a friend of mine. Come the election there'll be a **...** Glorious Summer by Ralph **Darlington** and Dave Lyddon ...

3.3.3 Wikipedia bias and data cleaning

Preliminary testing demonstrated that the most popular term was ‘village’ but perhaps more surprisingly the term ‘village and civil parish’ was second. Further investigation of the web addresses (uniform resource locators - URLs) containing references to the term ‘civil parish’ showed that many sites were simply reflecting exactly the same data as Wikipedia. Whilst it would be acceptable to include all Wikipedia ‘mirroring’ websites, their continued inclusion would give an increased weighting to a single source and privilege the Wiki-view of the world. As data were assumed to be some measure of a proxy for peoples’ perceptions of place, such sites were removed if the content were identical to that posted on Wikipedia. Also, sites that were not deemed accurate (for example gomapper.com which identified all different settlement hierarchies as ‘place, city/town’) were excluded.

The method draws on the strength that the web contains a multitude of different perspectives from a variety of people and sources. Ironically, however, this can also be one of the weaknesses of the method, in that those differing views of settlement might also contain historical (or aspirational) perspectives. This was partially overcome by the use of the search string “is a” where the present tense of the language would usually refer to more current viewpoints. Coupled with this – known genealogy and history websites (such as genuki.org.uk) were excluded within the rules from passing on terms.

The number of usable data returns was reduced from 27,531 records to 17,476 when duplicate data (same domain and content) or non-settlement references (identified manually[13]) were also removed (see examples in Table 3.4). Thus, only 6.8% of all data collated was used. Whilst this may seem a little wasteful, the effort was not in person time but computer processing and web traffic. The process was stable considering the number of total requests sent. With requests sequentially ordered on each settlement, the procedure was extremely suitable for horizontal scaling and the process was easy to recommence after any failure.

3.3.4 Validation method

Such an approach as outlined above would only be of merit if it were possible to demonstrate that the properties of ‘real space’ and virtual ‘web space’ were similar

[13] output were exported to a database and sorted on particular fields in order to identify groups of related entity, such as data relating to named Church(es) rather than settlement(s). Whilst such cases could in theory be extracted automatically, the ever-diminishing numbers of data records solved by an additional computational rule would require exponentially increasing effort and there would always be a requirement for some form of manual check to identify remaining false positives

Table 3.4 Examples of records with no settlement classification reference for the search query

Many of the world famous cruise liners will simply sail past your window! **Hythe is a** 12min Ferry trip away from **Southampton**. The Isle Of Wight ...
Ysgol Rhoscolyn. Ysgol **Rhoscolyn is a** School in Holyhead, **Gwynedd**. To contact Ysgol Rhoscolyn, please call their phone number on: Phone: 01407 860264
The **Beaulieu is a** wonderfully attractive tidal river tucked away within the confines of the **New Forest** National Park. The river itself is very unusual

(Pontikakis and Navratil, 2006). The aim of the validation procedure was to determine if people used similar terms to those found by trawling for data on the web. It might be the case that people who post on the web use a different language to describe settlement hierarchies, and thus, may not be representative of the population more generally. The research hypothesis was "Are vague settlement classifications generated from web data representative of those obtained through interviewing people?"

281 questionnaires were completed by Nottinghamshire residents. Participants were recruited through a variety of online sources (twitter accounts of Nottinghamshire County Council[14], Nottingham City Transport[15] and Reddit page for Nottingham[16]) and through attendance at local fitness classes held in a number of village locations within Nottinghamshire. Participants were asked to categorise twenty different places into one of the following classes:

City, Borough, District, Suburb, Neighbourhood, Town, Small Town, Large Town, Market Town, Former Mining Town, Village, Small Village, Large Village, Hamlet, None (I do not know the place), Other.

Each respondent was consulted on 20 places consisting of: 10 national and larger, well-known local settlements; along with a further 10 smaller settlements which were selected dependant on which area of Nottinghamshire each respondent was most familiar with (north, south, east, west or central). This was necessary as people would be unlikely to reliably know about the very small places other than those local to them. The settlements were chosen to give an even spread in terms of population size. This gave settlement classification data for a total of 60 different places (as listed in Table 3.11). The numbers of responses in each settlement category for each place were then compared between questionnaire output and web extraction data. In order to test the similarity of the two sets of data - linear regression was undertaken between the

[14]https://twitter.com/nottscc
[15]https://twitter.com/nct_buses
[16]http://www.reddit.com/r/nottingham

paired responses for each settlement classification term and place between the web extraction and questionnaire output.

3.4 Results and discussion

In total, 17,476 records were obtained containing 2,079 unique terms for the 2,426 settlements within the case study areas. There was, however, a large tail to the data where only 16 unique terms received over 100 occurrences within the web extraction output (as illustrated in the next section). This on average produced 7.2 terms for each settlement. The settlement information obtained clearly supported the frequently vague nature of classification, with individual places often having different views concerning how it was described. Table 3.5 illustrates the different perspectives obtained for selected settlements to the north of Leicester. Opinion was divided as to whether Syston should be described as either a village or town (a ratio of 0.6 to 1 respectively) - perhaps explaining why a number of data returns also simply referred to the place as a 'bustling community'. The duality of Thurmaston as both a 'village' and 'suburb' was of interest - as a historic village that has subsequently been incorporated into the wider urban area of Leicester. In such instances, it is easy to conceive how people could have a different perspective on how to classify the place. Hence, reinforcing the desirability of a vague classification where a place can be classified within a number of different categorises. Exploring the language used to describe places may also provide details about the nature of the area. For example, the reference to Rothley as a 'thriving community' could in part be a reflection on the affluence of the local vicinity.

3.4.1 Language

In addition to forming vague settlement classifications, the language and terminology used by people to describe settlements may provide new opportunities for exploring how place is described and aid our understanding of settlement semantics. It may also offer suitable terms for new settlement classifications with which the public would hold greater empathy. Table 3.6 illustrates the frequency of the top 35 settlement terms from the extracted 17,476 records. A number of important points were apparent including: prominence of specific terms, the use of words denoting size (large, small), the use of descriptive words (such as popular, picturesque and delightful for example), and the highly skewed frequency distribution.

The most popular terms were: village, hamlet, suburb, town, parish, and district. It should be noted that the term city only appears at rank 19 with 94 data returns. This

Table 3.5 Examples of web based settlement classification for selected settlements to the north of Leicester

Settlement name	**Web returned terms**	**Frequency**
Syston	bustling community	9
	town	8
	large village	4
	small town	4
	village	3
	suburb	1
Thurmaston	village	8
	suburb	5
Birstall	large village	10
	suburb	3
	village	3
Barrow upon Soar	large village	10
	village	9
	residential and industrial township	2
Quorn	village	26
	government town	1
	small hamlet	1
Rothley	village	18
	place	3
	thriving community	2
Sileby	village	9
	large village	9
	former industrial village	3
	parish	1
Mountsorrel	village	9
	popular residential location	6
	small village	1
	townhouse development	1
Anstey	large semi industrialised village	5
	semi industrialized village	4
	village	5
	large residential village	2

Table 3.6 The most common settlement terms extracted from the Bing API request data

Settlement term	Frequency	Frequency (%)	Cumulative frequency	Occurrences of term on the web	Normalised (per 10k web occurrences)	No. of difference settlements	Average no. of terms per settlement
village	3,160	18.1	18.1	1,180,000,000	0.027	1,059	2.98
small village	1,288	7.4	25.5	13,700,000	0.940	558	2.31
town	1,084	6.2	31.7	9,780,000,000	0.001	170	6.38
village and civil parish	1,076	6.2	37.8	287,000	37.491	394	2.73
hamlet	552	3.2	41.0	50,000,000	0.110	316	1.75
suburb	509	2.9	43.9	81,900,000	0.062	223	2.28
district	489	2.8	46.7	1,440,000,000	0.003	96	5.09
small town	388	2.2	48.9	59,300,000	0.065	125	3.10
market town	287	1.6	50.5	5,130,000	0.559	46	6.24
large village	229	1.3	51.9	390,000	5.872	79	2.90
small village and civil parish	188	1.1	52.9	54500	34.495	74	2.54
place	184	1.1	54.0	5,420,000,000	0.000	97	1.90
civil parish	184	1.1	55.0	474,000	3.882	78	2.36
popular village	175	1.0	56.0	478,000	3.661	69	2.54
small market town	169	1.0	57.0	295,000	5.729	40	4.23
parish	145	0.8	57.8	157,000,000	0.009	106	1.37
small hamlet	99	0.6	58.4	419,000	2.363	50	1.98
town and civil parish	94	0.5	58.9	152,000	6.184	12	7.83
city	94	0.5	59.5	7,090,000,000	0.000	11	8.55
village and parish	82	0.5	59.9	125,000	6.560	59	1.39
settlement	68	0.4	60.3	166,000,000	0.004	59	1.15
small rural village	63	0.4	60.7	201,000	3.134	29	2.17
delightful village	61	0.3	61.0	1,440,000	0.424	31	1.97
borough	60	0.3	61.4	91,900,000	0.007	17	3.53
rural village	56	0.3	61.7	774,000	0.724	28	2.00
large village and civil parish	55	0.3	62.0	21300	25.822	14	3.93
hamlet and civil parish	55	0.3	62.3	8880	61.937	16	3.44
picturesque village	53	0.3	62.6	2,170,000	0.244	37	1.43
neighbourhood	53	0.3	62.9	97,000,000	0.005	24	2.21
pretty village	46	0.3	63.2	433,000	1.062	25	1.84
thriving market town	43	0.2	63.5	111,000	3.874	14	3.07
market town and civil parish	43	0.2	63.7	62,000	6.935	5	8.60
part of	42	0.2	63.9	2,120,000,000	0.000	28	1.50
parish and village	42	0.2	64.2	78,700	5.337	29	1.45
thriving village	41	0.2	64.4	186,000	2.204	26	1.58

was partially to be expected, however, due to the low number of potential occurrences of such a high order settlement form within the case study sample. The highly skewed nature of the frequency data was prominent. Over 18% of data returns related to a single term (village) and the top ten terms accounted for over 52% of records. There was, however, also a heavy tail to the distribution with 28% of data having fewer than twenty occurrences. As anticipated, terms one might expect with substantial populations (such as cities and towns) tended to have more data returns (in the order of magnitude of 6-9 data returns), many of the smaller settlement terms were dependent on small data sizes (perhaps only 2 or 3 data returns). This may influence the stability of the settlement classification at the individual settlement level and was considered further within the validation section. In time, such data volume issues may be overcome due to both the growth of data held on the web and the increasing indexing of the web. Further search queries (in addition to "is a") could also be used to increase data volumes.

In contrast to the approach taken in the previous table, Table 3.7 portrays the frequency of the top twenty-five settlement *words* from the settlement *terms*. Thus, a term such as 'village and civil parish' was split into each word – each with a count of one. This 'bag of words' approach derives its name from the fact that the ordering of the words is not taken into consideration. The method was of use to ensure that specific words of interest were not lost within the collective term. Words like 'community' may appear within a wide range of different terms (such as 'thriving community', 'vibrant community', 'multi-cultural community' for example), none of which themselves appear extremely frequently but taken together the concept which they represent – 'community' was more prevalent as a whole. Whilst the term, community, does not feature within Table 3.6 (being ranked 56th with only 29 occurrences), it ranked 19th in Table 3.7 with 208 data returns. In total the word community was used within 51 different terms. Thus, investigation using the *combination* of *terms* and *words* was shown to be important. Also of note was the preference to call sub-parts of urban areas as 'suburbs' (or 'districts' in London) over other terms such as 'neighbourhoods'. The practice of differentiating the size of the object using the qualifier 'small' or 'large' was demonstrated to be relatively common (although places were five and a half times more likely to be defined as 'small' compared to 'large').

3.4.2 URL classification

It was necessary to demonstrate the type of websites from where data containing settlement classification information were derived. For example, the data might be heavily biased from one website or a group of websites (such as online encyclopaedias).

Table 3.7 The most commonly cited settlement words (splitting up the terms into individual words)

Frequency	Word	Freq. cont.	Word cont.	Freq. cont.	Word cont.
8,508	village	245	city	129	coastal
3,405	town	208	community	118	pleasant
3,071	small	206	thriving	116	base
2,667	and	191	picturesque	115	new
2,432	parish	189	very	115	busy
1,955	civil	185	after	108	highly
950	market	184	sought	107	local
890	hamlet	173	borough	106	well
863	suburb	164	location	104	mining
773	district	148	pretty	102	seaside
553	large	146	historic	98	government
531	popular	143	part of	96	little
482	area	143	delightful	94	suburban
409	rural	139	great	91	vibrant
342	residential	132	quiet	91	lovely
342	place	132	beautiful	82	resort
289	former	129	settlement	81	neighbourhood

Exploring the domain names associated with the data demonstrated that a small number of sites appeared to give a high number of returns. In total, data were drawn from 4,285 different website domains. The top ten sites accounted for over 30% of data collected (see Table 3.8). Wikipedia alone accounted for 12% of data returns – which was clearly a significant proportion. Estate agent sites such as estateangels.co.uk, rightmove.co.uk, zoopla.co.uk and findaproperty.com returned 3.8%, 2.2%, 1.0% and 1.0% of all data respectively. [17]

The data was highly skewed with a heavy tail, with many websites only providing a very small number of data returns. Nearly 40% of data were from websites that provide fewer than ten data returns to settlement classification terms. Such sites were important as they were more likely to provide local information (for example local council pages such as www3.hants.gov.uk; local tourist sites such as visitnorfolk.co.uk; or local businesses sites such as nottspubs.co.uk).

Youtube and Facebook domains were ranked 29th and 30th respectively – each providing approximately fifty returns (each representing 0.3% of total data). Somewhat surprisingly, no entries regarding settlement classification were found from Twitter.

[17] ejje.weblio.jp is a Japanese encyclopaedia whilst serol-ogy.gnorimiesgiasex.com is a biology orientated encyclopedia and although both sites often cite references from wiki – they also provide content from other sources

Table 3.8 The most common domain names posting settlement classification from the web extraction data

Domain Name	Frequency	Frequency (%)	Cumulative frequency (%)
en.wikipedia.org	2,109	12.1	12.1
ejje.weblio.jp	794	4.5	16.6
estateangels.co.uk	668	3.8	20.4
rightmove.co.uk	379	2.2	22.6
serology.gnorimiesgiasex.com	353	2.0	24.6
at1ce.org	333	1.9	26.5
factbites.com	322	1.8	28.4
duckduckgo.com	190	1.1	29.5
zoopla.co.uk	180	1.0	30.5
findaproperty.com	176	1.0	31.5
information-britain.co.uk	132	0.8	32.2
uk.ask.com	130	0.7	33.0
homes.trovit.co.uk	125	0.7	33.7
allhoteldeals.net	116	0.7	34.4
uk.epodunk.com	115	0.7	35.0
apling.freeservers.com	89	0.5	35.5
wordiq.com	77	0.4	36.0
primelocation.com	77	0.4	36.4
globrix.com	75	0.4	36.9
norfolk-norwich.com	72	0.4	37.3

3.4.3 Synonymy and related concepts within settlement terms

A synonym is a word or phrase that has the same or nearly the same meaning as another word. Synonyms with an identical meaning are termed cognitive synonyms whilst those with similar but non-identical meanings are coined near-synonyms (Stanojević, 2009). A harbour may also be termed a wharf, dock, seaport or dockyard by different people with differing perspectives.

One might expect certain settlement classes to be more closely related to some than others, for example a hamlet is more likely to be described by others as a small village than a city. It was possible to identify settlement classification synonymy by identifying the different terms used to describe the same settlement. Examples of related words are shown within Table 3.9. Where a place was called 'town' it was likely that other data may refer to the same place using such terms as 'small town' and 'market town'. One of the few surprises within the table was that the terms 'suburb' and 'neighbourhood' were not used to describe the same place interchangeably and people were more likely to call a 'suburb' a 'district', 'town' or 'village' than a 'neighbourhood'. This will be further explored within the next chapter.

3.4.4 Relationship with population

It was also possible to examine the association between terms and the underlying population. There are, however, few datasources that provide measures of population estimates at scales suitable for small spatial objects such as hamlets. Even Census data available at Output Area (OA) and one of the smallest statistical units for which data are commonly provided would be too large for small settlements, such as hamlets (OAs cover on average 129 households and are not constrained by settlement form). To overcome this issue, the number of delivery points from the Royal Mail's Postcode Address File (PAF) were used as a proxy for the number of households (as previously utilised by Bibby and Brindley (2006, 2007, 2010)). The boundary of the settlements were drawn from the same Ordnance Survey (OS) - MRDP data used for the Bing API search on settlement names. Whilst data for freestanding units such as villages and towns appeared unproblematic within MRDP, the extent of many of the places *within* urban areas raised some concerns. Whilst fit for its cartographic requirements within OS, the data (constructed from the plethora of place name annotation appearing on OS maps through time) appeared to contain a large number of historic names. This becomes problematic when, within the data, areas are not allowed to overlap - thus, the historic areas appear to 'push out' more current place names. This was the case within Figure 3.3 where the more historic names of Tapton Hill, Summerfield and Broomfield

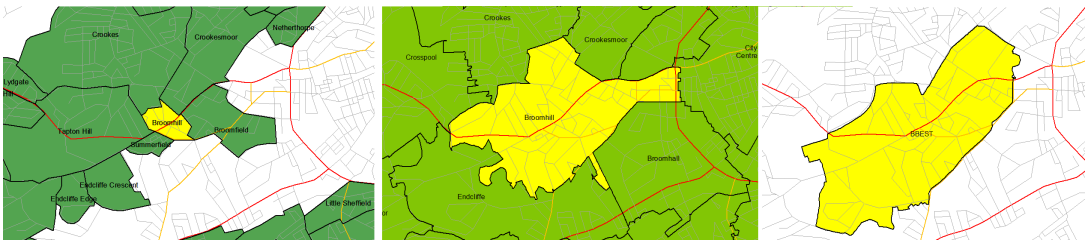
Table 3.9 Semantic Content: Relationship between Settlement Terms

Term	Frequency	Alternative terms and percentage of co-existing term frequency
district	19,927	[borough: 1.2], [central: 1.2],[neighbourhood: 1.1],[beautiful area: 0.6],[suburb: 0.5]]
town	8,594	[small town: 2.4], [market town: 1.2], [suburb: 1.1]]
small town	7,471	[town: 2.7], [market town: 1.9], [small market town: 1.8],[small village: 1.2]]
borough	6,573	[district: 3.7], [central: 1.4], [neighbourhood: 1.1]]
neighbourhood	6,279	[district: 3.4]]
market town	5,630	[historic market town: 2.8], [small town: 2.5], [small market town: 2.4],[busy market town: 2.1],[delightful market town: 1.8]]
small village	5,277	[village: 6.2], [hamlet: 1.9], [small town: 1.7]]
city	4,773	[coastal city: 1.3], [famous city: 1.2], [large city: 1.2],[historic city: 1.0],[large affluent city: 1.0]]
suburb	4,737	[district: 2.3], [town: 1.9], [village: 1.7]]
village	4,277	[small village: 7.7], [village and civil parish: 4.0], [large village: 2.1],[popular village: 1.9],[parish: 1.9],[suburb: 1.8]]
small market town	3,208	[market town: 4.2], [small town: 4.2]]
large village	2,729	[popular village: 3.4], [village: 3.2]]

within Sheffield limit the bounded extent of Broomhill. For this reason, populations were not estimated for those places within urban areas. This issue of bounded objects within urban areas is explored in greater detail in subsequent chapters. The output of tying population of MRDP objects and corresponding settlement classification terms extracted from the web can be found within Table 3.10 and Figure 3.4.

Reassuringly there were few surprises within the output. 'Cities' were, on average, said to be larger than 'large towns', which in turn had greater average populations than 'towns' and 'small towns'. 'Villages' were on average smaller than 'towns' and 'hamlets' were the smallest settlement category. 'Urban' areas in the UK are officially defined as settlements that exceed 10,000 population. With reference to Table 3.10 this would include places in order of magnitude of large villages and larger[18]. The most appropriate scale for an 'urban' distinction on the basis of Table 3.10 would appear to be around 13,800 population (6,000 households) and one might argue that with appropriate rounding, a 15,000 population threshold might have been a more realistic figure in relation to reflect towns and cities. The terms 'popular' and 'thriving' villages were more likely to relate to villages with higher populations than other terms such as 'rural' or 'pretty'. More importantly, however, were the variations and overlap *between* categorises, as clearly shown within the boxplots of Figure 3.4. The overlap between small town and large village was particularly insightful, where the median and lower-quartile ranges appeared near identical. The spread and overlap between the differing categorisations clearly supports the notion and requirement of vague settlement classification where a single population threshold cannot accurately reflect how people perceive and classify settlement. Critically, this also demonstrated that population size alone is not an accurate measure on which to base a settlement classification.

Fig. 3.3 Differing extents for Broomhill, Sheffield



a) MRDP *b) Sheffield Council defined* *c) Neighbourhood Forum*[19]

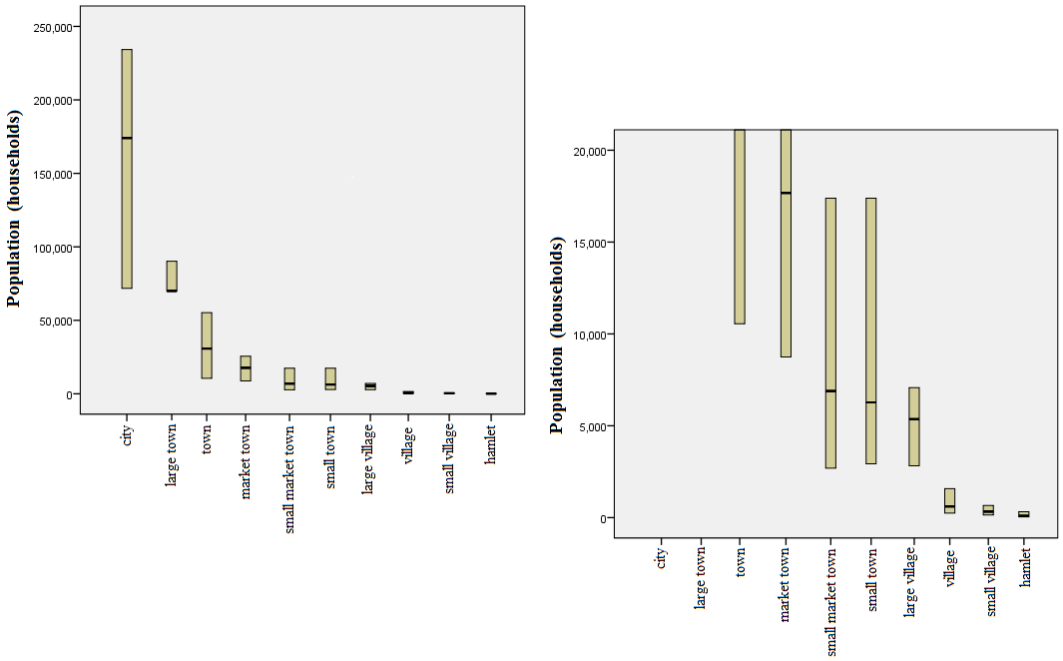
[18]Using an overall conversion ratio of 2.3 people for every household, as found within 2011 Census of Population data

[19]The BBEST area is "roughly defined as Broomhill" according to the BBEST website - http://bbest.org.uk/our-area

Table 3.10 Settlement Terms and Household Population

Term	Median population (households)	Mean population (households)	Standard Deviation of population (households)
city	174,047	163,797	87,193
large town	69,987	68,977	18,766
town	30,765	36,285	39,055
market town	17,675	21,867	21,251
town and civil parish	15,642	14,063	10,668
thriving market town	14,227	17,641	13,398
port	14,227	22,375	42,338
historic market town	13,387	17,038	10,314
market town and civil parish	10,058	12,262	5,670
small market town and civil parish	10,058	7,270	3,739
small market town	6,890	10,800	10,470
small town	6,272	12,169	14,184
large village	5,360	5,757	4,142
large village and civil parish	4,186	4,797	2,548
semi rural village	1,696	1,483	1,390
popular village	1,389	2,126	1,828
thriving village	1,094	2,017	2,091
quiet village	1,050	1,269	1,114
picturesque village	1,003	1,279	1,360
pleasant village	862	1,584	1,430
delightful village	779	1,179	1,304
beautiful village	708	1,130	1,413
sought after village	675	1,378	1,368
village	603	1,510	2,982
village and parish	465	1,015	1,379
small rural village	447	574	591
village and civil parish	443	981	1,438
rural village	426	614	956
pretty village	413	1,001	1,375
parish and village	400	520	629
parish	389	1,421	2,797
small village	328	649	1,028
civil parish	312	858	1,619
small parish	253	444	818
small village and civil parish	218	266	271
small picturesque village	126	301	338
hamlet	97	327	943
small hamlet	93	302	1,418
hamlet and civil parish	49	42	19

Fig. 3.4 Average Related Population for Selected Settlement Classification Terms



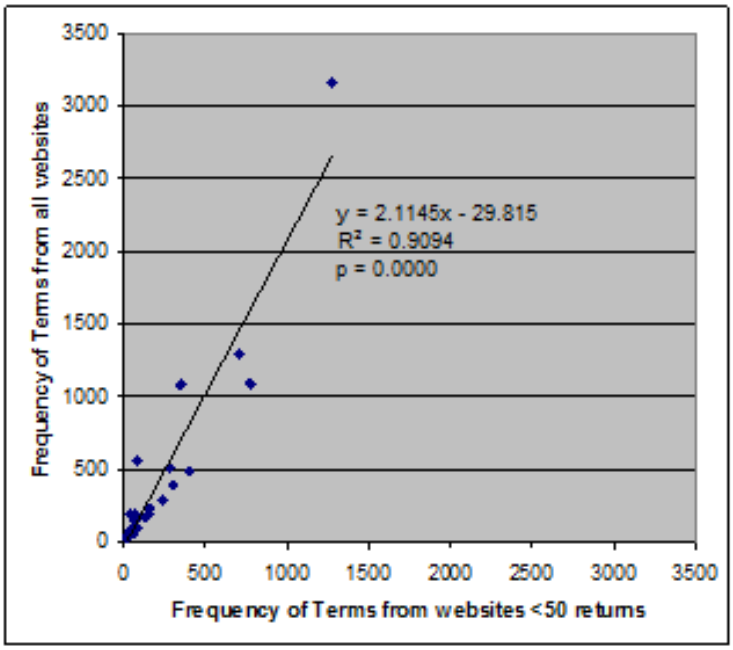
a) For all household sizes *b) Inset for under 20k Households*

3.4.5 Validation

In order to test the robustness of the output and ensure that data were not merely the representation of a select few web datasources, frequency tables were recalculated only for those websites providing fewer than fifty data returns. This was an important test as Wikipedia alone accounted for over 12% of data (Table 3.8) and it was felt that such data may bias the output. As shown within Figure 3.5, there was, however, a very strong relationship between the frequency of terms from the two datasets (from all websites and from websites with fewer than fifty returns) with an R^2 of over 90%. Thus, supporting the view that data were not merely a representation of a dominant small number of websites.

Secondly, output from the web extraction method and 281 questionnaire responses were compared for 60 different settlements. An example of output for 10 selected settlements can be found within Figure 3.6, whilst the full comparison can be found within Appendix A. The overall impression from Figure 3.6 was one of similarity, there were, however, also noticeable disparities. Carlisle and Salford, for example, were far more likely to be described as a 'town' and 'borough / district' respectively by Nottingham residents than in the web extraction output which had a greater likelihood of tagging the settlements in their official viewpoint as a 'city'. The figure also provides strong support for the notion of a vague settlement classification, with very rare occurrences of total agreement concerning the category a specific settlement

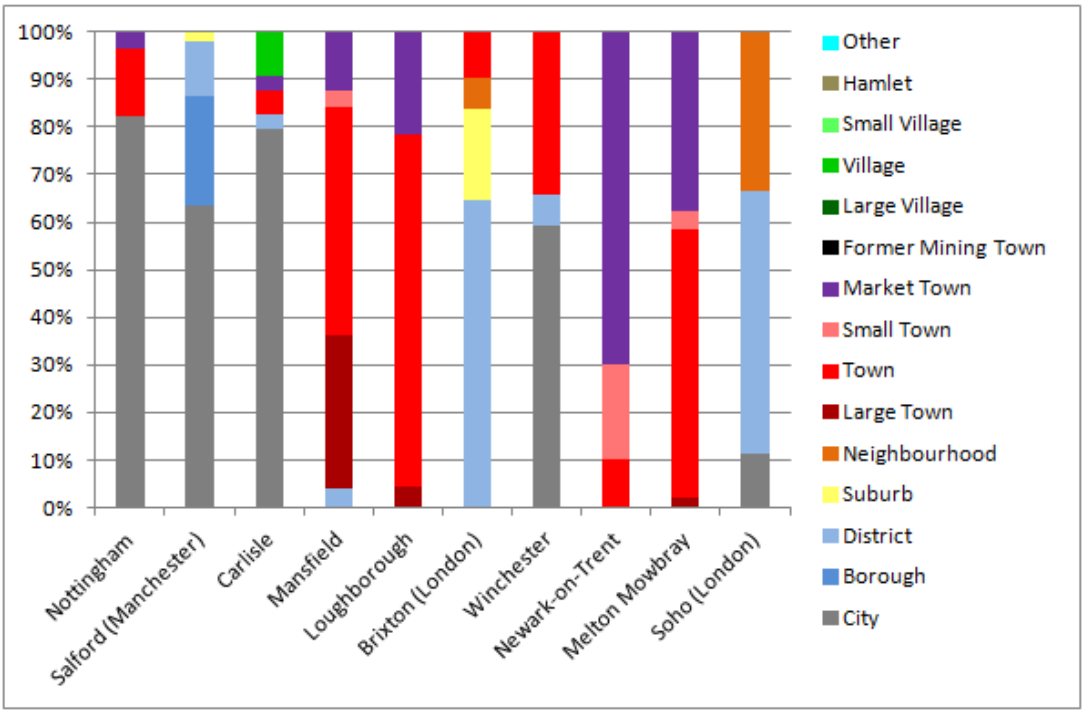
Fig. 3.5 Robustness test: Comparing output for all websites and only those with less than 50 data returns



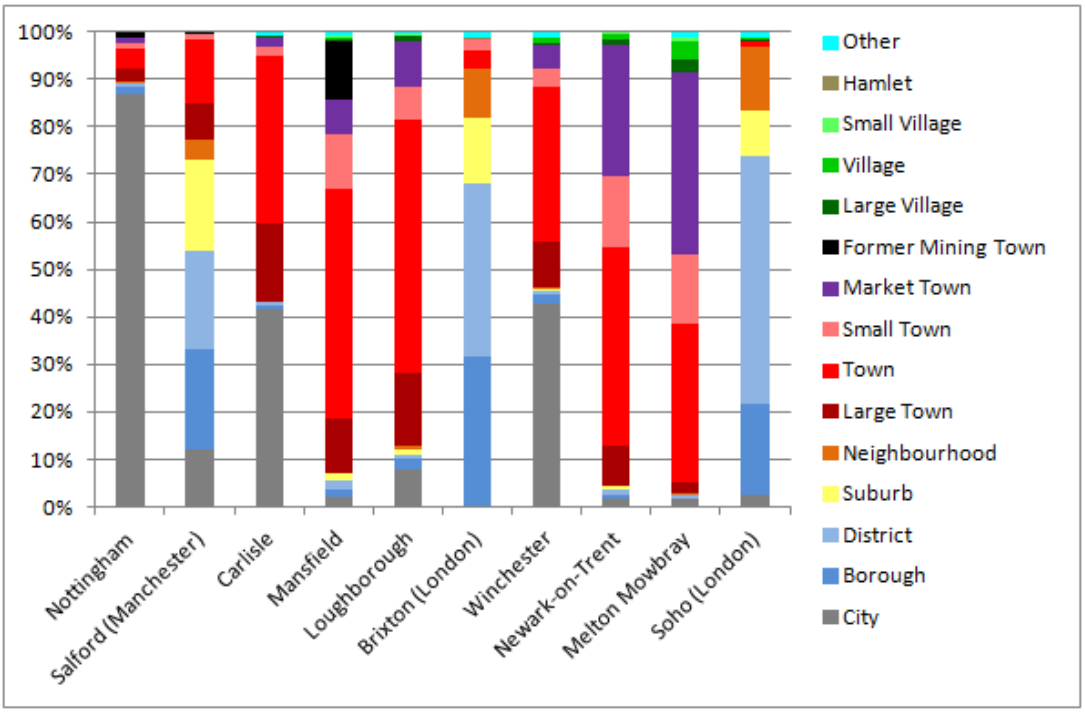
should be classified as. The strongest agreement within the figure was for Nottingham with 86% of questionnaire respondents reporting the place as a 'City', compared with 82% for the web extraction output. In contrast, Mansfield was described with a multiple of four different classes by at least 10% of respondents ('town', 'small town', 'former mining town', 'large town'), with the top category of town only accounting for 48% of responses. This compared favourably with three categories with over 10% of data within the web extraction data ('town', 'large town', 'market town'), with again 48% agreeing to classify the settlement as a 'town'.

In order to further validate the vague settlement classification extracted from the web, linear regression was undertaken for the paired responses - comparing each settlement classification term from web extraction with that for the same settlement term obtained from the questionnaire output. Analysis was undertaken both for all settlements as a whole and also for each individual settlement. For example, comparing the percentage of respondents who classified a settlement using each classification term (city, town and so forth) with corresponding data extracted from the web. Details of which can be found in Table 3.11. This shows that the web extraction output explained 46% of the variation in questionnaire responses ($y = 1.1073x - 0.6709, p < .001$). When size distinctions were removed (ie not distinguishing between 'small village' and 'village' categorises), the amount of variation explained increased to 72% - supporting the notion that the questionnaire responses were more likely to include a size distinction within their categorisation ($y = 1.0654x - 0.6132, p < .001$). This may be because

Fig. 3.6 Comparison between web extraction output and questionnaire responses for selected settlements



a) From web extraction output



b) From questionnaire responses

residents were asked to select the most appropriate category from a provided list which contained sized options. If residents had to propose their own categorisations they may well have been less likely to make size distinctions. Such a consideration, however, was outside the scope of this work, whereby the focus was on the creation of the vague

settlement classifications themselves. One issue with size distinctions is that they are dependant upon what the comparison is made against - in other words they are relative. Mansfield may be considered a small town by 12% of questionnaire responses but also as a large town by a further 11%. Both are true relative to the comparator - Mansfield is a large town compared to smaller places such as Melton Mowbray or Newark-on-Trent, but a small town when compared against Greater London or Nottingham.

The web extraction classification was closer to that obtained from asking residents in certain settlement types than for others. Generally, those places within urban areas (such as neighbourhood, suburb, district and borough classifications) compared less favourably than other 'free-standing' place types. The variance explained by free-standing settlements was three times that for places within urban areas (26%[20] compared with 83%[21] when ignoring size distinctions), perhaps due to the complexity of such areas, as parts within a larger entity. This complexity was hinted at by the fact that within the residents questionnaires, on average 10 of the 16 terms were used to classify places within urban areas compared to 7.6 terms on average to describe free-standing settlements. This demonstrates that some place types (such as those within urban areas) were harder to define, and hence, more vague than others.

The regression relationship between the questionnaire and web extraction data (excluding the size distinction) can be found within Figure 3.7 for both (a) all settlements and (b) only free-standing settlements (removing those places within urban areas). Whilst, the amount of variation explained increases from 72% to 83% the plot still exhibits noticeable scatter - especially whereby a reasonable number of residents described the place in a certain category but there were no matches at all in the corresponding web extraction data. Such was the case for Keyworth, East Leake, Blyth and Ruddington which although described by the majority of questionnaire responses as 'villages' were also deemed to be called 'towns' by a quarter of responses. The web extraction data, however, had no responses to the places as 'towns', instead with a far greater propensity to identify the settlements simply as 'villages'. Such an effect could potentially be partially aspirational where residents may aspire to the place being larger whilst the web extraction method might tend to focus on the dominant (and often official) perspectives.

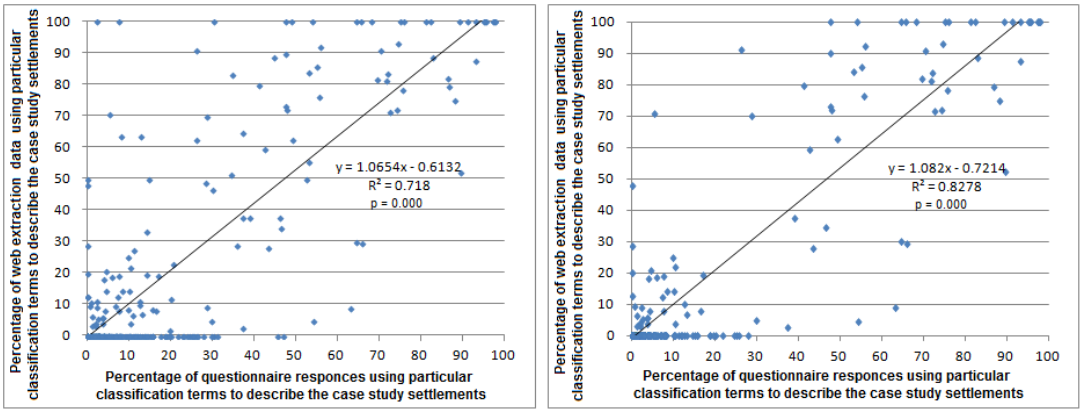
Concerns over the low number of data returns for small individual settlements has previously been raised. This said, however, validation did not prove any worse for smaller settlements (where the majority of residents recorded the place as either 'village' or 'small village') when compared to other settlements. In fact the average variance within the resident questionnaires explained by the web extraction method

[20] $y = 1.0927x - 2.0011, p < .001$
[21] $y = 1.082x - 0.7214, p < .001$

Table 3.11 Comparison of Settlement Classification between Questionnaire Responses and Web Extraction data at the individual settlement level (significance levels: * p < .05, ** p < .01)

Settlement name	Equation (all)	R^2	p value	sig	Equation (ignoring size)	R^2	p value	sig.
Nottingham	y = 0.9571x + 0.2683	0.982	0.000	**	y = 0.944x - 0.4915	0.965	0.000	**
Salford (Manchester)	y = 0.9575x + 0.2656	0.180	0.101		y = 0.6346x + 3.3218	0.104	0.333	
Carlisle	y = 1.2061x - 1.2883	0.495	0.002	**	y = 0.7104x + 2.6327	0.335	0.062	
Mansfield	y = 0.9963x + 0.0228	0.748	0.000	**	y = 1.16x - 1.4542	0.965	0.000	**
Loughborough	y = 1.3583x - 2.2392	0.918	0.000	**	y = 1.0527x - 0.4791	0.966	0.000	**
Brixton (London)	y = 1.1245x - 0.7781	0.570	0.001	**	y = 1.097x - 0.8815	0.548	0.009	**
Winchester	y = 1.3839x - 2.3995	0.829	0.000	**	y = 1.0445x - 0.4042	0.871	0.000	**
Newark-on-Trent	y = 1.0075x - 0.0469	0.415	0.010	**	y = 0.7451x + 2.3175	0.461	0.022	*
Melton Mowbray	y = 1.3095x - 1.9347	0.848	0.000	**	y = 1.1753x - 1.594	0.965	0.000	**
Soho (London)	y = 1.024x - 0.1501	0.757	0.000	**	y = 0.9989x + 0.0099	0.738	0.001	**
Worksop	y = 0.5192x + 3.0051	0.286	0.043	*	y = 0.528x + 4.291	0.483	0.018	*
Hucknall	y = 1.1414x - 0.884	0.788	0.000	**	y = 0.9373x + 0.5704	0.887	0.000	**
Alfreton	y = 1.2554x - 1.5962	0.631	0.000	**	y = 1.074x - 0.6728	0.990	0.000	**
Retford	y = 0.996x + 0.0248	0.238	0.059		y = 0.5302x + 4.2707	0.143	0.252	
Mansfield Woodhouse	y = 0.2838x + 4.4764	0.028	0.566		y = 1.1198x - 1.0888	0.534	0.011	*
Southwell	y = 1.3744x - 2.3402	0.650	0.000	**	y = 1.1493x - 1.357	0.785	0.000	**
Ravenshead	y = 2.1485x - 7.1782	0.758	0.000	**	y = 1.8166x - 7.4236	0.894	0.000	**
Lowdham	y = 1.6199x - 3.8744	0.851	0.000	**	y = 1.4151x - 3.7736	0.917	0.000	**
Blyth	y = 1.3044x - 1.9025	0.261	0.003	**	y = 1.3425x - 3.1133	0.798	0.000	**
Oxton	y = 1.6054x - 3.7838	0.810	0.000	**	y = 1.2531x - 2.3005	0.991	0.000	**
Keyworth	y = 0.822x + 1.1124	0.224	0.078		y = 1.1467x - 1.334	0.847	0.000	**
Ruddington	y = 0.7131x + 1.7929	0.295	0.038	*	y = 1.4208x - 3.8258	0.865	0.000	**
East Leake	y = 0.8229x + 1.1071	0.250	0.060		y = 0.855x + 1.3186	0.727	0.001	**
Tollerton	y = 1.7462x - 4.6638	0.886	0.000	**	y = 1.3546x - 3.2233	0.982	0.000	**
Gotham	y = 1.0023x - 0.0142	0.552	0.001	**	y = 0.8544x + 1.324	0.950	0.000	**
Sutton Bonington	y = 1.13x - 0.8124	0.719	0.000	**	y = 1.0316x - 0.2869	0.999	0.000	**
Kinoulton	y = 0.868x + 0.8251	0.298	0.014	*	y = 1.1055x - 0.9587	0.995	0.000	**
Stanton-on-the-Wolds	y = 1.3405x - 2.1283	0.921	0.000	**	y = 1.2158x - 1.9622	0.976	0.000	**
Plumtree	y = 0.7782x + 1.386	0.765	0.000	**	y = 0.9317x + 0.6209	0.976	0.000	**
Thrumpton	y = 1.4088x - 2.5551	0.611	0.000	**	y = 1.2581x - 2.3463	0.922	0.000	**
Bingham	y = 1.2963x - 1.8516	0.897	0.000	**	y = 1.1374x - 1.2487	0.840	0.000	**
Radcliffe-on-Trent	y = 0.318x + 4.2623	0.048	0.454		y = 1.2038x - 1.8524	0.687	0.002	**
Burton Joyce	y = 1.4734x - 2.9586	0.955	0.000	**	y = 1.2633x - 2.3936	0.988	0.000	**
Cropwell Bishop	y = 1.332x - 2.0748	0.872	0.000	**	y = 1.0516x - 0.4688	0.999	0.000	**
Aslockton	y = 1.514x - 3.2124	0.707	0.000	**	y = 1.0298x - 0.2714	0.999	0.000	**
Langar	y = 1.1463x - 0.9143	0.492	0.004	**	y = 1.0541x - 0.4919	0.999	0.000	**
Cropwell Butler	y = 1.0798x - 0.4986	0.464	0.005	**	y = 1.0257x - 0.2334	0.999	0.000	**
Hoveringham	y = 1.3461x - 2.1628	0.883	0.000	**	y = 1.1298x - 1.1802	0.996	0.000	**
Colston Bassett	y = 1.2417x - 1.5103	0.891	0.000	**	y = 1.082x - 0.7456	0.997	0.000	**
Granby	y = 1.4812x - 3.0074	0.767	0.000	**	y = 1.0584x - 0.531	0.999	0.000	**
Long Eaton	y = 1.4736x - 2.9598	0.710	0.000	**	y = 1.1492x - 1.3565	0.955	0.000	**
Beeston	y = 1.6283x - 3.9267	0.561	0.001	**	y = 1.5157x - 4.688	0.922	0.000	**
Ilkeston	y = 1.0079x - 0.0496	0.686	0.000	**	y = 0.9112x + 0.807	0.950	0.000	**
Eastwood	y = 0.4459x + 3.463	0.051	0.329		y = 0.386x + 5.5816	0.109	0.321	
Stapleford	y = 1.1195x - 0.7469	0.295	0.038	*	y = 1.4833x - 4.3941	0.877	0.000	**
Chilwell	y = 1.8026x - 5.0161	0.675	0.000	**	y = 1.8559x - 7.781	0.656	0.002	**
Kimberley	y = 1.384x - 2.4	0.218	0.052		y = 1.8462x - 7.6923	0.884	0.000	**
Trowell	y = 1.9261x - 5.788	0.661	0.000	**	y = 1.8138x - 7.3986	0.764	0.000	**
Awsworth	y = 1.4483x - 2.8017	0.406	0.000	**	y = 1.5189x - 4.717	0.941	0.000	**
Cossall	y = 0.7208x + 1.7451	0.142	0.014	*	y = 1.115x - 1.0453	0.880	0.000	**
West Bridgford	y = 1.3283x - 2.0521	0.613	0.001	**	y = 1.5943x - 5.4029	0.866	0.000	**
Arnold	y = 1.1958x - 1.2237	0.562	0.001	**	y = 1.0017x - 0.0154	0.661	0.002	**
Carlton	y = 0.7993x + 1.2544	0.424	0.007	**	y = 0.8551x + 1.3176	0.418	0.032	*
Gedling	y = 0.1222x + 5.4861	0.003	0.801		y = 0.1725x + 7.5226	0.007	0.808	
Bilborough	y = 2.0306x - 6.4412	0.842	0.000	**	y = 1.9485x - 8.6229	0.851	0.000	**
Hyson Green	y = -0.2508x + 7.8172	0.011	0.700		y = -0.4432x + 13.12	0.034	0.590	
Wilford	y = 0.5033x + 3.1042	0.041	0.416		y = 0.9123x + 0.7974	0.213	0.153	
Colwick	y = 1.948x - 5.925	0.779	0.000	**	y = 1.8655x - 7.8682	0.838	0.000	**
the Lace Market	y = 0.6993x + 1.8794	0.312	0.032	*	y = 0.6534x + 3.1508	0.273	0.099	
Hockley	y = 0.034x + 6.0373	0.000	0.960		y = -0.0902x + 9.9108	0.002	0.903	
Total	y = 1.1073x - 0.6709	0.459	0.000	**	y = 1.0654x - 0.6132	0.718	0.000	**

Fig. 3.7 Relationship between questionnaire output and web extraction data (excluding size distinctions)



a) All settlements included *(b) Free-standing settlements only*

was consistently higher for smaller settlements (67 and 68% for villages and small villages respectively compared with 47% for larger settlements for all data; and 94 and 98% compared with 64% when size distinctions were not included). Given this, it appears that data volumes within smaller settlements did not have a detrimental effect on the validity of the output. Overall, it has been shown that the research hypothesis can be upheld and that vague settlement classifications generated from web data were representative of those obtained through interviewing people, particularity when size distinctions were excluded.

3.4.6 Fuzzy-babilistic

The question concerning the extent to which people perceived settlement objects as either fuzzy or probabilistic in nature has remained so far unaddressed. The questionnaire purposefully facilitated but did not explicitly state that respondents could select more than one category for any given settlement. Viewing objects as a mix of multiple categories would fit closer to the fuzzy model than a probabilistic approach (the questionnaire specifically requests participants to answer how *they* personally would define the settlement). Whilst the vast majority only defined settlements belonging to a single class, 7.6% of responses recorded settlements in more than a single category. The most common co-category combinations were:

- Borough and District
- District and Suburb
- District and Neighbourhood
- Town and Market town
- Suburb and Neighbourhood

- District and Town
- Town and Small town

Co-category combinations were also more likely in places within or part of a larger settlement (11.9% of the 1,733 data returns) compared to freestanding settlements (5.0% of the 2,712 responses). Whilst the majority of responses did not necessarily conceive objects in a fuzzy approach, the possibility that objects can be a combination or mix cannot be denied. Under such a proposal, the settlement classification adopted might perhaps be best conceived as fuzzy in nature.

3.5 Discussion and settlement classification conclusions

Places can be complex entities that are constantly changing and tied to their historic past. Thus, it is quite easy to comprehend why a place such as Gedling, Nottingham was described in a multiple of manners. It was a historic village and maintains a number of listed buildings from this era (hence data relating to Gedling as a 'village') but has since become enveloped by the growing urban area of Nottingham (data for Gedling as a 'suburb') and has given its name to both the local borough council (in 1974) and Parliament constituency (in 1983) (hence references as a 'borough' and 'district'). Thus, Gedling *is* all of these terms at the same time. This gives weight to the rational of a vague settlement classification, as a place can be described in many different perspectives. Such an approach offers an alternative to the standard current settlement classifications which categorise the settlements purely based on some quantitative measure (such as population or other measurements of form or function) and that require discrete, mutually exclusive categorisation.

The importance and value of the concept of vagueness within settlement classification is dependent upon the extent to which people may perceive the same settlement in different classes of categorisation. On average, questionnaire responses classified a place in 2.8 different classes of the 15 available with at least a level of 10% agreement amongst participants - demonstrating the highly variable degree to which different people perceive and classify place. When size distinctions were ignored, the figure was still high, standing on average at 2.1 of the 9 classes with a minimum of 10% agreement. In all case study areas within the validation procedure, the highest level of agreement between participants stood at 87% for the description of Nottingham as a 'city'. In 46 of the 60 case study areas, however, participants failed to agree with 50% of responses or higher in a single category. Even when discounting for size distinctions, 15 of the 60 case study areas still failed to achieve a 50% level of agreement within a single category - supporting the divergent nature in which people classify places.

This chapter has demonstrated the high level of variations in how people classify places and as such has shown the importance of constructing vague attribute data for places (through the use of a vague settlement classification). The chapter has illustrated how this can be achieved - through the mining of passive-crowdsourced data held on the web and using very simple rules to extract the required content. The method performed well when validating the web extraction output with resident perceptions of settlement classification ($y = 1.0654x - 0.6132, p < .001, R^2 = 0.72\%$, when size distinctions were excluded). The validation, thus, supports the research hypothesis that vague settlement classifications generated from web data *are* representative of those obtained through interviewing people when size distinctions were excluded.

Output also demonstrated that some settlement types (those within urban areas) were harder to classify and define than others (such as those of freestanding settlements). Thus, some place types are more vague than others. The high degree of variations found within the levels of population associated with the differing settlement classes demonstrates that population size alone is not an accurate measure on which to base a categorisation that represents public perceptions of settlement classification. Purves et al. (2011) acknowledge that even substantial digital archives can be prone to issues of bias and contributor inequality and that such issues are frequently hard to disentangle from valid data. As demonstrated by the validation procedures, the method appears robust; even for smaller settlements which are likely to have fewer data-returns, and thus, be more susceptible to bias. Whilst there is still work required for the method to be fully automated - the chapter has demonstrated the proof of concept, viability and merit of undertaking such an approach.

In addition to the formation of vague settlement classifications, it has been shown that the method can also aid the exploration of how places are more generally described. Lupton et al. (2011) stress that the names of categories such as 'mining town' or 'multicultural inner city' are particularly valuable to policy users. There is scope for further investigating the language used to describe specific classes of settlement. For example, even though 'market towns' are often seen as important drivers within the rural economy (for example within LEADER which delivers Rural Development Programme funding at a local level[22]), defining them has proved more problematic. The 2005 Rural-Urban definition for Local Authorities (RERC, 2005) identified 'larger market towns' as settlements with populations of 10,000 to 30,000 with an agglomeration of specified services. In contrast, the 2014 update of Rural-Urban definitions for Local Authorities (Bibby and Brindley, 2014) adopted an approach closer to other elements of the Rural-Urban Classification by using dwelling and business densities to identify settlements larger than 10,000 population but with a significant role as a service hub

[22]https://www.gov.uk/government/publications/leader-approach-in-the-rdpe-national-delivery-framework/leader-approach-in-the-rdpe-national-delivery-framework

for their rural hinterland. It would be interesting in future work to compare towns referenced as 'market towns' within web data with those settlements other methods identified as such.

Moving to the next chapter:
Whilst this chapter has demonstrated the potential for vague attribute data (through the place type of settlement classification), the following chapter will explore how vagueness can be utilised within another of the three elements of place objects - the spatial extent (or boundaries). It will focus on defining those places *within* urban areas for two reasons. Firstly, it was these areas that appeared hardest to define - performing poorly within the model validation of this chapter and exhibiting the greatest variation of terms amongst residents. Secondly, as also demonstrated within this chapter, with no official boundaries for such places, it is problematic to derive further information from other sources - such as population or socio-economic information from Census data. For these reasons, the following chapter will investigate vague boundaries (as opposed to attributes / class discussed within this chapter) for the 'sub-parts' of urban areas.

Chapter 4

Vague Boundaries: Urban Locales - Existing work

4.1 Introduction

The previous chapter investigated the potential of *vague place types* through a vague settlement classification drawn from the differing perspectives of place held on the web. In contrast, the following set of Chapters (4-8) will investigate defining *vague boundaries* for the sub-divisions of our urban towns and cities, as well as exploring *vague place names* of such areas. This chapter reviews the current literature concerning those sub-divisions such as suburbs and neighbourhoods (collectively termed in this work - 'urban locales') including the motivations of their creation and methods for doing so. The subsequent chapter will then propose a method of modelling vague urban locales through passive crowdsourcing, with results and validation appearing in Chapter 6. Chapter 7 addresses a number of specific questions regarding urban locales (as outlined in Chapter 1 which sought to expand the method and investigate certain properties of these newly created spatial units). The main discussion concerning the method for vague urban locale identification and implementation can be found in Chapter 8, along with example case studies demonstrating the value of the work and potential further work.

Urban areas were chosen as the focus as there are no official boundaries of the vernacular sub-divisions *within* our urban areas and the previous chapter illustrated that the place type for these urban sub-divisions appear particularly problematic to define (with much variation in the terms used amongst residents to describe such areas). Cities are highly complex and dynamic built environments - "At every instance, there is more than the eye can see, more than the ear can hear, a setting or a view waiting

to be explored" (Lynch, 1960, p. 1). There can be no denying that there is a need to, in some manner, sub-divide these complex places to aid our awareness and assist in shared understanding and way-finding. The previous chapter has demonstrated the plethora of classes that can be used to describe named ‘parts’ or sub-divisions within our cities. Suburbs and neighbourhoods are the most frequent terms but also terminology such as: districts, residential areas, parts, (housing) estates, edge cities, localities and so forth, can all be used to describe places *within* our towns and cities.

Chapter 3 demonstrated that the most common place types within our urban areas were suburbs and neighbourhoods. This chapter will explicitly discuss the main motivation for defining suburb and neighbourhood boundaries and the current literature on the subject. Before doing so, however, attention must firstly be given to definition of the terms: neighbourhood and suburb.

4.2 ‘Hoods vs ‘Burbs: towards a definition of urban locales

Even amongst urban geographers there is little consensus as to what exactly constitutes either a ‘neighbourhood’ (Jenks and Dempsey, 2007; Kearns and Parkinson, 2001) or ‘suburb’ (Forsyth, 2012; Harris, 2010). Such terms are ambiguous with no agreement concerning their definition. Without such, it follows suit that there is also sparse accord on how to identify them spatially. To complicate matters further, different countries will also hold varying standpoints and interpret the words with differing meaning. This body of work, however, is only concerned with the definition within the UK context.

4.2.1 Neighbourhoods

Neighbourhoods are often expressed as the places where events of everyday life occur (de Certeau, 2000). They are geographical units to which people connect and identify. Thus, neighbourhoods become personally meaningful, creating the background for people’s life stories and experiences (Wilson, 2009). In the UK, they are at the heart of mainstream urban policy agenda (Smith et al., 2007) and have been described by a former UK Secretary of State for Communities and Local Government as “the building blocks of public services society” (Pickles, 2010). Over recent decades the concept of neighbourhoods has become increasingly central to government policy, notably in social, economic and political exclusion processes, as well as civil society initiatives and strategies of redevelopment and regeneration (Moulaert et al., 2007).

Although the debate concerning precisely how neighbourhood effects transpire is ongoing (see Galster (2012)), there is little doubt that such small area geographies heavily influence our lives (Sampson, 2012). The Office of the Deputy Prime Minister (ODPM) (2005) wrote that “People living in deprived neighbourhoods are less likely to work, more likely to be poor and have lower life expectancy... Living in a deprived area adversely affects individuals’ life chances over and above what would be predicted by their personal circumstances and characteristics”.

The term ‘neighbourhood’ means different things to different people. In this respect, there is a great deal of multiperson disagreement (see Chapter 2) within the interpretation of the concept. The term has a number of very different meanings or formulations, with no right or wrong, merely differing views. A selection of these perspectives are illustrated below:

1 neighbourhood units are conceived as building blocks of the city focused around the need for facilities (Perry, 1992).

2 regionalisation considers aggregation of smaller units into larger ones that are tied by similar attributes in some attempt at homogeneity (eg Martin (1998)).

3 individual, property centred (McKenzie, 1921) whereby every person has their own specific neighbourhood centred on their place of residence.

4 social networks: territorial groups who meet for primary social activities and organise spontaneous social contacts (Glass, 1948).

5 sense of social identify: Place is important for people’s identity (Blockland, 2001). People identify with the area in which they live and create a strong sense of personal belonging to it.

6 political rhetoric. The term was widely popular within ‘New Labour’ to utilise attachment to local space in order to assist involvement within, along with responsibilities and duties of individuals to their local areas (Fairclough, 2000). The strong rhetoric concerning neighbourhoods and local empowerment continued within the Conservative – Liberal Democrat coalition (2010-2015), as the earlier quote at the beginning of this section from Eric Pickles demonstrates. Neighbourhoods are seen as the suitable scale for delivery of government policy such as in the form of neighbourhood renewal, neighbourhood planning and neighbourhood policing.

What the majority of the formulations have in common is that they aim to group individuals together that share something in common (including shared values, charac-

teristics or perceptions), often with the purpose of bonding them together or forming an attachment to the area itself.

The Oxford English Dictionary generally defines a ‘neighbourhood’ as "A district or portion of a town, city, or country esp. considered in reference to the character or circumstances of its inhabitants" (OED Online, 2015). This view is supported by the Dictionary of Geography which specifies the concept as: "A district forming a community within a town or city... [which] evolve their own distinctive characteristics, or subculture" (Mayhew, 2009). In essence, such definitions rest on the *character*, or feel, of the area that make it stand out from the rest of the city. The weakness of such an approach is that the character of areas are constantly evolving (for example the gentrification of London’s East End) and yet we generally perceive the entities themselves as continuous; persisting through time (even if their boundaries may shift).

4.2.2 Suburbs

Whilst there is extensive literature concerning neighbourhoods - suburbs have received relatively less attention in recent years. There is also a general tendency to portray suburbs in a negative light. "In the West, suburbs could hardly be less fashionable" (The Economist, 2014b), they play "a secondary role in regeneration and urban policy" (Local Futures Group, 2007) and "the interests of the suburbs are systematically neglected by Whitehall, as money is diverted towards fashionable inner-city regeneration and the well-articulated special interests of the countryside" (The Telegraph, 2014). Although the suburbs have a greater presence in the academic literature than in the policy debate (Jones et al., 2008b), suburban studies are generally neglected as a distinctive field (Vaughan et al., 2009).

In contrast to the more generic spatial nature of the term neighbourhood, ‘suburb’ is usually cited as being spatially peripheral within an urban setting, defined as "the outlying parts of a city (either beyond or just within the city boundaries); the parts of a city outside the commercial and civic centre" (OED Online, 2015). Such concepts are naturally vague - *how far* from the city centre do the suburbs start and *how far* do they extend (Forsyth, 2012)? Historically, suburbs referred to the area immediately *outside* a town or city (OED Online, 2015), further adding confusion to the concept as, through time, urban spaces expand and encompass surrounding areas.

In an analogous manner, English Heritage (2007), in their report ‘Suburbs and the Historic Environment’, defined the suburbs as: "outgrowths or dependencies of larger settlements - somewhere with a clear relationship with a city or town, but with its own distinct character". The limitations of concepts based around a place’s *character*

have already been discussed in relation to neighbourhoods and such a vague definition makes it problematic to clearly identify suburban areas. Based on such a definition it is easy to understand how it is possible to estimate that over 80% of England’s population reside in areas that can be classified as suburban (English Heritage, 2007; Local Futures Group, 2007). Such a figure is overinflated as it only distinguishes between rural, suburban and a highly urbanised category (inner London Boroughs and the city cores of the Metropolitan areas) (Independent Transport Commission, 2004). As such, it combines neighbourhood and suburb estimates and current methods do not allow for any separation of the distinction. It is also based on 1991 Ward data which are likely to be too coarse a spatial resolution to define suburbs with any accuracy. Johar and Maguire (2007) argue that one of the key challenges affecting the understanding of our suburbs is the failure of definition and classification which fundamentally undermines the ability to collate statistical evidence in order to understand any changes occurring within the suburbs.

There is a growing recognition that "the distinction between central cities and their suburbs is blurring" (The Economist, 2014a) and that suburbs should not simply be defined in relation to the city itself (sub-urban - outside and in some capacity subordinate to the centre). Instead, suburbs are complex and dynamic places often containing a mixture of functions (Griffiths et al., 2009; Jones et al., 2008b; Vaughan et al., 2009). The question, however, of how to accurately identify the footprint of suburbs remains.

4.2.3 Urban locales

The distinction between suburbs (areas on the edge of the city) and neighbourhoods (general parts of our cities) is vague and open to interpretation, as has been argued within Chapter 2. One person’s suburb may well be another’s neighbourhood. Within Open Street Map (OSM)[1], a neighbourhood is a sub-category of suburb[2] denoting that perhaps there are size distinctions, with neighbourhoods being smaller. In the UK, there appears no clear cut distinction, frequently resulting in the terms being used interchangeably. Regardless of what class of object they may deemed to be called, according to Savage (2009) such small areas (neighbourhoods and suburbs) within our urban landscape offer a sense of place, social bonding and identity in an otherwise large, fast paced, rapidly changing and impersonal environment. More importantly still, is the fact that such areas do not pre-exist. Instead they are abstract constructs and clearly would not exist without us defining them in a prior manner. In this light, such a concept cannot be defined in isolation to the meaning that we place upon them,

[1] a collaborative project to create a free editable map of the world - http://www.openstreetmap.org.uk
[2] http://wiki.openstreetmap.org/wiki/Talk:Neighbourhood and as shown within Table 3.1

and should be defined by some theory before identification of the units themselves can transpire.

Generally neighbourhoods can refer to any area within a town or city (including those in both inner-city and suburban areas), whereas suburbs specifically relate to areas outside the core city and on the edge (Mayhew, 2009). Such definitions (as parts of urban areas) are however, naturally vague and open to multiperson disagreement. This research argues that it is the *name* that binds the space together and thus seeks to geographically locate named areas. Such a process should be central to the generation of the concepts, as opposed to attempting to investigate the character of place (and then attempting to name the generated units afterwards). For present purposes all named sub-divisions of towns and cities will collectively be termed '*urban locales*' as it circumvents pre-conceptions of using existing terms that are often laden with personal meaning. Thus, for the purposes of this work, an '*urban locale*' is defined simply as **any named area, or sub-part of our towns and cities**. This will, of course, include suburbs, neighbourhoods, housing estates, non-residential estates and business parks. The concept rests on the pragmatic premise that it is the degree of consensus concerning the placename that is the unifying aspect that binds the space together. Thus, it should be stressed that rather than forming aggregations of existing administrative areas that one then has to try and fit existing names to, any method sought within this work will be based on the assumption that the *name* that people call areas should be at the heart of urban locale boundary generation.

4.3 The motivation for urban locales

In line with the argument within Chapter 2, existing work has demonstrated that urban locales (such as neighbourhoods and suburbs) are vague, with people forming different perspectives of their boundaries (Coulton et al., 2011; Montello et al., 2003; Twaroch et al., 2008). Whilst such units hold extensive utility for researchers investigating a range of social processes, such approaches are limited by the fact that no official boundaries for urban locales exist in the UK. This is despite the fact that there would be a significant and wide ranging benefit of such boundaries. The following section will outline a selection of the potential advantages of defining the extents of urban locales.

- **Units of analysis**
  Despite their social importance, there are no official urban locality boundaries within the UK. Urban locales play an integral role in forming community and social networks (Wilson, 2009), thus many studies seek to analyse social issues at this

geographic scale. Without official boundaries, however, in the UK studies either have to define their own urban locale extents (necessitating that only a handful of urban locales could be analysed due to the extensive workload that be required to manually define such boundaries) or they have to resort to using administrative boundaries as proxies (Twaroch et al., 2008). Nevertheless, such administrative boundaries do not necessarily correspond with urban locale extents even with the same name (Flemmings, 2010; Schockaert and De Cock, 2007; Twaroch et al., 2009a).

An alternative to classifying data by urban locale, when comparing different geographic areas, would be to use geo-demographic categorisations[3]. For example, the Output Area Classification (OAC)[4] uses clustering algorithms to partition 2011 Census data into twenty-six groups (and seventy-six subgroups) sharing similar characteristics. Whilst such methods are especially useful at comparing across extensive geographic areas (such as national scales) they are less effective at a city scale where users may be more aware of local place names and may desire to compare specific known areas. For example, when comparing house price differentials within the city of Sheffield, the public and other audiences may find it easier to interpret output using well known and familiar named urban locales (such as 'Crookes' and 'Broomhill' in Sheffield) than if data were aggregated to OAC groupings such as 'Challenged Diversity' and 'Aspirational Techies' which are harder to interpret and may have less meaning to residents.

A key driver behind this work is the standpoint that a national set of urban locale boundaries would provide a robust foundation for a host of social applications, in domains ranging from criminology to epidemiology (for example Sampson (2012) discusses how neighbourhoods can influence a host of factors from juvenile delinquency, adult crime, teenage sexuality, substance abuse, and mental health among other outcomes).

Furthermore, a great deal of government policy is currently delivered at the Ward geographic scale[5]. If community engagement is to be effective - it should occur at the urban locale scale to which residents identify with and relate to (Taylor and Wilson, 2006). This is not to say that all policy should be focused at the urban locale scale which ultimately also needs to be linked effectively to wider strategic and especially economic agendas (Taylor and Wilson, 2006).

- **Neighbourhood planning**
  Defining urban locales is a key element within 'neighbourhood planning' – which

[3] the grouping of people in space according to socio-economic conditions

[4] available from http://www.opengeodemographics.com

[5] Although population counts of Wards vary substantially, the national average is approximately 5,500 - http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/administrative/england/electoral-wards-divisions/index.html

seeks to assist local communities to produce plans to shape development within their local area[6].

A Neighbourhood Plan applies to a Neighbourhood Area whose boundaries are subject to public agreement with the local authority. There can only be one Neighbourhood Plan covering any one area, so boundaries may not overlap with other Neighbourhood Plan areas. This may create a considerable tension given that it has been previously argued that the extents of geographic units such as urban locales are likely to be vague and overlap due to susceptibility of multipersonal disagreement. This is supported by Bailey and Pill (2014) who found that Neighbourhood Planning appears to work better in rural areas, such as parish councils, with clearer boundaries and argue that the policy itself was not really designed for an urban context. Although over 80% of England's population live in built-up urban areas (defined in line with the official designation as containing at least 10,000 people), less than one-third of the 1,671 neighbourhood planning areas (as of September 2015)[7] were within built-up areas - demonstrating the relatively poor take-up of the process in more urban environments. This view is supported by the London Assembly (2012) who felt that in London, Neighbourhood Planning may be hampered by the problems of defining the term 'neighbourhood'. Furthermore, the London Assembly (2014) suggest that the "slow" uptake of neighbourhood planning areas may partially be due to difficulties in defining boundaries within complex urban settings. It is clear that a methodology for generating urban locale boundaries would be beneficial not only to local forums, for the designation of areas within Neighbourhood Planning, but also to local council planning officers, who are required to oversee and approve the designated areas.

- **Statistical units**
  The need for 'neighbourhood' level statistics in the UK can be traced back to the Social Exclusion Unit's 1998 report on deprived neighbourhoods (Office for National Statistics (ONS), 2006). Available administrative spatial units vary considerably in size, as illustrated in Figure 4.1 and include:

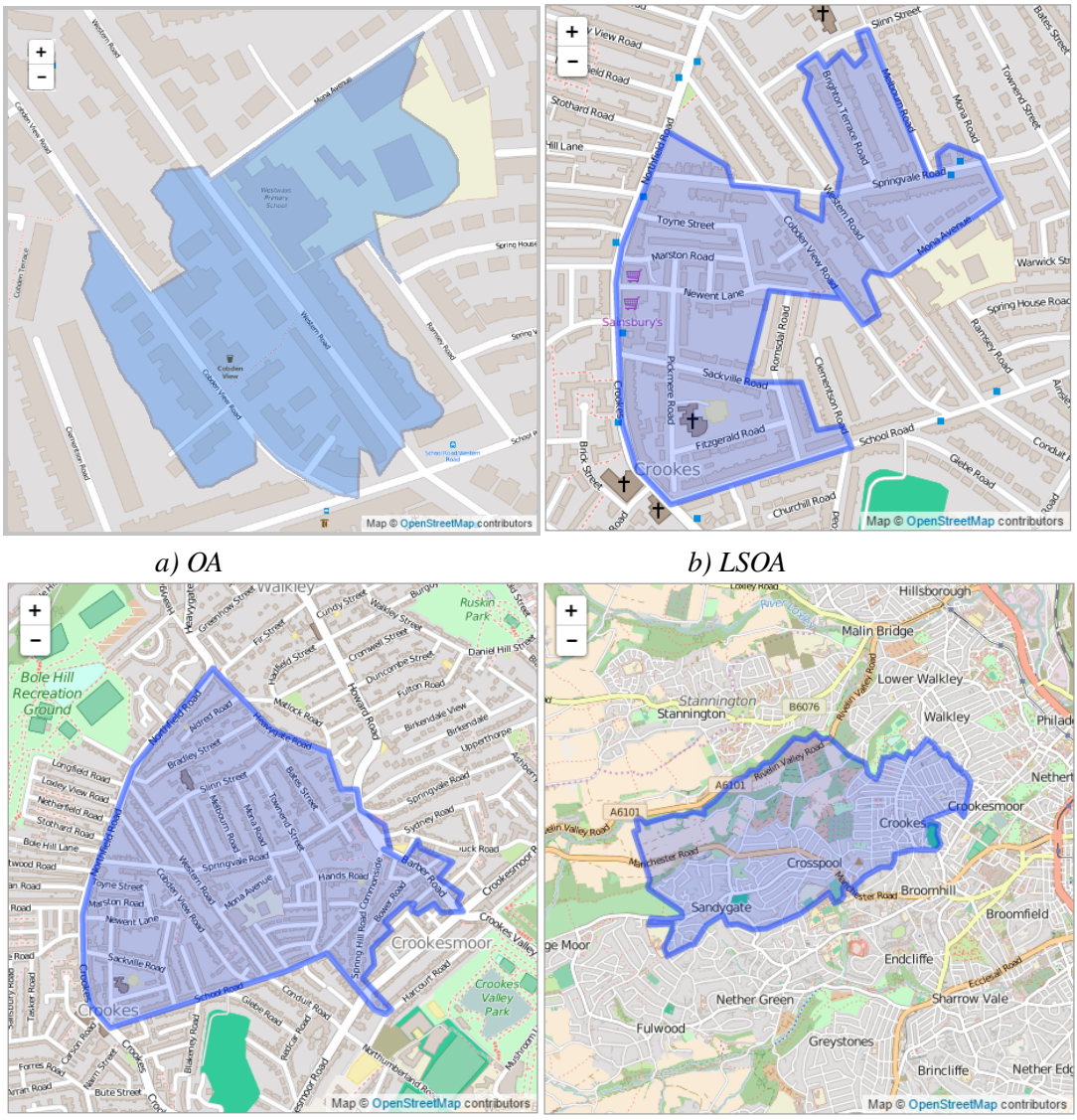
  - Output Areas (OAs - the lowest level of geography produced across Census of Population topics) containing approximately 300 people;
  - Lower layer Super Output Areas (LSOAs) which on average contain in the order to magnitude of 1,600 people;
  - Middle layer Super Output Areas (MSOAs) containing about 7,800 people; and

[6] https://www.gov.uk/government/policies/giving-communities-more-power-in-planning-local-development/supporting-pages/neighbourhood-planning

[7] http://neighbourhoodplanner.org.uk

– Electoral wards/divisions consisting of approximately 6,500 people. [8]

Fig. 4.1 Illustrative example of the scale of various administrative areas



a) OA *b) LSOA*

c) MSOA *d) Ward*

Source: Office for National Statistics, 2015

Even when data are available at extremely fine scales through official sources (for example, 2011 Census data available at OA level), the absence of urban locale boundaries themselves makes it impossible to obtain data for the vernacular units that people actually identify with. Without such data, people are forced to use administrative data as proxies. Ward level administrative data often share the same names as urban locale information but the boundaries can be quite different to those that people associate the name with (Flemmings, 2010; Schockaert and De Cock, 2007; Twaroch et al., 2009a). This can result in misleading data as the output

[8] http://www.ons.gov.uk/ons/rel/census/2011-census/population-and-household-estimates-for-wards-and-output-areas-in-england-and-wales/index.html

associated with the administrative name might not fit the extent people anticipate given the name. If urban locale boundary information were available, it would be possible to aggregate the smaller OA data to them in order to estimate official data for urban locales.

- **INSPIRE Directive**
  Under the European INSPIRE Directive, all EU member states should maintain a list of geographic names ("names of areas, regions, localities, cities, suburbs, towns or settlements, or any geographical or topographical feature of public or historical interest"[9]). INSPIRE aims to establish an infrastructure for spatial information within Europe in order to support environmental policies. The Directive addresses 34 spatial data themes - one of which is Geographical Names (within Annex 1 of the Directive). Within this context, a geographical name serves as a means to identify a location through some form of gazetteers containing the names with corresponding features such as locations (by means of co-ordinates), feature types and/or other necessary information. As previously stated, one aspect of these geographical names include 'suburbs', or urban locales as defined here.

  The Ordnance Survey 1:50000 (OS 1:50K) product has been used for the England and Wales' INSPIRE obligations. There are, however, a number of issues with this resolution. Principally, the OS 1:50K represents an index of map names for users to lookup which paper maps contain the specified place name and it is not specifically a gazetteer (Gittings, 2014). This has a number of implications for its use: the spatial referencing is imprecise (allocating place names only to a 1km square); it can be incomplete within built-up areas when there is no physical room on the paper map for further annotation; it contains numerous repeated names; and its feature types are limited (containing only 10 classes, with urban locales being grouped together within 'Other settlements' along with villages and other features) (Gittings, 2014). Whilst the OS 1:50K is an extensive and useful product, it does not contain a detailed account of urban locales as it is not possible to specifically distinguish urban locales from other settlement types and accurately position them in geographic space.

- **Semantic interoperability**
  There is a general need for semantic interoperability concerning urban locale areas, so that users of such named areas can be aware of the areal extents that other people may associate with. For example, if a local council adopts a particular place name for a specific scheme (for example for parking restrictions), it should be aware of the extent over which others may also associate with the same name.

- **Web search services**
  Output of bounded urban locales would also be of use within web search services

[9] http://inspire.ec.europa.eu/theme/gn

(Ahlers, 2012; Angelaccio and Buttarazzi, 2000) – allowing users to constrain queries to a specific geographic location, such as using Google Maps (Chen et al., 2006; Christoforaki et al., 2011). Currently the locational element within most online structured data entry search facilities (such as those within yellow pages) are handled using postcodes. Whilst such an approach is suited to residents, any visitors to the area may not necessarily be aware of specific postcodes but are more likely to know the place name.

The benefits of defining urban locales are wide ranging and certainly stretch beyond those selected and discussed above (other example beneficiaries include estate agents, police intelligence, social housing allocation and so forth). As previously mentioned a national set of urban locale boundaries could provide a framework for a plethora of social applications. The following quotation provides a good summary highlighting the need for appropriate urban locale units at this geographic scale but also the issues of a lack of both boundary data and methods for the generation of such units:

> *"For policymakers, practitioners, and funders committed to developing place-based initiatives in partnership with residents, the identification of a locally meaningful 'neighborhood' is a fundamental issue. But efforts to identify such a unit are often stymied by the lack of pertinent data and credible methods".* Cynthia Guy, Foreword in Coulton et al. (2010)

4.4 Current modelling methods

Whilst the majority of map output rests upon a physical existence (for example, buildings, roads, rivers and forests), in contrast, place names and their boundaries are human constructs. Thus, they cannot be simply measured through standard surveying techniques and instead require a different approach. Whilst much research emphasizes the merits of social units at the spatial scale of urban locales, the subjective nature of the concept and the difficulties in data collection that ensue, means that in many countries – including the UK – there simply are *no officially defined urban locales*, either in terms of names or boundaries. Whilst a number of urban locale gazetteers do exist in the UK (for example OSM, Yahoo! Geoplanet Places[10] or Geonames) they do not contain boundary information (indicating instead only single geospatial points of reference), are incomplete and, importantly, rarely agree (Brindley et al., 2014b). Historically, Ordnance Survey (OS) recorded place names (and their centre point locations) by asking privileged occupations within the vicinity (such as clergymen,

[10]https://developer.yahoo.com/geo/geoplanet

school masters and doctors) and would avoid asking residents who were considered unlikely to know the name and its spelling (Parker, 2015)[11]. Undertaking such an approach would not only be unfeasible for a national scale within the resources of this research, it would also place undue emphasis on the views of a select number of people.

4.4.1 Resident mapping

There has been a long tradition of geographers mapping urban locale areas, as shown by the work in the 1960s in America by Lynch (1960). This research has been exceptionally influential to latter studies investigating vernacular place names (Abdelhaq et al., 2013; Coulton et al., 2011; Dalton, 2007; Montello et al., 2003; Orford and Leigh, 2013; Talen, 1999; Vallee et al., 2014). It is generally acknowledged that there is considerable variation within the perception of boundary extents of urban locales amongst residents (Talen, 1999). Such work frequently highlight the need for vague / overlapping geographies and embracing the varying perspectives held by different people. This is well demonstrated by the boundaries drawn by twelve responses for Hamstead Garden Suburb, London in Figure 4.2. Although such works have immense value, they usually utilise a small number of urban locale case studies due to the nature of data collection (such as requesting residents to draw perceived boundaries). Consequently, such methods are not only time consuming but also not viable for a wider scale such as a national coverage.

One exception of this, however, is Wikimapia[12] - an online editable map, which aims to map and describe geographical objects through crowdsourcing. The data, however, are not primarily concerned with urban locale level places and also include many other features (such as boundaries for cemeteries, allotments, educational facilities, sewage treatment works, individual business premises and so forth) resulting in less than 20% of polygons for the Nottingham Urban Area relating to urban locales and less than 12% in Sheffield. With a reliance on other people to actively input the information, achieving complete data coverage can be problematic and inconsistent. For such reasons, whilst nearly 15% of the Nottingham Urban Area has no urban locale information within Wikimapia the figure rises to 50% in Sheffield. The representativeness of the output may also be questionable as it may be the view of a very small number of people. In the case of Nottingham Urban Area, only six different profiles were involved with the urban locale boundary creation and editing.

[11] Most current OS maps still rely on these historically defined urban locale place names, although the degree of consensus concerning such names is unclear and will be explored further within the validation section of Chapter 6

[12] http://wikimapia.org

Fig. 4.2 Twelve Resident views of Hamstead Garden Suburb Boundaries



Source: Dalton (2007)

4.4.2 Space syntax

Space syntax is the name given to a body of theory and analysis which aims to understand the role of configuration of a built environment within social processes. It examines the potential flows through the built environment (for example the number of intersections within a road network). Dalton (2007, 2011), Hillier and Vaughan (2007) and Vaughan et al. (2010) all use space syntax in order to identify urban locales. Whilst the method is good at identifying local centres, it appears more problematic to actually delineate boundaries and also may not necessarily construct extents that correspond to named areas as residents perceive them. Additional reference to place name gazetteers would be necessary in order to name the areas defined by space syntax methods. For example, the area circled within Figure 4.3 may appear structurally segregated from the surrounding streets but it is difficult to interpret exactly which streets may form the boundary (especially to the south west) and it is impossible, from this information alone, to acknowledge the area as 'Maiden Lane housing estate', Islington. Imagine this difficultly for an area more connected to the surrounding streets where it becomes nigh on impossible to disentangle adjacent urban locales. Furthermore, the method does not inherently allow for overlapping of urban locale units.

Fig. 4.3 Space syntax: Maiden Lane housing estate



Source: Hillier and Vaughan (2007)

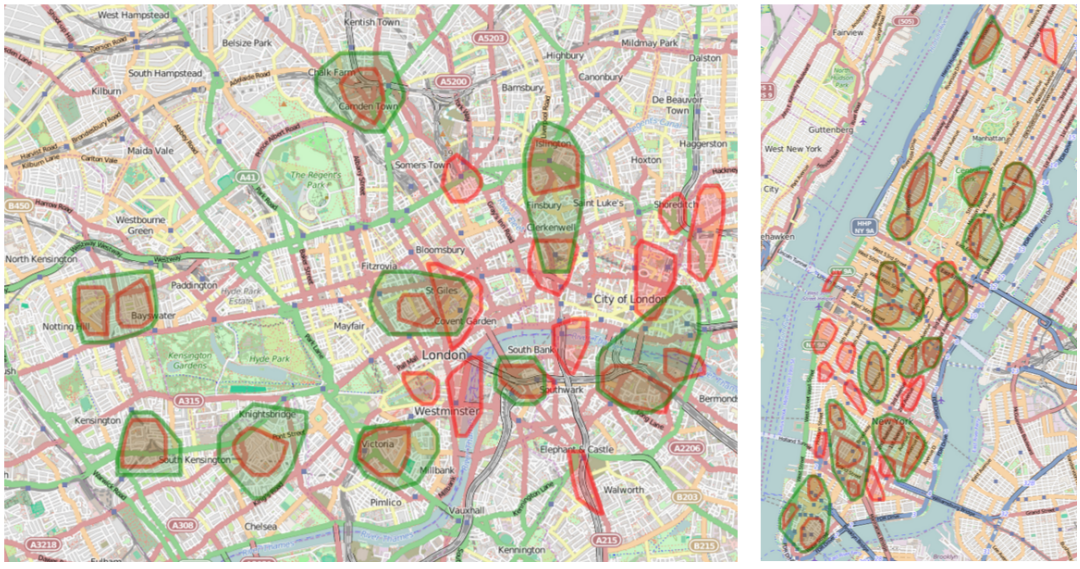
4.4.3 Web derived data

With the growing data volumes held on the web, Chapter 2 outlined the potential for harvesting data from the web such as urban locale level information. Such an approach has been applied to a variety of different sources as discussed below. In recent years research in vernacular geography has increased due to the growing availability of mass datasets on the web and social media; and in parallel, interest in the mapping of geographic objects via mining web data has steadily grown. Whilst the notion of mining the web for geospatial features is not new, much of the prior research has either focussed on either specific geographical objects (e.g. landmarks) or acted on much larger geographical scales (e.g. large regions such as the the 'north' of England or the West Midlands) (Jones et al., 2008a; Pasley et al., 2008; Purves et al., 2005b, 2007). This stems from the fact that the larger geographic scale of an object, the broader the range of co-location entities that may be used to georeference or ground it – features such as cities, towns, villages and streets can all be utilized to represent points within the larger target feature. In contrast, the choice of co-location objects for smaller units such as urban locales is far more restrictive, making their extraction from web based data more problematic due to finer geographic scales and lower data returns that many residential urban locales produce (Schockaert, 2011). The following two sections will discuss two different approaches to using web derived data to map urban locales: (1) Clustering social media data and (2) Cross referencing with existing gazetteers.

4.4.4 Web derived data: Clustering social media data

There is a body of research that utilises clustering of georeferenced social media data, such as that generated by geo-tagged micro-blog "tweets" (Arribas-Bel, 2015), or Foursquare "check-ins" (Cranshaw et al., 2012; Noulas et al., 2011; Zhang et al., 2013). A variety of clustering methods have then been applied to the resulting point data, ranging from bespoke spectral approaches to traditional k-means clustering. Similarly, automated zoning procedures (AZP) have been applied in order to optimize objectives such compactness or homogeneity, by iteratively recombining a larger set of areal units into a smaller set of output areas (Martin, 2003). Methods show particular strength in being able to isolate functional areas but due to data coverage rarely appear able to provide a complete coverage of urban locales for a city, with frequent gaps occurring where no urban locale is assigned. This is demonstrated in Figure 4.4, with the output of clustering Foursquare check-ins for London and New York identified through the Hoodsquare project[13] (Zhang et al., 2013). A further significant issue is that the functional units created, like those within Space Syntax, will not create associated names and may not be a suitable representation of named areas as perceived by people.

Fig. 4.4 Hoodsquare zone identification through Foursquare check-ins (the project identified both macro and micro zones, shown in green and red respectively)



a) London *b) New York*

Source: Zhang et al. (2013)

[13] http://pizza.cl.cam.ac.uk/hoodsquare

4.4.5 Web derived data: Cross referencing with existing gazetteers

An alternative to such clustering-based approaches is to extract boundary information from the web based on urban locale names held within a gazetteer. Gazetteers are essentially lists of known places, containing attributes such as geographic coordinates, place types and alternative names. Examples of gazetteers include Yahoo! GeoPlanet and Geonames. The data within gazetteers are mostly manually selected and reviewed by domain experts in a time-consuming and cumbersome task (Laere et al., 2014). The process of cross referencing with existing gazetteers can be separated into two components: firstly the retrieval of point information relating to each urban locale gazetteer entry; and secondly the boundary creation which requires transference of the point locations to some areal measure. Both processes will be discussed in turn.

4.4.5.1 Obtaining point data

The current literature has used a range of web based sources to obtain co-ordinates for places related to urban locales, as demonstrated below.

- **Tags within photography sites:**
  A number of studies (Feick and Robertson, ming; Grothe and Schaab, 2009; Hall et al., 2011; Hollenstein, 2008; Hollenstein and Purves, 2010; Keßler et al., 2009b; Li and Goodchild, 2012; Purves et al., 2011; Rattenbury and Naaman, 2009) have identified urban locale names from existing gazetteers appearing within user generated metadata, captions and tags[14] of georeferenced photographs on photograph-sharing sites such as Flickr, Panoramio[15] or Geograph[16]. The existing gazetteers provide the name to look for within the photograph tag and the corresponding photograph then provides the co-ordinates to position or map the name. The photographs' coordinates are either created automatically by a built-in GPS within the camera or provided manually by the user by locating photographs using a map interface (Hollenstein and Purves, 2010). Whilst such a method shows promising potential for the identification of urban locale locations, as with any method, there are a number of limitations that are either endemic of the cross referencing procedures (discussed at the end of the section) or specific to using photograph tags to identify urban locale locations (as outlined below).

  According to Hollenstein (2008) and Hollenstein and Purves (2010) up to 35% of photograph tags include at least one place name tag of some granularity. The vast

[14] a keyword users can attach to an image in order to aid other users searching
[15] http://www.panoramio.com
[16] http://www.geograph.org.uk

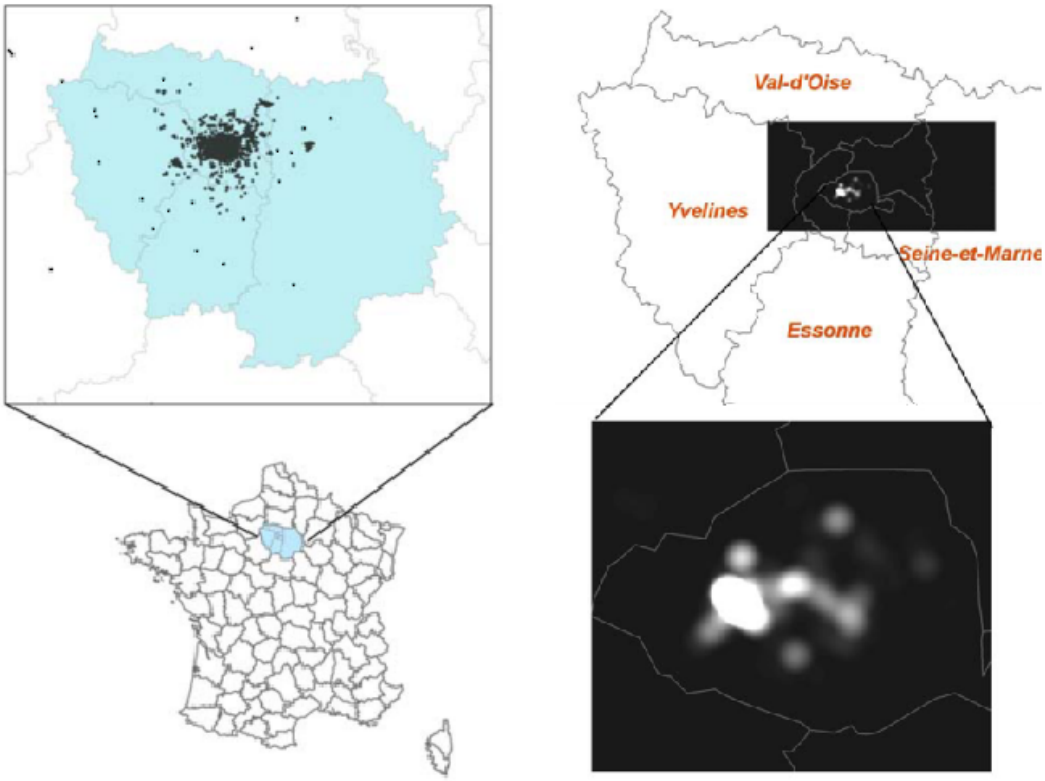
majority however, relate to the city scale or higher geographies (such as regions and counties) and in a study of Zurich, Hollenstein and Purves (2010) found that only 1.2% of tags made mention of urban locales and similar scaled features. Due to the high data volumes of photographs involved this may not be an issue for prominent areas where there are likely to be a high number of photographs taken (for example tourist areas such as Mayfair, London where 212 tagged photographs were identified by Hollenstein (2008)). Data volumes within residential urban locales of small towns, however, remain unclear but are discussed further within Chapter 8.

There will frequently be erroneous outlier datapoints that require removal (which will be discussed in detail in due course), however, there are a number of data accuracy issues specifically concerning the use of tags within photographs, on the web, due to the inaccuracy and imprecision of people geo-referencing and tagging photographs (Vasardani et al., 2013). This is evident within Figure 4.5a which shows photographs tagged as Paris where a number of outlying points should not be associated with being within 'Paris'. This may be in part due to a number of factors. References to an urban locale within a photograph tag does not mean that the photograph was necessarily taken within the said place but may be a distant shot of the urban locale, thus, the geo-coordinates of the photograph and the urban locale are not directly related. Figure 4.5b demonstrates this well, illustrating Flickr tag locations for the Eiffel Tower. Alternatively, the photograph's geographic position and the urban locale within the tag may not be related at all. For example, a person may tag a photograph as "Catching the train to Paris" or "Looking forward to my trip in Paris next week". In such cases it is wrong to attribute the co-ordinates at where the photograph was taken with *Paris* directly. Users may at times add the same tags to a whole set of photographs that they upload simultaneously (Hollenstein and Purves, 2010). Spatial references within photograph captions are frequently complex and vague (see Hall et al. (2011)). It should also be noted that due to geo-referencing issues and multiple photographs from a single user, a relatively small number of users can exert a significant influence on the output of data, even on seemingly very large data sets (Hollenstein and Purves, 2010). For example, Hollenstein and Purves (2010) found that 82.6% of photographs with the tag "innercity" were from just 10% of the users.

Finally, there may be future uncertainty concerning the degree to which photographs would represent user generated content. Flickr have demonstrated an interest in the mapping of urban locale boundaries in their work on alpha shapes[17] through to quattro shapes[18]. If Flickr's intention is to suggest urban locale names to users on uploading geo-referenced images, such actions may then alter the manner in which users choose to name places using their own tags (Hollenstein and Purves, 2010).

[17] http://code.flickr.net/2008/10/30/the-shape-of-alpha
[18] http://quattroshapes.com

Fig. 4.5 The spatial footprint of Flickr photograph tags



a) Photo locations with the tag "Paris" *b) Photo locations with the tag "Eiffel Tower"*
Source: Li and Goodchild (2012)

- **Search Engine requests:**
  A different approach to defining point locations with associated urban locale names is to perform specific queries involving urban locale names from existing gazetteers using web search engines. The work of Jones et al. (2008a) and Arampatzis et al. (2006) were principally concerned with larger geographical entities than urban locales - such as the Cotswolds, Mid Wales and the Highlands of Scotland. The procedures used, however, could equally be applied to smaller geographic units with only slight adjustments. Jones et al. (2008a) investigated the quality of different search terms and found that queries based on the placename and an associated concept (such as "hotels in [placename]" or "places to visit in [placename]") returned the most geographically-rich pages. Such searches resulted in a high number of results from directory-style web-pages which were found to contain a high degree of finer grained geographic elements (such as postcodes) which could then be used to geo-reference and map place names.

  Flemmings (2010) used ten different search query phrases (such as "Clifton in Bristol"; or "* is located in Clifton") within the Google search engine in order to locate the Bristol urban locales of Clifton and Bedminster. The returned snippets were then interrogated in order to identify any postcodes within them which could

be used to geo-locate the data returns. The advantage of such a method is that it draws data from a potential *range of different sources* rather than relying on a single source which could run the risk of introducing bias. The output appears to work particularly well in areas of higher data returns (populated areas with a high number of facilities, such as hotels, which reflect a good degree of web-visibility). It is less clear how well the output (especially for urban locale identification) might work for more suburban urban locales, especially when expected data volumes may be sparse. Whilst search engines are a good source of data (as they aim to index all web-pages that might be of interest), it must be recognised that a large part of the web is not found by search engines, as it remains un-indexed (the Hidden or Deep Web might be as much as 400 to 550 times larger than the commonly defined World Wide Web (Bergman, 2001)) and search engines themselves can introduce biases of an unknown nature (Twaroch et al., 2009a).

- **Single websites:**
  Instead of utilising search engines to find the relevant web-pages, it is also possible to target specific web sites or resources. For example, Schockaert and De Cock (2007) sent queries to Yahoo! local API[19] containing the city name (to reduce ambiguity) and urban locale name from a gazetteer, in order to return lists of businesses within the locale. Yahoo! local was a mapping service and so would return geo-coordinated points as output. Using the same principles, Twaroch et al. (2009a) sent requests to the Google Maps API in order to map the urban locales of Cardiff. The main limitations of such an approach is that there is a lack of data within more residential areas and only urban locale centres are adequately mapped (with the higher data returns due to business web presence). Also, the method is completely dependant on a single source. The fate of the now defunct Yahoo! local demonstrates the potential issue. Also, as a single source, the number of data returns might be restricted. Schockaert and De Cock (2007) admit that there were too few businesses returned by Yahoo! local in lesser known urban locales.

  Alternately, Twaroch et al. (2008, 2009a,b) use data from the classified advertisement site - *Gumtree*, in order to provide urban locale references for the city of Cardiff. When users upload an advert they must provide a postcode which Gumtree then converts to geo-ordinates when they serve data requests. Users may also optionally provide a local place name with their address within the classified advertisement. Thus, it is possible to search on a named urban locale (the name coming from a gazetteer) and obtain a list of advertisements that users have described as being within that urban locale. One advantage of this datasource is that, in contrast to other approaches, it should provide output for more residential areas. One limitation of the method is that specifying a "local area" or including a map when placing

[19] the service was closed on September 28th 2013

an advertisement are optional and both elements are necessary in order to extract urban locale and co-ordinate information. Whilst not a problem in larger urban locales (for example of the 49 advertisements within Crookes, Sheffield - 43 had map information), this can affect data volumes in smaller locales (in Broomhall, Sheffield of the 4 returns, only 2 had map information).

4.4.5.2 Bounding points

The second requirement (after obtaining point data) for retrieving urban locale data through cross referencing with existing gazetteers is to bound the collected points. Once point data have been retrieved from some specified source, it is necessary to aggregate the points in some manner to create bounded areal units for each urban locale. This process can be undertaken in a number of different approaches as outlined subsequently. There will frequently be errors within the data whatever source was used to derive the urban locale information (for example the stray points relating to 'Paris' within Figure 4.5a). Such outliers can either be removed prior to the bounding process or incorporated as an element within it. What is important, however, is that data errors are recognised and treated. Differing methods for bounding the point data are described below.

- **Clustering**
  Both k-means and hierarchical clustering were used by Rattenbury and Naaman (2009) in order to cluster Flickr tags. One issue of the approach was that a single urban locale could be split into multiple components, as was the case when photos with the tag 'Golden Gate Bridge' were split into spatially disjoint areas corresponding to popular viewing locations (Rattenbury and Naaman, 2009).

- **Convex hulls**
  Convex hulls can be used to bound the points (such as used in Twaroch et al. (2009b)) but such methods can be heavily influenced by any remaining data outliers. Alpha shapes are generalizations of the convex hull and have been used by Flickr in their work on alpha shapes[20] through to quattro shapes[21].

- **Voronoi diagrams and Delaunay triangulations**
  Alani et al. (2001) constructed a Voronoi diagram from the points that were described as either internal or external to the desired feature (for example in this context each urban locale name). The approach was also implemented by Lam et al. (2002) and Wilson et al. (2004) in their work defining urban locales in Los Angeles. One

[20] http://code.flickr.net/2008/10/30/the-shape-of-alpha
[21] http://quattroshapes.com

limitation of the method is that it is sensitive to the underlying pattern of data points and areas with sparse data will form larger polygons as part of the Voronoi diagram simply due to the lack of data (Wilson et al., 2004). Arampatzis et al. (2006) modified the approach using Delaunay triangulations and took account for the issue that error may exist within the encoding of points as either internal or external to the feature of interest. The method adjusts the boundary between the two sets to eradicate outlier points. The techniques are based purely on geometric criteria and do not take account of actual likelihood measures of a point being inside the target region (although the authors note that the method could be adapted to use such information). Keßler et al. (2009b) identified Soho both in London and New York from Flickr photo tags using Delaunay triangulation, although the authors noted that the areal extents were frequently larger than common perceptions and that thresholds within the method would need adjusting to account for the issue.

- **Kernel density estimation**
  Density estimation techniques, within GISc, allow the interpolation of a continuous density surface from a set of point data with numeric attributes (Jones et al., 2008a). In its most basic form, a density estimation involves converting the area of interest into a raster grid of uniform cells. A moving window (with a specified search size or bandwidth) is then centred on each of the grid cells in turn and the point values within the search window are summed. This summed value is assigned to the grid cell of interest before the search window is then centred on the next grid cell and the process repeated. Kernel density estimation (KDE) methods incorporate a weighting to the point values according to their distance from the starting cell. This has an effect of smoothing the influence of points with distance, so that points nearer the starting cell have a greater influence on the density value (O'Sullivan and Unwin, 2003). KDE has been applied in the transforming of point data into vague areal features in a variety of contexts including those of: landscape forms (such as mountains, lakes and hills by Fisher et al. (2007, 2004) and Purves et al. (2011)), geographically extensive vernacular places (such as The Midlands, Highlands or Cotswolds by Purves et al. (2005a,b)), city centres (Thurstain-Goodwin and Unwin, 2000), and vague village extents (Almadani et al., 2013). It has been the most widely used method of transferring the point tuples (name of area from gazetteer, X-coordinate, Y-coordinate) into urban locale areal features and has been utilised by Clough and Pasley (2010); Grothe and Schaab (2009); Hollenstein (2008); Hollenstein and Purves (2010); Purves et al. (2005b); Twaroch and Jones (2010); Twaroch et al. (2008, 2009a,b).

  Three decisions need to be made within the KDE process: (1) the choice of kernel function; (2) the bandwidth (the search window size, also known as the kernel radius or search radius); and (3) the cell size (or resolution) of the uniform grid. All of these choices will influence the shape of the surface in terms of the smoothness

generated (Bowman and Azzalini, 1997; Silverman, 1986). The weighting often takes the form of a quadratic *kernel function* with a value of zero at radius *r* (Jones et al., 2008a). O'Sullivan and Unwin (2003) found that the choice of bandwidth (or search size) was more important than the kernel function decision, although it is acknowledged that there is no general agreement on how to best approach the choice of bandwidth (Hollenstein and Purves, 2010). In general, the *bandwidth* is arrived at experimentally (O'Sullivan and Unwin, 2003) but needs to be small enough to capture local variation within the feature of interest but large enough to capture multiple point locations within the search window (Almadani et al., 2013; Jones et al., 2008a). Too large a search size tends to 'oversmooth' and spread the pattern beyond where the boundaries are most likely to be perceived, while too small a radius will result in strongly focussed and disjoint surfaces (de Smith et al., 2007; Li and Goodchild, 2012). Li and Goodchild (2012) tested a number of bandwidths ranging from 500 metres to 1000 metres but settled on 800 metres for "better visual representation". Mackaness and Chaudhry (2011) used a bandwidth of 500 metres - based on an estimate of tolerable walking distance and Flemmings (2010) also used a 500 metre search radius. Thurstain-Goodwin and Unwin (2000); Twaroch et al. (2008, 2009a), however, used a 300 metre bandwidth within the urban environment. In relation to *cell sizes* used within current literature, 100 metre resolutions have commonly been used on a national scale for raster grids (Bibby and Brindley, 2006, 2007, 2013; Bibby and Shepherd, 2004) and 50 metre grids have often been used within individual cities (Hollenstein, 2008; Li and Goodchild, 2012).

Crisp and sharp boundaries can be subsequently obtained from KDE surfaces by selecting only those grid cells that exceed a specified threshold KDE value and vectorising the output. As Jones et al. (2008a) argues, however, the selection of a threshold value in this manner removes vagueness from the representation, leaving locations to only be considered as inside or outside of the vague object. As reasoned in Chapter 2, there is no single threshold value to distinguish the vague object as the feature will be Sorites susceptible (Fisher, 2000). This has led some studies to follow a process similar to that employed by Montello et al. (2003) in creating multiple contours in order to represent differing confidence levels in the object's extents, which are often based simply but effectively on data volumes (such as Hollenstein and Purves (2010); Jones et al. (2008a); Thurstain-Goodwin and Unwin (2000); Twaroch et al. (2009a)). Using such a process, outliers can also be eliminated by removing those areas with less than a certain proportion of the data covered. For example, Hollenstein (2008) and Twaroch et al. (2009b) excluded the lowest 10% and 20% of density values respectively. Such an approach, however, can suffer from errors of omission in areas of low data volumes (such as on city edges) and be particularly fragile to endemic data errors.

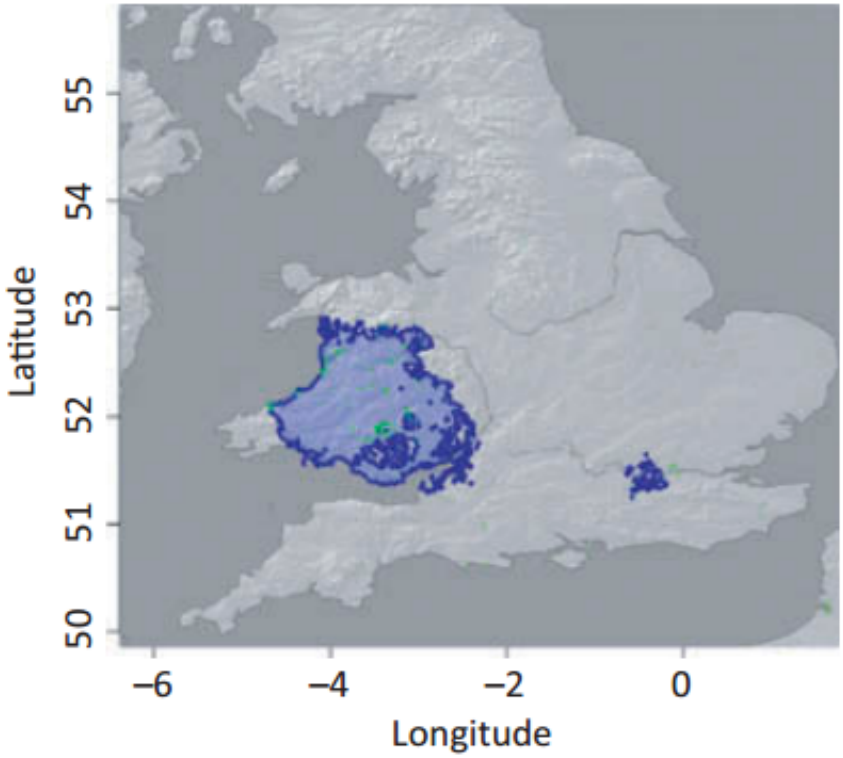
- **Support Vector Machines**
  Support Vector Machines (SVMs) are a set of supervised learning models that have been applied in the field of classification (Grothe and Schaab, 2008). Generally the desired goal is the decision concerning which one of two classes any observation belongs to. The method, however, can also be applied in a modified form to identify outliers within a single class. Such a process is termed one-class SVM, whereby it is possible to find the smallest sphere that contains data, allowing for a free parameter concerning the maximum proportion of outliers permitted (see Schölkopf et al. (2001)). Thus, the procedure is dependant upon the setting of this factor, although all methods will, to some degree, be conditional on such parameters. The function used to identify outliers is the same as for KDE (and requires the same parameters for bandwidth and cell size), the only difference is that weights are obtained through an SVM optimization step (see Grothe and Schaab (2008) for details).

  SVMs have been used to identify large scale vague geographic areas appearing in Flickr tags, such as the Alps, Black Forest or Rocky Mountains (Cunha and Martins, 2014; Grothe and Schaab, 2008, 2009). Although Grothe and Schaab (2009) found that SVM gave results comparable and in many circumstances preferable to KDE, the output of some examples appear questionable. For example, in defining the extent of Mid-Wales, Cunha and Martins (2014) included substantial sections of Greater London within the bounded area (as illustrated in Figure 4.6) without discussion concerning the apparent error. Furthermore, the method appears very susceptible to data volume issues with Cunha and Martins (2014) stating that the low recall (0.13) for the output in the Ukraine was attributable to the low number of photos being 903 in number. Given the small spatial scale of urban locales, numbers of data returns are likely to be relatively low and potentially far fewer than 900 in number.

4.4.5.3 General limitations of cross referencing with existing gazetteers

Whilst limitations concerning specific methods and sources have already been discussed, there remains a number of problems that will occur irrespectively of the method or source when using existing gazetteers to aid the identification of urban locales. Unfortunately, the urban locale gazetteers required for such an approach, seldom provide a definitive source of urban locale names. This is certainly the case in the UK, where for the city of Sheffield there was only agreement between five sources (Sheffield City Council official records, OSM, Yahoo! Geoplanet, Geonames, and OS) in 6% of urban locales (Brindley et al., 2014b). Furthermore, only 8% of urban locales were found in at least four of the sources, and in 52% of cases the urban locale name was found in only a single source. Such disparity between urban locale naming schemes is unlikely to reflect actual usage.

Fig. 4.6 The spatial footprint of Mid-Wales



Source: Cunha and Martins (2014)

Cross referencing methods can also suffer from issues of place ambiguity (Overell, 2011), the most common form being referent ambiguity (Wacholder et al., 1997) whereby confusion occurs due to multiple entities in different locations possessing identical names (for example is 'Cambridge' a reference to the Cambridge in the UK or Massachusetts, USA?). Whilst referent ambiguities are relatively straightforward to ameliorate via rule bases (Buscaldi, 2011), other forms of place ambiguity are more problematic. For example, a web query amassing documents for the urban locale 'Manor, Sheffield UK' not only returns data pertaining to the place itself, but also a wide range of information pertaining to other, subjectively-related urban locales containing that named element - entities such as 'Manor Park' and 'Manor Top'. This generates significant and hard-to-rectify noise in output sets, a problem exacerbated when combined with incomplete gazetteer information (for example out of the five sources of Sheffield gazetteers previously mentioned, all three of "Manor", "Manor Park" and "Manor Top" only occurred in a single source). Building rules to resolve these subtle ambiguities is non-trivial given their subjective and overlapping nature, resulting in significant skewing of geographical representations (such as in Flemmings (2010) who failed to separate points relating to Bedminster Down, Bristol from those of Bedminster, Bristol). The issue is of particular relevance to mapping urban locales as the degree of ambiguity associated with a place has been shown to be proportional to its depth within a geographic ontology (Buscaldi, 2011) and given the relatively

small scales of urban locales, place ambiguity is likely to be highly prevalent. As Vasardani et al. (2013) argues, currently no service exists, without semantic or context, than can unambiguously communicate about place names with users, similar to the ability of human communication.

4.5 Summary

For the purposes of this work, an urban locale is defined simply as any named area, or sub-part, of our towns and cities. This will, of course, include suburbs, neighbourhoods, housing estates, non-residential estates and business parks. The concept rests on the assumption that it is consensus concerning the '*name*' that critically binds the space together and that any method of deriving urban locale boundaries should seek the names and extents that differing people believe exist.

The documented related work illustrates that traditional research interest in vague 'neighbourhood' level objects is still maintained in recent work, with growing exploration of how mass datasets and web-based information can help identify and map urban locales. There is, however, a reliance on existing gazetteers (which frequently contain numerous historic features and are not complete), issues of ambiguity are prevalent and current methods are usually only tested on single area or substantial world cities and not may not necessarily be representative of all settlements. The following chapter will aim to build vague urban locales from web derived data in order to gain differing perspectives on the geography of the objects and also account for the issues raised above.

Chapter 5

Building Vague Urban Locales

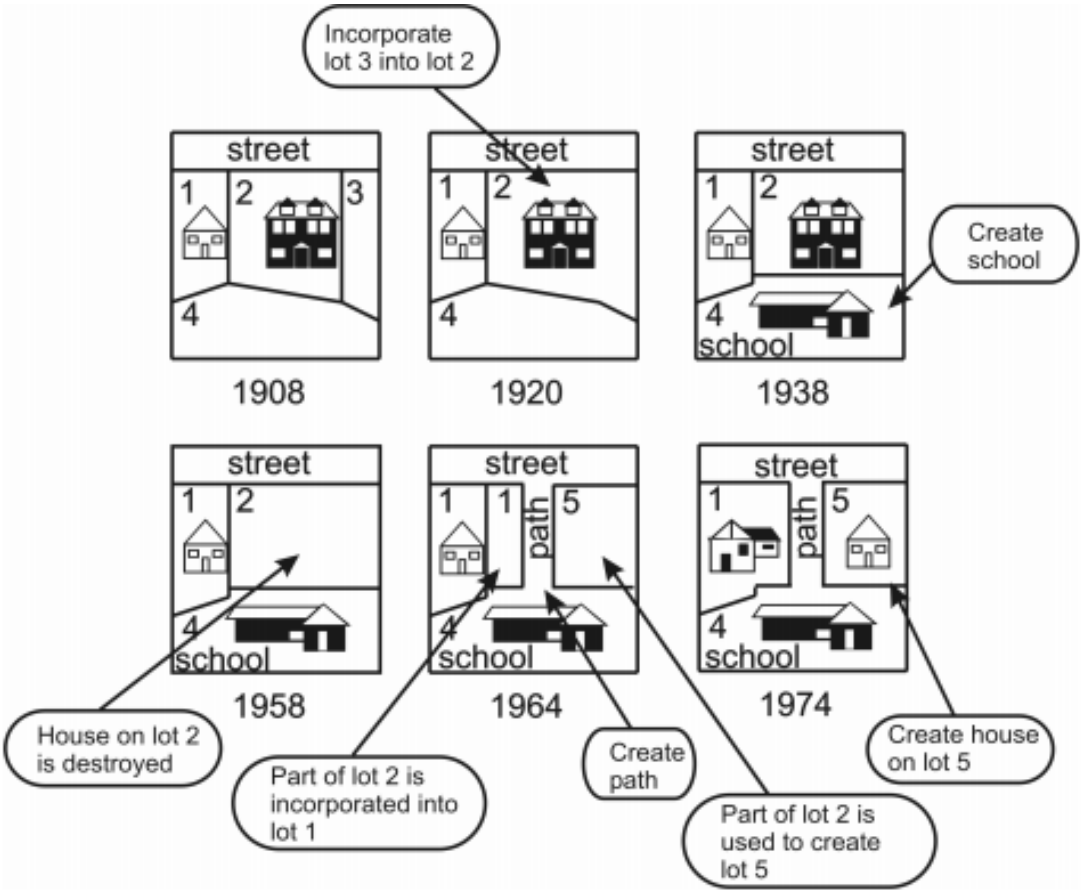
This chapter develops the work in Chapter 4 – the aim here is to build algorithms in order to automatically create the output data that would represent vague urban locale objects, potentially containing a wide range of different peoples' interpretations (through multiperson disagreement). The research does not intend to form a fully national model but provide the proof of concept that could be implemented as required. It should be stressed that whilst the method appears to be a relatively simple sequential approach, this is not the case and the method has been arrived at through an extensive iterative process. In particular, the various parameters used within the method have been derived through extremely intensive testing using local knowledge of the urban locales within the city of Sheffield before being applied more widely.

5.1 Towards an urban locale framework

As the previous chapter argued - urban locales do not pre-exist, they are human constructs. As such, they are what Smith and Varzi (2000) term 'fiat boundaries', which exist only in virtue of human effect (for example national borders, property demarcation). In contrast, Smith and Varzi (2000) define 'bona fide' as boundaries that exist independently of human act (such as shorelines). Fiat boundaries are more likely to be vague with multiperson disagreement and, are thus, generally better represented by a field (or raster) approach within GISc - whereas the more crisp bona fide boundaries are well handled using vector objects. The argument concerning whether an object essentially ceases to exist if the property of that said object changes (mereological essentialism) has been heavily debated within philosophy (see Chisholm (1973)). Whilst a detailed engagement with such theory is beyond the scope of this research, it is important to appreciate the difficulty of defining objects, be they

geographic or not. Of course, generally we accept that objects persist even when they are continually changing (degrading, growing and so on) and thus retain the classifications and naming of what we term them. This is a fundamental simplification of the objects that we take for granted (just like the argument concerning the problem of definition in Chapter 2, whereby a line is in itself indefinite). In line with the argument concerning the 'ship of Theseus' (discussed in Chapter 2) - how much change is required for an object to stop being defined in the way it is? It is easy to conceptualise when objects change dramatically and rapidly - for example a caterpillar into a butterfly. The situation is less clear when faced with gradual or more minor modifications - when the change becomes vague and a matter of personal opinion; forming the precise objects that this body of work is concerned with. The key element of the approach adopted here is the level of consensus that forms supporting the *name* of the object. The example of Worboys' (2005) urban locale evolution in Figure 5.1 demonstrates how the spatial configuration of an urban locale can completely alter through time and yet the same urban locale object remains. This can be achieved as the contents of the space are tied by the level of consensus within the urban locale name.

Fig. 5.1 Urban locale evolution



Source: Worboys (2005)

The solution proposed by Bibby (2005) is to shift the focus away from the objects to the geographic space, forming objects from the sum of small areal units. Much of this work (for example Bibby (2009); Bibby and Brindley (2006, 2007, 2013); Bibby and Shepherd (2004)) centres on 'space with attributes' where the space is formed from collections of hectare cells (from a raster grid) which are aggregated in terms of the attributes (for example named areas or other classifications). Similar to the described approach, this work employs a framework where objects can be formed out of aggregations of raster cells which are binded together by some level of consensus of the *place name*.

5.2 Detailed method

Bringing together the separate strands, the method aims to draw differing perspectives of urban locale names from passive crowdsourcing of the web, to map the output as represented as raster grids and to construct aggregations of names where a level of agreement within the data is maintained. In order to derive the urban locale names, web scraping of postal addresses held on the web is proposed in order to identify any user generated content within the address. The methodology implemented to dynamically generate urban locale areas is composed of the following steps:

- Data collection,
- Misspelling identification,
- Urban locale derivation,
- Splitting over linear features,
- Post-processing: Synonym identification, and
- Effective storage.

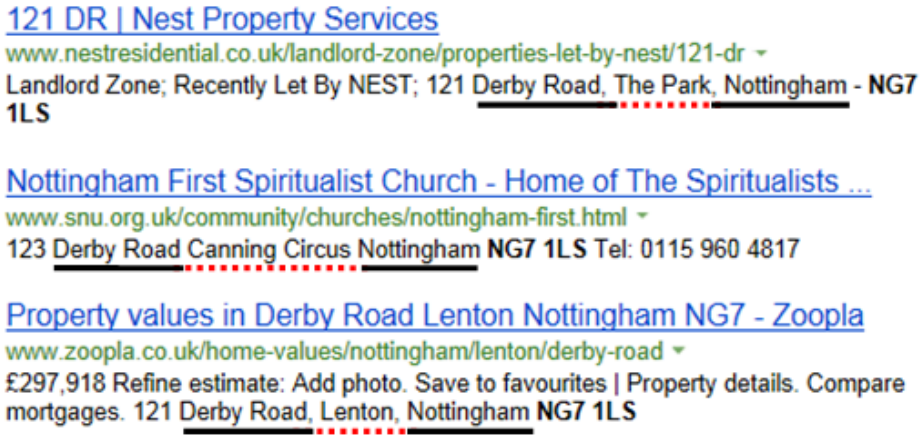
The validation of the method is considered in the following chapter. Each of the stages of the methodology will be addressed in turn.

5.2.1 Data collection

The first requirement of the method is to extract postal addresses from mass text corpora that reveal common usage (in the thesis this aggregated corpus is the web,

but it could be extended to include many more forms of datasource). Within the UK, an official postal address in an urban area is defined as: *building number and street; city; postcode*[1]. The underpinning assumption is that, even though official urban addresses in the UK do not require any information to be entered between the street and city elements, people often interleave urban locale names within that structure (see Figure 5.2) – this is the source of the urban locale knowledge. It should be stressed that such practices are not limited solely to postal addresses appearing *on the web* and that the same convention, of including urban locale information within postal addresses more generally, is a relatively common practice. In this regard, the use of the web is merely as a tool to assist in data collection. A benefit of this approach to generate urban locales is that it specifically identifies objects at the neighbourhood level order of magnitude - appearing between street and urban settlement spatial scales. One might consider it analogous to responses to the question - "Where, in *x*, do you live?", where *x* represents any significant sized urban settlement such as Birmingham or Lancaster. As such, the process is likely to exclude smaller named urban places (such as urban squares) which might not be considered appropriate for neighbourhood level places.

Fig. 5.2 Identifying urban locale terms between street and settlement content



The main exception to the addressing format as discussed above are for what the Royal Mail term "Dependent Localities". These take two forms:

1 peripheral adjoining areas where a Locality name is specified by Royal Mail within the official address in order to aid the sorting and delivery of mail[2]. The official postal address in such cases becomes: Building number and street,

[1] As defined by Royal Mail, the main UK postal service company. e.g: “312 Sackville Road, Sheffield, S10 5GY”

[2] Such as parts of Ecclesfield, Owlthorpe, Stannington and Woodhouse in Sheffield or Beeston, West Bridgford or Carlton, for example, in Nottingham

dependent locality, city, postcode, such as in the case: 15 Enfield Street, Beeston, NOTTINGHAM, NG9 1DN where 'Beeston' is the additional element.

2 to aid ambiguity where there are a number of streets with the same name within a settlement. Similar to the example above, the official address format becomes: Building number and street, *dependent locality*, city, postcode. The inclusion of a dependent locality (second line of the address) in the following example would disambiguate between the three *different* streets with the same name:

1 Back Lane	1 Back Lane	1 Back Lane
Holmfirth	Clayton West	Shelley
HUDDERSFIELD	HUDDERSFIELD	HUDDERSFIELD
HD7 1HQ	HD8 9PP	HD8 8LD

Thus, in these two cases, urban locale information can occur within addresses that might not necessarily be generated from users but may simply be a reflection of Royal Mail practise. The scale and implications of these cases are discussed in more detail in Chapter 8. Such practices, however, are relatively rare and in the urban area of Sheffield account for just over five percent of postcodes.

The strength of the method lies in the fact that although urban locale information is usually not required within postal delivery within the UK, people still frequently include it. Whilst it is still perfectly possible to undertake the method for rural areas, the same user generated content would not be possible. This is because addresses in such areas will most likely reflect those officially defined by Royal Mail - with no opportunity to incorporate user generated content at an urban locale spatial scale. As such, drawing distinctions as in the work of Almadani et al. (2013), who modelled village extents from Royal Mail data, are merely a reflection on the Royal Mail's perception of the world and may not be an accurate reflection on how others may view the naming of space. In the context of this work - 'urban' was defined in line with official Government definitions of Urban areas[3].

It is anticipated that by dealing with specific point locations (postal addresses), people may be able to more accurately reflect on what urban locale they fall within, as opposed to the highly detailed information that would be required to draw precise delineated boundaries of urban locales. Given that people have posted the address information on the web, it is likely that whoever created the information (given that they are not official postal addresses) probably knows the area reasonably well (for example home or work location). Furthermore, given the likely range of datasources, different sources may

[3]built up areas identified by Ordnance Survey with resident populations greater than 10,000 people according to the 2011 Census

provide information within specific geographic spaces (for example estate agents in residential areas, business listings for local shopping centres and so forth), potentially supporting comprehensive geographic coverages.

The Bing Search API[4] was chosen as the 'search engine' in preference to other options (such as Google) as it offers a stable API, a number of free searches per month[5] (which is useful especially in testing phases) and allows fully automated retrieval within the terms and conditions. In contrast, the Google Custom Search[6] is limited to just over 3,000 free searches a month and only provides a maximum of 10 returns for a single search request, compared to Bing's maximum of 50 per request. Thus, the Bing Search API can in theory provide nearly 17 times more returns per month. Testing demonstrated that the Bing Search API produced good levels of data volumes[7]. Automatic interaction with the Bing Search API was achieved through coding written in the open source Python programming language[8].

A case study approach was adopted in order to test the efficiency of the method (validation is contained within Chapter 6). A total of twenty-three different case study 'settlements' were selected covering a wide geographic coverage of the UK (see Figure 5.3) and a range of varying population sizes (see Figure 5.4). Whilst a national coverage of England and Wales urban areas would have possible, this work has been based on the premise of using the free monthly Bing Search API search allowance. The financial cost of running the method nationally depends on the speed of delivery that is required (based on the number of requests sent to the API per month). Full national coverage of England and Wales urban areas could be obtained within a four month period for a cost of just over £4,000[9]. Whilst such an approach would be desirable, the aim of this work is to demonstrate the potential of such a method, as opposed to directly delivering it.

To determine what to search for, the method first obtains a set of postcodes via Ordnance Survey's (OS) Code-Point Open dataset[10]. These are automatically iterated through the Bing Search API which returns snippets (brief details of the web-pages returned from the search string including a title, URL and description). Relatively simple linguistic pattern matching techniques can be applied to the snippets in order

[4] https://datamarket.azure.com/dataset/bing/search

[5] 10,000 if using both the Bing Search API and Bing Search API – Web Results Only

[6] which replaced the Google Web Search API which was officially deprecated in November 2010

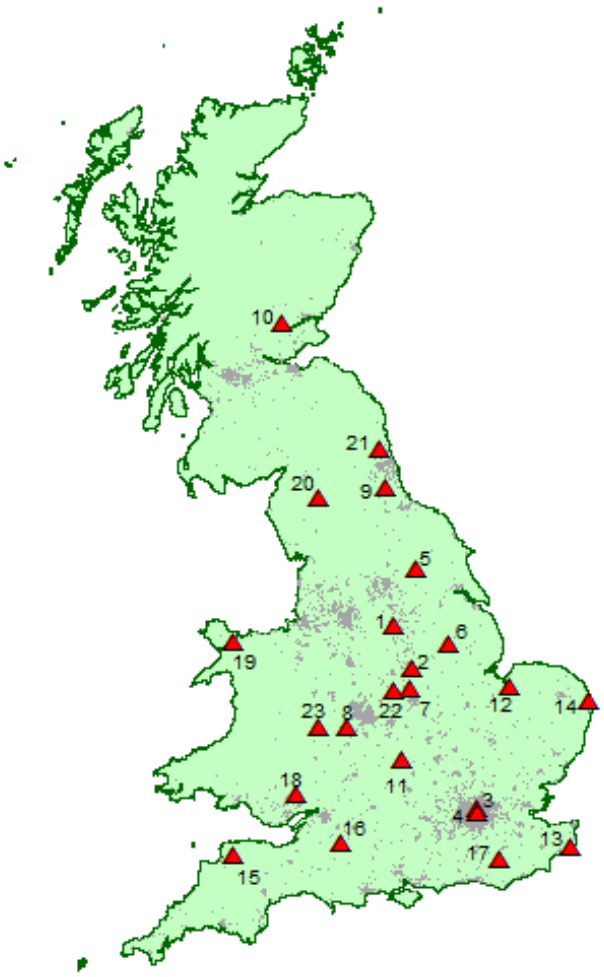
[7] For example, for the postcode 'S10 1GR' - the Google Custom Search API produced 18 returns, 4 of which referred to an urban locale name within the address data, whilst the Bing Search API returned 39 web-pages which contained an urban locale reference in 11 instances

[8] https://www.python.org

[9] based on information from the Bing Search API correct on 06/05/2015

[10] OS Code-Point Open contains both current and historic postcodes which are both included within the method here as people may not necessarily only use current postcodes. About one third of all postcodes in urban areas are historic in nature

Fig. 5.3 Case study settlements

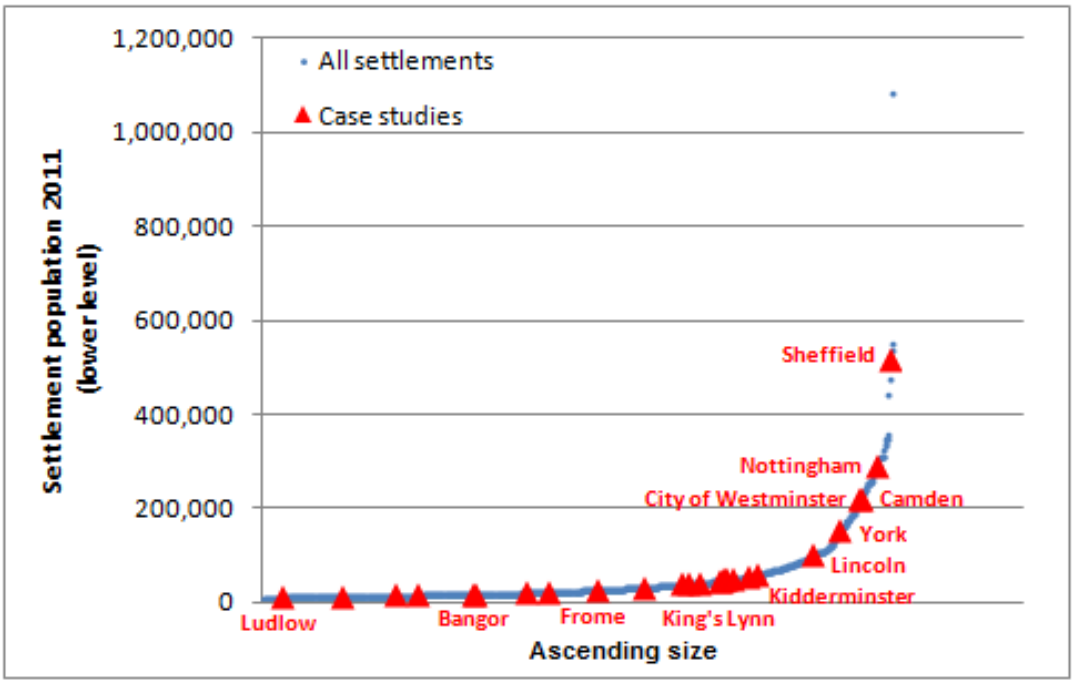


Identifier	Settlement
1	Sheffield
2	Nottingham
3	Camden
4	Westminister
5	York
6	Lincoln
7	Loughborough
8	Kidderminster
9	Durham
10	Perth
11	Banbury
12	Kings Lynn
13	Dover
14	Great Yarmouth
15	Barnstable
16	Frome
17	Crowborough
18	Pontypool
19	Bangor
20	Penrith
21	Morpeth
22	Ashby-de-la-zouch
23	Ludlow

to: (1) find the input postcode; (2) apply rule based filters to identify street and city names (either via identification in Royal Mail's Postal Address File lists, or through detection via common suffixes and abbreviations, such as "rd" or "st"); (3) to then extract the text between these two entities, pairing the resulting urban locale candidate names with the current postcode. This process produces a set of <postcode, urban locale> pairs.

For such operations to be efficient, the snippets require standardisation and cleaning. The returned snippets often contain three dots (ellipses: '...') to show where text has been omitted. Such omissions are commonly relatively short but can also be of considerable length, in which cases the text before and after the ellipse may not necessarily be related. In order to remove such potential false positives, the text of the snippets are cut into segments divided by the ellipses and treated separately. Text is also converted into lower-case and punctuation is removed. Whilst it would have been possible to go to the original web-pages of the URL provided by the Bing Search API returns, the benefit of doing so was small, as generally the data required could be

Fig. 5.4 Case study settlements: Relationship with population size



found within the snippets and the number of data returns were high (as demonstrated subsequently).

This data collection method builds on that of Schockaert and De Cock (2007) and Twaroch et al. (2009a) who extract urban locale terms from postal address information found within Yahoo! Local and Gumtree respectively. The method here, however, differs by not searching on the urban locale name (as supplied by a gazetteer) and then attempting to identify an associated postcode, but instead by undertaking searches based on the *postcode* and then detecting urban locale information that co-occurs between the street and city elements of postal addresses. This allows the method to be non-reliant on gazetteers, and due to the tight delimitation between street and city elements with the formatting of postal addresses, ameliorates issues of place ambiguity (it is, for example, straightforward to separate references to 'Manor Top' from those explicit to 'Manor').

5.2.2 Misspelling identification

Eradicating noise within the set of <postcode, urban locale> is an important issue. The crowd-generated nature of the raw data from which the set is extracted means that it will likely contain numerous variations of spelling referring to the same entity. To address this a data cleansing process is applied, which attempts to identify cases of misspelling via exhaustive application comparing all urban locale name candidates

using Ratcliff/Obershelp (which computes the doubled number of matching characters divided by the total number of characters in the two strings, as used by Schuth et al. (2007); Valls and Ontañón (2012)) and Levenshtein Distance algorithms (which computes the minimum number of edits needed to transform one string into the other, as used by Abdelhaq et al. (2013); Hahmann and Burghardt (2010)). To augment this procedure a distance decay effect was also exploited that enabled a greater likelihood that near things would match with a close spelling than more distant objects. Thus, the calculation of the threshold becomes a weighted measure of both the string distance and geographic distance. Only cases where the mean centres of the two sets of points relating to the urban locale names being compared that were within 1km of each other were considered as potential misspellings (on the premise that an urban locale with a similar name would be unlikely to refer to the same object if it were greater than 1km away).

Both distance metrics utilised return a score between 0-1 (with 1 being a perfect match). Cases where candidate names returned a similarity of over a 0.8 threshold on either of the two measures are deemed misspellings and their associated postcode sets merged. Table 5.1 contains a number of spelling examples. The threshold value is adjusted considering the geographic distance between the mean centres of the two sets of points associated with the urban locale name (with this additional element being termed the 'geographical contribution'). The geographical contribution may add up to an additional 0.2 to the threshold value, where the 0.2 contribution is itself scaled on the geographic distance between the two sets of points. This scaling is achieved by dividing the geographic distance between the two sets of points by 5000 (given the maximum geographic distance could be 1km, resulting in an additional 0.2 contribution to the threshold, whilst a geographic distance of only 10m would incur an additional 0.002 to the threshold). The formula for calculating the final threshold value is thus:

$$0.8 + (\textit{geographical distance}/5000)$$

For example, if the mean centres of Broomhill and Broomhall were only 10m apart, a final threshold value of 0.802 would be required (0.8+(10/5000)). If however, they were 300m or 800m apart – the threshold would rise to 0.86 (0.8+(300/5000)) and 0.96 (0.8+(800/5000)) respectively. Hence, in this instance, Broomhill and Broomhall would only be considered a spelling issue if the mean centres of the two objects were within 400m of each other.

Once this process is complete, it is then possible to geo-locate the surviving urban locales based on geo-referencing of their associated postcodes. This results in a set

Table 5.1 Example spelling comparisons

Word 1	Word 2	Levenshtein algorithm	Ratcliff/Obershelp algorithm
broomhill	bromhil	0.71	0.88
broomhill	bromehill	0.78	0.89
broomhill	broomhall	0.89	0.88
broomhill	brighton	0.25	0.35

of geospatial points labelled with urban locale names and a frequency indicating the number of times that particular name was observed.

5.2.3 Urban locale derivation

In order to construct sets of urban locale surfaces, spatial kernel density estimates (KDE) can be applied to the remaining set of <urban locale, frequency, geocoded pointset> triples. Such an approach has been widely adopted within the literature (including the work of Clough and Pasley (2010); Grothe and Schaab (2009); Hollenstein (2008); Hollenstein and Purves (2010); Purves et al. (2005b); Twaroch and Jones (2010); Twaroch et al. (2008, 2009a,b)). Certain parameters are required for the KDE procedure as discussed within the previous chapter. Firstly, as a raster process, the area of interest (for which the raster grid will be created) needs to be defined. This is undertaken automatically by defining the maximum extent of the point output for each settlement and rounding to 100m to fit with standard British National Grid co-ordinates. Secondly, a quartic approximation to a Gaussian kernel was applied (see Silverman (1986)) using Python scripting and ArcGIS. Finally, the KDE measures used here rely on a cell size of 50m (as recommended in other existing work (Hollenstein, 2008; Li and Goodchild, 2012)) and utilise two different search radii: 300m and 1600m. The use of the two separate search radii is required in order to tease out differing underlying data structures.

A larger geographic smoothing scale will identify broad patterns of data points and thus can be used to remove data outliers beyond suitable concentrations of data. Such a scale, however, would be wholly unsuitable for generating the fine level of detail required for capturing the variability within the urban locale boundaries themselves. In contrast, a smaller scaled search radii would be required that depicted the variations of urban locale boundary at a suitable scale and that would be able to identify presence or absence of the locale, for example considering if particular streets or small groupings of houses might be inside or outside the urban locale. The use of different scales of measurement draws influence from Bibby and Shepherd (2004) where varying scales were used to identify types of settlement pattern. The validity of these parameters are considered in the next section. The 300m parameter is commonly used within the

literature (Thurstain-Goodwin and Unwin, 2000; Twaroch et al., 2008, 2009a) - whilst the 1600m search radius was used within Bibby and Shepherd (2004) and Bibby and Brindley (2013) to identify contiguous masses analogous to urban agglomerations.

Previous work using KDE surfaces for urban locale generation use the absolute numbers of data points. Such an approach works well at defining urban locale (shopping) centres as these areas frequently return the highest data volumes. This body of work, however, also makes use of the proportion of data assigned to particular urban locales (for example cells within 300m KDE may concur that 80% of data identify the cell with a specific urban locale name)[11]. This measure (termed here the urban locale percentage grid) is especially important in edge locations (such as on the urban fringe or next to large parks) or in areas of relatively sparse population density, where absolute values will be *expected* to be low. Such local variations in density are characteristic of first order effects like population density (O'Sullivan and Unwin, 2003) and may not serve to aid the estimation of the extent of the urban locale (Grothe and Schaab, 2008), merely reflecting on the underlying population distribution. Lower data volumes in such cases should not necessarily mean that the area is in any way less associated with a particular urban locale name, it may simply be a reflection of the context in which the place is located.

5.2.3.1 Pre-filtering rules

Before generating the urban locale KDE surfaces, an output filter is applied to remove the appearance of stray non-urban locale terms. Due to the noisy and unstructured nature of web data these are relatively common, with a multitude of incongruous entries appearing between street and city elements[12]. The sheer number of different practices and oddities render application of text specific rules intractable, so these residual texts are handled via the following *generalized* rules (which utilise both the absolute data volumes for an urban locale *and* the percentage of data returns at the two search radii):

- number of data points > 40; and maximum value of [KDE with 1600m search radius] > 15; and at least 30 cells where [KDE with 300m search radius] > maximum value of [KDE with 300m search radius]/2

[11] calculated by dividing each grid cell of the 300m KDE surface of absolute values for the given urban locale by the corresponding grid cells for the total of 300m absolute value KDE surfaces for all urban locales

[12] Example data entities include "coldwell lane *: city:* sheffield: post code: s10 5tj", "tom lane *dental practice in* sheffield s10 3pb", "manchester road *registrant address: sandygate registrant address: registrant address:* sheffield registrant address: south: s10 5ps", etc

- more than 30 cells where the urban locale percentage grid value (with 300m search radius) is at least 70% and at least 30 cells where [KDE with 300m search radius] > maximum of the [KDE with 300m search radius]/2

If the urban locale candidate passes either of these two tests, the urban locale and associated point data are collated and progress to the next stage in order to create a continuous surface.

5.2.3.2 Urban locale surface generation

Unlike previous studies which at this juncture produce a single KDE surface of the point data for each urban locale, this work constructs a number of computational rules based on the KDE measures. This is undertaken in order to further reduce noise / error and is necessary to ensure that endemic data errors are accounted for[13]. Like the filtering rules in the previous section, a combination of absolute and percentage of data are used at the two search radii. A KDE with a 300m search radius was produced in those cells that passed one of the following conditions (providing that there were at least 10 cells in total):

- substantial in terms of the number of data returns: [KDE with 1600m search radius] > maximum of the [KDE with 1600m search radius]/2; or
- locally significant with a majority percentage but also sustained over a specific area: [percentage grid of KDE with 300m search radius] > 50 and [KDE with 300m search radius] >70.

One criticism of the KDE approach is that it has a tendency to over-smooth at the edges (Purves et al., 2005a; Twaroch et al., 2009a) and to help compensate for this a 2% cut-off is applied in order to 'trim the edges'. In such cases grid cells with less than or equal to 2% of the maximum grid cell value were reclassified as zero and hence removed.

It should be noted that the method is based on deriving data from postal addresses and as such is centred upon populated areas. Unpopulated areas (such as parks and cemeteries) are not of core concern but due to the KDE smoothing processes may be incorporated within the urban locale if they are smaller than 600 metres in diameter

[13] Such as in Sheffield, UK where data for 'Norton' urban locale was incorrectly assigned across the city due to a 3rd party provider for Google incorrectly labelling all city postcodes as 'Norton' instead of 'Sheffield'. Even with a 10 or 20% cut-off (Hollenstein, 2008; Twaroch et al., 2009b), existing methods of data cleansing would incorrectly deem Norton to exist across the entire city

and surrounded by data points for the given locale, or if they are smaller than 300 metres and are located on the edge of a locale – as defined by the data. If the purpose of urban locale generation remains with populated quanta (for example tying Census data to locales), such matters are of little consequence. In other cases, however, the identification of unpopulated space may be of concern. It may be possible to combine data from different sources (such as urban locale names appearing within Flickr tags) to alleviate the issue – as discussed within the Further Research in Chapter 8.

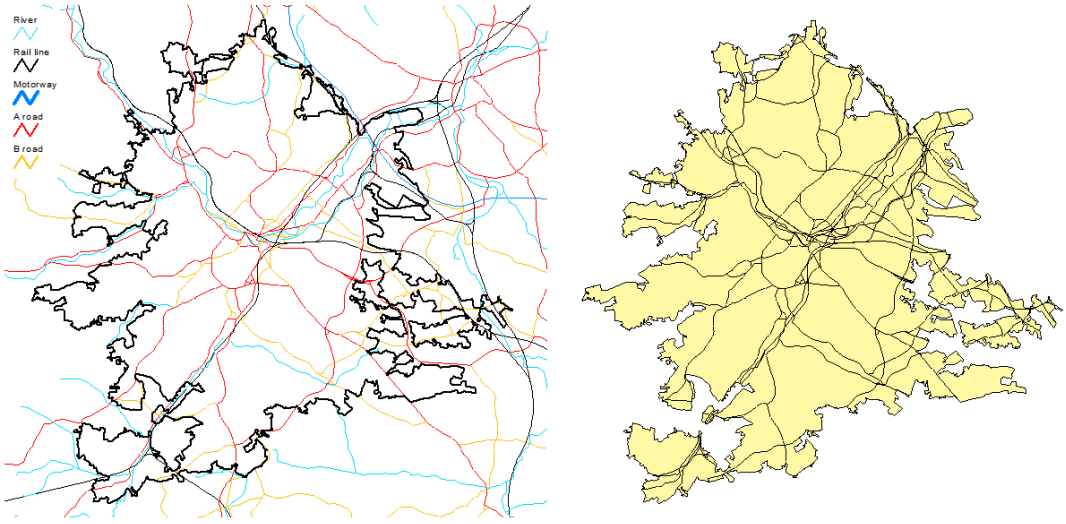
5.2.4 Linear features as breakline constrains for smoothing

On many occasions (but not always), urban locale boundaries may adhere to 'hard' boundaries such as rivers, railway lines, motorways and other main roads. On the other hand, KDE output pays no regard to such features, resulting in urban locales spreading across these constraints. The use of hard breaklines to manipulate the behaviour of raster surfaces is relatively common within terrain modelling and network analysis (see de Smith et al. (2007)). The problem with urban locale surfaces, however, is that at times the locale will cross such features, but at others - it will not. This body of work introduces a simple solution to this issue based on the premise that crossing specified linear features is only allowed to occur if the underlying data supports the notion - whereby data (extracted from the web containing the urban locale name within postal addresses) also traverse the linear feature. Firstly, a small pre-processing step for data preparation is required whereby the polygon of each urban settlement study area is converted to a line feature and merged with the other linear features that the user desires to use as hard breakpoints (for example rivers, railways, motorways and so on[14]). The lines are then converted back into a polygon theme whereby each bounded area of space created from the combined line-set represents a new separate polygon feature. An example of this pre-processing is illustrated within Figure 5.5.

Within the automated scripting process, before final KDE surfaces are created for an urban locale, only polygons containing at least one data point (after the data cleaning and noise reduction processes have been implemented) are available to the KDE process. This is handled through the use of a raster grid mask where the pre-processing polygons with relevant point data are assigned a value of 1 and those with no data are assigned a value of 0. The KDE surface is multiplied by the mask to remove those areas where smoothing should not transpire if the underlying point data does not support it. An example of the method is illustrated in Figure 5.6. Figure 5.6 reinforces the importance of using the percentage of data returns as a measure, as without such regard absolute data volumes alone (shown in the figure) highlight an apparent 'hole'

[14] such data can be freely downloaded as OS Open data such as OS Open Roads, OS Open Rivers, Meridian^{TM}2 from https://www.ordnancesurvey.co.uk/opendatadownload/products.html

Fig. 5.5 Pre-processing: conversion of aggregated linear features to polygons



a) Example of linear features, Sheffield *b) Conversion to polygons, Sheffield*

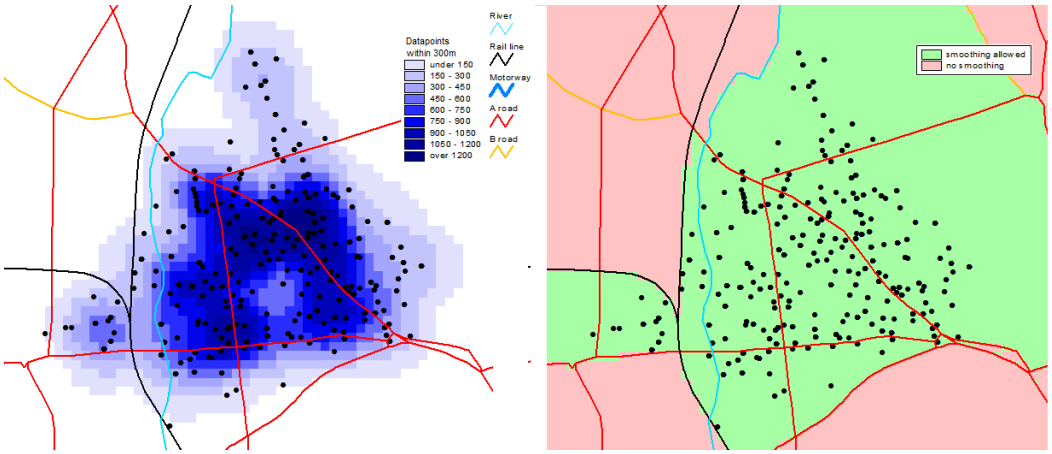
at the centre of the urban locale. The area corresponds to Mellers Primary School. This should not be interpreted that the 'hole' is less like the Radford urban locale than surrounding spaces but simply reflects that there are fewer data returns around the school. If the percentage of data returns for Radford urban locale were illustrated, there would be no hole evident as although there are fewer data returns around the school, there is complete agreement that the area is termed 'Radford'. The use of linear features as breaklines is represented as an user specified option within the overall automated scripting process.

5.2.5 Post-processing: Synonym identification

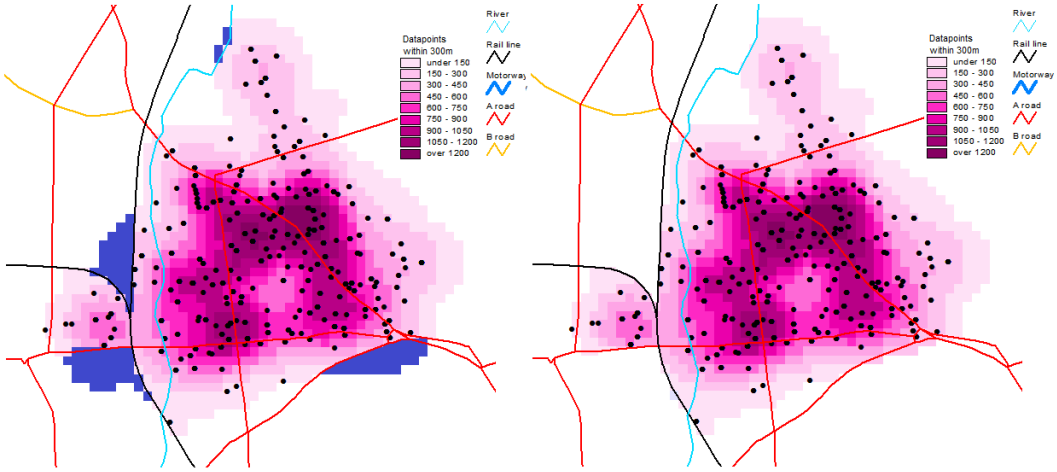
Issues related to spellings are dealt with in Section 5.2.2. There are, however, cases where synonyms may exist - where a completely different word is used to describe the same place (for example New York and the Big Apple or Sheffield and the Steel City). Variations of place names due to different languages can be identified through the synonym process, as an example the postcode CF3 5EA in Cardiff returns data relating to both the English naming of 'St. Mellons' and the Welsh version of the same place - 'Llaneirwg'. Synonyms would not be identified within the spelling process as the two words may bear no, or very little, similarity between them. In this regard, the comparability is measured not in terms of the place name but rather the spatial footprint of the object. Synonyms should have near identical geographies. Based on this, a post-process has been implemented that performs linear grid regression[15] comparing the data values for every grid cell for the two urban locale raster grids, in

[15] Using the Python numpy.polyfit function

Fig. 5.6 Linear features as breakline constrains for smoothing: Example of Radford, Nottingham



a) KDE (Absolute data, 300m bandwidth) *b) Mask: green = where smoothing allowed*



*c) Fig(a)*Fig(b): blue = areas masked out* *d) Final modified KDE (300m bandwidth)*

order to identify those with great similarity. Cell values are only compared if data exists in one of the two grids being analysed. The output is not only of interest to synonym identification but also as a test of general similarity between urban locales.

5.2.6 Effective storage and distribution

None of the existing literature discusses the issue of storing the data once collected. This may at first appear a trivial matter but if the procedures were run on a national scale, the problem of dealing with thousands of raster grids (one for each urban locale) could become cumbersome and lead to issues of ambiguity. The following procedure can help alleviate this issue. Each settlement (as defined by the OS built up areas) is initially treated as a separate entity. The final set of raster grids for the settlement can then be tied back to the postcode point locations (OS Code-Point Open dataset), with

membership of every urban locale becoming a column attribute. After-all, the original input data for the process were postcodes and so it seems sensible (once data have been cleaned, spread and manipulated) that data revert back to their original form for effective storage and a state that is more suitable for further analysis. Allocating the gridded data back to postcodes also highlights that the process is concerned only with populated areas and does not attempt to infer urban locales within unpopulated spaces. When data are held within Python, this is achieved using a tuple list structure for each postcode:

[Postcode, [[Urban Locale Name, Percent], [Urban Locale Name, Percent]...]]

For example:

[S2 2AT, [[Arbourthorne, 29.2], [Gleadless, 70.8]]] [S10 3DF, [[Broomhill, 100.0]]] [S2 2AT, [[Brightside, 24.4], [Grimesthorpe, 69.3], [Wincobank, 6.3]]]

Within a GIS structure, data storage is less efficient, requiring a variable for each urban locale within the settlement. A map view containing each raster grid can also be automatically converted to Keyhole Markup Language (KML)[16] forming a single KML or compressed KMZ file that can then be easily shared with other people and viewed free in popular applications such as Google Earth or ArcGIS Explorer.

5.3 Parameter justification

A comprehensive validation procedure is undertaken within Chapter 6. This section, however, explores the suitability of the various main parameters used within the method. Whilst the parameters appear, within the previous section, to have been artificially produced, it should be stressed that they have been derived through an extensive period of testing using local knowledge of the urban locales in Sheffield.

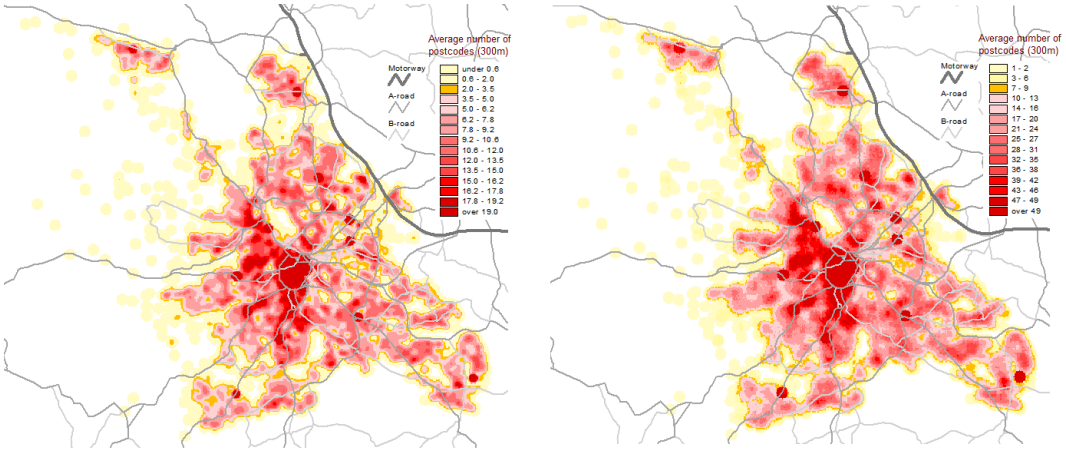
5.3.1 KDE

As discussed in the previous chapter, the choice concerning the kernel function is less critical than that of the search size (also known as the bandwidth) (O'Sullivan and Unwin, 2003). It should, however, not be completely disregarded. The choice predominately rests upon the level of completeness and accuracy within the data. This is because the main difference between KDE and other similar measures, such as simple density or moving averages, is that KDE gives a greater importance to

[16] Using Map to KML within ArcGIS

close-by data compared to that further away. If data represent a complete and accurate population, then the moving average approach (such as adopted within the work by Bibby (2009); Bibby and Shepherd (2004) and Bibby and Brindley (2006, 2010, 2013)) would be most appropriate. If data, however, were a sample with inherent uncertainty, giving greatest weight to closest data would appear sensible. The similarity of the geographic distribution of KDE and moving average techniques is evident in Figure 5.7 which shows the number of postcodes for the city of Sheffield. Given the overall similarity of the two measures, this method followed the majority of similar studies (Clough and Pasley, 2010; Grothe and Schaab, 2009; Hollenstein, 2008; Hollenstein and Purves, 2010; Purves et al., 2005b; Twaroch and Jones, 2010; Twaroch et al., 2008, 2009a,b) and utilises the KDE approach.

Fig. 5.7 Differences between KDE and simple moving average: number of postcodes, Sheffield



a) KDE (300 metres) *b) Moving average (300 metres)*

5.3.2 KDE: Bandwidth

The choice of the search size (also known as bandwidth) is the most critical function within the KDE process (O'Sullivan and Unwin, 2003), although it is acknowledged that there is no general agreement on how to best approach the bandwidth choice (Hollenstein and Purves, 2010). The bandwidth is usually derived experimentally (O'Sullivan and Unwin, 2003) but needs to be small enough to capture local variation within the feature of interest but large enough to capture multiple point locations within the search window (Almadani et al., 2013; Jones et al., 2008a). The potential of automating the choice of bandwidth is considered subsequently within Section 5.3.5. The purpose of KDE is to fill-in data in areas of low data volumes. As such, selection of a suitable bandwidth within KDE is dependent upon the distribution of data points within the study area. If the data volumes and spatial footprint are densely spread, then the bandwidth can be smaller than if data are sparsely distributed and the KDE needs

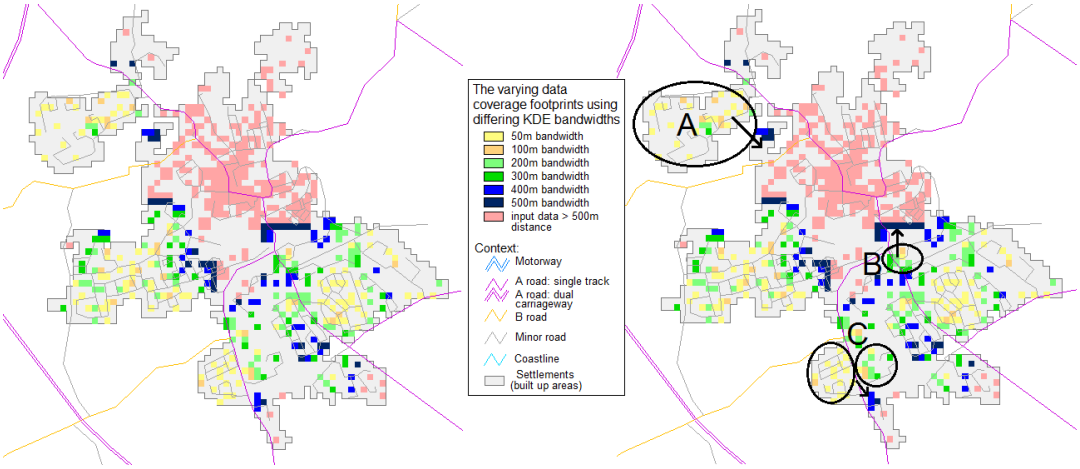
to infill between data points. Table 5.2 outlines the data coverage of the case study areas using a variety of bandwidths.

Not all spaces within urban areas, however, may necessarily have urban locales associated with them. Although when assigning other socio-economic data to urban locale units it may be advantageous to have a near complete coverage. It is expected that larger regional cities (such as Nottingham, Southampton, Liverpool) are more likely to have more complete coverages of urban locales than smaller towns. Within the larger case study environs (Camden and Westminster, Sheffield, Nottingham and York), there appears little benefit of using bandwidths greater than 300 metres, which would tend to oversmooth data rather than simply infilling data voids (an increase of less than 1% coverage in both Nottingham and Camden / Westminster was recorded when increasing the search radius from 300 to 500m). Bandwidths smaller than 200m would appear to result in extensive areas of populated land not being assigned urban locales (with 37% of area not being covered in Sheffield and 39% in York). Thus, bandwidths of 200m to 300m would appear most ideal. The 300m search radius increased data coverages in Sheffield and York to over 90% (from 85 to 92% in Sheffield and 83 to 91% in York).

It is, however, not solely a question of maximising the increases in coverage but also about the spatial footprint of smoothing with different bandwidths. The key is to enable infilling of those parts of the settlement without data, whilst not over smoothing and excessively expanding data values into areas that may not necessarily be associated with the spread urban locale name. The yellow cells within Figure 5.8 represent the presence of urban locale names tied to the raw 50m grid cells, whilst in contrast those cells coloured light green, green and dark green represent the populated cells which are only assigned urban locale names due to the smoothing process at 100m, 200m and 300m respective bandwidths. Similarly, blue and dark blue cells are those cells that did not contain references to urban locale names at the postcode or 50m level but could be assigned locale names during the KDE process if smoothed to a corresponding 400 or 500m. The rationale is to explore the suitable bandwidths which appear to allow for data infilling but does not extend the urban locale information beyond apparently coherent spatial configurations of space. For example, data around point A within Figure 5.8b starts to spread beyond the coherent mass into a separate urban area to the south when using bandwidths greater than 300m, whilst data around point B expands into a different area to the north using 400 and 500m bandwidths. Similarly data around C in Figure 5.8b expand to the south into what might be a separate area of the settlement when the bandwidth exceeds 300m. Thus, Figure 5.8 demonstrates that generally bandwidths greater than 300m appear to spread data excessively and 200-300m bandwidths would be most suitable.

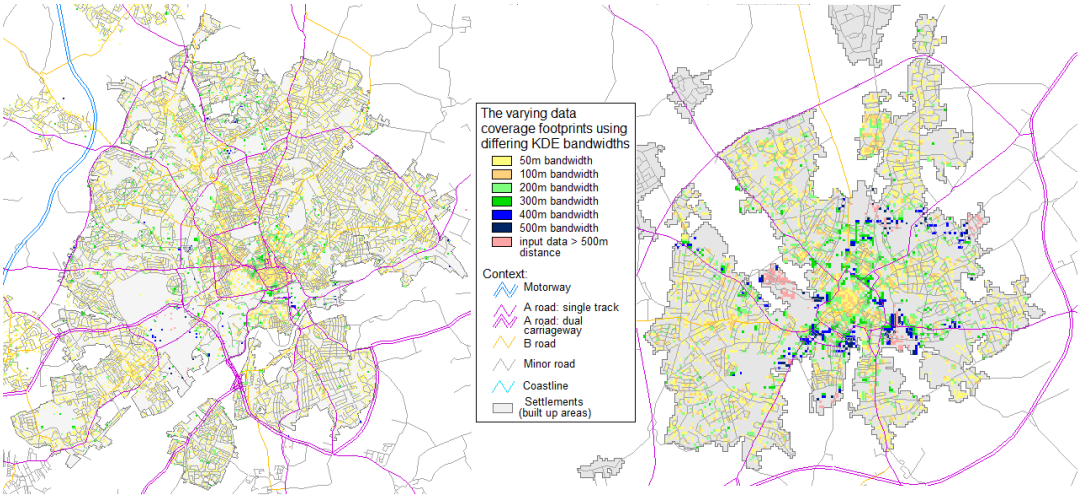
Such a proposition is supported by further case examples (such as the spanning across rivers in Great Yarmouth and King's Lynn and crossing dual carriageways in Bangor, Penrith, Perth when using bandwidths greater than 300m). Figure 5.9 shows the spread of data coverage using a range of bandwidths from 50 to 500m for Nottingham and York, whilst the other case study areas can be found within Appendix B. A 300m parameter was utilised for the reasons cited above and the fact that a number of previous studies have used the 300m bandwidth (Thurstain-Goodwin and Unwin, 2000; Twaroch et al., 2008, 2009a), whilst none had previously employed a 200m radius.

Fig. 5.8 Spreading of data beyond coherent areas: within Morpeth



a) Raw output *b) Spreading of coherent areas*

Fig. 5.9 Spread of data coverage using different KDE bandwidths



a) Nottingham *b) York*

Table 5.2 Data coverage using differing bandwidths

Settlement	Populated cells (50m resolution)	Data coverage (%) 50m search radius	Data coverage (%) 100m search radius	Data coverage (%) 200m search radius	Data coverage (%) 300m search radius	Data coverage (%) 400m search radius	Data coverage (%) 500m search radius
Camden and Westminster	17,314	58.5	86.3	97.7	99.1	99.7	99.9
Sheffield	13,550	46.9	62.6	84.5	92.3	95.4	97.2
Nottingham	12,542	69.6	83.9	96.6	99.1	99.6	99.7
York	4,288	43.4	61.2	82.8	90.8	94.9	96.9
Lincoln	3,242	19.2	23.6	32.4	39.0	43.5	47.6
Great Yarmouth	1,611	47.1	54.7	61.6	66.4	71.4	76.6
Kiddeminster	1,475	12.4	17.2	27.2	36.6	42.8	48.3
Banbury	1,429	7.0	9.7	16.7	21.8	26.9	31.9
Loughborough	1,308	5.9	10.5	20.2	28.3	38.1	47.9
King's Lynn	1,299	29.1	33.0	41.8	48.3	52.4	56.5
Durham	1,269	42.6	50.9	61.5	66.1	69.5	72.3
Dover	1,163	23.5	26.9	33.4	38.6	42.5	46.9
Pontypool	965	72.2	80.1	91.6	96.6	98.5	99.4
Barnstaple	929	35.7	46.4	64.6	75.9	84.8	89.9
Frome	761	2.2	3.5	5.0	7.8	10.8	13.0
Crowborough	673	6.2	8.2	12.9	15.3	16.8	18.7
Penrith	587	2.7	3.7	5.6	8.9	12.4	16.9
Morpeth	551	18.7	23.2	35.2	43.9	51.5	58.1
Bangor	476	5.0	5.9	8.8	9.7	10.7	11.1
Perth	467	36.4	40.3	53.3	60.6	64.0	68.5
Ludlow	345	1.2	1.2	1.4	3.5	5.5	8.1
Ashby-de-la-Zouch	341	0.0	0.0	0.0	0.0	0.0	0.0

5.3.3 KDE: Cell size

Ideally the cell size should be as small as possible but account for issues within the postcode accuracy and spread of the households represented by a single postcode centroid. The postcode accuracy is not a significant factor, being accurate to a metre resolution[17]. The area an average postcode might cover can be estimated on the following basis. There are on average 15 households per postcode[18]. Given that housing is being constructed at an average density of 43 dwellings per hectare[19], this equates to roughly 2.9 postcodes per hectare. Thus, one postcode covers approximately an area of 69.8 x 69.8 m^2. It is desirable to have a rounded figure and the estimate is close to the 50m cell size used by Li and Goodchild (2012), Hollenstein (2008) and the OS (who use a 50m grid for the creation and representation of the UK's official built up areas, 2011[20]).

5.3.4 Data point frequency cut-off

One of the more critical pre-filtering rules entailed having a minimum of 40 data points for a named area to be considered for inclusion as an urban locale. The logic being that fewer than such a number would not produce a reliable and robust KDE surface. The best evaluation of this measure is gained through the validation process where urban locales from the final web extraction output can be compared to names appearing within other urban locale gazetteer sources (see Chapter 6 for details). For illustrative purposes, however, Table 5.3 demonstrates some examples of the sort of places that are excluded based on the numbers of data points. Whilst the table shows evidence of smaller urban locales (such as High Storrs and Carbrook both with 36 data points), it appears that a number of non-urban locale place types (streets, squares and facilities) start to appear at fewer than 40 data points.

5.3.5 Adaptive parameters

A number of previous studies (Jones et al., 2008a; Twaroch et al., 2008) have highlighted the potential use of adaptive kernel density estimation (Brunsdon, 1995) for defining KDE parameters automatically, based on the original underlying point data

[17] ONS Postcode Directory (ONSPD) User Guide

[18] using the postcode headcounts http://www.nomisweb.co.uk/census/2011/postcode_headcounts_and_household_estimates

[19] Land use change statistics: https://www.gov.uk/government/statistical-data-sets/live-tables-on-land-use-change-statistics

[20] www.ons.gov.uk/guide-method/geography/products/census/key-statisics-for-built-up-areas-user-guidance.pdf

Table 5.3 Examples of places excluded on the basis of having fewer than 40 datapoints

Place	Datapoint Frequency
leopold square	38
greasbro road	37
high storrs	36
tech street2 street3	36
carbrook	36
admin street2 street3	34
endcliffe	34
university	33
little matlock stannington	31
page hall	31
s10 property rent search	31
bamford street tram stop	29
road	29
herdings	27
farm	27
lower walkley	27
manchester road	25
park hill	25
wicker	25
court	24
primary school	23
batemoor	23
entrance	18
tram stop	16
hemsworth	16
lowfield	15

distribution or through machine learning and additional data (use as land coverage) (Cunha and Martins, 2014). Thus, it would be possible to increase the search radius if data points were sparse or within unpopulated areas. Such methods are particularity suited where:

- the objects of concern vary in geographic scale (for example using the same process to map vernacular names of the Rocky Mountains, USA and the Cotswolds, UK, as in Jones et al. (2008a), or where:
- there is considerable variation in the context in which the objects of concern are found (that is to say where the underlying numbers of population vary greatly). In such cases, it is often the variation in population rather than the objects of interest that drive the geographic pattern, for example KDE mapping of crime levels in both urban and rural settings.

Given that neither of these issues is of primary concern (the focus remains urban locales within an urban context), the benefits of using adaptive parameters are unclear. There would, however, be the danger that adaptive methods could be susceptible to over smoothing in areas of low data. Furthermore, Chapter 1 raised the additional research question of comparing the frequency of urban locale data to the underlying population in order to potentially identify areas with low/high data returns. Such an undertaking would be hindered if the denominator of the density function was varying in addition to that of the numerator. Given the already relatively complex series of parameters for the method and that results obtained were considered to be good with the methods applied, it was felt that an adaptive process was not required and, in this instance, would be more likely to add uncertainty rather than clarity.

Whilst it would be entirely possible to increase the search radius in unpopulated areas, doing so without any other information may well introduce error as areas are blindly incorporated. Consider the example of a large park between two urban locales. Increasing the search radius in such a case would simply prescribe areas of the park to the nearest locale. In contrast to being split by nearest locale, one may well consider that the park is more likely to either fall entirely within one, both or neither of the urban locale names. In such a scenario, it would seem more fitting to explore the potential of incorporating alternative data (such as references to urban locales within data relevant to unpopulated areas, such as Flickr tags) than forcing names upon unpopulated areas where there is no underlying support in the data. The potential for incorporating other data within the method is discussed in more detail within the Further research section in Chapter 8.

Although the various parameter testing undertaken within this section is critical in developing a suitable methodology, the real testing of the parameters manifest them-

selves in how well the output performs within the validation process (as discussed in the following chapter).

5.4 Summary

This chapter outlines the method undertaken in order to map vague urban locale objects in a fully automated manner. The aim of the approach developed is to implement the findings of the previous chapter in order to construct vague spatial boundaries and vague place names of urban locales using passive data mined from the web. The process turns multiperson disagreement into a positive phenomenon by capturing the extensive differing views of named space that people may hold and express by uploading them within postal address information on the web. Such output can be mapped in a vague approach, as discussed, in order to form vague geographic objects. One potential strength of the approach is that people are more likely to know if a point location relating to them (such as a home, place of work and so forth) is within a given named area rather than having to draw exact boundaries of an urban locale where greater uncertainty is likely to exist. Whilst the method may appear to be a linear process, it should be stressed that both the approach, and in particular the various parameters used within it, have been derived through an intensive iterative process using extensive testing and local knowledge. The following chapter illustrates example output from this method along with providing validation of the process.

Chapter 6

Results and Validation of Vague Urban Locales

6.1 Results

Following application of the vague urban locale methodology as set out in the previous chapter, this chapter demonstrates example output and proceeds to validate the work. Urban locales were generated for the twenty-three case study areas. In total 140,856 postcodes were sent to the Bing API, producing 2,182,558 data returns. Of these, 519,799 contained additional information between the street name and settlement, which were used to generate a total of 550 urban locales within all twenty-three case studies. Summary statistics can be found in Table 6.1, demonstrating that data volumes varied considerably - predominately as a reflection of the population size of the case study areas. For example, output ranged from the 64,349 postcodes (including historic postcodes) within the study areas of Camden and Westminster, London to just 540 postcodes in Ludlow which produced 1,097,345 and 7,416 data returns respectively. Of these records, 126,236 and 143 contained urban locale information, which were then used to identify 143 urban locales in Camden and Westminster and just a single locale in Ludlow. Although each urban locale was generated from an extremely varied number of data returns, on average a locale contained 666 data points.

By using the Bing API, the method draws information from a wide range of different sources of data. These can be assessed through the domain names of the data points. The most frequently occurring domains can be found in Table 6.2 for a selection of the case studies, whilst a more comprehensive investigation is provided in Appendix C. Frequent classes include *real estate* sites (such as zoopla.co.uk, rightmove.co.uk and primelocation.com) and *business directory information* sites (including local-

Table 6.1 Summary statistics for data output

Settlement	Postcodes	Requests sent	Total data returns	Data returns with urban locale address element	Number of urban locales	Urban locale datapoints	Area of settlement with locale data (%)	Requests sent per postcode	Urban locale datapoints as a percentage of Total data returns	Average final datapoints per locale
Camden and Westminster	64,349	119,480	1,097,345	198,111	143	126,236	99.1	1.86	11.50	883
Nottingham	18,946	25,953	312,760	139,472	102	121,154	99.1	1.37	38.74	962
Sheffield	18,244	28,001	288,071	84,245	121	63,871	92.3	1.53	22.17	528
York	9,393	14,811	105,342	18,388	37	14,825	90.8	1.58	14.07	401
Great Yarmouth	2,089	3,551	32,191	22,412	10	8,440	66.4	1.70	26.22	844
Lincoln	4,351	6,848	62,194	9,026	16	6,971	39.0	1.57	11.21	436
Pontypool	1,339	1,804	13,168	7,124	14	6,465	96.6	1.35	49.10	462
Durham	1,980	3,193	22,683	17,769	15	3,640	66.1	1.61	16.05	243
King's Lynn	1,769	2,619	23,301	4,186	9	3,328	48.3	1.48	14.28	370
Barnstaple	1,429	2,571	20,577	4,024	13	3,194	75.9	1.80	15.52	246
Dover	1,620	2,580	22,032	4,022	7	3,140	38.6	1.59	14.25	449
Banbury	2,688	3,770	27,579	2,072	6	1,165	21.8	1.40	4.22	194
Kiddeminster	1,937	3,061	29,159	2,201	8	1,002	36.6	1.58	3.44	125
Perth	1,833	3,114	25,999	1,734	6	830	60.6	1.70	3.19	138
Morpeth	912	1,263	7,556	777	5	445	43.9	1.38	5.89	89
Frome	956	1,576	14,452	896	2	418	7.8	1.65	2.89	209
Penrith	910	1,449	10,740	736	3	359	8.9	1.59	3.34	120
Loughborough	2,194	3,484	28,594	1,163	4	333	28.3	1.59	1.16	83
Crowborough	1,045	1,572	13,684	506	2	211	15.3	1.50	1.54	106
Bangor	1,540	2,002	7,895	469	2	146	9.7	1.30	1.85	73
Ludlow	540	924	7,416	410	1	143	3.5	1.71	1.93	143
Ashby-de-la-Zouch	792	1,154	9,820	56	0	0	0.0	1.46	n/a	n/a
TOTAL	140,856	234,780	2,182,558	519,799	550	366,316	n/a	n/a	n/a	n/a
AVERAGE	6,403	10.67	99,207	23,627	25	16,651	80.97	1.67	16.78	666

life.co.uk, smilelocal.com, bizzy.co.uk, ufindus.com, yell.com) - which could be case study specific (such as nottingham.enquira.co.uk, londontown.com, viewlondon.co.uk, allinlondon.co.uk, ilondon.co.uk). Whilst real estate and business directory information account for the most frequently occurring domains, a host of user generated content from social media can also be found within the data. For example, in Nottingham there were 231 data points from facebook.com, 160 from linkedin.com, 147 from gumtree.com and 131 from ebay.co.uk. The tail of the domain name distribution should not be disregarded and in the case of Nottingham over 17% of data came from domains that produced 10 or fewer data points. A fuller investigation of datasources and domain names can be found both in the classification of URLs (Chapter 7.1) and in Appendix C.

This chapter will continue with the distinctions set out within the concept of place as outlined in Chapter 2 and separately consider output concerning spatial boundaries and place names. It will firstly examine the vague urban locale extents before turning attention to vagueness arising in the place name from synonym identification and issues of spelling, which also derive from the method as outlined in the previous chapter.

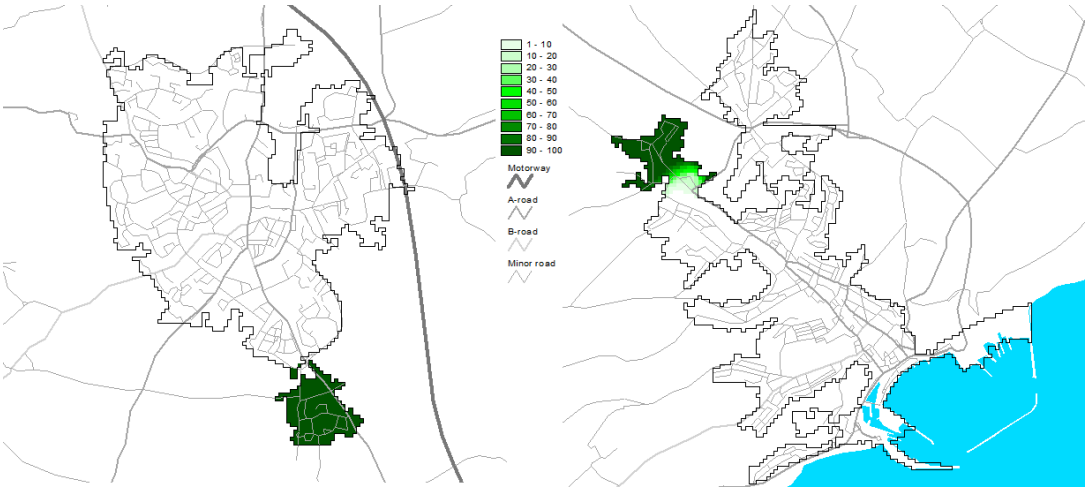
6.1.1 Spatial boundary results

For purposes of illustration six examples of the grids produced by the automated procedures are shown in Figure 6.1, whilst output for data from the other case study areas can be found in Appendix D. In contrast to the individual urban locale grids within Figure 6.1, Figure 6.2 illustrates the output when urban locale KDE surfaces are allocated back to postcode locations (as described within Section 5.2.6) for the city of Sheffield - where each *combination* of urban locale is shown in a different colour (for example in the north-west corner: areas defined as Stocksbridge are coloured green, areas called Deepcar are in red and those areas that are called either Stocksbridge or Deepcar are coloured blue). Using the same process, output for urban locales in and around the London Districts of Camden and Westminster are shown in Figure 6.3. In addition, this figure scales the size of the urban locale labels by the number of data returns, so that more prominent locales exert greater visibility. Examples of the output for Sheffield as KML data in Google Earth can be found in Figure 6.4.

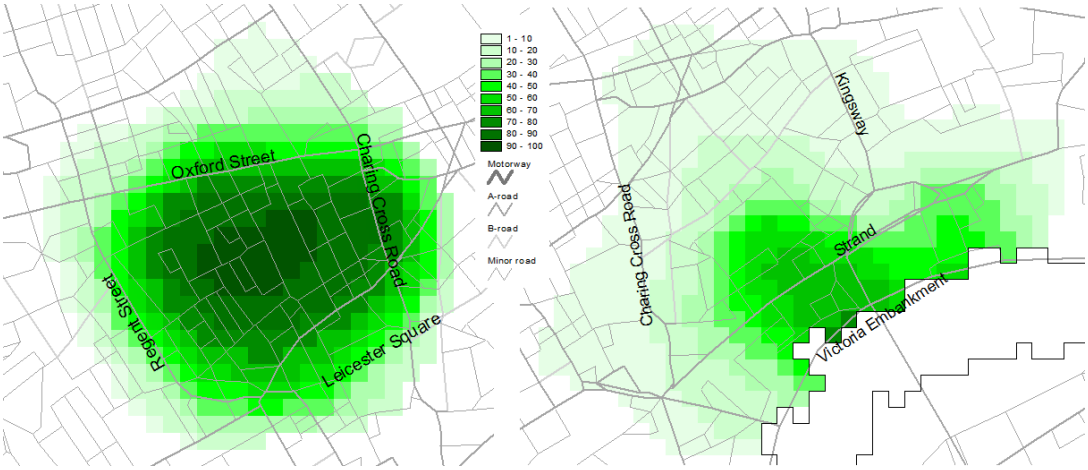
Table 6.2 Domain name frequencies for selected case study areas

Domain name frequencies for **Nottingham**:				Domain name frequencies for **Camden and Westminster**:				Domain name frequencies for **Dover**:			
Domain name	Frequency	Percent	Cumulative %	Domain name	Frequency	Percent	Cumulative %	Domain name	Frequency	Percent	Cumulative %
www.zoopla.co.uk	6,371	5.3	5.3	www.londontown.com	4,583	3.6	3.6	www.zoopla.co.uk	294	9.4	9.4
nottingham.enquira.co.uk	3,162	2.6	7.9	www.viewlondon.co.uk	4,359	3.5	7.1	www.rightmove.co.uk	115	3.7	13.0
www.smilelocal.com	3,036	2.5	10.4	www.allinlondon.co.uk	4,073	3.2	10.3	www.dover.gov.uk	89	2.8	15.9
bizzy.co.uk	2,930	2.4	12.8	www.ilondon.co.uk	3,161	2.5	12.8	www.yell.com	87	2.8	18.6
www.ufindus.com	2,738	2.3	15.1	www.sugarvine.com	2,597	2.1	14.9	www.streetcheck.co.uk	75	2.4	21.0
www.rightmove.co.uk	2,737	2.3	17.3	www.locallife.co.uk	2,367	1.9	16.7	www.primelocation.com	73	2.3	23.3
www.yell.com	2,200	1.8	19.1	www.london-eating.co.uk	2,314	1.8	18.6	www.thomsonlocal.com	72	2.3	25.6
www.zettai.net	2,129	1.8	20.9	london.enquira.co.uk	1,996	1.6	20.2	www.bigreddirectory.com	70	2.2	27.9
www.cylex-uk.co.uk	2,066	1.7	22.6	www.visitlondon.com	1,741	1.4	21.5	www.britishpostcodes.info	64	2.0	29.9
www.ukcities.co.uk	2,060	1.7	24.3	www.londonnet.co.uk	1,647	1.3	22.8	ukpostcodedata.com	57	1.8	31.7
www.yelp.co.uk	1,856	1.5	25.8	homes.trovit.co.uk	1,645	1.3	24.1	www.zettai.net	50	1.6	33.3
www.moreuk.com	1,799	1.5	27.3	www.192.com	1,281	1.0	25.2	www.smilelocal.com	45	1.4	34.7
www.nottinghamonline.co.uk	1,737	1.4	28.7	www.localdatasearch.com	1,217	1.0	26.1	www.checksure.biz	43	1.4	36.1
www.192.com	1,682	1.4	30.1	www.checksure.biz	1,094	0.9	27.0	www.118118.com	39	1.2	37.4
company-director-check.co.uk	1,645	1.4	31.5	www.fluidlondon.co.uk	1,094	0.9	27.9	directory.independent.co.uk	38	1.2	38.6
www.mandis.co.uk	1,528	1.3	32.7	www.freeindex.co.uk	1,084	0.9	28.7	www.shiftbusiness.co.uk	37	1.2	39.7
www.freeindex.co.uk	1,394	1.2	33.9	www.mandis.co.uk	1,055	0.8	29.6	homes.trovit.co.uk	37	1.2	40.9
directory.independent.co.uk	1,383	1.1	35.0	www.yell.com	952	0.8	30.3	www.mandis.co.uk	35	1.1	42.0
www.checksure.biz	1,357	1.1	36.2	www.rightmove.co.uk	871	0.7	31.0	www.ufindus.com	32	1.0	43.1
www.touchlocal.com	1,335	1.1	37.3	www.qype.co.uk	848	0.7	31.7	plus.google.com	30	1.0	44.0
www.thomsonlocal.com	1,318	1.1	38.4	www.zettai.net	804	0.6	32.3	www.moreuk.com	30	1.0	45.0
www.london-gazette.co.uk	1,272	1.0	39.4	www.118.com	798	0.6	32.9	www.london-gazette.co.uk	29	0.9	45.9
www.118.com	1,137	0.9	40.3	www.bizwiki.co.uk	780	0.6	33.6	www.housepricepaid.com	27	0.9	46.8
www.nottingham.co.uk	1,114	0.9	41.3	bizzy.co.uk	768	0.6	34.2	www.192.com	27	0.9	47.6
homes.trovit.co.uk	1,076	0.9	42.1	www.fancyapint.com	766	0.6	34.8	www.nhs.uk	22	0.7	48.3
TOTAL	121,154	100.0	100.0	TOTAL	126,236	100.0	100.0	TOTAL	3,140	100.0	100.0

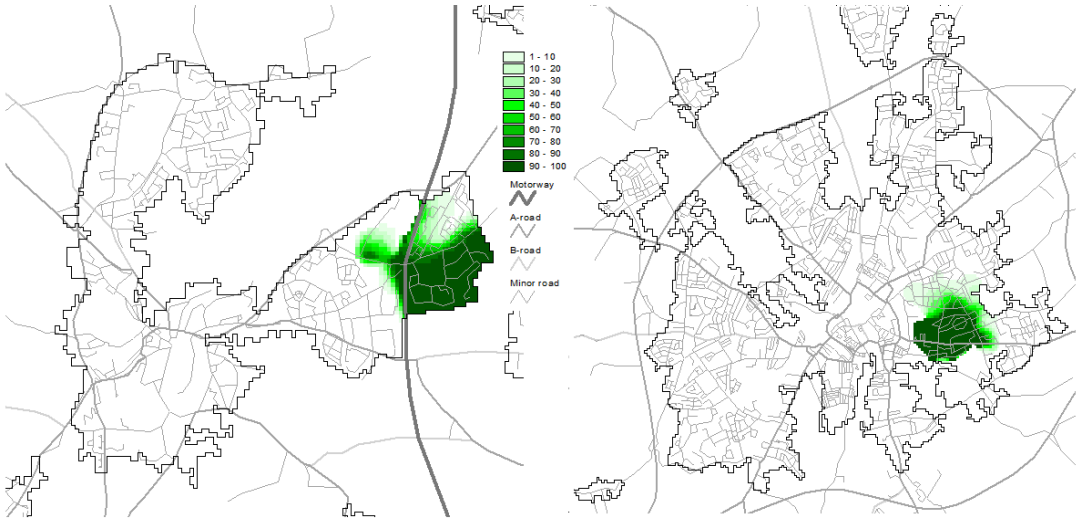
Fig. 6.1 Example data output (percentage of data)



a) Bodicote, Banbury *b) Temple Ewell, Dover*

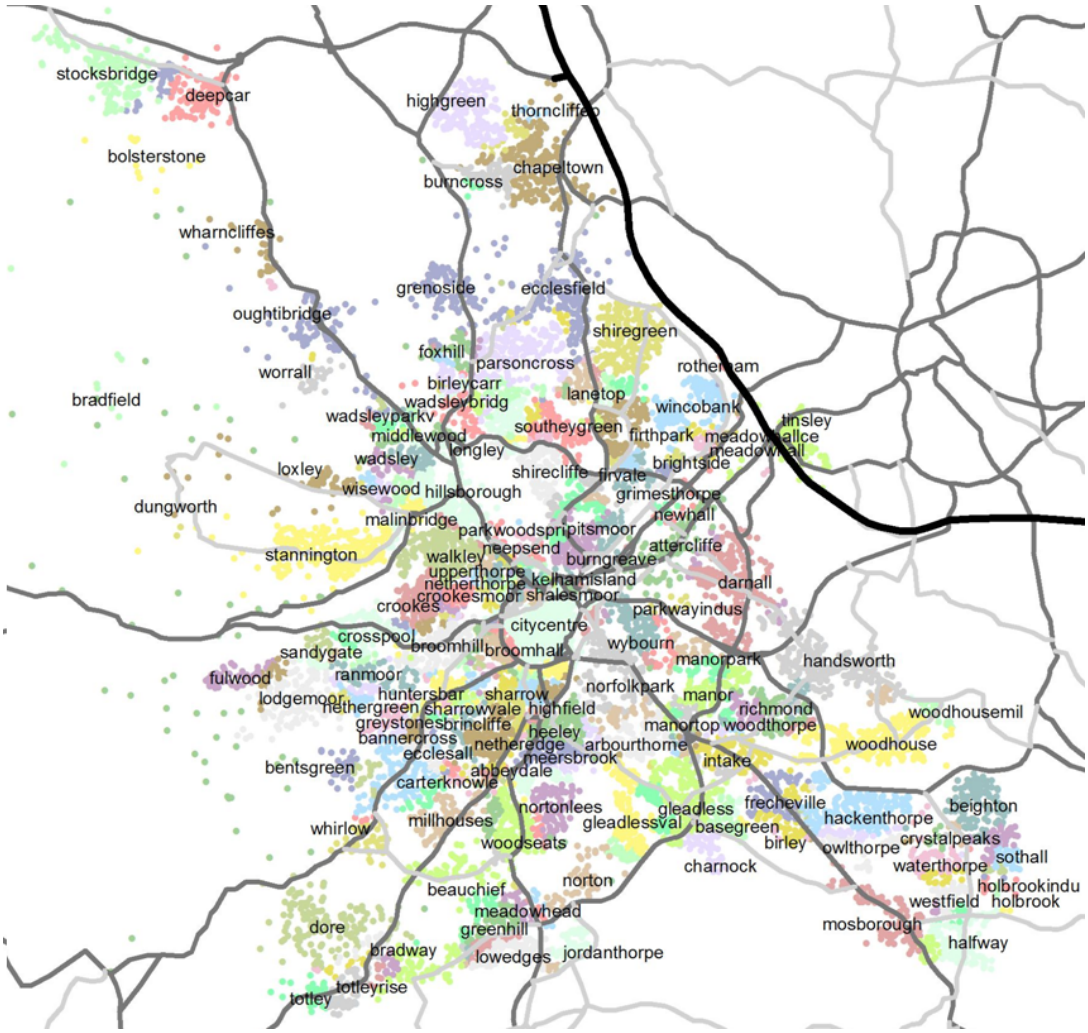


c) Soho, London *d) Strand, London*



e) Belmont, Durham *f) Tang Hall, York*

Fig. 6.2 Example data output: Urban Locales in Sheffield



6.1.2 Place name results: Synonyms and misspellings

Examples of vagueness appearing within place name elements are found in the form of variants in spelling and synonyms. Illustration of the sort of spelling issues that the method identified can be found in Table 6.3. Similarly, examples of the post-process synonym identification using linear grid regression analysis are illustrated within Table 6.4.

6.2 Validation

The validation seeks to explore the extent to which urban locales generated from web based content are representative of those derived from asking people directly. The research hypothesis is "Are vague urban locales generated from web data representative of those obtained through asking people directly?" Ideally validation would compare

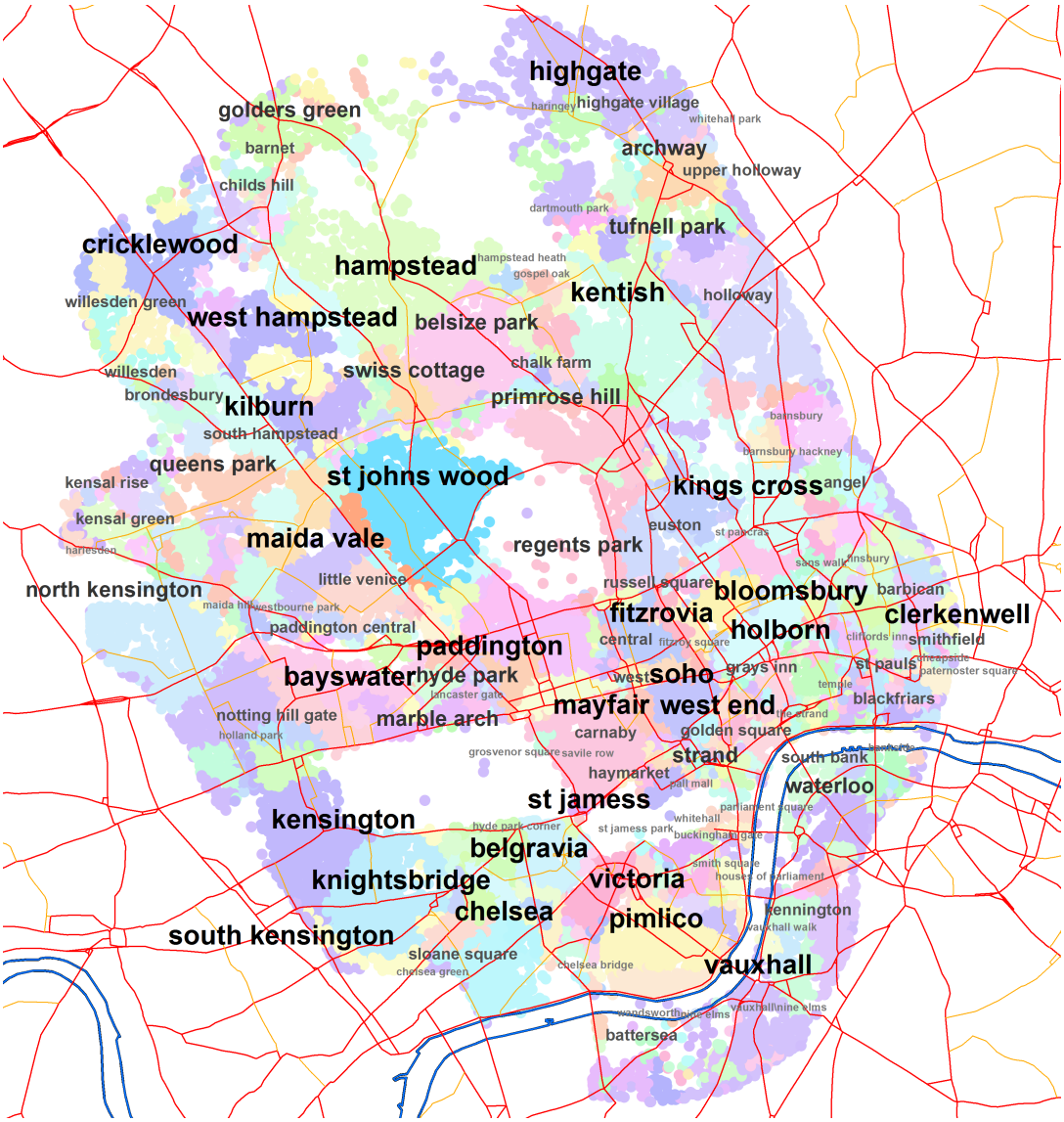
Table 6.3 Examples of output cleaning within the 'spelling' identification process

Class	Spelling 1	Spelling 2	Frequency	Case area	% of the urban locale records
Abbreviation examples:					
	high green	high grn	57	Sheffield	4.14
	west bridgford	w bridgford	52	Nottingham	0.39
	st johns wood	saint johns wood	40	London	1.83
Punctuation examples:					
	st johns wood	st john's wood	106	London	4.86
	regents park	regent's park	101	London	10.12
	st anns	st ann's	33	Nottingham	6.20
Plural form examples:					
	st jamess	st james	478	London	38.58
	highfield	highfields	40	Sheffield	38.46
	parson cross	parsons cross	31	Sheffield	18.13
Whitespace examples:					
	bobbers mill	bobbersmill	93	Nottingham	42.27
	south bank	southbank	63	London	25.30
	west bridgford	westbridgford	40	Nottingham	0.30
	nether edge	netheredge	38	Sheffield	7.35
	fir vale	firvale	34	Sheffield	47.89
Spelling examples:					
	west bridgford	west bridgeford	307	Nottingham	2.28
	mapperley	mapperly	59	Nottingham	1.95
	blenheim industrial estate	bleneim industrial estate	56	Nottingham	32.18
	bulwell	bullwell	56	Nottingham	1.51
	huntington	huntingdon	50	York	2.23
	st anns	st annes	43	Nottingham	8.08
	piccadilly	picadilly	35	London	3.72
	carrville	carville	33	Durham	11.66
	ecclesall	eccleshall	31	Sheffield	5.31
	chapeltown	chapletown	29	Sheffield	1.17
	frecheville	frechville	29	Sheffield	12.24
	darnall	darnell	23	Sheffield	3.89
	south wootton	south wooton	22	Kings Lynn	1.81
	hampstead	hamstead	21	London	0.43
	chilwell	chillwell	21	Nottingham	1.15
	marylebone	marleybone	20	London	0.25
Miscellaneous examples:					
	the lace market	lace market	236	Nottingham	32.96
	lenton lane industrial estate	lenton industrial estate	58	Nottingham	33.53
	clifton moor	clifton moorgate	34	York	3.16
	castle marina	castle marina park	28	Nottingham	28.28
	victoria business park netherfield	victoria park netherfield	16	Nottingham	32.00
	business park nether poppleton	business park poppleton	10	York	6.58

Table 6.4 Examples of output from synonym identification (only those with an R^2 greater than 90% were considered as synonyms)

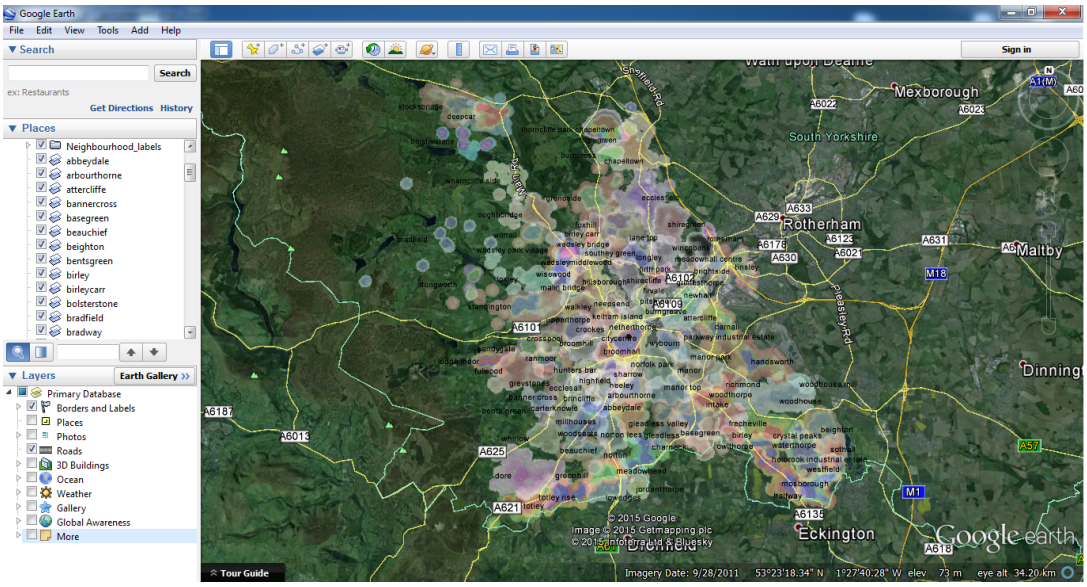
Urban locale 1	Urban locale 2	Case study area	R-squared (%)	Slope	Intercept	p value
pottington business park	pottington industrial estate	Barnstaple	96.9	0.136	0.020	0.000
archway	upper holloway	London	85.2	0.334	0.466	0.000
west central	west end	London	78.7	1.364	2.452	0.000
chelsea bridge	pimlico	London	76.3	11.849	1.177	0.000
bestwood	bestwood estate	Nottingham	76.1	0.411	-0.007	0.000
gorleston	gorleston-on-sea	Great Yarmouth	74.7	0.181	-3.708	0.000
holbrook	holbrook industrial estate	Sheffield	66.6	2.018	-0.044	0.000
roundswell business park	roundswell commercial park	Barnstaple	62.0	0.143	-0.086	0.000

Fig. 6.3 Example dataoutput: Urban Locales in and around Camden and Westminster, London

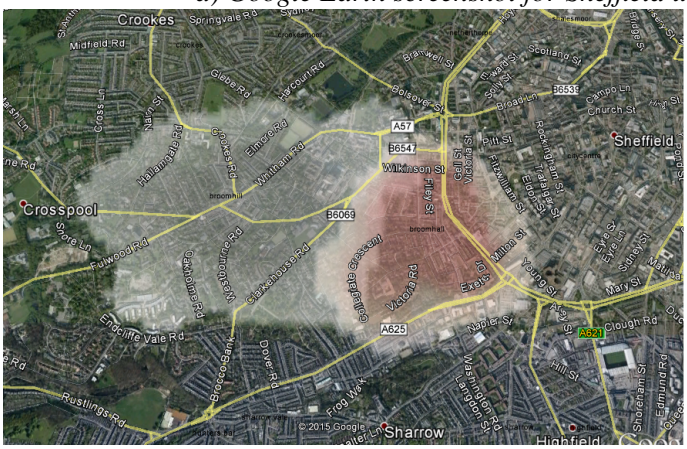


output against a complete coverage of resident views. Collecting such data, however, for any extensive geographic area would be pragmatically impossible. Furthermore, any *sample* of resident opinions in itself may not accurately reflect the true underlying representation of urban locales. Whilst other digital sources of urban locale data exist at scale, as stated in Chapter 4 such representations rarely include geographic extents and most gazetteers, such as Geonames, reference urban locales as single 'centre' points. This is due to the difficulties in data collection and the subjective nature of urban locale geographies but leads to the issue that there are no true "ground truth" representations to compare against at scale. As such, it is not possible to undertake a traditional approach whereby the urban locales extracted from web sources are compared against some definitive benchmark datasource, using terminology such as false positives (incorrectly included within the new data) and false negatives (in the

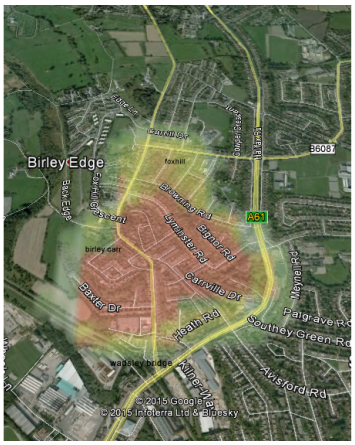
Fig. 6.4 Example of urban locales exported to KMZ for easy distribution to non-GIS users



a) Google Earth screenshot for Sheffield urban locales



b) Inset - Broomhill & Broomhall urban locales, Sheffield



c) Inset - Birley Carr, Sheffield

definitive data but missed in the new datasource). Thus, traditional validation of the urban locales is not just difficult but impossible to test over a sustained geographic area with sufficient sample numbers. At best, one might consider the various different datasources to offer *alternative perspectives* rather than a definitive viewpoint. Under such a condition, validation is best considered along the lines of set theory (such as in Venn diagrams) where *agreement* is considered as the intersection between data and false positives / negatives as relative complement - the differences between the various datasets. This is not to say that error will not be contained within the data, just that disagreement between various sources is not necessarily a result of *error*.

The situation is further exacerbated by the fact that the efficacy of any validation can only be accurately assessed in reference to some specific end application. In aiming to derive urban locale extents that have general utility there must be an attempt to compare

outputs with some form(s) of alternative representation. For this reason, the method undertaken makes a distinction between (i) validation against existing gazetteers based solely on urban locale *name* and (ii) validation of the urban locale *geographic extents*. The spatial boundary information can be tested against both a *sample* of resident views and the limited number of alternative digital datasources that include locale extents. Validation will centre upon output for the cities of Nottingham and Sheffield where the author has a good local knowledge of the urban locales. Such qualitative information can prove insightful when drawing distinctions between differing perspectives of the various datasources and potential 'errors'.

6.2.1 Comparison of urban locale names

It is possible (and relatively simple) to compare the urban locale *names* against a composite picture derived from the following multiple gazetteer sources:

- OSM[1];
- OS data[2];
- Yahoo! Geoplanet[3];
- Geonames[4]; and
- local council derived urban locale names[5].

The ease of comparison is due to the fact that comparisons are solely based on the presence and absence of urban locale names within the different sources. Urban locales that are cited by a number of different sources are more likely to be the more significant named areas for a particular town or city. As such, the combination of different source agreement between locale names potentially becomes an important measure. Similarly, those urban locales which are only referenced in one of the plethora of gazetteer sources are more likely to include potential data errors. Given that the sources are known to be incomplete (Brindley et al., 2014b) and derived in a relatively ad-hoc fashion, the degree to which the web extraction method can generate a superset of their contents is also of interest. Such an approach was implemented in the cities of

[1] using the suburb place type features that fall within the urban boundary extent

[2] compromising of a combination of the 50k gazetteer, VectorMap Local and data extracted from the internal company data of the Multi Resolution Data Programme (MRDP)

[3] using the suburb place type features that fall within the urban boundary extent

[4] using populated places (coded PPLX or PPL) features within the urban boundary extent

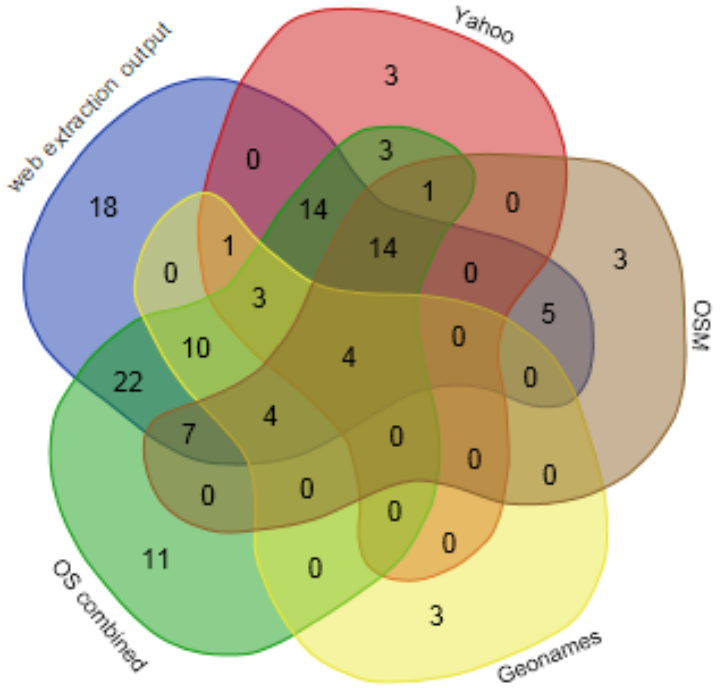
[5] delineated by Sheffield City Council for policy analysis (Thomas et al., 2009) and thus only available for the city of Sheffield case study

Nottingham and Sheffield where the author has a significant qualitative knowledge of local urban locales. The number of urban locales each source contained can be found within Table 6.5. Tabular output of the differing combinations of source can be found in Appendix E, whilst a visual interpretation[6] is illustrated in Figures 6.5 and 6.6[7] and demonstrate the overall poor agreement between the various datasources.

Table 6.5 Number of Urban Locales identified by different sources

Datasource	Number of urban locales: Sheffield	number of urban locales: Nottingham
Council defined	103	n/a
Geonames	50	25
OS combined	235	93
OSM	123	38
Yahoo! Geoplanet	42	43
Web extraction output	121	102
Unique elements	**283**	**126**

Fig. 6.5 Venn diagram of Nottingham Urban Locales by datasource



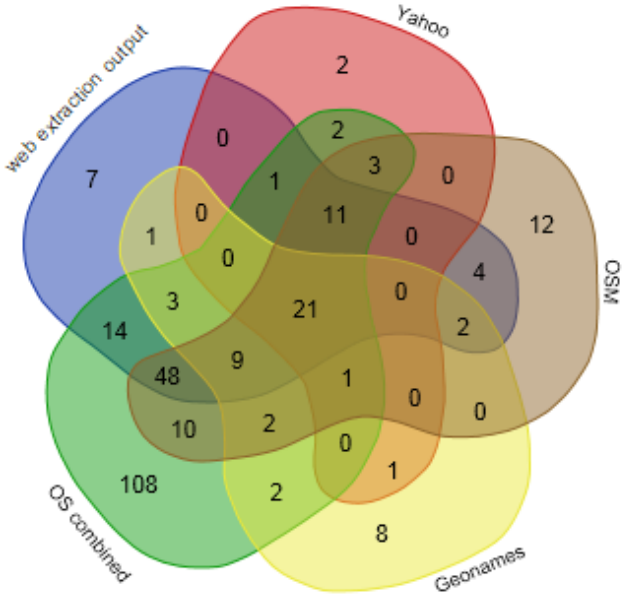
6.2.1.1 Gazetteer disparity

In the case of Nottingham - only 3.2% of the combined total of 126 urban locale names were contained within all the different sources, 46% were in three or more sources and only 69.8% of urban locales were identified by at least two of the five sources. In

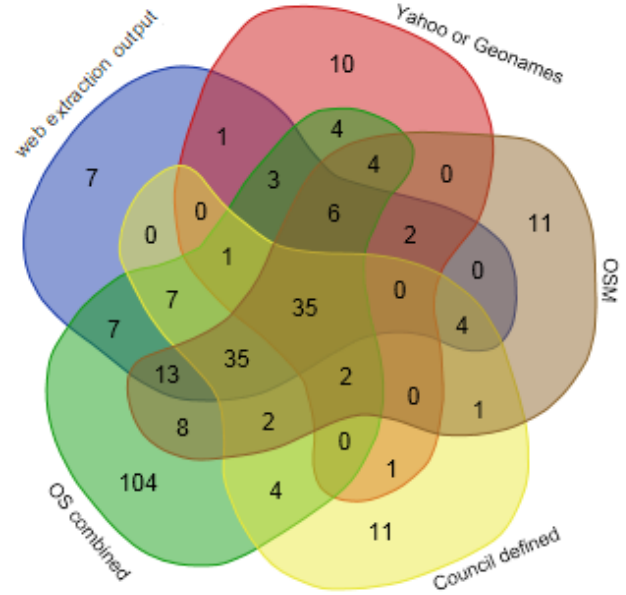
[6] constructed using http://bioinformatics.psb.ugent.be/webtools/Venn

[7] due to the difficulties in visualising 6-way Venn-diagrams, two separate figures are produced for the Sheffield output

Fig. 6.6 Venn diagram of Sheffield Urban Locales by datasource



a) Comparison between web extraction data, Yahoo! Geoplanet, OSM, Geonames, and OS data



b) Comparison between web extraction data, Yahoo! Geoplanet or Geonames, OSM, OS data and local council defined data

comparison, within Sheffield - only 6.0% of the combined 283 urban locale names were contained within all six different sources[8], 40.3% were in three or more sources and only 49.8% of urban locales were identified by at least two of the six sources. This means that in Nottingham and Sheffield, 30.2 and 50.2% of urban locales respectively,

[8]Local council defined urban locale names were only available for Sheffield - representing the extra source of data

were found only in a single source. Such figures demonstrate the low level of general agreement between the different datasources for urban locales.

In the city of Nottingham nearly half of the urban locales identified in only a single source were found only in the web extraction output (n=18) with close to a further third contained only within OS data (n=11). A significant proportion of names found only within web extraction output were industrial estates and business parks (n=6). The remainder represent relatively well known areas such as Canning Circus and Hockley along with common smaller areas such as Clifton Estate and Wilford Village and contain no surprises or questionable output, demonstrating the incompleteness of many existing urban locale gazetteers. These 'omissions' within current gazetteers are discussed in greater detail in the next section and have important implications for the INSPIRE Directive (as discussed in Chapter 4). In contrast, there were over 3.7 times more locales identified by only a single source in Sheffield than compared to Nottingham. The vast majority (73%) were found solely in OS data (n=104). Whilst some represent relatively well-known urban locales not included in the other non-OS data (such as Owlerton and Templeborough), a substantial proportion appear to relate to very small geographical places that nest within larger entities. It is highly unlikely that if residents were asked which part of the city or town they lived in that they would provide such names as an answer and would probably cite the larger urban locales that these minutiae nest within. Examples of such cases (in Sheffield) include Den Bank, Wadsley Grove, Carsick, Stephen Hill and Howard Hill. Whilst these do not represent data errors, they are not necessarily of a suitable scale to (1) be well known to other residents of the town or city to have general applicability and (2) might not be robust enough (due to their size) to accurately map and aggregate other socio-economic data to. A number of these smaller areas also appear to relate to historic names that people might not necessarily use in reference to their current urban locales. Examples in the west of Sheffield include Steel Bank and Summerfield.

6.2.1.2 Comparing web extraction output

Disparities apart, the question remains whether it is possible to determine how well the web extraction output performs compared to the other urban locale datasources. It could be argued that those urban locales with greater levels of agreement between the various sources are potentially more likely to be the larger, better known and most critical ones. Given the poor level of comparability between existing gazetteers as previously discussed, perfect agreement is unlikely. This said, the degree of overlap is particularly encouraging. Table 6.6 illustrates that the web extraction output accounted for 84 of the 88 total number of locales identified in Nottingham (95.5%) when agreement was limited to only locales which were found in at least two of the five

different gazetteer sources. In comparison, of the 121 urban locales defined by the web extraction output in Sheffield – 106 (93.0%) were also found in at least three of the six existing sources of urban locales. Table 6.6 also contains the level of comparability between web extraction output and all urban locales identified with a varying degree of agreement between the different datasources.

Table 6.6 Comparability of web extraction output with other gazetteers with varying level of agreement between the datasources

No. of sources of agreement:	1	2+	3+	4+	5+	6
Nottingham:						
Urban locales identified by web extraction output	102	84	57	25	4	n/a
Total number of Urban locales identified	126	88	58	25	4	n/a
% of locales identified by web extraction output	81.0	95.5	98.3	100.0	100.0	n/a
Sheffield:						
Urban locales identified by web extraction output	121	114	106	77	39	17
Total number of Urban locales identified	283	141	114	80	39	17
% of locales identified by web extraction output	42.8	80.9	93.0	96.3	100.0	100.0

A complete list of the 25 urban locales identified by the web extraction output (in Nottingham or Sheffield) but not found in the other gazetteer sources can be found in Table 6.7. Other evidence of the urban locale's existence was gathered for the majority of cases: from within a different class within OSM (for example as an 'Industrial Area' rather than specifically as a 'suburb'); as ancillary features in OSM (occurrences of near-by objects, including for example, bus stops, village greens, named building and so on, that share the urban locale name); or from references in historical OS maps. This provided further evidence for all but seven of the urban locales - shown in red in Table 6.7 (five in Nottingham and two in Sheffield). There is little doubt that the **City centre** actually exists and is not a data error issue. The problem is that most urban locale gazetteers do not include such a feature. **Clifton Estate** was at one time the largest council estate in Europe[9] - again there can be no denying its existence and its presence on the list is due to the fact that many may refer to the area generically as Clifton rather than making a distinction between Clifton and Clifton Estate. There is a substantial body of online material relating to **Heronridge** (also called Heron Ridge) compromising of house details on estate agent websites, local businesses carrying the name, along with mentions on Wikipedia[10] and council documentation[11]. Similarly, there are a number of references to '**Victoria Business Park, Netherfield**' on the web, including references in local and national government documentation[12] and local press

[9] http://en.wikipedia.org/wiki/Clifton,_Nottingham

[10] http://en.wikipedia.org/wiki/Bulwell

[11] http://open.nottinghamcity.gov.uk/comm/download4.asp?dltype=inline&filename=Response Upload130731.doc

[12] https://democracy.gedling.gov.uk/documents/s1693/Application%20No.%2020130614.pdf and https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/343224/ad-6154-13-august-2014.pdf

material[13]. Although no references to '**Warren Hill**' can be found in historic OS maps - there is text annotation citing The Warren and Warren House. There are also a large number of websites listing Warren Hill, including a residents' association for Warren Hill which is recognised by the local council[14].

In Sheffield, there are no references to '**Parkway Industrial Estate**' within old OS maps, although there are references to an 'industrial estate' along with descriptions of 'Parkway Market' from the 1950s and 'Parkway Works' from the 1990s. Whilst **Wadsley Park Village** (built on the former site of Middlewood Hospital between 2001 and 2006) does not appear in any of the existing neighbourhood gazetteers, the area does have its own community website and forum (http://www.wpvonline.co.uk). There appears supporting evidence for all the urban localities that were identified by the web extraction output but not found within other gazetteer sources - further demonstrating the incompleteness of existing sources of urban locales within current gazetteers and supporting the need for additional methods of identification.

Table 6.8 contains the 12 urban locales in Nottingham or Sheffield that were found in a number of existing sources of urban locales[15] but that were not in the final output of the web extraction data. Whilst references to all of these locales were found within the raw data (as expressed by the frequency column within Table 6.8) they contain very low levels of data returns and were, therefore, not considered robust enough to be included in further analysis. The rationale being that any map produced using such small numbers may not be robust or particularly meaningful. It should be stressed that this represents a very small proportion of the total number of urban locales that the method identified.

Table 6.9 applies the same principle as within Table 6.6 for all the different datasources and demonstrates that the two most comprehensive sources are the combined OS and web extraction output. Given that the OS data are actually a combination of three different data, if a single gazetteer were to be used - the web extraction output would be the best choice as it appears to consistently contain the highest number of urban locales in agreement with the other sources.

In summary, the urban locales within web extraction output compare favourably with other gazetteer sources and whilst differences exist, the vast majority of data identified by the web extraction output was generally supported by the other existing gazetteers. The majority of urban locales that were found in the other gazetteers but not through

[13]http://www.nottinghampost.com/D-day-pound-96m-plan-build-830-new-homes/story-20395558-detail/story.html

[14]http://warrenactiongroup.org and https://democracy.gedling.gov.uk/mgOutsideBodyDetails.aspx?ID=249

[15]Due to the differing number of datasources for the two case areas, agreement of three of the six sources was required in Sheffield but two of the five within Nottingham

Table 6.7 Urban locales identified by web extraction output but that were not included in other gazetteers

Urban Locale Name	Frequency in web extraction output	Found in OSM	Found in Historic OS maps
Nottingham:			
Blenheim Industrial Estate	182	Yes - different class	No
Boulevard Industrial Park Beeston	49	Yes - different class	No
Canning Circus	122	Yes - ancillary features	Yes
Castle Boulevard	51	No	Yes
Chetwynd Business Park Chilwell	86	Yes - different class	No
City Centre	274	No	No
Clifton Estate	135	No	No
Clifton Village	91	Yes - ancillary features	No
Colwick Park	58	Yes - ancillary features	Yes
Heronridge	50	No	No
Hockley	190	Yes - ancillary features	Yes
Lenton Lane Industrial Estate	173	Yes - different class	No
Old Market Square	92	Yes - ancillary features	Yes
Papplewick	135	Yes - ancillary features	Yes
Queens Drive Industrial Estate	225	Yes - different class	No
Victoria Business Park Netherfield	50	No	No
Warren Hill	42	No	No
Wilford Village	50	Yes - ancillary features	No
Sheffield:			
Holbrook Industrial Estate	347	No	Yes
Hunters Bar	323	Yes - different class	Yes
Meadowhall Centre	228	Yes - different class	Yes
Newhall	75	No	Yes
Parkway Industrial Estate	61	No	No
Thorncliffe Park, Chapeltown	44	Yes - different class	No
Wadsley Park Village	61	No	No

Table 6.8 Urban locales not identified by web extraction output

Urban Locale Name	Frequency	Urban Locale Name	Frequency
Nottingham:		Sheffield:	
Bulwell Forest	9	Batemoor	21
Highbury Vale	30	Carbrook	36
Porchester	12	Greenland	20
Standard Hill	1	Hemsworth	16
		Herdings	29
		Hollins End	10
		Lowfield	11
		Park Hill	25

Table 6.9 Comparability of urban locale output for the different gazetteers with varying level of agreement between the datasources

Sources in agreement:	1	2+	3+	4+	5+	6+	average (excluding 1 source only)
Nottingham:							
web extraction output	81.0	95.5	98.3	100.0	100.0	n/a	98.4
OS combined	73.8	93.2	98.3	100.0	100.0	n/a	97.9
Yahoo! Geoplanet	34.1	45.5	63.8	84.0	100.0	n/a	73.3
OSM	30.2	39.8	51.7	88.0	100.0	n/a	69.9
Geonames	19.8	25.0	37.9	44.0	100.0	n/a	51.7
Sheffield:							
OS combined	83.0	92.9	94.7	100.0	100.0	100.0	97.5
web extraction output	42.8	80.9	93.0	96.3	100.0	100.0	94.0
OSM	43.5	79.4	90.4	98.8	100.0	100.0	93.7
Council defined	36.4	65.2	75.4	91.3	89.7	100.0	84.3
Yahoo! Geoplanet	14.8	28.4	32.5	42.5	79.5	100.0	56.6
Geonames	17.7	30.5	33.3	41.3	74.4	100.0	55.9

the web extraction output were cases of very small areas where the scale and number of data returns may not be suitable for detailed mapping or to aggregate other socio-economic data to. Where additional urban locales were identified by the approach, the names appear to represent areas 'missed' by the existing gazetteers - demonstrating the positive practical implications of the proposed method. It is important to stress that whilst the source of the web extraction output is from postal address information harvested from the web, the high degree of similarity of names with the other urban locale datasources demonstrates that the data are representative of urban locales as a whole and are not merely a specific 'kind' or sub-set of urban locales.

6.2.2 Comparison of urban locale geography

Validation of the *geography* of the urban locales is far more problematic due the subjective nature and general lack of data for comparisons. For these reasons, three separate procedures were undertaken. Firstly, the web extraction output could be compared with other urban locale geographies for complete settlements. Such an approach is extremely insightful but limited by the sparse available data. Secondly, opinions of the urban locale boundaries from the web extraction output could be sought from selected stakeholders. Whilst such qualitative information is important, it would only be limited to a small number of specific locales as such stakeholders who would rarely have detailed knowledge outside their main activity space. Furthermore, for each stakeholder, such analysis would be the views of a single person and may not be representative of the population more generally. Finally, a sample of people could be asked which urban locale they were in / lived in and the output compared to the

different extents provided by the various urban locale datasources. Each approach was undertaken and will be discussed in greater detail.

6.2.2.1 Comparisons with areal data

As specified within Chapter 4, there are currently few sources available that contain boundary details of urban locales at a national scale:

1 Wikimapia, an online editable map generated through crowdsourcing[16],

2 Quattroshapes[17] constructed using existing gazetteers and matching of textual descriptions within Flickr photos, and

3 the OS Multi Resolution Data Programme (MRDP), data held internally within OS and not publicly available.

Fortunately, however, within the city of Sheffield it was possible to source a unique datasource of 100 local urban locales constructed by Sheffield City Council[18]. It was thus possible to compare these different urban locale geographies with each other in order to identify the level of agreement. Given the levels of divergence between the contents of urban locale *names* within the different sources of gazetteer, as discussed in the previous section, comparison of extents were only made in those 19 cases where the urban locale was included within all five sources containing boundary information (Wikimapia, Quattroshapes, OS MRDP, Sheffield Council defined areas and the web extraction output).

Figure 6.7 illustrates an example of boundary comparison for the urban locale of Highfield, Sheffield. Whilst Figure 6.7a-f demonstrate the geographies for the differing sources, Figure 6.7g shows the level of agreement between the sources. On face value, comparisons appear good with the exception of the Quattroshapes data where the extent appears far too large compared to the other sources. Only Quattroshapes and the Council defined area extend north beyond the ring road (A61, St. Mary's Gate / Road).

When such data are aggregated for all 19 comparator urban locales, the level of agreement between the sources can be assessed. Before doing so, however, it is

[16] Although it should be noted that as stated in Chapter 4 - Wikimapia is not intended to specifically map urban locales, although it does include many such features. Coverage, however, is not complete and nearly 50% of the Sheffield Urban Area has no urban locale information

[17] http://quattroshapes.com

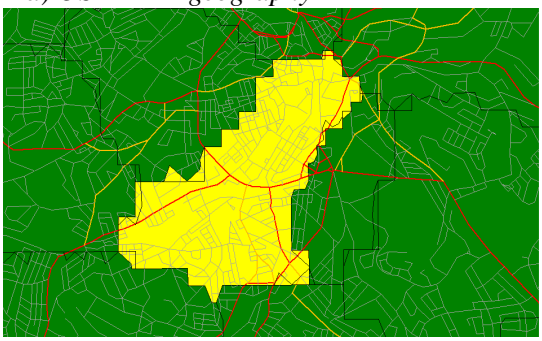
[18] These were derived through an extensive process of manually grouping administrative data (Census Output Areas) using local knowledge and lengthy consultation, see Thomas et al. (2009) for further details

Fig. 6.7 Comparing the geography of Highfield urban locale, Sheffield

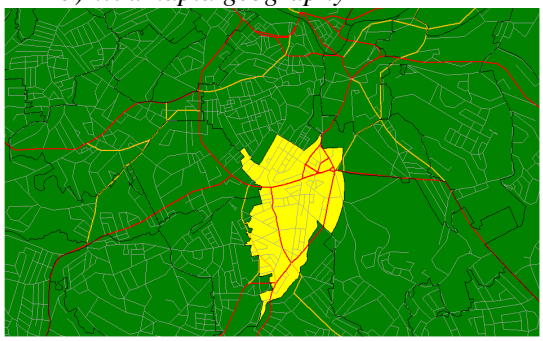


a) OS MRDP geography

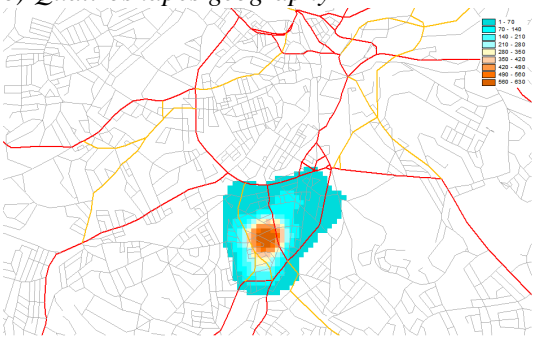
b) Wikimapia geography



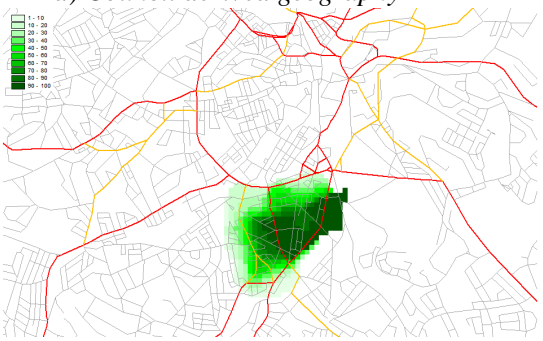
c) Quattroshapes geography



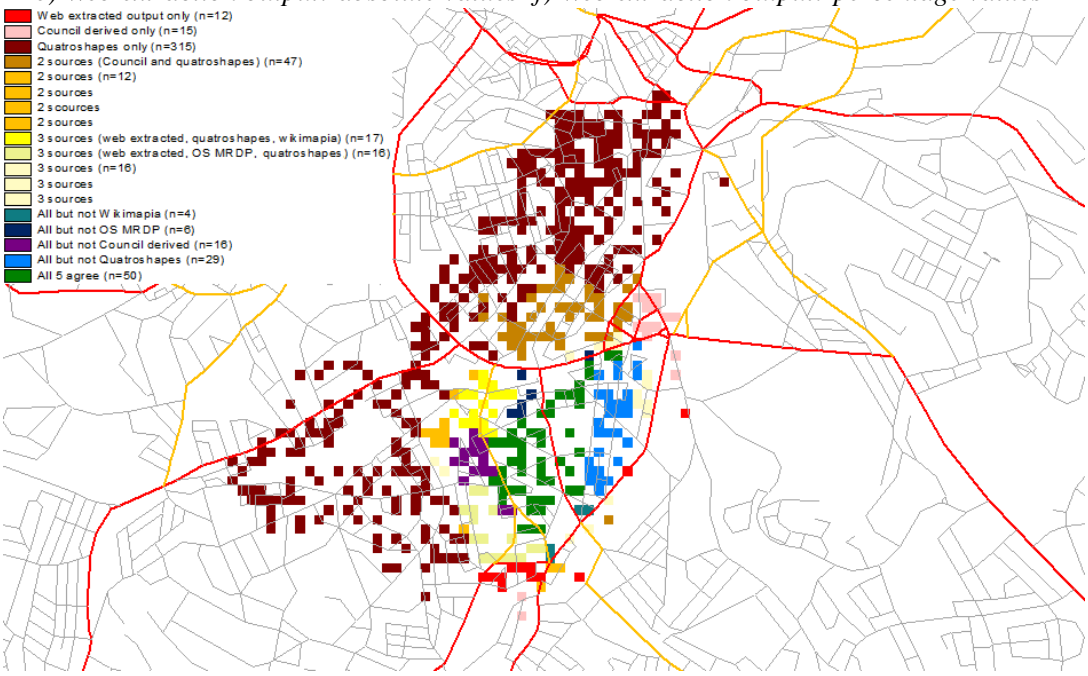
d) Council derived geography



e) Web extraction output: absolute values



f) Web extraction output: percentage values



g) Combined geographies

important to reflect on the nature of units in extreme cases in order to fully understand the underlying process. If a datasource contained locales with extents greater than expected, by default - it would be likely to agree with the other sources but also have a large number of cells without agreement. In contrast, a datasource that had particularly small urban locale extents, would have a low number of cells not in agreement but also a small number of cells in agreement with the other sources. Thus, it is the combination of both the numbers of cells in agreement and not in agreement that is key and picking just one of the measures alone would bias results. Three measures are important:

1 a high number of cells in agreement between the differing sources (similar to 'True positives' if the combination of sources actually represented a ground truth)

2 a low number of cells identified by the method but not supported by evidence in any of the other sources (similar to 'False positives' if the combination of sources actually represented a ground truth)

3 a low number of cells not identified by the method but where evidence from the other sources supported existence (similar to 'False negatives' if the combination of sources actually represented a ground truth)

The level of cells in agreement can be found in Table 6.10 and shows that OS MRDP had the highest agreement in cells supported by three or more sources. The web extraction output did not perform particularly well. This may partly be an effect of the method allowing for overlapping areas which potentially generates larger areal extents than compared against other methods. Such analysis, however, privileges those areas with a small areal extent (and thus likely to have a high proportion of agreement). It should also be remembered, that the level of agreement might not be a good indication of ground truth as all those sources in agreement may be incorrect. The average number of (non-zero) cells for each source, for the 19 case areas, are as follows:

- Quattroshapes: 230 cells
- web extraction output: 154 cells
- Wikimapia: 107 cells
- Council defined areas: 101 cells
- OS MRDP: 82 cells

As demonstrated within Table 6.11, Quattroshapes has a high number of cells where support for the urban locale geography was not evident within the other datasets. The

Table 6.10 Average number of cells identified by the source for the 19 comparator urban locales with agreement with geographies from other data sources

	5 sources		4+ sources		3+ sources		2+ sources		Average rank
	%	rank	%	rank	%	rank	%	rank	3+ sources
web extraction method	35.9	4	62.0	4	78.5	4	93.9	4	4.0
Council defined	50.3	2	75.3	2	87.8	3	97.2	3	2.5
OS MRDP	61.5	1	83.7	1	91.3	2	98.7	1	1.0
Quattroshapes	22.1	5	32.1	5	40.8	5	53.4	5	5.0
Wikimapia	47.6	3	73.9	3	92.1	1	98.3	2	2.5

web extraction method had the second highest number of cells that were unsupported by the other methods. This may again be because the method oversmooths data or it may simply be a reflection on the process, which unlike the other methods with their discrete boundaries, allows overlapping multiple membership. The high degree of inclusion of the web extraction output as illustrated within Table 6.12 is clear, and at all three levels of agreement (2 of the 4 other sources; 3 of the 4; and all 4 datasets) the method had the lowest number of cells not included (or 'missed') when evidence from the other sources suggested presence of the urban locale(s). The OS MRDP data is ranked last in this measure, perhaps due to the often small geographic coverages of its areas. Such a process, however, will privilege areas of extensive extent - and thus more likely to have fewer cells not included).

Table 6.11 Number of cells not in agreement with any of the other four sources

	Number of cells	Percentage of datasource	Rank
web extraction output	178	7.7	4
Council defined	53	2.3	3
OS MRDP	20	0.9	1
Quattroshapes	2,035	87.7	5
Wikimapia	34	1.5	2

Table 6.12 Percentage of cells 'missed' by the method where evidence from the other four sources would suggest the locale's presence

	2 or more (of the other 4) sources in agreement	3 or more (of the other 4) sources in agreement	all 4 of the other sources in agreement	Average rank of the 3 calculations
web extraction output	1.28	0.16	0.01	1.0
Council defined	10.33	4.18	1.20	4.0
OS MRDP	16.47	8.18	3.04	5.0
Quattroshapes	2.49	1.40	0.70	2.3
Wikimapia	8.56	2.25	0.62	2.7

Combining the ranks (see Table 6.13) from the measures is not straightforward as two of the three privileges small areal extents, whilst only one benefits from larger extents. If this fact is ignored and no weighting applied, the OS MRPD has the highest level of agreement, with the web extraction method third. If weighting is distributed proportionally to measures that are likely to be biased to large and small

areal extents[19], the OS MRPD fairs less well and Wikimapia is the source with the greatest agreement in the geography of the 19 case areas. The web extraction method is ranked second, with OS MRDP third. Whichever weighting method applied, the web extraction output fairs favourable against the Council defined geographies in terms of the level of agreement between available sources. It should also be noted that whilst Wikimapia had the highest levels of agreement with the other data, as a method it is reliant on willing and knowledgeable contributors and geographical coverage is frequently incomplete.

Table 6.13 Combining ranks from Table 6.10, Table 6.11 and Table 6.12

	Table 6.10 rank	Table 6.11 rank	Table 6.12 rank	Combined Rank (equal weighting)	Combined Rank (adjusted weighting)
web extraction output	4.0	4.0	1.0	3.00	2.50
Council defined	2.5	3.0	4.0	3.17	3.38
OS MRDP	1.0	1.0	5.0	2.33	3.00
Quattroshapes	5.0	5.0	2.3	4.11	3.67
Wikimapia	2.5	2.0	2.7	2.39	2.46

In the absence of ground truth, the level of agreement with other datasources is an interesting method for investigating differences between the different data. It should be noted, however, that the level of agreement with other datasources might not necessarily be a good indication of greater accuracy. Whilst probability might inform us that a greater likelihood is present if two sources are in agreement than one source alone, the outcome is not definitive. Furthermore, there is nothing to say that in such cases both sources might be wrong and the single source (without agreement) might better reflect the unknown true representation. Such are the difficulties when evaluating data where there is no explicit ground truth.

6.2.2.2 Stakeholder views of urban locale output

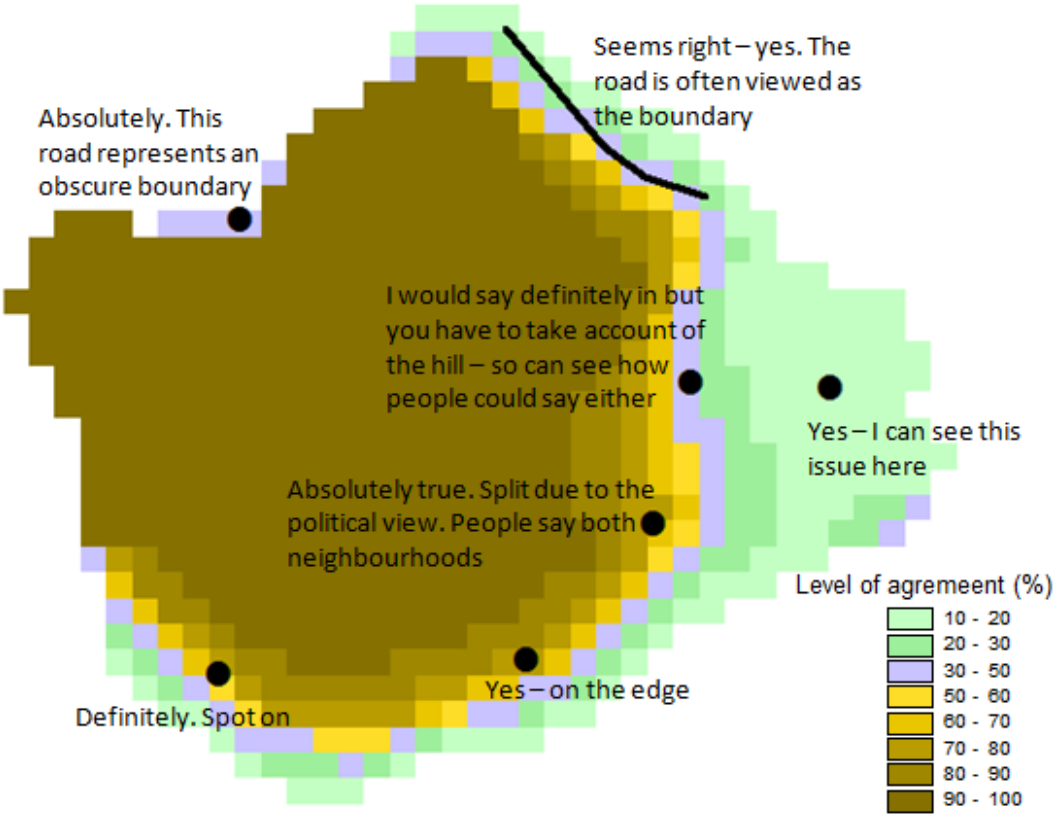
Whilst it is entirely plausible to seek opinions of the boundaries formed by the web extraction output from selected stakeholders (such as local council officers or members of local community groups), the detailed knowledge required for such comparisons would only exist for a small number of urban locales for each stakeholder. Such validation would place a significant weight on the interpretation of a single perspective which might not necessarily be representative of the population as a whole. Given such limitations, only a small number of stakeholder views were sought - principally those of the committee of a local community group within the city of Sheffield and three local planning officers of a local council.

[19](Table 6.10 rank+ Table 6.11 rank)/2 versus Table 6.12 rank

When extents of a number of locales were presented to three local planning officers of a local council, the vague boundaries of locales shown were not disputed - demonstrating either a high level of agreement or lack of local knowledge to be able to adequately assess the boundaries.

The boundaries of an urban locale were also discussed with the chair of a local community group who had lived in the area for nearly 20 years. The feedback from the chair was positive, as shown within the comments in Figure 6.8 where named roads were discussed and opinions sought as to whether the chair felt that the road would be considered inside or outside of the named area.

Fig. 6.8 Exploring the edge of an urban locale with a chair of a local community group



It is pertinent and interesting that when asked (as part of the separate resident validation in Sheffield discussed below), most residents claimed to be unable to map and define the extent of their urban locale with any accuracy. Those that did claim able to do so were acutely aware that it may not reflect the opinion of others. This demonstrates the weakness in method of asking stakeholders' opinions of areal boundaries for urban locales. Such views represent merely a single view which might not be more generally applicable and upheld by others.

6.2.2.3 Comparison with samples of residents' views

Perhaps the most illuminating validation can be achieved through asking residents their view on what the area in which they live is called. This approach is the most suitable within the validation process for considering the extent to which urban locales generated from web based content might be representative of those derived from asking people directly. Data were collected for three different scenarios, each with increasing spatial resolution:

- resident views from an on-line questionnaire for places within Nottingham;
- information collected from residents by a Neighbourhood Forum; and
- resident views in targeted streets where web extraction output portrayed evidence of differing views of urban locale name within the same local space.

Each of these different forms of resident validation will now be discussed in turn.

- **Residents' views in Nottingham:**
  This validation explores resident responses for urban locales in general across the city of Nottingham. Sixty-eight resident views of urban locale were collected through asking people who responded to the survey used within the validation process for the settlement classification in Chapter 3. Residents were simply asked "the name of the area" in which they lived and to provide their postcode. Thus, the point location of a resident's address can be compared to the different underlying geographies of urban locale that the various sources (Wikimapia, Quattroshapes, OS MRDP, web extraction output) produce. Additionally, resident views were also compared with the Voronoi Diagrams constructed from the other existing urban locale centre-point gazetteers (Geonames, OSM and Yahoo! Geoplanet). Such a procedure conceptually compares the resident's reported urban locale and the closest centre-point from each of the gazetteers. A slight difficulty is encountered due the vague nature of classification within the web extraction output: for example a resident calling the area Bulwell matches even if the web extraction output declares that 99% of data relate to the area as Arnold and 1% as Bulwell. In order to demonstrate the level of likely consensus, cases were identified as a 'minor' match if less than 50% of data within the web extraction output were found to agree with the resident perspective. This represented 10% (n=7) of resident surveys and details of which can be found in Table 6.14. A record was also attributed as a 'minor' match if the two sources were only different due to a qualifier such as Lenton and *New* Lenton.

  A total of sixty-eight resident views were obtained for a total of sixty-six unique postcodes covering thirty-four different urban locales. The output of the resident

Table 6.14 Cases where resident view matched with web extraction data but not where data were in majority agreement

Resident stated area:	Corresponding web extraction output:
Chilwell	53% Beeston, *37% Chilwell*, 9% Chilwell-Beeston
Radford	67% Hyson Green, *28% Radford*, 3% Forest Fields
Basford	75% Bulwell, *21% Basford*, 3% Old Basford
Canning Circus	70% Radford, *19% Canning Circus*, 11% Arboretum
City centre	89% The Park, *11% City Centre*
West Bridgford	90% Gamston, *10% West Bridgford*
Beechdale	93% Wollaton, *7% Beechdale*

view comparisons can be found within Figure 6.10 and demonstrate the high level of agreement found between residents and the web extraction output, with over an 98% agreement (comprised of 88% where data specified a supporting majority view with the resident and a further 10% agreement but as a minority representation). There was no agreement in just one of the sixty-eight cases - where the resident postcode and specified urban locale was actually over 3km away from the corresponding locale identified by the web extraction output. Given this large distance and the fact that similar disparities were also found between this particular resident view and the other sources, the postcode supplied by the resident in this case is likely to be incorrect[20]. The web extraction output outperformed all of the other sources in terms of agreement with resident views. Both the Wikimapia and OS MRDP in comparison only agreed in approximately 70% of cases, whilst the remaining sources fared worse with agreement in only 30-40% of residents surveyed.

The Voronoi Diagram output are built from sets of discrete points that result in discrete boundaries. In contrast the web extraction output are represented by a vague representation (300m KDE smoothings). If a similar KDE measure was applied to the centre-point data of the gazetteers used within Voronoi Diagram output, the level of matching (between the Voronoi Diagram output and 68 resident views) would actually worsen. This is primarily due to the small geographic scale of areas within 300m of the gazetteer centre-points. Only 8 of the 68 resident views were found within 300m of OSM gazetteer centre-points (and of the 8 resident views, only 75% were correctly predicted by the OSM gazetteer). Yahoo! Geoplanet and Geonames performed similarly, with again only 8 resident views found within 300m of Yahoo! Geoplanet centre-points and only 2 resident views found within 300m of Geonames centre-points (for both sources only 50% of resident views were correctly predicted by the gazetteer). Thus, the poor agreement between resident views and Voronoi Diagram output from gazetteer centre-points, compared to the levels of agreement from the web extraction output, do not necessarily appear to be due

[20] a correction of one digit from a 4 to a 9 could potentially explain the effect.

Fig. 6.9 Approximate location of resident perspectives of home urban locale name within Nottingham

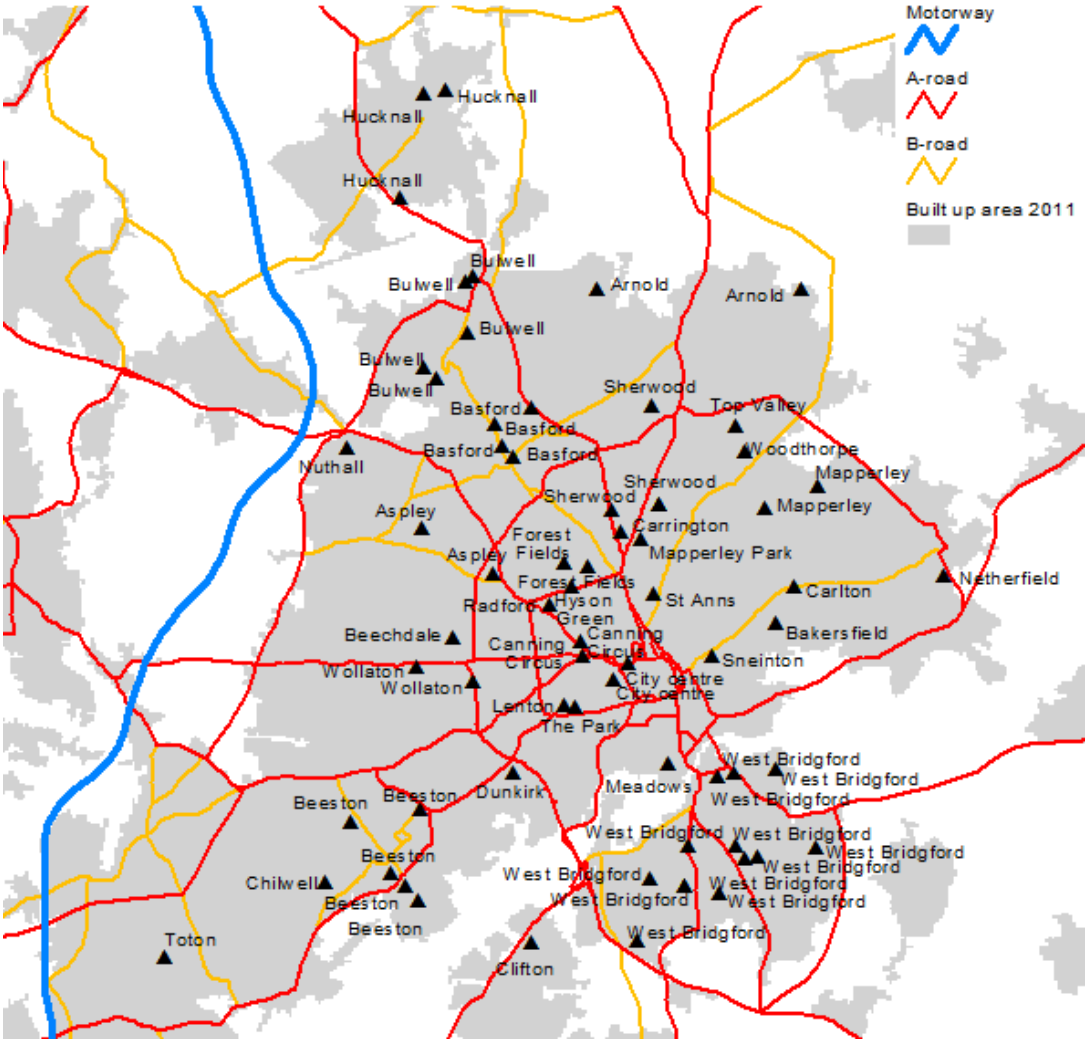
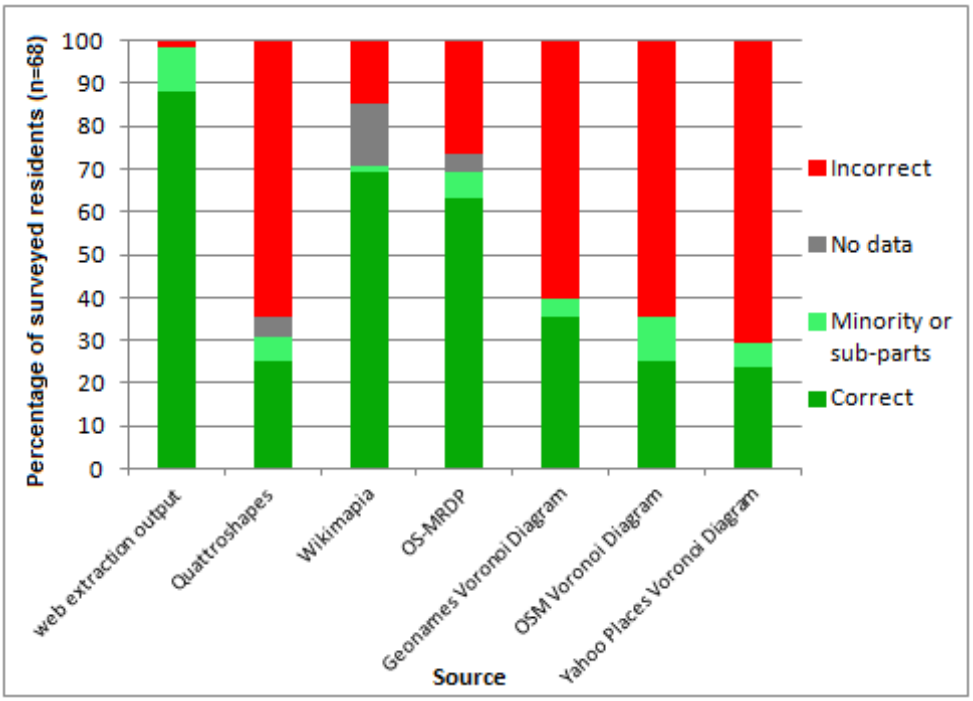


Fig. 6.10 Comparison of resident views of urban locale and different sources of urban locale geography



to methodological differences (where one set of data is derived from 300m KDE smoothings in contrast to gazetteer centre-points formed from Voronoi Diagrams).

The poor performance of the Voronoi Diagram output may in part be a reflection on the centroid locations being inadequately positioned relative to real centroids of the urban locale, which in turn would affect the Voronoi region that is simply a function of the input centroid and, hence, takes little account of the complex nature of urban locale boundaries.

A large proportion of resident data (nearly 15%) found no comparison with Wikimapia due to lack of urban locale data within Wikimapia, demonstrating the difficulty in achieving full coverage through active crowdsourcing. It should also be noted that in a number of cases, residents' views were to be found in close-by urban locales within Wikimapia. In this regard, there were 9 cases where agreement with the resident view was found within 300 metres. Even allowing for these relatively 'near-misses', however, the level of agreement would still only stand at 87% - some way shy of the 98% found for the web extraction output.

- **Information collected by a Neighbourhood Forum:**
  It is possible to specifically examine the detailed geography of an individual urban locale if enough 'resident' responses can be collected. This comparison is based on perceptions of neighbourhood boundaries obtained by members of a local Neighbourhood Forum. In order to maintain confidentiality, the Forum within this work will be referred to by the pseudonym of 'Crespmoor'. Within Neighbourhood

Planning (as outlined within Section 4.3), in order to gain recognition as a Neighbourhood Forum a discrete boundary needs to be agreed between the Forum and local council. Crespmoor Neighbourhood Forum sought views of the boundary of Crespmoor from those who lived or worked in the area through both responses to an on-line questionnaire[21] and at a local community arts festival through people placing coloured stickers on a physical map. In total just under 400 replies to an on-line questionnaire asking people for their postcode and response to the single question: "Do you live in Crespmoor?" Also, 329 data points were gathered from the field survey where people at the Crespmoor local community arts festival were asked to mark their home location with a colour coded sticker representing different neighbourhood names within the area. The two sources of information collected data from both participants who viewed themselves as within Crespmoor and also from those who believed they were outside of Crespmoor[22]. The raw point data from the two sources of information collected can be seen in Figure 6.11.

Comparisons between the web extraction output and the data collected by Crespmoor Neighbourhood Forum after applying KDE methods to the point data are illustrated in Figure 6.12 and show strong similarities. The main disparity is that the resident views collected by the Neighbourhood Forum appear to have higher data returns northwards of the A-road running east to west across the area. Whilst data are found within the web extraction output for this area, they are slightly lower in data volumes.

Linear grid regression was used to explore the relationship between the resident data (y - dependent variable) and the web extraction output (x - independent variable). Comparison is based on the two variables for each corresponding cell of the KDE grids[23]. Linear grid regression analysis showed that 77.0% of the variance (R^2) between the field survey absolute values and those from the web extraction method could be explained[24], whilst the R^2 between the on-line questionnaire and web extraction data was 75.9%[25]. The main differences were not in the geography but levels of intensity with the web extraction output generating more returns in non-residential areas, in contrast to the resident opinions which were highest in residential areas. It is possible that many of the participants from the field survey might also have taken part in the on-line questionnaire which might explain the high similarities between the two sources (R^2 = 95.8%[26]).

[21] distributed through a combination of sites on social media, publicity in a local newspaper and selected targeting of known email lists of the forum

[22] in contrast to the on-line questionnaire (which didn't collect the alternative urban locale names from participants), the field survey identified six alternative urban locale names

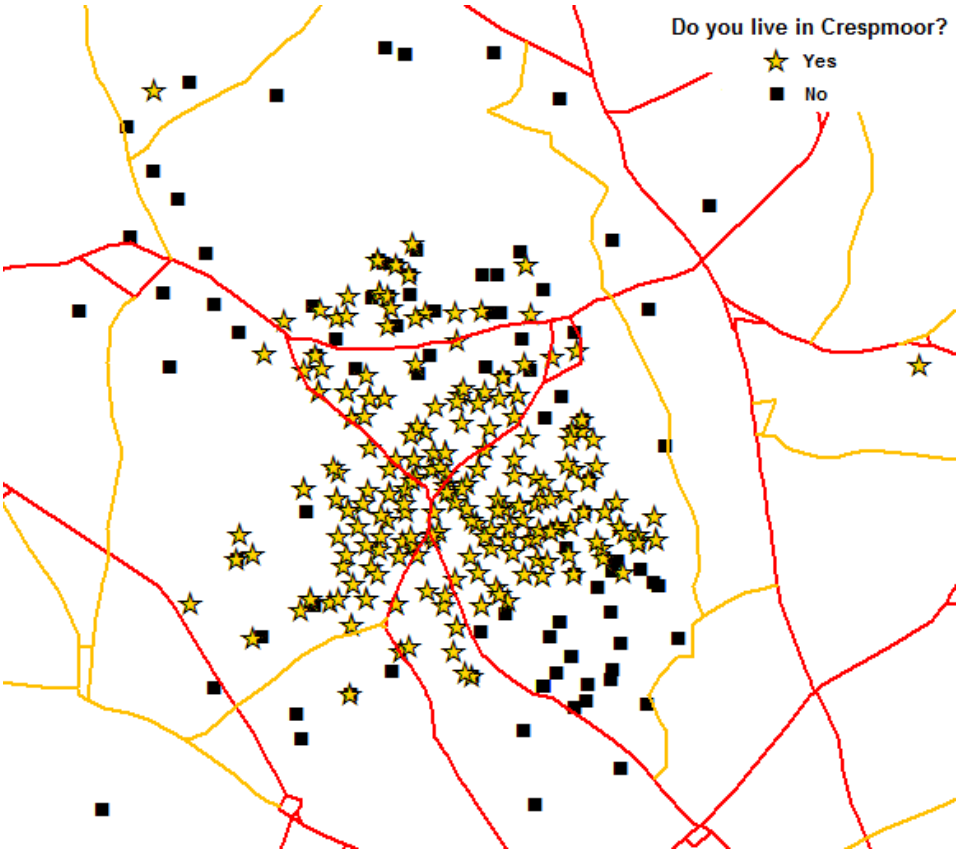
[23] Only cells where either of the KDE grids were greater than zero were compared

[24] $y = 0.0975912x + 0.0010709, p < .001$

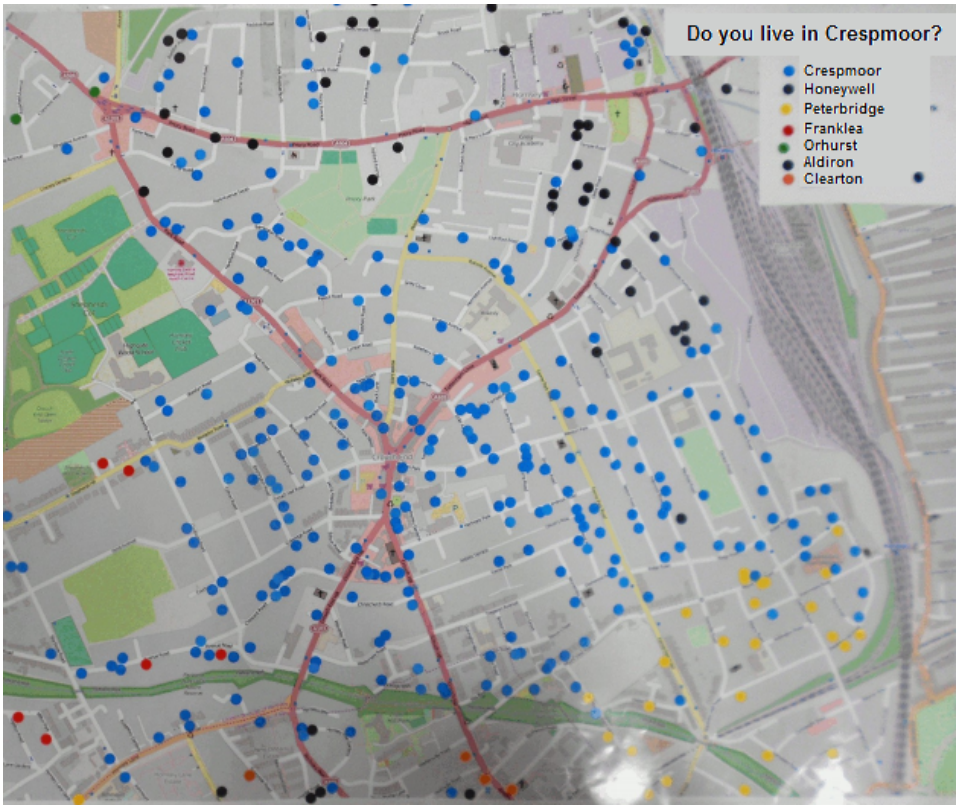
[25] $y = 0.1005764x + 0.0012664, p < .001$

[26] $y = 0.9426515 * [\text{on-line responses}] - 0.0000586, p < .001$

Fig. 6.11 Raw point data collected by Crespmoor Neighbourhood Forum



a) On-line questionnaire responces



b) Field survey responces

It is difficult to fully compare the geographies of the urban locale percentage grids[27] as the data contain few references to urban locales other than Crespmoor. As demonstrated within Table 6.15, in comparison to the web extraction data, the number of data returns for other neighbouring urban locales in the area were low and this produces the high percentage values appearing at the edge of the extent in Figures 6.12d and 6.12e without gradual decreases in percentage terms as one moves into adjoining urban locales (as evident within Figure 6.12f). This illustrates the difficulty in gaining resident survey data points not only within Crespmoor itself but also continuing to gather survey responses into adjoining locales, in order to accurately map the areas of transition. It also demonstrates the strength of the web extraction method in achieving this. The important implication is that whilst such resident mapping, collected as part of the neighbourhood planning requirement, can adequately map the geography of their urban locale using the absolute data approach, they will rarely be able to also comprehensively map the alternative perception by using the urban locale percentage grids. These distinctions (between absolute and percentage of data returns) were first outlined in Chapter 5 and are debated further within Section 8.1.1.2.

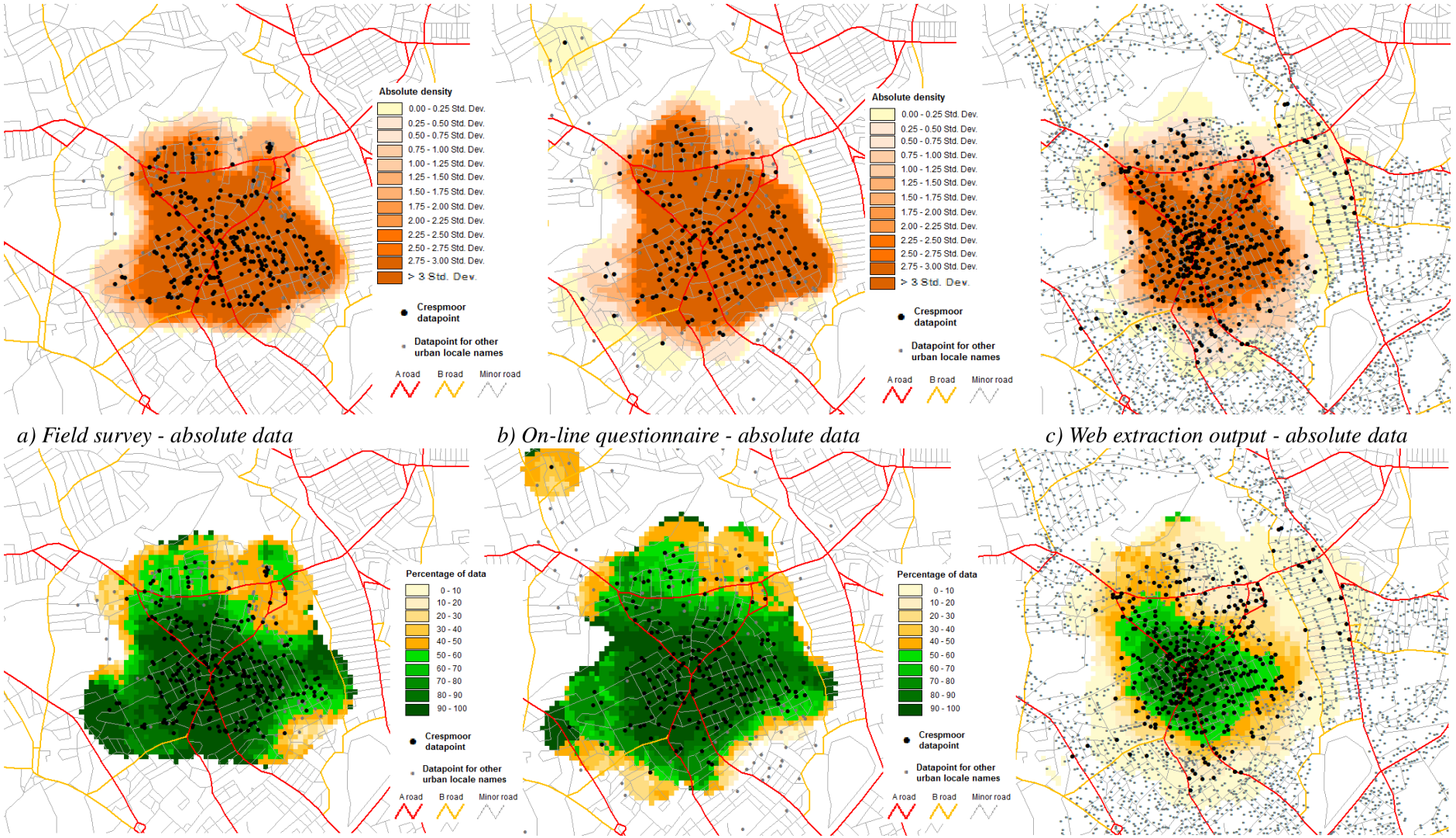
Table 6.15 Comparing data volumes for Neighbourhood Forum resident views and web extraction output

	Total number of data returns	Data returns relating to Crespmoor urban locale	Data returns relating to other urban locales
Field survey	329	264	65
On-line questionnaire	396	285	111
Web extraction output	7,000	1,834	over 5,000

It is possible to compare the resident perceptions of the Crespmoor urban locale with the set of alternative data extents (as applied within Section 6.2.2.1 of this chapter). In this regard, boundaries from Wikimapia, OS MRDP, Quattroshapes and the web extraction output were all compared with the resident data collected by the Neighbourhood Forum. Figure 6.13 maps out the comparison whilst Table 6.16 provides a summary view and demonstrates that the web extraction data was considerably closer to the resident perceptions than any of the other data. Not only did the web extraction output produce the highest level of agreement with resident perspectives, by a significant margin (72% compared to the next highest datasource of only 32%), but also returned the smallest level of cells identified by the residents but not included within the output from the specified method (ie cells that were 'missed' by the method). It should be noted that in all cases the extents from the differing methods produced (as shown in Figure 6.13) were smaller than

[27] calculated by dividing each grid cell of the 300m KDE surface of absolute values for the given urban locale by the corresponding grid cells for the total of 300m absolute value KDE surfaces for all urban locales

Fig. 6.12 Comparing the geography of resident perceptions for an urban locale with web extraction output



a) Field survey - absolute data

b) On-line questionnaire - absolute data

c) Web extraction output - absolute data

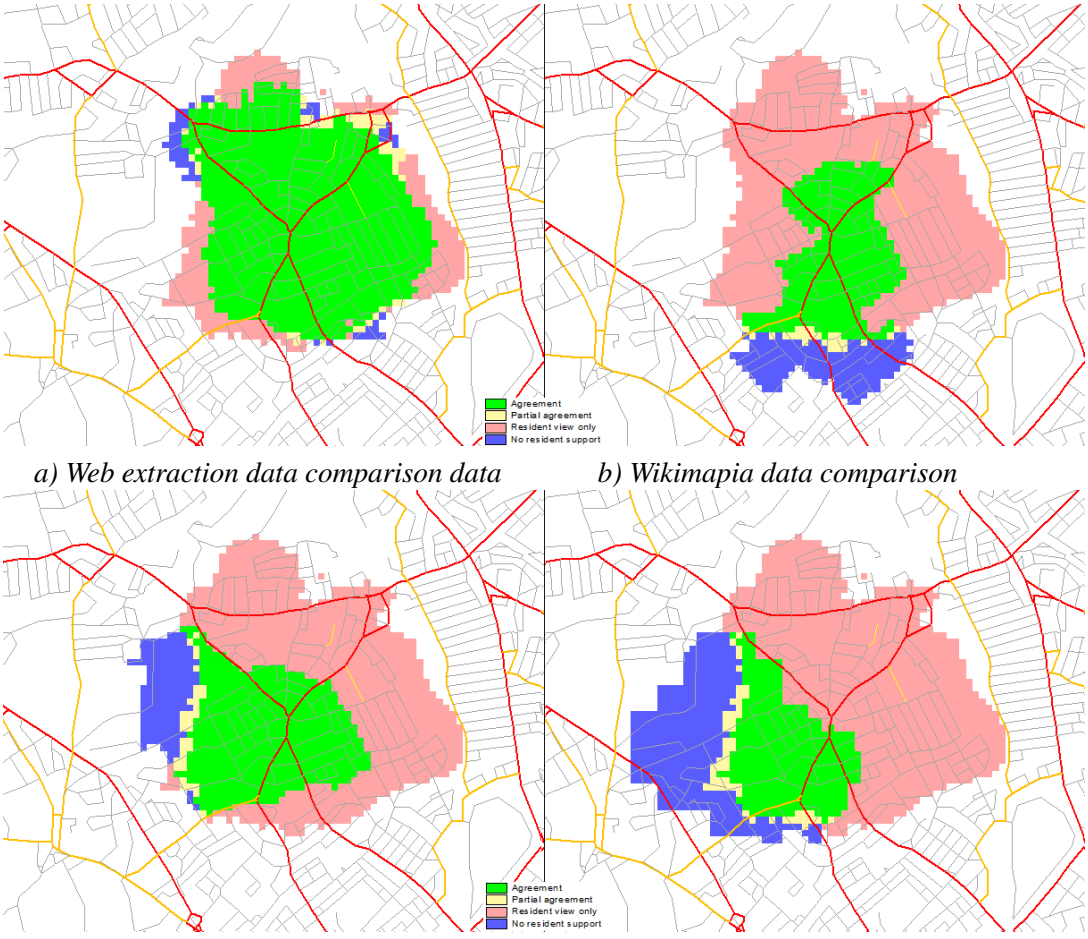
d) Field survey - percentage data

e) On-line questionnaire - percentage data

f) Web extraction output - percentage data

the proposed size of Crespmoor urban locale as defined in either data collected by the Crespmoor Neighbourhood Forum (approximately 315 hectares). The closest was the web extraction method (282 has), whilst the other three datasources were substantially smaller (Quattroshapes: 183 hectares; OS MRDP: 162; Wikimapia: 146) perhaps due to the fact that they do not allow for overlapping areas.

Fig. 6.13 Level of geographic agreement between resident perceptions for a single urban locale with a range of other data output



a) Web extraction data comparison data *b) Wikimapia data comparison*

c) OS MRDP data comparison *d) Quattroshapes data comparison*

- **Testing differing perspectives within the same localised space:**
  By targeting areas where the web extraction output purports to show divergent variations in the locale name for a given space, it is possible to validate the data by comparing the name of the urban locale when asking residents their views and opinions. In order to validate the data, one would expect the same differences in opinion to occur in both the web extraction output and resident responses. To this end, opinions were sought by visiting residential properties in three postcodes in Sheffield (S10 1FA - Roslin Road; S10 1BJ - Crookesmoor Road; and S10 5BD - Crookes Road) and residents were asked the name of the area that they lived in. The three streets were chosen as they appear (in the web extraction output) to sit at the intersection of three different urban locales (Broomhill, Crookes and Crookesmoor).

Table 6.16 Comparing resident boundaries for a selected urban locale with varying other datasource methods

	Web extraction data		**OS MRDP**		**Quatroshapes**		**Wikimapia**	
	No. of cells	% of cells	No. of cells	% of cells	No. of cells	% of cells	No. of cells	% of cells
Agreement (both sources of resident data support output from the method) Similar to 'True positives' if sample of residents was ground truth	1,016	72.1	482	32.4	331	19.4	389	25.5
Partial support (one of the resident datasources is in agreement with output from the method but not both resident sources) Similar to 'True positives' if sample of residents was ground truth	56	4.0	34	2.3	54	3.2	25	1.6
No resident support (method identifies area but no supporting resident evidence) Similar to 'False positives' if sample of residents was ground truth	54	3.8	132	8.9	348	20.4	170	11.1
Resident view only (method 'misses' resident defined areas) Similar to 'False negatives' if sample of residents was ground truth	283	20.1	839	56.4	970	57.0	941	61.7

The responses are mapped in Figure 6.14, demonstrating the highly variable nature of urban locales with frequently neighbours possessing different views with little spatial cohesion. The data are also tabulated in Table 6.17 along with the comparative output from the web extraction method. It is interesting that in one instance even people living in the same property had different opinions on the name of their local area.

It would be possible to compare resident views with those obtained through the web extraction output using a Chi-square test in order to examine if there were any statistical difference between the level of responses. In such a case, the resident views would constitute the expected values whilst the web extraction data would form the observed values. The appropriateness of a Chi-square test, however, is dependant on having enough data points. Given the small numbers associated with individual postcodes (approximately 15 properties), it is highly likely that, given response rates and dividing data points between a number of differing urban locales, there would be insufficient data points for a Chi-square test at the postcode level. Thus, in order to meet the assumptions of the test, data were *aggregated* for all three postcodes in order to investigate any differences between the output from the web extraction method and resident responses. Data aggregation in this case was acceptable due to the close proximity of the postcodes, with them covering an area of just over one hectare in size. No statistically significant difference was found between the web extraction output and resident responses ($X^2 = 0.09, df = 2, p < 0.05$). This has two important implications as it not only aids the validation of the web extraction output (being similar to that when asking residents directly) but also more fundamentally supports the notion and requirement of vague urban locale boundaries. If a more accurate representation of urban locales is required - the fact that people even within the same property can have such varying views[28] demonstrates that discrete boundaries can never capture such complexity and vague notions are necessary.

6.3 Summary

The sample results shown within the chapter demonstrate the potential of the method as described in the previous chapter. The process does not only produce vague urban locale boundaries but also generates output concerned with vague place names - through the content of lists of common misspellings and potential synonyms.

[28] In this instance the residents stressed the importance of status of belonging to certain urban locales as the rationale for differences of opinion within the household. Whilst of pertinent interest to future work, the investigation of *why* people identify with different urban locale names was considered outside the scope of this work

Fig. 6.14 Study areas for testing diversity of opinion in selected streets

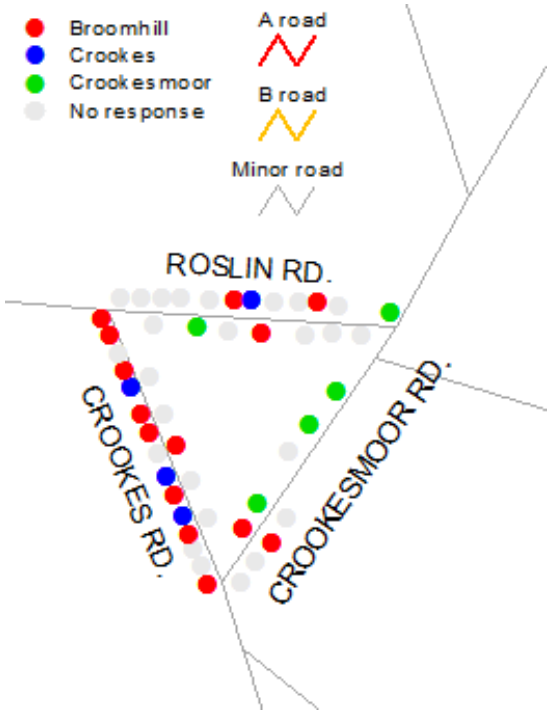


Table 6.17 Testing the level of variation within the urban locale name in selected streets, Sheffield

a) Roslin Road:

	Web extraction output	Asking residents	% Web extraction output	% Asking residents
Broomhill	3	3	**60.0**	**50.0**
Crookesmoor	2	2	**40.0**	**33.3**
Crookes	0	1	**0.0**	**16.7**
Total response rate:		*5 from 19*		*26.32*

b) Crookesmoor Road:

	Web extraction output	Asking residents	% Web extraction output	% Asking residents
Broomhill	3	3	**50.0**	**60.0**
Crookesmoor	3	2	**50.0**	**40.0**
Total response rate:		*5 from 9*		*55.56*

c) Crookes Road:

	Web extraction output	Asking residents	% Web extraction output	% Asking residents
Broomhill	15	9	**75.0**	**75.0**
Crookes	5	3	**25.0**	**25.0**
Total response rate:		*10 from 20*		*50.00*

The usual validation process of comparing generated output with some ground truth is not possible as no such fundamental truth exists. Instead, a range of different comparisons were constructed - testing both urban locale names and boundaries.

Comparison of urban locale names between the different existing gazetteer sources demonstrated the lack of consistency between the sources. In Nottingham and Sheffield, 30.2 and 50.2% of urban locales respectively, were found only in a single source. This inconsistency is further illustrated by the over-representation of OS place names within Sheffield which is in stark contrast to the number of OS place names identified in Nottingham. In total the web extraction method identified 18 locales that were not included within the existing gazetteers. Whilst many represented named industrial estates (n=11), the remainder consisted of areas that ideally would have been found in the other sources. Their absence has important ramifications for completeness and the INSPIRE Directive (as discussed in Chapter 4). Overall, in terms of urban locale name, the web extraction output appeared to be the most comprehensive source - with the highest level of agreement with the other sources.

The closest measure to a ground truth for urban locale boundary information is gained through samples of resident responses. The web extraction output was found to be a good indicator of resident perceptions of urban locale when tested using three different scenarios as summarised below:

1 When tested against other available sources of urban locale boundary information in a sample of 68 resident views in Nottingham, the web extraction output proved to be closest measure to the views expressed by residents. Comparison with resident views was extremely high with over 98% level of agreement. It was noteworthy that the sources with the next highest level of accord, the level of agreement dropped to just 70% (Wikimapia and OS MRDP).

2 A detailed comparison of a single urban locale geography using 400 resident opinions from data collected by a Neighbourhood Forum revealed that 77.0% of the variance could be explained by linear grid regression between the field survey absolute values and those from the web extraction method ($y = 0.0975912x + 0.0010709, p < .001$). Furthermore, the geography of resident views was matched closer to the web extraction data than any of the other alternative sources of bounded urban locale (72% agreement of cells, compared to the next highest datasource of only 32%). The web extraction method also had the lowest proportion of cells both identified by the method but without the support of resident views (analogous to 'False positives' if the sample of residents' views were considered representative of the ground truth) and where

the method did not identify areas that the resident view had identified (Similar to 'False negatives' if sample of residents was ground truth).

3 Variations of urban locale perceptions for three streets in Sheffield supported the notion that there was no statistically significant difference between resident views and the data extracted from web output ($X^2 = 0.09, df = 2, p < 0.05$). The output also demonstrated that residents in close proximity (and even within the same household) could hold different views as to the name of the urban locale in which they lived - supporting the requirement of vague urban locales.

Under all three different approaches, the web extraction output performed closer to residents views than any of the other existing urban locale datasources. The resident validation, thus, upheld the research hypothesis - that vague urban locales generated from web data *are representative* of those obtained through asking people directly.

Chapter 7

Analyses of Vague Urban Locales

Following validation and once confidence in the data has been assured, it is possible to pose and answer certain questions concerning the urban locales. As such, this chapter represents a critique and reflection upon the method (Chapter 5) and results (Chapter 6). Chapter 1 outlined a number of sub-research questions relating specifically to urban locales to be investigated, including:

1 **Web groupings:** *To what extent is it possible to define urban locale areas for different user groups?* Addressing such a question is of value in exploring if varying user groups might generate different urban locale geographies. Furthermore, greater confidence in the output of data may be gained if the geographies of urban locales from different user groups are in consensus.

2 **Temporal aspects:** *To what extent is it possible to explore temporal change in urban locales?* Given the dynamic world in which we live, the degree of change in urban locale boundaries through time is of general interest.

3 **Tying other data to urban locale units:** *How feasible is it to tie other data (either point or area measures such as Census data) to the newly formed geographies of urban locale?* Aggregating other socio-economic data to urban locales may add value to both source and input data - creating output for the areas people connect and associate with in everyday life.

4 **Homogeneity:** *How homogeneous are the urban locales in terms of socio-economic characteristics?* One measure of defining 'neighbourhood' is through the aggregation of data to form homogeneous areas. In this light, exploring the degree of homogeneity/heterogeneity of urban locales is of interest.

5 **Comparison with underlying population geography:** *How well does the frequency of urban locale data compare to the underlying population? Are there*

unexpected areas with low/high data returns? Such a question may reveal interesting insights about the urban locale data and also aid validation of the output.

6 **Differing perspectives:** *To what extent do the urban locale spaces overlap or nest, creating differing perspectives? Can the hierarchical nature of space be examined?* The degree of overlapping of urban locales is critical in establishing the extent to which locale boundaries should be *discrete* or represented as *vague* entities. Furthermore, it may provide insight into the complex hierarchy and relationships between places within our urban areas.

The first three of these queries represent further advancement and refinement of the general method, whilst the last three items comprise of specific questions that the output of the work can address concerning the nature of urban locales. Each will now be explored in turn.

7.1 Web groupings

To what extent is it possible to define urban locale areas for different user groups (for example: an estate agents view of urban locales versus a resident's perception)?

The basis of the urban locale method has shown that it has been possible to extract various perspectives of spatial objects from web based data. Different groups perceive place with different meanings, uses and values (Holloway and Hubbard, 2001). With this in mind, it might also be the case that *differing groups of people* may form urban locales with different distinct geographic extents. The rationale behind performing such a procedure is that a greater confidence in output might be achieved if many user groups are all in agreement. There might also be certain groups where public confidence in the geographies they produce are frequently questioned. The classic example is the common image that estate agents are likely to twist and distort the geographies of certain place names (Flemmings, 2010; Fumia, 2010; Knox and Pinch, 2009). Thus, the usefulness of the approach is in the identification of distortions within the geography of specific urban locales by certain groups. The research question explores this issue by attempting to classifying the URLs associated with each data record and then mapping the data points for each URL group separately. Whilst it would be possible to perform co-training techniques to classify the URLs based on the text within each snippet, such a process was considered outside the scope of this work which instead undertook a simple rule based approach as outlined below.

7.1.1 Web groupings - method

Manual inspection of the data identified that a large proportion were derived from lists of business listings such as Yellow Pages or company information sites (such as pages for company reports, credit scores or company director information). Additionally, estate agents and letting agents represented a further significant grouping. A third category consisted of the data not in either of the other two previous groupings and contained a range of documents including social media sites, personal blogs and other user content. This then gave the following typology:

- 'Business information sources'
- 'Real estate sources'
- 'Other sources (including user content)'

Classification was undertaken by the following sequential procedure:

1 **known domains**: due to the frequency distribution whereby a large number of records originate from a relatively small number of website domains - it seems sensible to manually assign a small number of domains where the classification was clear given the domain URL (for example either well know sites such as yell.com[1]. Estate agent sites were added from The Property Ombudsman (TPO) website[2]. Only 175 domains were manually assigned but this accounted for approximately half of the 121,146 URL data records for the Nottingham case study area.

2 **keywords in URL**: words within the URL might indicate the source categorisation - for example references to 'estateagent' or 'accommodation' would indicate that the data belonged to a real estate source, whilst references to 'businesspages' or 'companysearch' would result in a business source classification. The full list of keywords used in this manner can be found within Appendix F.

3 **metadata info**: classification was based on whether the web-page metadata title, description or keywords contained either any of the business directory oriented words (for example: business directory, local directory, business listing, yellow pages, business search, company credit and so forth) or at least 10% of content from the estate agent oriented words[3] (such as property, estate agent,

[1]Existing sources such as DMOZ (https://www.dmoz.org) could be used but were not employed in this study as classifications were found to not always be consistent

[2]http://www.tpos.co.uk/search-results.php? Note: this is a useful source as all residential estate agents in England and Wales who are marketing homes requiring Home Information Packs are required by law to be registered with TPO

[3]Real estate categorisation here needs to be tighter than for business information as false positives are more common with references of sale, renting, property etc. not necessarily relating to real estate

real estate, renting, sale, and so forth). The full list of terms can be found within Appendix F.

4 **terms within the web-page**: the text of the web-page itself is assessed and data assigned respectively if 2% or more of the words are either estate agent terms or business directory terms. The full list of terms used can be found within Appendix F. In order to classify lists of businesses - data were also classified as business oriented if the web-page contained at least ten different postcodes and telephone numbers or at least ten different company numbers and ten company suffices (such as ltd, plc etc).

The process is ordered - so that if classification is achieved through the first stage, further analysis is not required and the remaining steps are not necessary. If all the above tests fail, then the page is classified in the 'Other' category.

7.1.2 Web groupings - validation

In order to validate the process, a comparison was made between the automated classification of the URLs and a manual assignment of all 121,146 URLs for the Nottingham case study[4]. As shown in the 'confusion matrix' of Table 7.1, a correct assignment was made in 106,451 records, accounting for a 87.87% successful categorisation. Of the records that were miss-classified, the most common form was where the automated procedure classified pages as 'Other' but the manual check identified them as 'Business Information'. This form of miss-classification accounted for nearly 9,000 records and represented close to one third of the differences in classification. It included cases of business directory information where only a single business was shown on a page[5] or where a more general website hosted specific business information but again only contained a single business listing on the page[6]. Whilst it would be possible to modify the rules in order to identify such cases (perhaps looking for business names within the URL) this was considered outside the scope of this exploratory work.

In case the manual check was biased due to being performed by the same person who designed the original rules for classification, two independent members also classified 100 URLs into the specified three categorises. The 100 URLs were chosen randomly from amongst the web-pages where both the manual and automated

[4]categorised by examining a list of all the URLs and determining the most appropriate class. This required visiting the webpage if domains were not known

[5]for example: http://www.whatclinic.com/beauty-clinics/uk/nottingham/the-meadows/the-harley-medical-group-nottingham or http://www.inapub.co.uk/pubs/bistro-live/nottingham/ng11js/25071

[6]for example: http://www.tipped.co.uk/listings/192039/mayfair-school-of-motoring or http://www.theinvestingsite.com/banks/nottinghamshire/nottingham/barclays-NG9-1EF.htm

Table 7.1 Validation of web group assignment using manual checking

	Manual assignment:		
Automatic assignment:	business information sources	real estate sources	other sources
business information sources	**63,717**	290	11,622
real estate sources	237	**16,188**	311
other sources	1,622	613	**26,546**

process matched. This was undertaken in order to ensure that differences should be solely due to perception of the individuals classification as opposed to how individual data records might be best classified. There was a 94% and a 93% agreement between the author's classification and that of the two independent members. There was only one case where the two members both assigned a different class to the author and automatic classification procedure. Both members classified the URL 'http://www.londonrenting.co.uk/profiles/2494' as a 'Real estate source' whilst the author and rule based method identified it as a 'Business information source' due to the fact that the page is not performing real estate activities through renting opportunities but is instead listing contact details of letting agents.

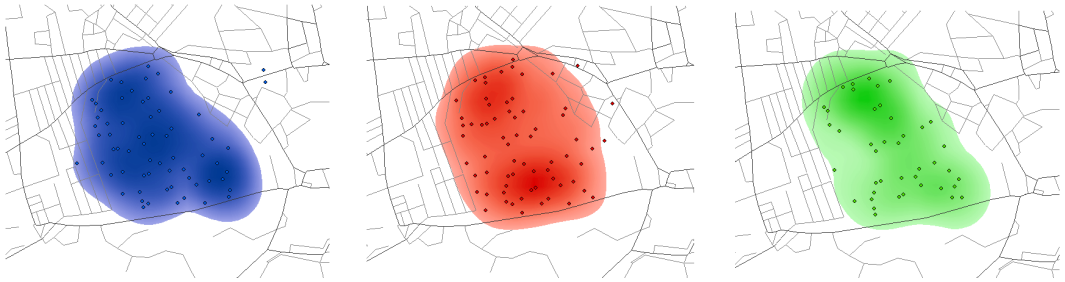
7.1.3 Web groupings - results and discussion

The frequency of data records falling in the three web categorisations within the case study area of Nottingham can be found in Table 7.2 and demonstrate the prominence of business information within the data. This supports the finding of Jones et al. (2008a) who found that search engines using queries such as "hotels in [placename]" returned a high number of results from directory-style web-pages. A list of the most frequently occurring domains can be found in Table F.1 in Appendix F. An example of different urban locale boundaries (for The Park, Nottingham) formed from the three web categorisations can be found in Figure 7.1. The figure helps demonstrate the fact that business information sources are not solely limited geographically to non-residential areas (such as shopping centres) and include many small businesses registered at home addresses (for example plumbers, electricians and so forth) occurring within more residential orientated urban locales.

Table 7.2 Web grouping classification output

	Business Info.	%	Real estate	%	Other	%
1. Manually assigned known domains	46,244	38.2	13,707	11.3	5,237	4.3
2. Keywords in URL	6,879	5.7	777	0.6	n/a	n/a
3. Metadata info.	5,181	4.3	2,046	1.7	n/a	n/a
4. Words within web-page	10,941	9.0	1,467	1.2	n/a	n/a
Residual					28,667	23.7
TOTAL	69,245	57.2	17,997	14.9	33,904	28.0

Fig. 7.1 Example map of web categorisation: The Park, Nottingham



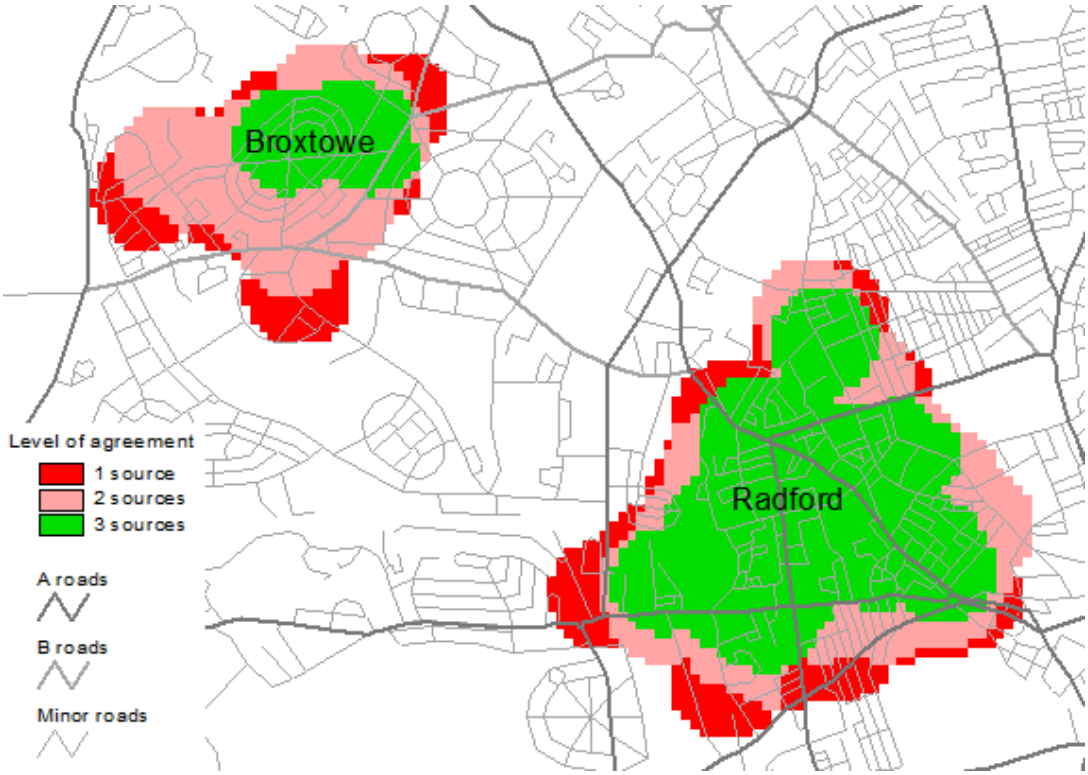
a) Business information sources b) Real estate sources c) Other sources

Once assignment of classification has occurred, it is possible to explore if differences exist between the geographies of the urban locales for the three different web categorisations. This was possible by simplifying the complex overlapping nature of locales to a form more suitable for comparison where each grid cell consisted of a *majority function* for each of the three categorisations based on the most frequently occurring urban locale name within 100m (generated as a moving average). Overall, within the Nottingham case study there was a strong replication of urban locale extents by the three groupings - with *97% of cells maintaining the same majority urban locale within 100m between business information and real estate sources*. Such a level of consensus not only demonstrates the overall high degree of similarity between user groups but also illustrates the extent to which the user groups are not geographically isolated (and coverage of each group appears relatively universal across space). This said, there are, however, areas of divergence between the groupings.

The degree of agreement between the three different user groups can easily be mapped, as in Figure 7.2. The greatest level of confidence in the data might be considered in those areas where all three sources are in agreement. Inversely, the more questionable areas of coverage might be those where only a single source identifies the space with a particular name. This may not always necessarily be the case, as data omission might simply be an artefact of the underlying population but the procedure may help identify areas warranting further investigation.

There are other alternatives in order to attempt to identify those urban locales where differences in geography between user groups may exist. The distribution of data record counts between the three different web groupings could be compared to those *expected*, given overall data records in the urban locale. This would be able to identify potential locales that were over or under-represented within the data. Such a method would be able to identify urban locale names that estate agents might avoid using or inversely might over-sell. For example, in total 57.2% of data related to the Business information class, 14.9% to Real estate and 28.0% Other. Thus, of the 1,018 data records in Bilborough, all things being equal, one might expect 581 to be

Fig. 7.2 The degree of support between the three different web groupings for Radford and Broxtowe, Nottingham



from Business Information sources (57.2%), 152 from Real estate (14.9%) and 285 from Other (28.0%). The actual composition for Bilborough, however, is 484 from Business (48%), 55 (5%) from Real estate and 479 from Other (47%). Conversions to z-scores demonstrates an under-representation for Business Information (-0.22) and Real Estate (-0.60) and over-representation within data from Other sources (0.58). Z-scores from the web classification composition can be found for a small number of urban locale examples within Table 7.3. Care must be taken with the interpretation of output, however, as distribution between the categorisations might in part be merely a reflection of the underlying structure of the urban locale. For example, an over-representation of data from the Real estate class might be due to the locale being a student area - and thus more likely to have a high turnover of property. Also, in an extreme case, one would not expect to find many references from the Real estate category for an industrial estate. Although outside the scope of this exploratory work, to overcome these issues, the expectations could be weighted by the numbers of residential and non-residential property. Even without such adjustments, however, the data may give an indication of interesting urban locales to further explore in order to identify areas where the differing web groups provide distinct geographies.

In order to further explore geographic differences at the urban locale level, linear grid regression analysis was undertaken between the raster grid output for the different web categorisations. Such an approach can identify those areas where the underlying

Table 7.3 Web grouping classification z-scores for selected urban locales in Nottingham

Urban locale	Number of Data points	Z-score Business information	Z-score Real estate	Z-score Other
Arboretum	152	-2.41	5.67	-1.27
Bilborough	1,018	-0.22	-0.60	0.58
Broxtowe	90	-0.13	-0.35	0.34
Canning Circus	122	-1.33	0.87	0.72
Carlton	6,892	0.67	0.34	-0.86
City centre	252	0.54	-0.21	-0.38
Colwick industrial estate	399	1.96	-0.97	-1.26
Daybrook	1,382	0.98	-0.06	-0.90
Forest Fields	583	-0.60	1.98	-0.68
Hyson Green	740	-0.37	-0.29	0.54
Lenton	1,902	-1.18	2.48	-0.43
Radford	1,246	0.07	-0.56	0.28
Sneinton	977	-0.86	0.19	0.70
St Anns	531	-0.61	-0.18	0.70
The Lace Market	716	-1.12	-0.42	1.34
The Park	705	-0.17	1.73	-0.93

geography of a locale is different between the web categorisations. Whilst Table 7.3 indicates an over-representation of data from Real estate sources for Forest Fields, linear grid regression analysis indicates that the geography of output produced from both the Business Information and Real estate sources are similar (R^2 = 97%[7]). In contrast, the geographies of the Broxtowe urban locale are more dissimilar, as shown in Figure 7.3, with only 25% of the variation within the map derived using Real estate sources, explained by the map from Business information sources[8]. The mapped residuals illustrate a distinct lack of data from real estate sources to the south of the locale (Figure 7.3c). Interestingly, much of the same area exhibits higher than expected rates of data from real estate sources within references to Strelley urban locale (Figure 7.3d). This might indicate that real estate, unlike other sources of data, may be more likely to name the particular area as Strelley over the term Broxtowe. The differences also incorporate the multiple dimensions frequently associated with place names. For example, the place name Broxtowe can refer to the urban locale but also gives rise to the name of the local administrative area (or Borough[9]). Furthermore, the perceptions of place names carry significant weight and names can conjure particular images associated with them. For example, the place name of Strelley frequently embodies the notion of a 'nice village' but the name is used for both the historical village core and the suburban former council estate area. Such are the complexities of place.

[7] $y = 1.4415573x - 0.0331582, p < .001$

[8] $y = 2.4742710x + 0.0702157, p < .001$

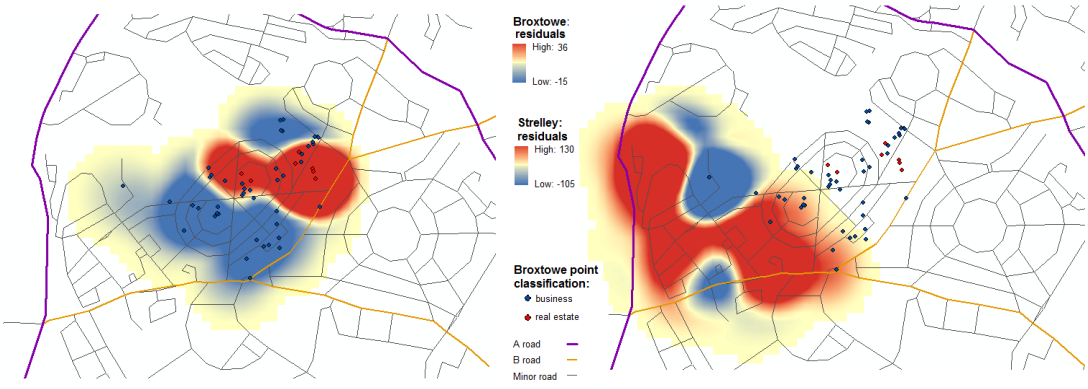
[9] https://en.wikipedia.org/wiki/Borough_status_in_the_United_Kingdom

The R-squared values from linear grid regression for the urban locales within the previous table can be found in Table 7.4. Generally, the geographies of urban locale from Business Information and Other sources were closer (average level of variance explained by linear grid regression was 83%) than comparisons made with Real estate views (average R^2 of 67% between Real estate and Business Information and 72% between Real estate and Other sources). This demonstrates that although generally Real estate viewpoints were consistent with other sources, there were noticeable variations. Given that it was previously noted that there was a high degree of similarity in the 100m majority function of urban locale[10] between business information and real estate sources (with 97% of cells maintaining the same majority), most differences are likely to be due to boundary distinctions. Whilst further research concerning differences in Real estate sources is of both value and interest, it is beyond the scope of this work to develop it further, as the aim here is to demonstrate the feasibility of defining differing geographies of urban locale based on user groups.

Fig. 7.3 Example map of web categorisation: Broxtowe, Nottingham[11]



a) Business information sources b) Real estate sources



c) Regression residuals, Broxtowe d) Regression residuals, Strelley

Even using this simple approach, this section has demonstrated that different groups of perspective regarding urban locale extents can be formed and compared. The strong

[10] the 100m moving average of the most frequent urban locale name as previously discussed on page 149

[11] Jittering (slight randomisation in co-ordinates) has been applied to Broxtowe point data shown here, as a post-processing step, in order to be able to visualise multiple points at the same location

Table 7.4 R-squared output from regression for web grouping classification, selected urban locales in Nottingham (significance levels: * p < .05, ** p < .01)

	Business info. and Real estate	Business info. and Other	Real estate and Other
Arboretum	0.63**	0.26**	0.41**
Bilborough	0.20**	0.84**	0.97**
Broxtowe	0.25**	0.77**	0.90**
Canning Circus	0.77**	0.89**	0.83**
Carlton	1.00**	0.87**	0.65**
City centre	0.40**	0.90**	0.68**
Colwick industrial estate	0.74**	0.83**	0.92**
Daybrook	0.77**	0.91**	0.74**
Forest Fields	0.97**	0.99**	0.76**
Hyson Green	0.74**	0.97**	0.71**
Lenton	0.65**	0.97**	0.81**
Radford	0.50**	0.88**	0.99**
Sneinton	0.61**	0.99**	0.94**
St Anns	0.29**	0.75**	0.96**
The Lace Market	0.89**	1.00**	0.90**
The Park	0.92**	0.71**	0.70**
Average	0.67	0.83	0.72

general association found between the differing web groupings adds confidence to the geographies formulated (97% of cells maintaining the same majority urban locale within 100m between business information and real estate sources). It was, however, also possible to identify variations at the individual locale level, such as in the case for Broxtowe, Nottingham.

7.2 Temporal aspects

To what extent is it possible to explore temporal change in the urban locale units?

Whilst the principal concern of the work has been spatially orientated, there is also an interest in the temporal nature of the constructed spatial objects (as briefly discussed within Chapter 2). There is a recognition in the literature that urban locales will not be static and that their boundaries will shift over time (Martin, 1998). In fact van Ham (2012) argues that the role of the temporal dimension is largely overlooked within the neighbourhood effects literature. With this in mind, this section tests the potential of the current method to explore if it is possible to assess the levels of change within urban locales. Ideally it would be possible to measure change continually and to be able to mine archived material of the web for historic perspectives. Unfortunately, this has not been possible as direct searching on archives such as the Wayback machine[12] are not currently possible. Therefore, the approach adopted here is one of comparing

[12]http://archive.org/web

timestamp data runs. Both long-term trends and more undesirable short-term instability were investigated and are considered below.

7.2.1 Short-term change: Monthly

Whilst mapping long-term change may be desirable, it is also important that output of the method is robust in the shorter-term. There would be little point in undertaking such methods if the geography produced was dramatically different from one day to another (baring change due to long-term trends). In order to investigate the short-term temporal stability of the method - requests were sent to the Bing API for a series of four monthly episodes (February to June 2015) for four urban locales within Sheffield (Carterknowle, Crookesmoor, Crosspool, Park Hill). These were chosen to reflect a variation in both geographic size (with Carterknowle and Park Hill being smaller locales) and expected stability (with Crookesmoor and Park Hill anticipated to be more transient in nature: Crookesmoor is an area with many students and Park Hill is undergoing substantial redevelopment).

Table 7.5 demonstrates that only 14-19% of data URLs re-occurred within all four monthly runs and that importantly a significant proportion of data (46-54%) was only present once. With such low replication in the content of the data, one might have concerns about the stability of the geography that such inconsistent data may produce. As Table 7.6 and Figure 7.4 illustrate, however, the outputted geographies of these urban locales appear very stable during the four months, with between 94-100% of (weighted) cells recurring in the same urban locale in all four monthly data runs.

Table 7.5 The number of times data records occurred within the four monthly runs (%)

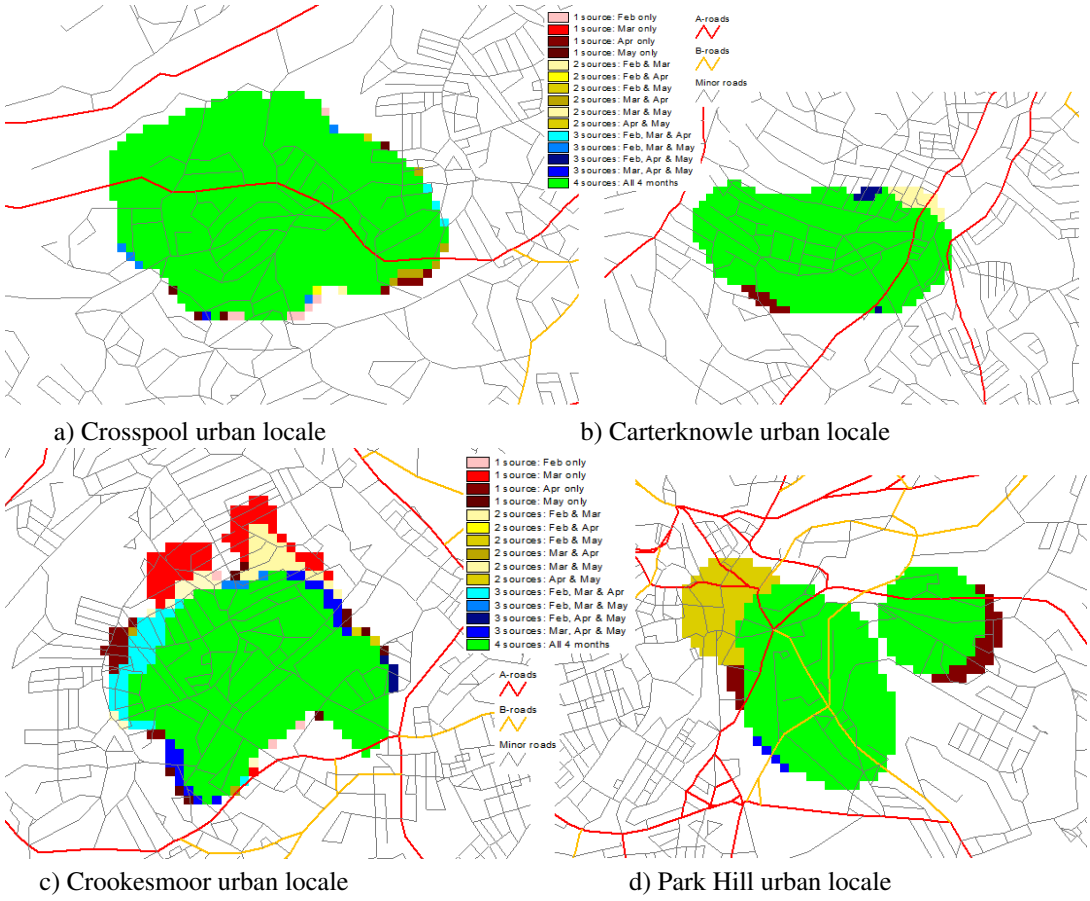
	Carterknowle	Crookesmoor	Crosspool	Park Hill
Only once	53.8	45.7	48.7	50.0
Twice	12.5	22.7	19.4	19.0
Three times	17.5	17.7	15.5	12.1
All four times	16.3	13.9	16.4	19.0

Even though the data themselves are likely to fluctuate month to month, reassuringly, the geography that the data produce remain highly stable. Thus, this raises the possibility of performing additional searches on the same set of postcodes within a limited time period in order to potentially further increase the number of data returns. Unsurprisingly, change is most likely to occur at the boundaries of urban locales. Using a probabilistic approach such as adopted here softens the impact of change as it is more likely to only represent a small proportion of weighted data. For example, the apparent extension northwards of Crookesmoor in Figure 7.4c accounts for only

Table 7.6 Level of agreement with monthly data for selected urban locales in Sheffield

	Months data available	Crookesmoor		Crosspool		Carterknowle		Park Hill	
		Cells (%)	Weighted (%)	Cells (%)	Weighted (%)	Cells (%)	Weighted (%)	Cells (%)	Weighted (%)
Feb only	1	0.5	0.3	0.9	0.4				
Mar only	1	10.1	1.2						
Apr only	1	2.1	0.1	0.6	0.1	1.9	0.1	6.1	1.0
May only	1	1.7	0.8	0.5	0.1				
Feb & Mar	2	2.3	0.2			3.3	0.0		
Feb & Apr	2			0.1	0.1				
Feb & May	2			0.1	0.0				
Mar & Apr	2	0.3	0.0	0.9	0.2				
Mar & May	2	4.3	0.7	0.1	0.1				
Apr & May	2	0.3	0.2					14.8	0.8
Feb, Mar & Apr	3	5.8	0.5	0.4	0.0				
Feb, Mar & May	3	0.7	0.1	0.6	0.7				
Feb, Apr & May	3	0.7	1.0			1.2	0.0		
Mar, Apr & May	3	3.3	0.9	0.1	0.0			0.7	0.0
All four months	**4**	**67.9**	**94.0**	**95.6**	**98.3**	**93.5**	**99.8**	**78.4**	**98.2**

Fig. 7.4 Urban locale data reproduction from four different monthly data runs



a) Crosspool urban locale

b) Carterknowle urban locale

c) Crookesmoor urban locale

d) Park Hill urban locale

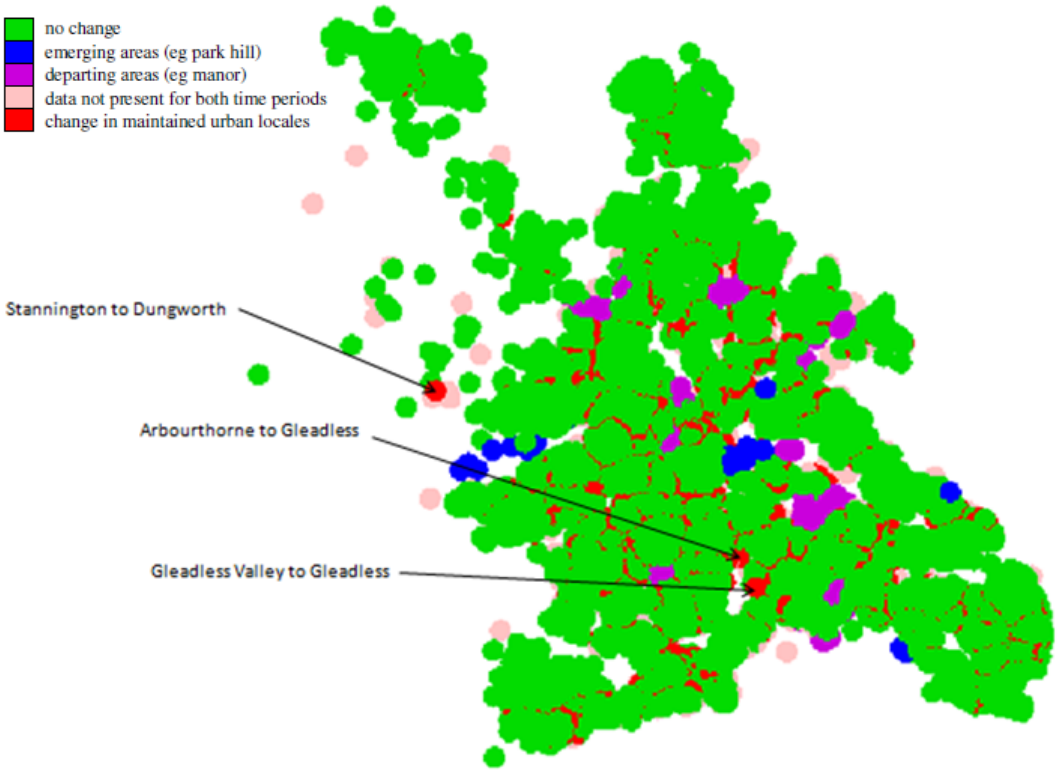
between 3-6% of the total data in the area. The question remains if it is possible to identify changes in longer-term trends of urban locale geography.

7.2.2 Long-term change: Two year span

Output for Sheffield was compared for two time periods: October 2012 and October 2014. The two year lag should give a valuable insight into the extent to which place names may alter through time. From the output of the KDE smoothings (as outlined in Chapter 5), nearly 87% of 50 metre cells in the study area were unchanged in the majority urban locale function between the two time periods (as shown in Figure 7.5 and Table 7.7). In this regard, these cells maintained the same urban locale with at least 50% of data (ie the majority) referring to the cell with the same urban locale name. This does not mean, however, that there was *real change* in urban locale of the remaining 23% of cells. Nearly 5% of cells reflect 'change' simply due to either 'new emerging' urban locales, or from areas no longer appearing within the data. These predominately represented small urban locales where data thresholds were passed resulting in areas being in/ex-cluded from the analysis. For example, Base Green

had 48 data points in 2012 but this fell to 38 by 2014, and hence, was excluded from analysis (as discussed within the rules of the method in Chapter 5). In contrast, the area Park Hill only produced 25 data returns in 2012 (as indicated back in Table 6.8) and having too few data points was not included within the analysis. By 2014, however, there were 47 references to Park Hill within postal addresses on the web, passing the threshold value of 40 and being included as an urban locale. Interestingly, Park Hill (a housing estate opened in 1961) has been undergoing phased redevelopment during this time and is not due for final completion until 2016. The lists of those urban locales passing the threshold values and either materialising or dematerialising between 2012 and 2014 can be found in Table 7.8. A further 4% of cells (predominately in areas of low population) were where data was not present for both years and hence cannot be directly compared. This leaves a 5% actual change in the majority urban locale name assigned to the cells and are shown coloured in red in Figure 7.5. These mostly consist of edge effects where small shifts in proportions within the data create a change in the majority urban locale. There are, however, some more sustained areas of change as labelled within Figure 7.5.

Fig. 7.5 Map of temporal change: majority function



To further explore changes at the individual urban locale level, linear grid regression analysis was undertaken between the output grids of each urban locale for the two different timestamps. Urban locales with the greatest differences in their geographic extents include Meadowhead ($R^2 = 43\%, p < 0.001$); Grimesthorpe ($R^2 = 68\%, p < 0.001$); Manor Park ($R^2 = 74\%, p < 0.001$) and Gleadless Valley ($R^2 = 76\%, p <$

Table 7.7 Temporal change: change in urban locale majority

Colour in Figure 7.5	% of total cells	Category
Green	86.6	no change
Blue	1.7	emerging areas (eg park hill)
Purple	2.8	departing areas (eg manor)
Light red	3.9	data not present for both time periods
Red	5.0	change in maintained urban locales

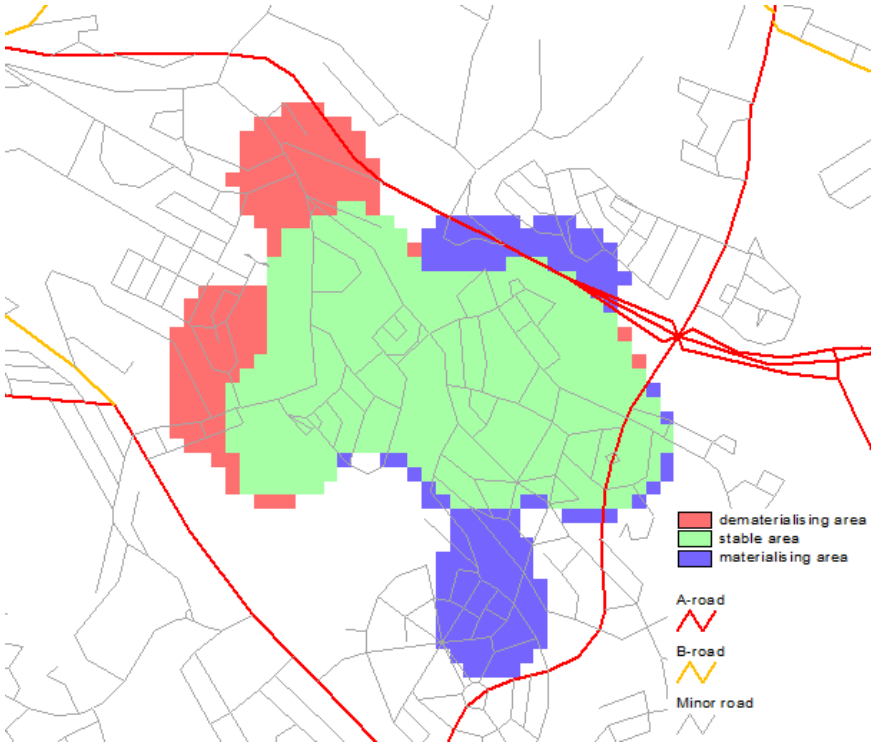
Table 7.8 Urban locale change, Sheffield 2012-2014

Materialising locales	Dematerialising locales
Park Hill	Abbeydale
Atlas North	Base Green
South Dore	Birley Carr
Dorehouse Industrial Estate	Brightside
Ridgeway	Charnock
Rivelin	Lane Top
	Manor
	Meadowhall Centre
	Middlewood
	Netherthorpe
	Parkway Industrial Estate
	Parkwood Springs
	Thorncliffe Park
	Wadsley Park Village

0.001). In contrast, the areas with greatest geographic similarity were Broomhill ($R^2 = 99\%, p < 0.001$); Worrall ($R^2 = 99\%, p < 0.001$); Woodhouse ($R^2 = 98\%, p < 0.001$) and Stocksbridge ($R^2 = 98\%, p < 0.001$). An example of change within the urban locale of Manor Park is also illustrated in Figure 7.6, where there appears to be a shift eastwards - being displaced in the west by Wybourn and forcing out the locale of the Manor in the east.

This section has shown that even with the monthly high turnover of data records found within postal addresses on the web obtained through the Bing API, the resulting geographies that can be formed from them appear highly stable in the short-term. The difficultly of urban locale data passing threshold vales and either being included or excluded from analysis is an issue for the method. Such problems, however, are relatively simply to rectify with the possibility of further rules to allow previously mapped locales to require fewer data records as a one-off basis for temporal change analysis. Such an adoption would aid to cushion the effect of shrinking locales with a less abrupt change in circumstances. A limitation of the method is that many web sites contain information that persists over long periods, or that is copied and duplicated from other web-pages. Such issues aside, the potential of mapping longer-term trends in urban locale extents has been demonstrated and adds weight to the benefit of running the proposed method for a complete national coverage.

Fig. 7.6 Example of temporal change in Manor Park urban locale, Sheffield



7.3 Tying other data to urban locale units

How feasible is it to tie other data (either point or area measures such as Census data) to the newly formed geographies of urban locale?

Once urban locale extents are generated, in order to use the new geographies to explore socio-economic issues such as those raised in Chapter 4, there is a need to tie other data to them. Areal interpolation is the process by which data with incompatible boundaries can be compared, with data for one zonal system being estimated from another. Goodchild and Lam (1980), Lam (1983) and Flowerdew and Openshaw (1987) give accounts of various methods of areal interpolation. The question of how to tie other data to these urban locale extents depends on the type of data being joined: point or polygon data types.

7.3.1 Joining point data

Point data are relatively straight-forward to assign to urban locale extents[13] and can be achieved by common functions within GISc software to assign a raster grid value to point data (for example such as the Extract Values to Points (Spatial Analyst)

[13] the start point should be the final KDE raster grids for each urban locale - being the most complete geographic coverage

function within ArcGIS). The points themselves may represent a vast array of social processes such as the number of crimes, occurrence of people with specific cancers, or applications of planning permissions. In this case, the raster values represent the percentage KDE density surface for each urban locale. The process needs to be automated[14] to iterate through all urban locale grids for the area of interest, providing a new column of attribute data for the point set with each urban locale comparison. The point data of interest can then be aggregated for each urban locale, using the weight of the KDE density surface. Using such a process, it is possible to count the number of (weighted) points falling within each urban locale. Weighting is based on the percentage of data in the cell allocated to a particular urban locale on the basis of the KDE grids. Consider the case where a point falls within a grid cell with the urban locale allocation of 60% to Crosspool and 40% to Crookes. Any attribute associated with the point data will be assigned weights on that basis (0.6 to Crosspool and 0.4 to Crookes).

This method was applied to geo-referenced Twitter data in order to investigate the varying sentiment within different urban locales. Figure 7.7 contains word clouds[15] from tweets for two contrasting urban locales in Sheffield for a single week in January 2014. Alternatively, point values could represent house prices which could then be averaged for urban locales[16], as shown in Table 7.9.

Fig. 7.7 Term frequency–inverse document frequency (tf–idf) of Twitter sentiment for two urban locales in Sheffield (positive sentiment words in green and negative in blue)



7.3.2 Joining polygon data

Joining polygon data to the urban locales is slightly more complicated and is best illustrated through use of an example: 2011 UK Census of Population data. Given that such Census data are predominantly tied to people's homes, the degree to which data can be tied to residential property is of interest. Relatively recent point data (for

[14] using a programming language such as Python

[15] word clouds were created using http://www.wordle.net/compose. Lists of positive and negative sentiment from Linguistic Inquiry and Word Count (LIWC) - http://www.liwc.net

[16] using freely available house price paid data from the Land Registry - http://landregistry.data.gov.uk

Table 7.9 House Prices (June 2012-July 2013) for selected areas in Nottingham

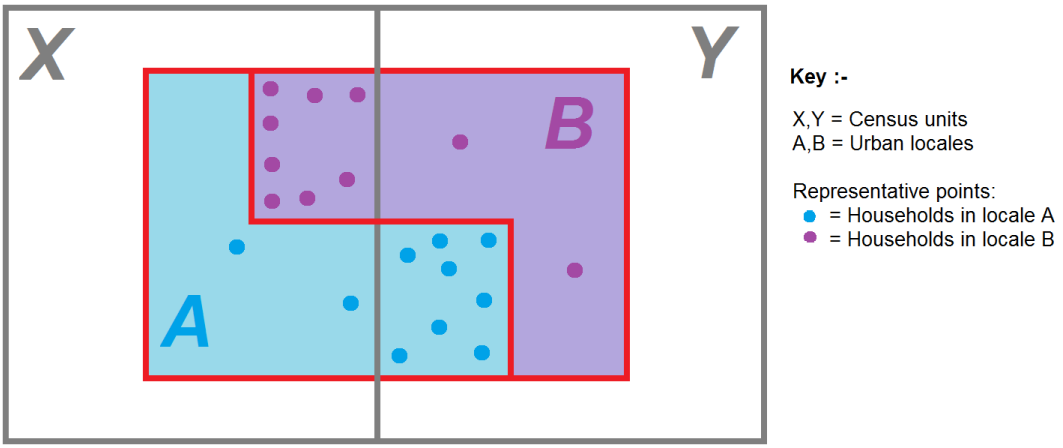
Urban Locale Name	Number of Transactions	Average Price (£)
The Park	161	276,303
Mapperley Park	104	211,438
Clifton Village	10	197,326
Wilford Village	14	189,108
Wollaton Park	44	178,890
Clifton Grove	6	164,585
Mapperley	307	150,815
Hockley	49	136,434
Canning Circus	79	129,332
Carrington	55	125,753
Dunkirk	27	113,865
City Centre	241	105,031
Sherwood Rise	53	101,826
Basford	211	99,702
Beechdale	44	99,271
Aspley	160	98,682
Old Basford	111	96,615
Bobbers Mill	61	90,823
Clifton	189	88,739
Radford	201	86,270
Hyson Green	90	84,998
Clifton Estate	129	84,377
Forest Fields	72	83,852
St Anns	141	81,864
Sneinton	133	72,298
Broxtowe	44	63,171

example representing individual buildings or postcodes) provide a rich datasource to facilitate point-in-polygon areal interpolation techniques. Point-in-polygon methods assign the source data to representative points which thereby 'summarise' the area data (Bracken and Martin, 1989). This has great value because the more accurately one can represent the underlying population, the finer grained the zonal region, and therefore, the more applicable overall results should be (Brindley et al. 2005).

Whilst it would be perfectly possible to estimate data based on the areal extent of each grid cell that fall within the urban locale boundary (termed area weighting or polygon overlay), given that Census data relate to the population, the use of representational points at a fine spatial scale (such as individual property level) to reflect the population's location have been found to improve estimates (see Brindley et al. (2005)). This is especially true in areas where the population are sparse, as Census data are centred on the populations' location rather than being skewed by the large areal extent of unpopulated spaces. Consider a simplified scenario shown in Figure 7.8. where there are two urban locales (A and B) falling within two Census unit (X and Y). Whilst

the majority of the *area* of urban locale A falls within Census unit X. In contrast, the majority of the *population* of the locale reside within Census unit Y.

Fig. 7.8 Areal interpolation: area weighting and point-in-polygon techniques



Once urban locale areas have been defined it is then possible to apply point-in-polygon areal interpolation techniques to tie official Census data (with different, incompatible boundaries) to the urban locale units. The definition of urban locale used is purposefully broad and as such encompasses many non-residential areas including business parks and industrial estates. For many purposes, data should not be estimated for such non-residential areas. For example, Census of Population data relate predominately to households and thus to residential areas only. Although the geographic units that Census data are provided for, such as Census Output Areas (OAs)[17], assume complete geographical coverage and so do not make distinctions between residential and non-residential areas. For such reasons, the easiest approach is to manually remove non-residential urban locales and re-estimate the percentage of data coverage KDE grids, thus, ensuring only residential urban locales are assigned residential related Census data. Once accomplished, the main data assignment method can be undertaken.

Firstly, the source data, urban locale KDE grids, are tied to point data indicative of the underlying population (ideally expressed by individual residential addresses but alternatively by postcodes, as used in the example in Figure 7.9a for simplicity) – termed the representative points. The values assigned to the representative points are weights of the density of the urban locale, for example if a particular representative point is located in a grid cell with a KDE value (with the 300m search radius) of 66 to belonging to Urban Locale A and a value of 33 to Urban Locale B - then it would

[17] the lowest geographical level at which census estimates are provided in England and Wales and containing approximately 130 households - http://www.ons.gov.uk/ons/rel/census/2011-census/population-and-household-estimates-for-wards-and-output-areas-in-england-and-wales/stb-population-and-household-estimates-for-small-areas-in-england-and-wales.html

be assigned proportionally with 0.66 to Urban Locale A and 0.33 to Urban Locale B. This step is highlighted within Figure 7.9b.

Secondly, using the common GISc technique of 'point-in-polygon' analysis, it is possible to assign each representative point to the target zone (OAs) it falls within (Figure 7.9c). Once undertaken, it is then feasible to aggregate the urban locale density values within each OA by urban locale and calculate the ratio of representative points that belong to each urban locale (Figure 7.9d). The census data for each OA can then be distributed according to the urban locale likelihoods (Figure 7.9e). Finally, data can be aggregated for each urban locale (Figure 7.9f). Using such a procedure any variable within the Census can be translated from Census units to urban locales.

Sample output from such a process is evidenced in Figure 7.10, whilst change over time is illustrated in Table 7.10[18]. In those cases where urban locale coverage is not complete (which is particularly likely in smaller towns) it would be possible to also include a 'Rest of town/city' category along urban locale data, in order to ensure that totals of the input data sum to that for the settlement as a whole.

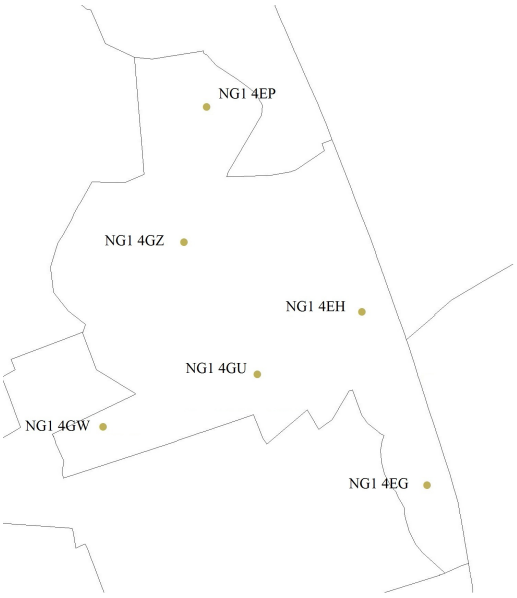
Table 7.10 Change in Census 2001 to 2011 for selected areas in Sheffield

	Crookes			City centre		
	2001	2011	Change	2001	2011	Change
All Households	4,518	4,526	8	2,437	5,727	3,289
Households with no cars or vans (%)	32.3	29.8	-2.4	65.4	67.8	2.4
Households with one car or van (%)	47.2	47.5	0.3	27.5	26.8	-0.7
Households with two cars or vans (%)	16.8	18.9	2.1	5.1	4.3	-0.8
Households with three cars or vans (%)	2.9	2.9	0.0	1.5	0.7	-0.8
Households with four or more cars or vans (%)	0.8	0.9	0.1	0.5	0.3	-0.1
All cars or vans in the area	4,195	4,424	229	1,083	2,256	1,173
Ave. number of Cars or Vans per household	0.93	0.98	0.05	0.44	0.39	-0.05

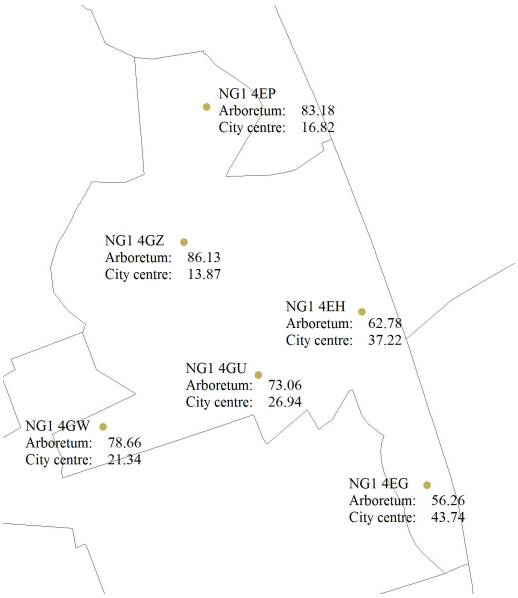
Such output demonstrates the importance of creating urban locales so that other data can be tied to them in order to perform a vast range of socio-economic research and analyses. Whilst some may argue that care should be taken with such an approach for polygon units, due to ecological fallacy with Census data being treated as homogeneously spread within the Output Areas, the scale of the units involved severely diminishes the issue. The allocation is from small Census Output Areas (median size of 5.4 hectares in the study area of Nottingham and Sheffield) to larger urban locales (151.5 hectares for the locales in Nottingham and Sheffield). Hence, on average for the case areas of Nottingham and Sheffield there are over 15 Output Areas within an urban locale. The case of ecological fallacy is less of a concern, due to the extent of homogeneity, when a number of small units are aggregated within a larger entity, than if estimations are made between two similar sized units or worse still, measures

[18] Note the geographic areas of urban locales remain static in this comparison as historical perspectives cannot be collected as discussed previously in this chapter

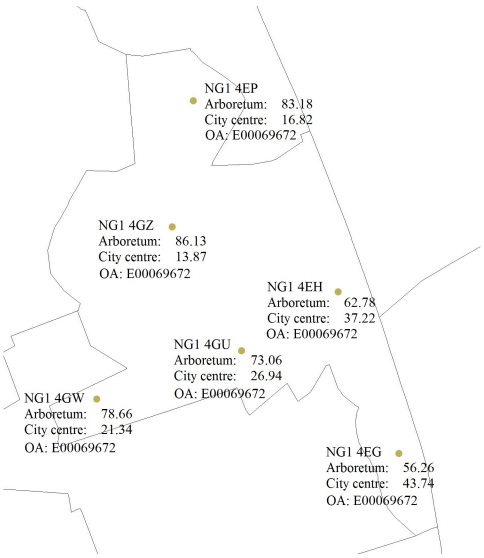
Fig. 7.9 Areal interpolation example for Output Area E00069672



a) Postcode Representative points



b) Assign KDE densities



c) Assign points to OAs

Output Area	Postcode	Arboretum	City centre
E00069672	NG1 4EP	83.18	16.82
E00069672	NG1 4GZ	86.13	13.87
E00069672	NG1 4EH	62.78	37.22
E00069672	NG1 4GU	73.06	26.94
E00069672	NG1 4GW	78.66	21.34
E00069672	NG1 4EG	56.26	43.74
Total		440.07	159.93
Ratio		**0.73**	**0.27**

d) Aggregate over each OA

Official census data:	Total residents	Males	Females
OA E00069672	275	159	116

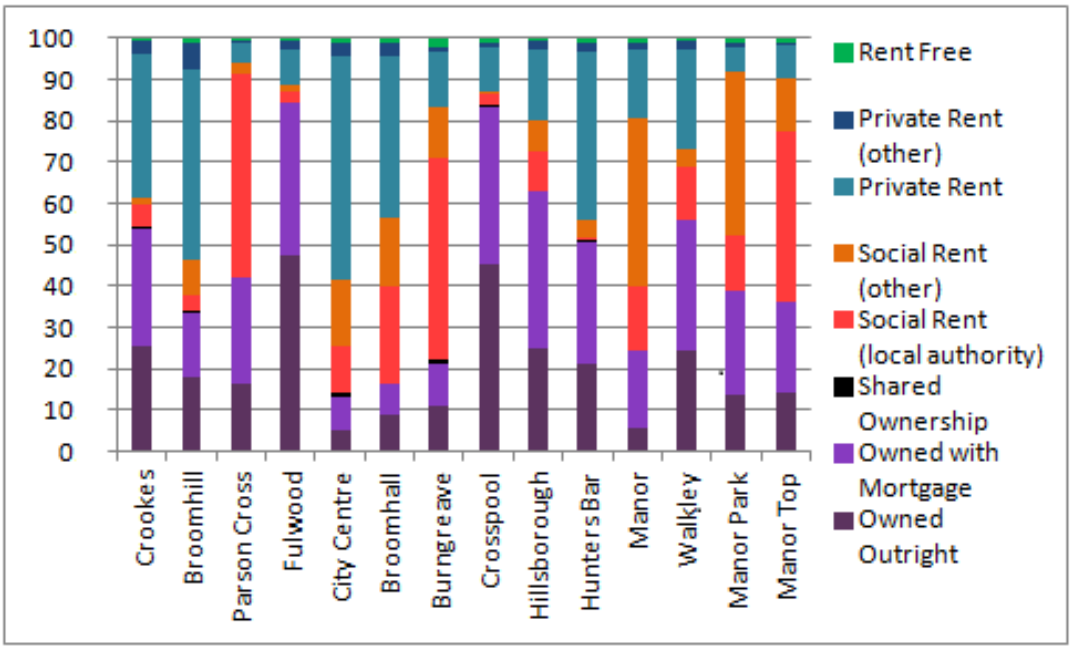
	Total residents	Males	Females
Arboretum	202 (0.73*275)	117 (0.73*159)	85 (0.73*116)
City centre	73 (0.27*275)	42 (0.27*159)	31 (0.27*116)

e) Assign Census data to OAs

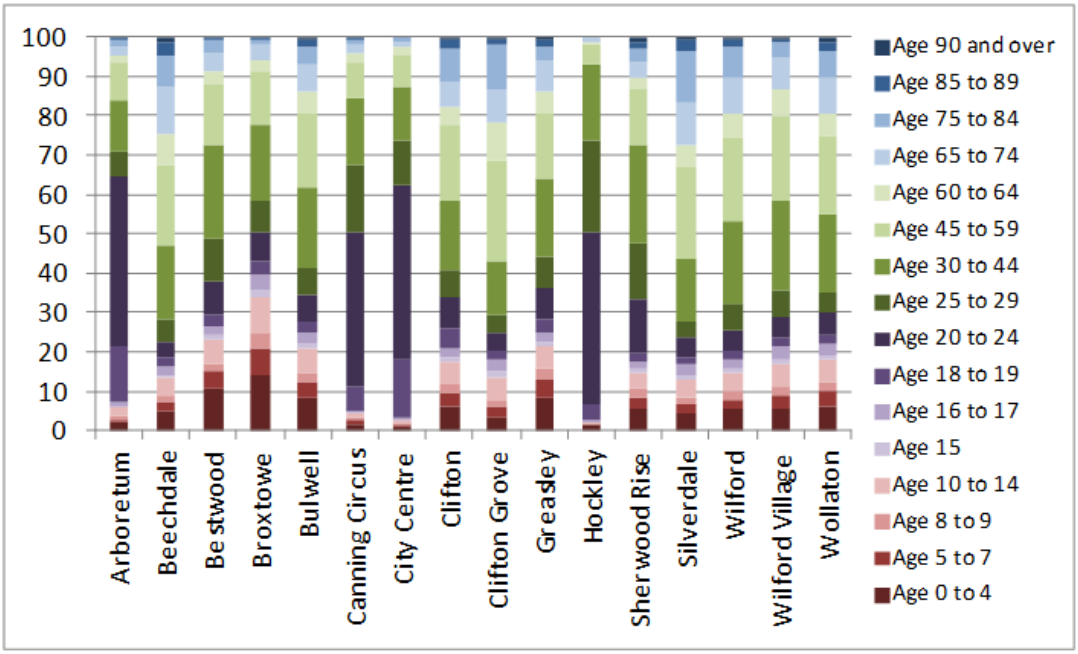
	Total residents	Males	Female
Arboretum	4,083	51.20%	48.80%
City centre	6,216	55.10%	44.90%

f) Aggregate OAs over urban locales

Fig. 7.10 Example Census of Population output



a) 2011 Census data for selected urban locales in Sheffield: Housing Tenure (%)



b) 2011 Census data for selected urban locales in Nottingham: Age Composition (%)

for a large unit are split between smaller output units. Given that Census Output Areas are the smallest statistical units for which Census data are available, there are no alternatives at such a fine level of spatial resolution. It is highly likely that aggregation from Census Output Areas to urban locales is far more in line with how residents perceive the named area representation than using other larger geographical spatial units such as Wards. Comparisons of homogeneity within various units of analysis is discussed within the next section.

7.4 Homogeneity

How homogeneous are the urban locales in terms of socio-economic characteristics?

One formulation of the concept of 'neighbourhood' referred to in Chapter 4 is based on regionalisation where small units are aggregated together based on attempting to form *homogeneous* units (for example see Martin (1998)). Knox and Pinch (2009) explicitly define 'neighbourhood' as "Territories containing people of broadly similar demographic, economic and social characteristics". Similarly, the OSM definition of 'suburb' includes a recognition that areas should have a distinct identity[19] - again implying some measure of homogeneity within but differences between other 'suburbs'. This research question seeks to explore the extent to which urban locales might be considered homogeneous in terms of their socio-economic characteristics. It is worth remembering that without the generation of suitable urban locale areal units it would be impossible to test any such hypotheses. Thus, again further demonstrating the requirement for a national set of urban locales.

The method outlined in the previous section of this chapter makes it possible to tie UK Census of Population data to urban locales. Using a single socio-economic variable from the 2011 UK Census of Population, it is possible to illustrate the degree to which data vary at a smaller spatial scale (for example Census Output Areas - OAs) within the larger urban locale units. For homogeneity, one might expect to find greater variation *between* urban locales than *within* urban locales. Figure 7.11 demonstrates the degree of variation at the OA level in the percentage of people aged 30-44 within Sheffield urban locales. The prominent visual impact in the figure lies not with the variation *between* urban locales (the median percentage value is portrayed by the bold black lines in each box-plot) but with the high degree of variation *within* locales, evidenced by the large interquartile ranges (shown by the lengths of the red boxes). Boxplots were generated for a range of variables for both Sheffield and Nottingham, a sample of which can be found in Appendix G and output was broadly similar. The key message

[19] http://wiki.openstreetmap.org/wiki/Tag:place%3Dsuburb

being that Nottingham and Sheffield urban locales do not appear at face value to be strictly homogeneous in terms of their socio-economic characteristics. The argument remains however, that some boundary units of analysis might be more homogeneous than others.

A more comprehensive examination of the homogeneity of Nottingham urban locale units needs to measure output between different sets of boundary data. It is possible to measure the homogeneity of Census data in urban locales compared against three different units of analysis commonly used with Census data: Lower level Super Output Area (LSOA), Middle level Super Output Area (MSOA) and Ward data. These datasources are referred to as the source datasets. Analysis was undertaken for the administrative area of the district of Nottingham. Table 7.11 contains summary information relating to the average size of these units of analysis within Nottingham district. Any assessment of homogeneity can also only be made by reference to some data variable. For this work, the homogeneity of units was tested against a range of socio-economic data drawn from OA data for 387 variables from the 2011 Key UK Census of Population. The OAs could then be aggregated to each of the source datasets and the variation within the OAs for each source assessed.

Table 7.11 Size of source datasets used within testing homogeneity of units

Source dataset	n	Average number of OAs per unit
Lower level Super Output Area (lsoa)	361	5.2
Urban locales	109	17.4
Middle level Super Output Area (msoa)	74	25.6
Wards	69	27.4

Although statistical tests of homogeneity exist (for example Levene's, Brown–Forsythe's and Bartlett's tests or F-test of equality of variances), they are computationally intractable for large datasets such as those used here to explore the variation of socio-economic Census data within urban locales[20]. In this light, a more informal exploration of the variations within the source datasets was undertaken. One might consider that a homogeneous unit of analysis may be of most analytical value if the variation is greater between the different units than the variations within each unit. The variation *within* units is simply measured by the standard deviation (square root of the variance) of Output Area Census values within each target unit. The variation *between* units is measured as the standard deviation of mean values for each unit within the datasource and calculated for each Census variable. For example, firstly the percentage of people aged 30 to 44 is measured for each urban locale as the weighted mean of Output Area values (weighted by the number of people within each urban locale contained in the

[20]For example the Levene's test within SPSS is limited to 50 groups - thus it is only possible to explore the variation of any variable within 50 different units from a source dataset. This represents only 14% of LSOAs within the case study area and results were not robust

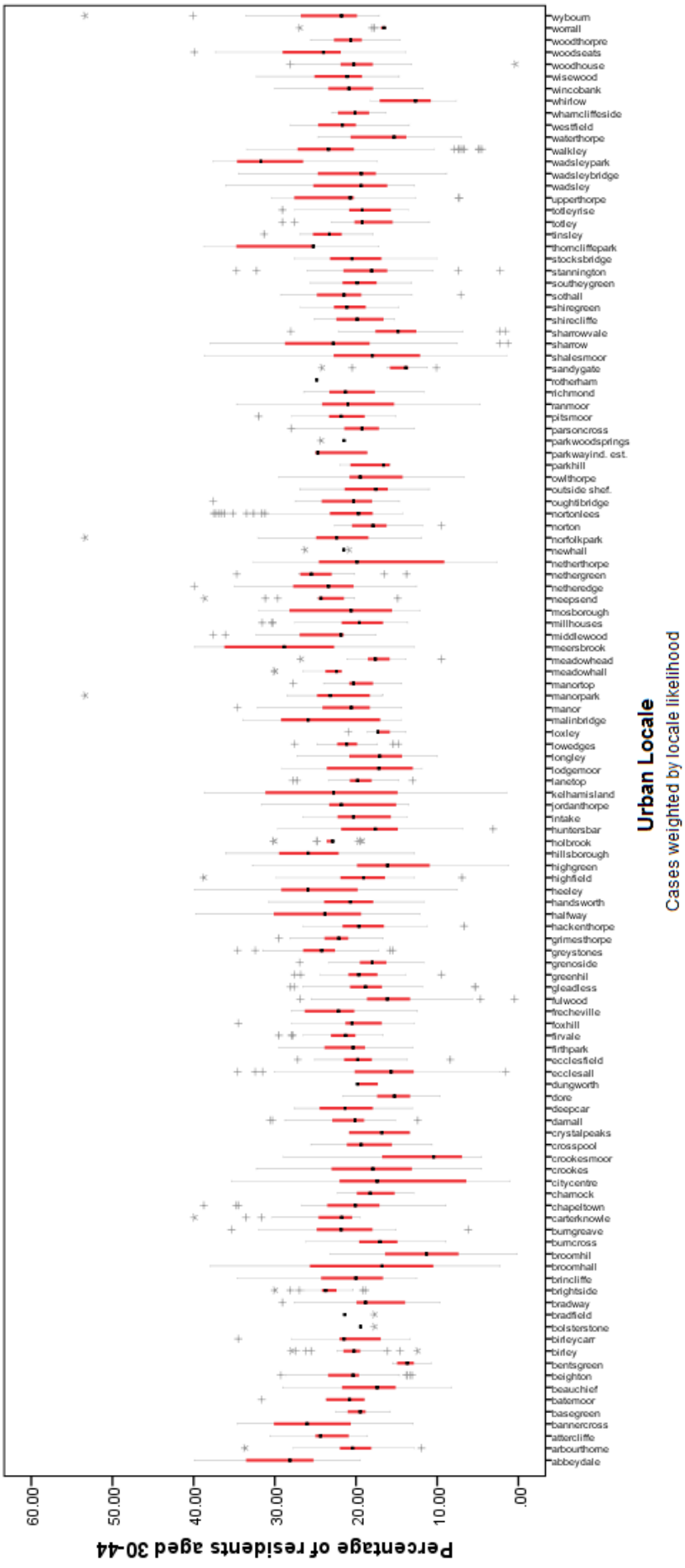
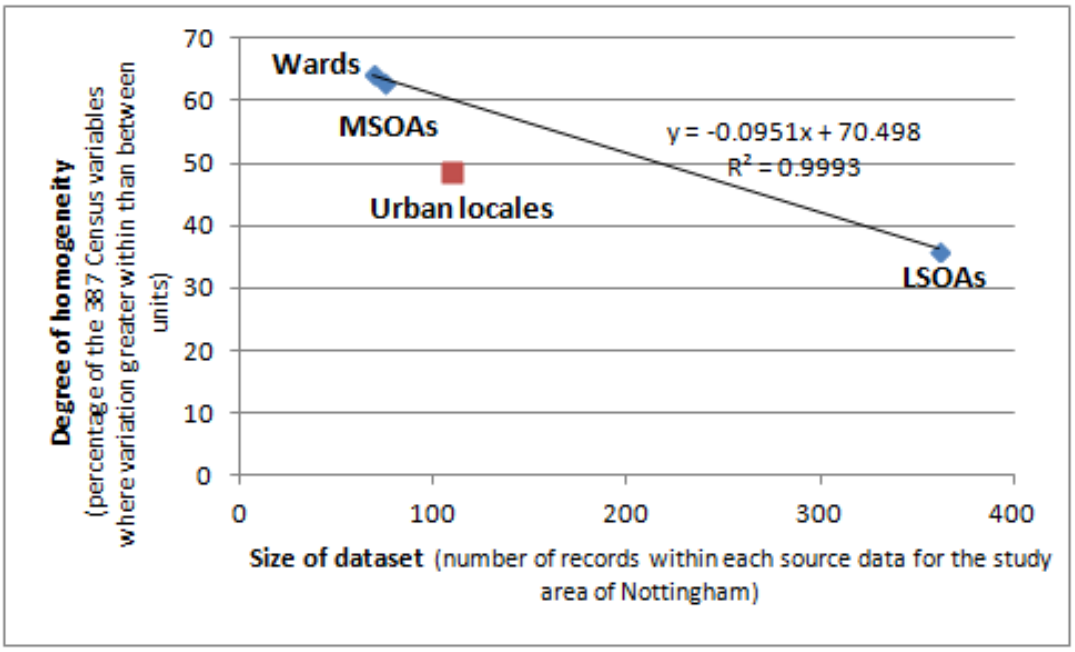


Fig. 7.11 Example of homogeneity for a single Census variable (Percentage of people aged 30-44) within Sheffield urban locales

OA). Secondly, the variation between units is calculated as the standard deviation of urban locale values for the variable of interest.

Once such calculations have been made - it is possible to use these to derive the proportion of variables (from the total of 387) where the variation within the unit is greater than variation between units. The lower this figure the more homogeneous the source unit might be considered to be. The proportion of the 387 Census variables where variation was greater within than between units was 64% for Wards, 63% for MSOAs, 49% for urban locales and 36% for LSOAs. One would, however, expect smaller areal units to be more homogeneous simply due to there being fewer aggregating OAs. To overcome this, assessment is made against the size of the source datasets (measured by the total number of data records within each source dataset for the study area of Nottingham[21]). Figure 7.12 demonstrates, one might expect that with 109 units (as in the Nottingham district case study), on average 60%[22] of the 387 Census variables would exhibit greater variations within than between units. In contrast, however, the actual number for urban locales is just under 50% - demonstrating that the units appear more homogeneous (in terms of socio-economic characteristics) than expected, given the size of the units of analysis and the level of homogeneity within LSOAs, MSOAs and Ward data.

Fig. 7.12 Comparing variations of homogeneity with differently sized areal units



[21] as indicated within Table 7.11 - for example there are a total of 361 LSOAs in Nottingham but only 109 urban locales for the same geographic area

[22] based on the assumption of the expected degree of homogeneity and size of dataset for urban locales as depicted by the straight line relationship within Figure 7.12

Whilst a great deal of variation in terms of socio-economic data exists within Nottingham urban locales and the units should not be seen as strictly homogeneous areas, these unique areas of interest appear more homogeneous than other administrative data alternatives (such as LSOAs, MSOAs and Ward data). The extent of heterogeneity within Nottingham urban locales gives support to the definition of urban locales based on identification of areas sharing the same *name*, as opposed to those based on some measurement of homogeneity of units. Whilst this exploratory research highlights a number of key points, future work would be desirable to extend the geographic scope and further investigate the degree of homogeneity in socio-economic characteristics within urban locale units.

7.5 Comparison with underlying population geography

How well does the frequency of urban locale data compare to the underlying population? Are there unexpected areas with low/high data returns?

Exploring the relationship between the urban locale and underlying population data may help us better understand the locale data and how it has been formed. It may also be able to help identify types of geographic area where data returns are particularly high or low. A series of linear grid regression analyses were undertaken between the numbers of web pages containing urban locale references within postal addresses for the case study area of Sheffield (the urban locale data returns) and numbers of domestic and non-residential property[23]. Output was generated for the case study cities of Sheffield and Nottingham where the author had a detailed knowledge of the spaces, essential for understanding the social processes that might influence the findings.

It should be clarified that urban locale information relating to domestic addresses (with the identification of domestic defined through the Royal Mail data) are unlikely to originate from the residents themselves. In fact, whilst it is possible for some resident views to be generated (for example through selling items on web based classified advertisements, people finding and dating sites), the majority of such data derives from non-resident sources (for example estate agents or through business directories where small companies, such as plumbers, use their residential homes for their registered office address).

[23] as contained within CodePoint - accessed through the Digimap service. Royal Mail define addresses categorised as domestic and non-residential property

7.5.1 Comparison with all data returns

The first regression explores the total number of data returns in Sheffield - regardless of if urban locale information were present. This could be thought of as the general visibility of the postal addresses on the web (or web-visibility). A strong relationship was found (R^2 = 87.5%) where the dominant influence was close to 10 data returns for each non-residential delivery point[24] and just under 1 data return per residential delivery point, as indicated by the linear grid regression equation below and in Table 7.12:

$$D = 0.0345 + 0.7258 * [res300] + 9.7136 * [nres300]$$

Calculated for each grid cell, where:

D = the number of total data returns within a 300m radius from the grid cell (regardless of urban locale content within the data returns)

res300 = average number of residential delivery points within 300m, for the grid cell

nres300 = average number of non-residential delivery points within 300m, for the grid cell

Table 7.12 Regression output: relationship between all data and delivery points: Sheffield

Variable	Coefficient	Std. Error	t-Value	Significance
[Intercept]	0.0345	0.0026	13.08	<0.00001
[res300]	0.7258	0.0012	610.37	<0.00001
[nres300]	9.7136	0.0134	726.23	<0.00001

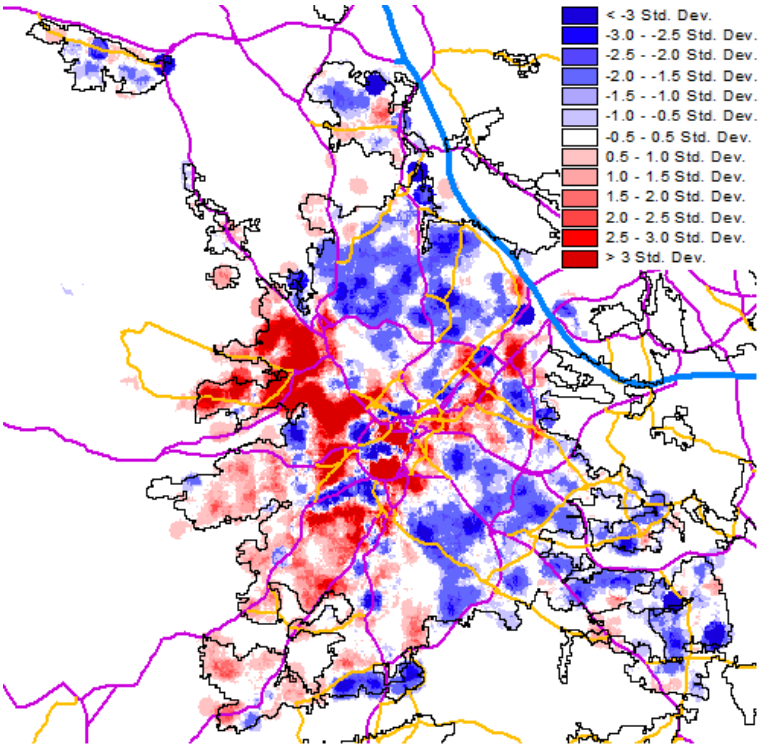
The residuals from this relationship are illustrated in Figure 7.13 and appear to indicate a higher degree of web-visibility for postcodes in the more affluent western side of Sheffield along with a slightly less pronounced industrial corridor spreading through Tinsley from the city centre to the M1. A similar relationship was found for Nottingham (see Appendix H for full details).

7.5.2 Comparison with urban locale data returns

Similar analysis can be undertaken for those data records that were found to explicitly contain an urban locale name within the address. The relationship between the numbers of data containing urban locale names and the underlying delivery points, however, is not as strong as the one found above - for all data (as shown in Appendix H). This is because of the influence of data within Royal Mail defined 'Dependent Localities'

[24] a delivery point is a premise level addresses (or commonly, in other words a letterbox) as defined within Royal Mail data

Fig. 7.13 Residuals for the relationship between the total number of data returns and numbers of residential and non-residential properties



(refer back to Chapter 5) where one would expect to find references to urban locale names. Once accounted for (by adding variables for delivery points from Dependent Localities separately), the variance explained increases (R^2 = 72.7%) and the following equation derived:

$$UD = 0.0405 + 0.1120 * [res300] + 0.4567 * [nres300] + 0.3912 * [dlres300] + 2.9630 * [dlnres300]$$

Calculated for each grid cell, where:

UD = the number of data returns containing an urban locale name within the postal address (smoothed to 300m)

res300 = the average number of residential delivery points within 300m

nres300 = the average number of non-residential delivery points within 300m

dlres300 = the average number of residential delivery points within 300m in cells containing Dependent Locality postcodes

dlnres300 = the average number of non-residential delivery points within 300m in cells containing Dependent Locality postcodes

As shown within the equation and Table 7.13, there were approximately 0.5 references to an urban locale name for each non-residential delivery point, 0.1 references for a residential delivery point, an additional 3 references for non-residential property in

Dependent Locality areas and an additional 0.4 references for residential property in Dependent Locality areas. The residuals from the model are displayed within Figure 7.14 and show some similarity to those from the residuals for all data records (shown back in Figure 7.13). Areas to the west of Sheffield along with more peripherally located areas appear more likely to cite an urban locale name within their address, when posting information on the web. Perhaps understandably, some of the largest negative residuals appear concentrated in the city centre (where one might expect fewer people to specifically refer to the 'City Centre' within the postal address between street and settlement elements). To explore such patterns, the residuals can be aggregated to the urban locales as shown within Table 7.14. From the table, less affluent areas appear less likely to generate data points. More surprisingly is the reluctance of data carrying references to Westfields, Waterthorpe and Crystal Peaks to the south-east of the city. There is also a strong contrast between Netherthorpe (north-west of the city centre) - which are less likely to have data containing references to the locale name than the neighbouring Upperthorpe. Further investigations of the reasons behind such factors are outside the scope of this work. Again, similar relationships were found for Nottingham (see Appendix H for full details).

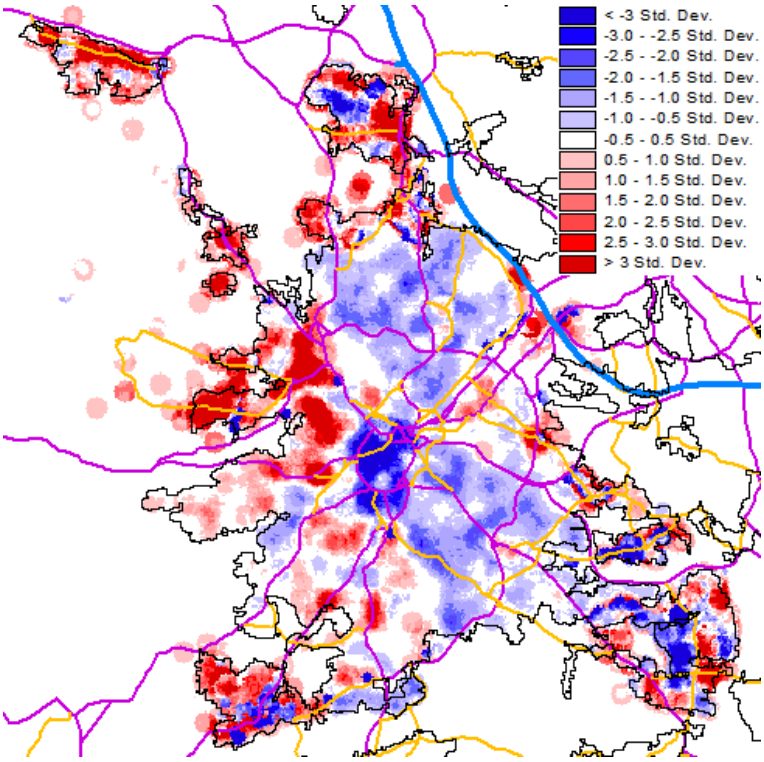
Table 7.13 Regression output: relationship between named urban locale data and delivery points: Sheffield

Variable	Coefficient	Std. Error	t-Value	Significance
[Intercept]	0.0405	0.0009	44.79	<0.00001
[res300]	0.1120	0.0004	254.71	<0.00001
[nres300]	0.4567	0.0048	95.33	<0.00001
[dlres300]	0.3912	0.0010	391.36	<0.00001
[dlnres300]	2.9630	0.0168	176.30	<0.00001

This section has illustrated the strong associations that are found between the underlying population distributions (as measured by numbers of residential and non-residential delivery points) and address information appearing on the web. The term web-visibility was introduced (the number of all occurrences of postcode data appearing on the web - irrespective of whether the information contains a reference to an urban locale name or not) and appears of considerable interest when compared against the underlying population distribution. The spatial pattern of residuals from this relationship provides a powerful and insightful perspective on the geographic footprints of information on the web. The lack of access to the web in deprived areas is well documented (for example see Crang et al. (2006); Department of Trade and Industry (2000); Gardner and Oswald (2001)). Additionally, this work highlights the notion that deprived areas/places themselves also appear less prevalent on the web (with a lower web-visibility).

Exploring the relationships with the underlying population also serves to demonstrate one limitation of the method - in that references to Dependent Localities may be merely a reflection of Royal Mail's official view, rather than measures from more desirable

Fig. 7.14 Residuals for the relationship between named urban locale data and numbers of residential and non-residential properties



sources, such as residents, local businesses and so forth. The use of Dependent Localities within Royal Mail addressing practices, however, is relatively uncommon, accounting for just over five percent of postcodes in the Sheffield case area. This is in strong contrast to the work of Almadani et al. (2013) where the views would entirely constitute those of official Royal Mail data. The strong association between the underlying population distribution and volume of data records containing an urban locale name is encouraging and, with a few exceptions (such as the city centre), tend to reflect a similar pattern in the residuals found for all data records (web-visibility).

7.6 Differing perspectives

To what extent do the urban locale spaces nest, creating differing perspectives? Can the hierarchical nature of space be examined?

Importantly, the method as outlined within Chapter 5 allows for the investigation of those areas where perceptions vary and *different* people or organisations may call the area by *different* names. Underlying the principle of the work is that people may well have varying opinions of what any given geographic space is called. Exploring this further is a critical aspect of the requirement of urban locales as vague and potentially

Table 7.14 The fifteen highest and lowest average regression residuals when aggregated to urban locales

Urban locale	Average residual score
Westfield	-0.924
Crystal Peaks	-0.874
City centre	-0.736
Waterthorpe	-0.514
Netherthorpe	-0.443
Wybourn	-0.435
Jordonthorpe	-0.427
Norfolk Park	-0.383
Woodhouse	-0.365
Highfield	-0.360
Arbourthorne	-0.357
Batemoor	-0.356
Gleadless Valley	-0.348
Longley	-0.344
Shalesmoor	-0.337
...	...
Parkwood Spring	0.398
Holbrook industrial estate	0.431
Grenoside	0.467
Deepcar	0.478
Oughtibridge	0.487
Loxley	0.554
Stocksbridge	0.713
Walkley	0.766
Hillsborough	0.784
Crosspool	0.785
Broomhill	0.800
Meadowhall	0.815
Worrall	0.869
Chapletown	0.997
Stannington	1.147

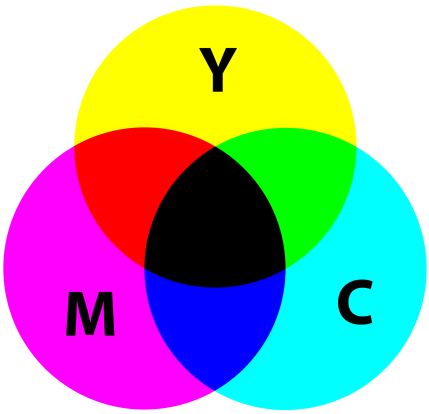
overlapping areas, as opposed to simply defining discrete objects. This section seeks to explore the extent to which differing perspectives of urban locale are formed.

7.6.1 Differences within a specified urban locale

It is possible to explore the differing perspectives between three urban locales using simple cartographic approaches. Applying Cyan-Magenta-Yellow-Key (CMYK) composites of the KDE surfaces for urban locales is a useful indicator of the extent to which neighbouring locales may overlap. Such a method colours three urban locales according to the mix of different locales within each cell of the study area. For example, as shown within Figure 7.15, areas in each of the three locales with no mixing (where the urban locale is not contested) will be shown in one of the three principle colours: Cyan, Magenta, or Yellow. Where differing perspectives occur, however, a mix of the pertinent principle colours are applied as indicated within Figure 7.15, with

black (key) being where all three urban locales co-exist to an equal degree. Using such an approach it is easy to identify urban locales with tighter boundaries, without mixing (such as between the London areas of Bloomsbury, Fitzrovia and Soho in Figure 7.16a) and those with differing levels of mixing. There are places, such as China Town in Figure 7.16c, where the entire area of the locale is considered to completely nest within other urban locales, with no parts considered to solely constitute only China Town.

Fig. 7.15 CMYK color model[25]



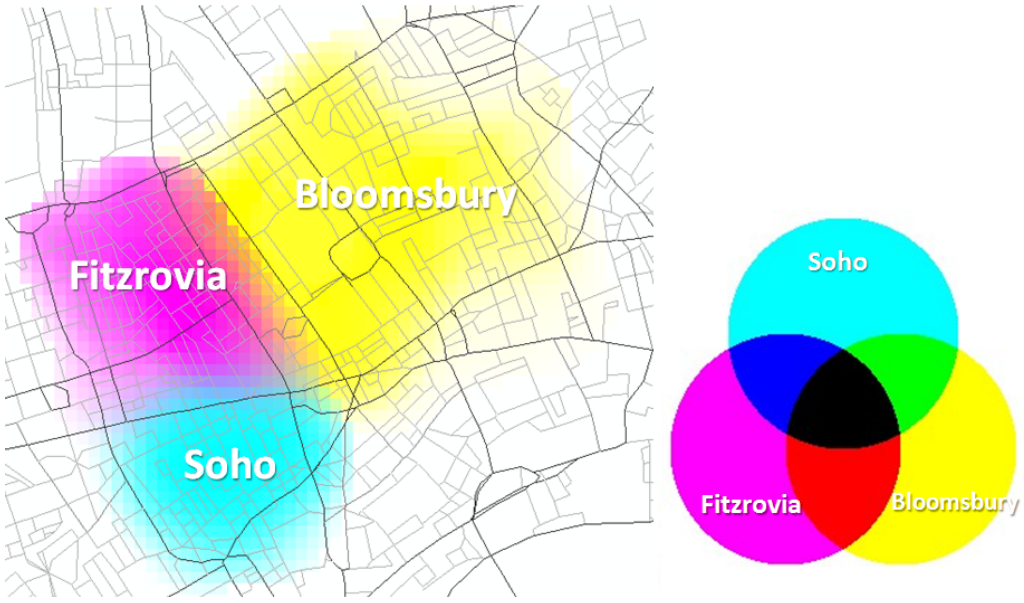
Whilst the benefits of such a technique is clear, there is one substantial limitation in that it is only possible to compare the geographies of three different urban locales at any one time. Whilst this would not be an issue for the majority of cases, there may well be occasions when more than three urban locales co-exist within the same local space. The extent to which this may occur will be discussed within the following section.

7.6.2 The extent of differing perspectives

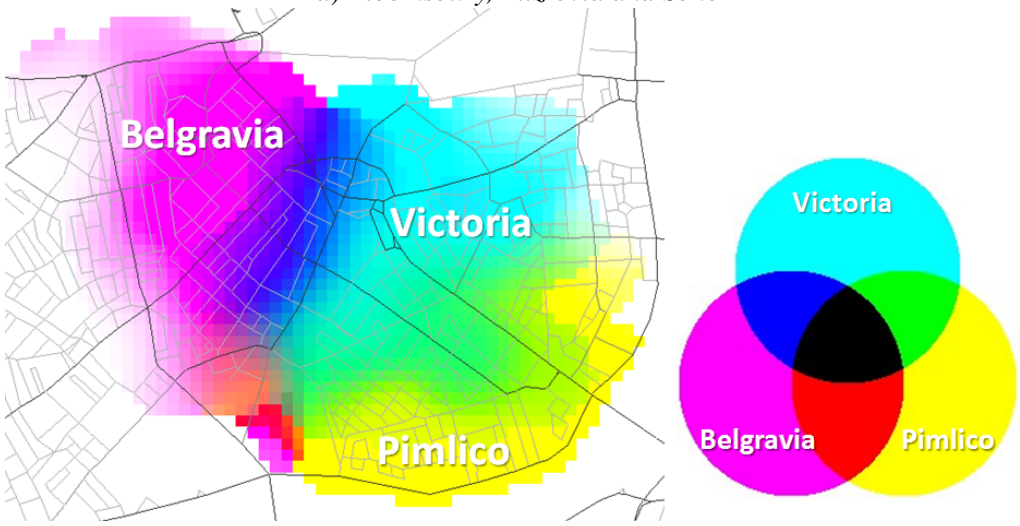
Determining the exact degree to which the same spaces might be considered to contain multiple urban locales is not as straightforward as it first seems. It would be possible to calculate the number of 50 metre cells outputted by the method (as described in Chapter 5) where there were more than one urban locale. Although validation (Chapter 6) supports the output of the method, the degree of *KDE smoothing* (even when only including populated cells) may be slightly exaggerated and it cannot be certain that areas with multiple names are not entirely due to the smoothing process. Such an approach is likely to over-estimate the degree of differing perspectives. Alternately, it is possible to estimate the numbers of cells where the cleaned *input data points* (prior to any KDE smoothings) provide differing views on what the space may be termed. Essentially, this second approach represents the degree to which variation

[25] http://en.wikipedia.org/wiki/CMYK_color_modeln

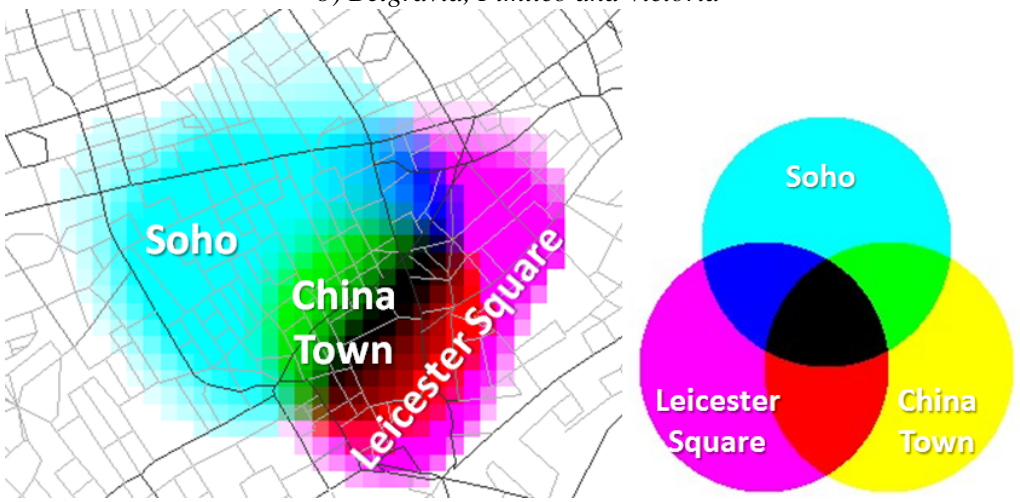
Fig. 7.16 Cyan-Magenta-Yellow-Key (CMYK) composite for example urban locales in London



a) Bloomsbury, Fitzrovia and Soho



b) Belgravia, Pimlico and Victoria



c) Soho, China Town and Leicester Square

occurs within the same space, prior to smoothing, and thus demonstrates the minimal extent of vagueness within the data that discrete boundaries must inherently miss. Such an approach represents the minimal extent as it is reliant on the raw input data which is known to be a sample and might exclude perspectives on the following basis: (1) some cells will return no data at the raw 50m grid cell level and (2) even cells with data might include additional opinions concerning the name of the space if all viewpoints were collated. Whilst such a method avoids any issue regarding smoothing, the degree of differing perspectives will be under-estimated due to the fact that input data are only a proportion or sample of total urban locale perceptions. This said, it is clear that the true degree of differing perspectives will lie somewhere between these two estimates.

For the case study areas of Sheffield and Nottingham, as shown within Table 7.15, areas that are seen as having different perspectives, with multiple urban locale names, range considerably depending on the method used for their estimation. They vary from 8 to 40% in Sheffield and between 14 and 43% for Nottingham. Even at the lower, modest estimate, however, they represent a considerable proportion. The geographical footprint of the areas with more than one urban locale identified are illustrated in Figure 7.17. Two classes or types of area with different urban locale perspectives are evident:

1 Boundary edges: where the edges of urban locales meet. These might be the simple case of two boundaries crossing, such as at the location annotated as Sn1 within Figure 7.17 for Lodgemoor / Fulwood and Sn2 for Hackenthorpe / Owlthorpe in Sheffield or En1 Bilborough / Wollaton and En2 Gedling / Carlton in Nottingham. Alternately, the boundaries of a number of different locales may all converge at one point, for example S1 in Figure 7.17 representing the convergence of Ecclesall, Hunter's Bar, Brincliffe, Nether Edge, Sharrow and Sharrow Vale; S2 for Kelham Island, Neepsend, Shalesmoor, Netherthorpe; N1 for Basford, New Basford, Forest Fields, Hyson Green and Radford or N2 for City Centre, The Park, Canning Circus, Lenton, New Lenton, Radford.

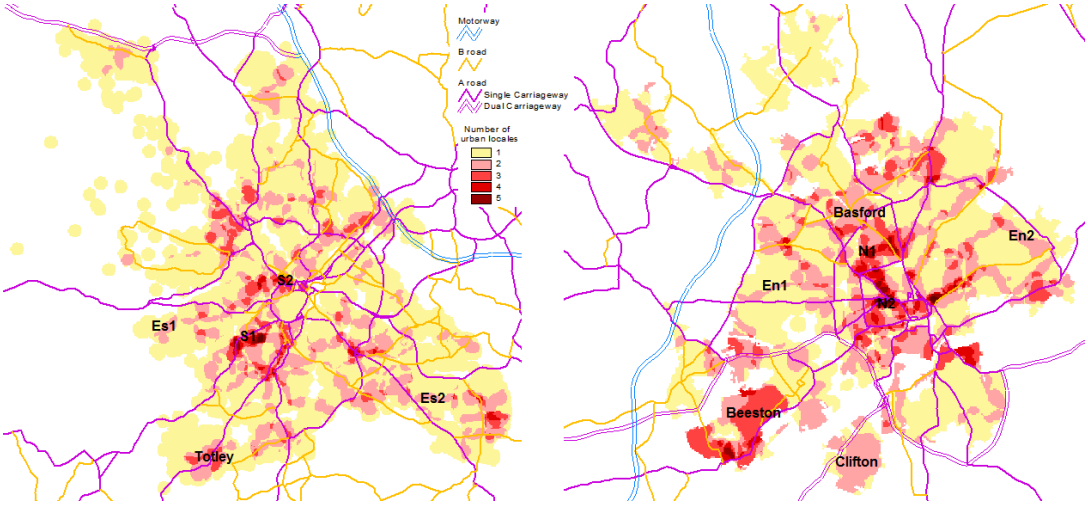
2 Extensive named areas: Alternately, large areas may overlap which are certainly not attributed boundary effects. This happens for the following areas in Nottingham: Basford which in parts can also be termed Old and New Basford; Clifton which can be split into Clifton estate, Clifton Village and Clifton Grove; and Beeston which in parts also contains Chilwell and Toton. Further investigation of such relationships or hierarchies are considered subsequently.

It is, however, important to recognise that the level and extent of overlap within urban locale boundaries will be dependent upon the context in which it is found. In a smaller

Table 7.15 Degree of differing perspectives of urban locale: within Sheffield and Nottingham

Number of urban locales	**Sheffield:** 300 smoothing (%)	Input data (%)	**Nottingham:** 300 smoothing (%)	Input data (%)
1	60.2	91.7	57.3	86.0
2	31.1	7.7	28.4	13.5
3	7.2	0.6	11.9	0.5
4	1.3	0.1	2.1	0.0
5	0.2	0.0	0.2	0.0

Fig. 7.17 Geographic footprint of the differing perspectives of urban locale: the number of different locales existing and potentially overlapping



a) Sheffield b) Nottingham

urban settlement (such as Ludlow) there are unlikely to be 'competing' overlapping urban locales, whilst in the densely populated areas of London (such as Camden), there is more likely to be a multitude of opinions concerning urban locales. Whilst it is beyond the scope to further investigate this limitation, such an issue could be pursued if a national set of urban locales were generated.

7.6.3 Urban locale inter-relationships

Using the methods as described above, it is possible to further explore the degree of hierarchy and overlap with other urban locales. Tables 7.16 and 7.17 illustrate the degree of differing perspectives between urban locales for the 15 locales with the greatest overlap in both the study areas of Nottingham and Sheffield. For example, Beeston Ryelands urban locale in Nottingham consists of 13 ($50m^2$) cells, all of which are contained *within* Beeston urban locale. Canning Circus has 80% of its cells (accounting for 19 cells) that are also considered to be within other urban locales, mostly in the locales of Radford, Lenton and Arboretum. An example of

the relationships that exist between urban locales is illustrated within Figure 7.18 [26] - which shows the locales that overlap with the geographic space shared by Beeston urban locale in Nottingham. It shows that there are 212 cells where people could call the area either Beeston or Chilwell, which account for 25.7% of the total area of Beeston and 86.9% of the area of Chilwell. The figure simplifies circumstances by only extracting to a single level of hierarchy, for example Chetwynd business park is completely contained within both Chilwell *and* Beeston before the city level of Nottingham.

Table 7.16 Degree of differing perspectives: The 15 urban locales in Nottingham with the greatest level of overlap

Urban locale	Cells with more than one urban locale		Overlap with the following locales (including cell count)
	%	Count	
beeston rylands	100.0	37	beeston 37
bramcote hills	100.0	14	bramcote 14
chetwynd business park	100.0	3	chilwell 3
lady bay	100.0	33	west bridgford 33
wilford village	100.0	12	wilford 12
clifton estate	93.0	40	clifton 40
warren hill	92.3	12	arnold 11; top valley 1
chilwell	88.5	222	beeston 211; toton 4; chetwynd business park 3; attenborough 2; bramcote 1; hucknall 1
lenton abbey	87.5	14	beeston 14
mapperley plains	86.7	13	mapperley 11; arnold 2
old basford	84.4	77	basford 75; whitemoor 2
toton	84.2	84	beeston 80; chilwell 4
canning circus	80.0	19	radford 8; lenton 5; arboretum 3; city centre 1; new lenton 1; the park 1
sherwood rise	80.0	39	forest fields 14; new basford 11; sherwood 10; carrington 4
new lenton	75.0	39	lenton 28; the park 5; dunkirk 2; radford 2; canning circus 1; castle marina 1

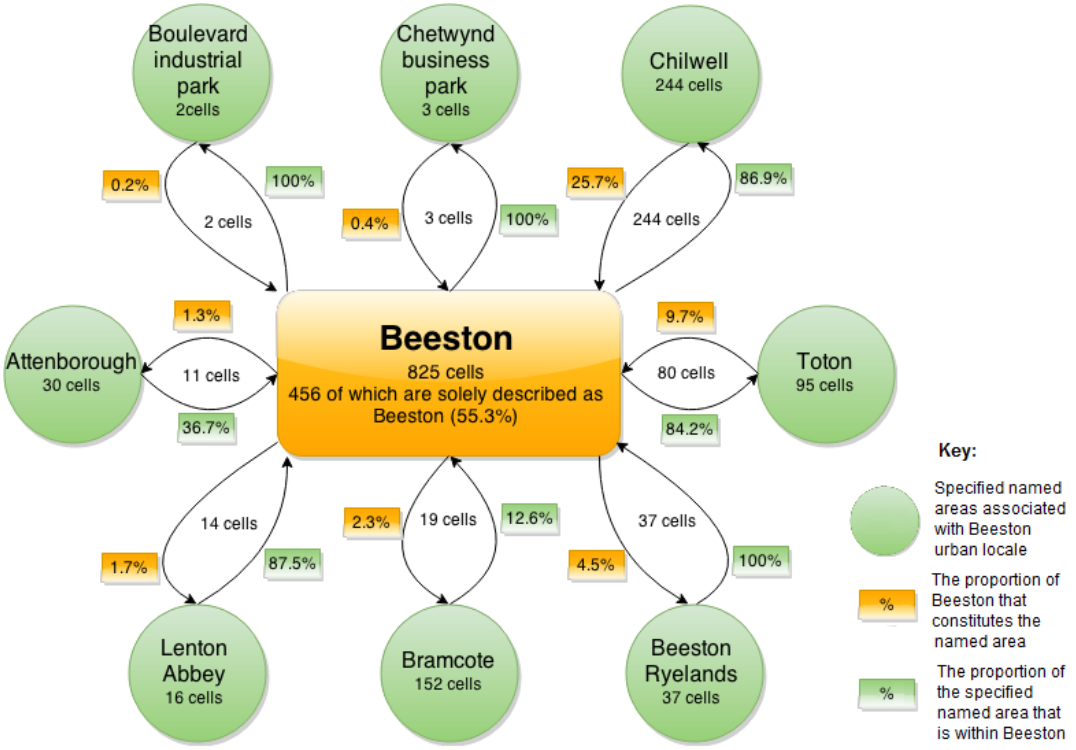
This section has illustrated the use of simple Cyan-Magenta-Yellow-Key (CMYK) composites in order to portray the geographic overlap between differing perspectives of urban locale where the same space might be deemed to be termed a number of different names by differing people. The section has also demonstrated that it is possible to use the data outputted from the method in order to investigate the degree of containment and overlap between locales in order to gain insight concerning the hierarchies of space.

[26] Drawn using draw.io: https://www.draw.io

Table 7.17 Degree of differing perspectives: The 15 urban locales in Sheffield with the greatest level of overlap

Urban locale	Cells with more than one urban locale		Overlap with the following locales (including cell count)
	%	Count	
holbrook industrial estate	100.0	19	holbrook 14; halfway 5
meadowhall centre	100.0	2	meadowhall 2
thorncliffe park chapeltown	100.0	4	chapeltown 4
sharrow vale	91.7	33	ecclesall 15; hunters bar 14; sharrow 2; brincliffe 1; nether edge 1
holbrook	86.7	17	holbrook industrial estate 12; halfway 5
crystal peaks	83.3	5	waterthorpe 4; sothall 1
totley rise	76.8	46	totley 32; dore 11; bradway 3
middlewood	65.0	13	hillsborough 6; wadsley park village 4; wadsley 2; oughtibridge 1
crookesmoor	63.0	34	crookes 21; walkley 10; broomhill 2; upperthorpe 1
hunters bar	62.1	52	ecclesall 31; sharrow vale 14; brincliffe 3; banner cross 2; greystones 2
woodhouse mill	60.0	9	woodhouse 9
brincliffe	59.4	23	nether edge 13; ecclesall 4; hunters bar 3, banner cross 1; sharrow 1, sharrowvale 1
burncross	59.0	23	chapeltown 23
abbeydale	55.0	13	nether edge 10; carterknowle 2; heeley 1
malin bridge	54.5	12	hillsborough 7; stannington 2; wisewood 3

Fig. 7.18 Urban locale relationships for Beeston, Nottingham



7.7 Summary

This chapter has been able to address the research questions as raised in Chapter 1. The first three of these represented advancement and refinement of the general method (forming web groupings; exploring temporal aspects; and tying other data to urban locale units), whilst the last three items comprised of specific questions that the output of the work could address concerning the nature of urban locales (assessing the level of homogeneity within urban locale units; comparing the volumes of data output with the underlying population geography; and exploring the extent to which differing perspectives of urban locale are formed). The fact that data were drawn from a range of various sources meant that it was possible to extract different urban locale geographies from the sources and then compare them. Using such an approach it was possible to demonstrate that urban locale geographies from Real estate sources were generally similar to output from Business information data (97% of cells in Nottingham had the same majority urban locale when smoothed to 100m), although noticeable differences for specific urban locales were also evident. Whilst exploring areas of divergent perspectives is of value, areas of agreement between the different web groups was able to identify spaces where a greater confidence in the data may be maintained. It was also possible to explore change in urban locales in Sheffield by running the analysis for two separate time periods (2012 and 2014). Comparison of the majority urban locale function revealed a 5% change in cells within the two year period, demonstrating the high degree of temporal stability.

Tying other data to the urban locale units (such as Twitter, house price or Census of Population data) demonstrated the potential of urban locales to form the basis for important business, policy and social decisions, adding value to both sets of data. The aggregation of Census of Population data to the urban locales also allowed for further analysis to be undertaken on the urban locale data units, including testing homogeneity. Results demonstrated that whilst urban locales in Nottingham do not appear homogeneous in terms of socio-economic characteristics, they were more homogeneous than other alternative administrative data units at a similar spatial scale. Such questions demonstrate potential research areas that can be investigated once urban locales have been constructed but that were previously difficult or impossible to address. As such, constructing a UK wide urban locale geography would open up many different avenues of research potential.

Comparisons were also made between the level of data returns and the underlying population - in order to identify geographic areas with higher or lower than expected urban locale data output. Generally, it was found that less affluent areas generated lower numbers of data points appearing on the web. Finally, the degree of overlap

between urban locales was investigated and found to be *at least* 8% of area in Sheffield and 14% in Nottingham. *Critically, this illustrates the degree of error that any discrete formation of urban locales could not avoid and demonstrates the importance of the principle of vague urban locales*. The following chapter forms a discussion concerning the vague urban locale method and output, including the merits and limitations of the approach. It will also illustrate through the use of a small number of case studies the type of application that the methodology can assist, and highlight a number of avenues to pursue for future research in the area.

Chapter 8

Discussion and Case Study Application of Vague Urban Locales

This chapter discusses the main advantages and limitations of the method for generating vague urban locales. Whilst the locale units hold general utility, this chapter, through the use of three case studies, also discusses example application areas. The three case studies include (1) the designation of boundaries for more in depth qualitative research; (2) the use of specific locale boundaries and associated Census data to assist local forum groups; and (3) the utilisation of the method to support neighbourhood planning within a local council. Finally, this chapter will also discuss the potential future work that could be applied either directly to improve the method or in applying similar principles to other data and contexts.

8.1 Discussion

Before examining the case studies in depth, this section discusses some of the main strengths and weaknesses of the proposed method that have not already been considered in the previous chapter.

8.1.1 Advantages of method

A number of key advantages of the method have been debated in the previous chapter - notably:

- the ability to produce probabilistic vague maps of urban locales from web derived data;
- the opportunity to distil such information into different groups of data (business information, estate agents and other);
- the capacity to explore change in urban locale boundaries over time;
- the potential to tie other socio-economic data to these social units; and
- the capability to further analyse the urban locales in order to learn more about these place names and extents. Such analysis has previously been impossible to undertake in a large scale approach (for example investigating the hierarchical nature and relationships of urban locales, or the homogeneity of locales in terms of socio-economic data).

This section will now discuss other advantages of the method not previously considered.

8.1.1.1 Data volumes

The median number of data points associated with an urban locale in Sheffield was 217 (mean of 528) and 209 for Nottingham (mean of 954). Other existing work rarely explicitly discuss the number of data returns the technique produces (for example see Flemmings (2010); Twaroch et al. (2009a)), making it difficult to directly compare with the method developed here. The geographies of urban locales generated from Flickr photographs (through Quattroshapes) and other sources were compared within the validation section in Chapter 6. This section explicitly concerns the number of data points that differing methods can draw upon. A small number of case example urban locales were selected in order to explore the number of data returns resulting from differing methods (Gumtree, Flickr API[1] and web extraction output) - as shown within Table 8.1. Generally speaking, whilst a high number of data points may not necessarily result in a 'better' and more accurate urban locale geography, low data returns potentially could result in a unrepresentative perspective.

The number of data returns, however, are heavily dependant on the context of the urban locale and uneven distributions of data points across space should be further investigated. Whilst the Flick API generally provides high data volumes for tourist areas, they are far lower in more 'ordinary' suburbs. For example, in the tourist areas of Soho, London a total of 19,438 Flickr photographs were returned compared to 14,161 references from the web extraction output; whilst Wollaton, Nottingham (home of Wollaton Hall and Wollaton Park) produced 2,726 Flickr photographs and 1,840

[1]Thus, using the raw Flickr photographs as input point data as opposed to using pre-processed output such as from Quattroshapes

data records from the web extraction output. In contrast, the more suburban areas of Chilwell, Nottingham returned 347 Flickr photographs but 1,852 records from the web extraction output and 49 Flickr photographs for Crosspool, Sheffield compared to 542 data points from the web extraction output. Data returns for Gumtree were significantly lower than those found using the web extraction method (a factor difference of 15 in the example locales investigated).

Table 8.1 Comparing data volumes with differing approaches

Urban Locale	Gumtree returns	Geotagged Flickr photos	web extraction data
Broomhall, Sheffield	10	158	168
Carter Knowle, Sheffield	0	2	55
Crookes, Sheffield	59	300	827
Crosspool, Sheffield	10	49	542
Parson Cross, Sheffield	25	14	349
Walkley, Sheffield	47	128	1,428
Bobbers Mill, Nottingham	1	4	220
Chilwell, Nottingham	140	347	1,852
Clifton Estate, Nottingham	0	0	122
Dunkirk, Nottingham	49	99	295
Radford, Nottingham	379	149	1,252
Wollaton, Nottingham	1,320	2,726	1,840
Chalk Farm, London	84	1,682	174
Fitzrovia, London	133	7,250	3,344
Pimlico, London	315	3,254	1,195
Soho, London	296	19,438	14,161
North Hykeham, Lincoln	257	93	3,965
Waddington, Lincoln	49	83	1,199

A more comprehensive comparison between Flickr data and web extraction output for the city of Sheffield can be found in Appendix I. In summary, the web extraction output returned a higher number of data points for Sheffield - although it is recognised that in more tourist areas (such as central London), the picture will be very different. The median number of data returns from Flickr photographs for Sheffield urban locales was 41 (mean of 131), compared to 213 from the web extraction output (mean of 520). It is, however, also important to recognise the potential spatial footprint of data returns. The web extraction output are tied to postcodes and so data points will, by default, consist of fewer *unique* sets of coordinates than Flickr photographs (where georeferencing is not constrained in such a manner). This said, the web extraction output for the Sheffield case study continues to produce a greater number of unique data points for mapping with a median of 43 unique data locations per urban locale (mean of 59) compared to only 19 from Flickr (mean of 50).

This section demonstrates that the number of data records from the web extraction method is at least comparable with alternative methods used in previous literature. The potential of combining data from differing sources is discussed in more detail in the 'Future work' section.

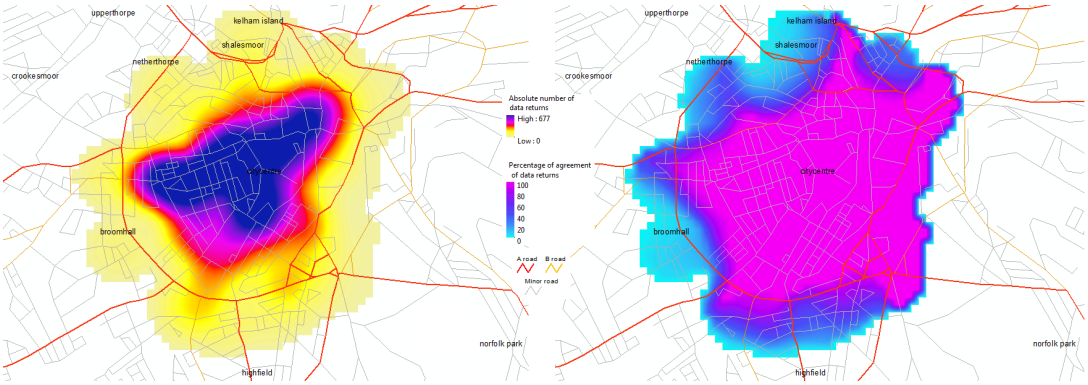
8.1.1.2 Use of percentage data

As discussed within Chapter 5, current literature frequently use KDE surfaces for urban locale generation (Clough and Pasley, 2010; Grothe and Schaab, 2009; Hollenstein, 2008; Hollenstein and Purves, 2010; Purves et al., 2005b; Twaroch and Jones, 2010; Twaroch et al., 2008, 2009a,b). They, however, only use the absolute numbers of data points. Whilst such an approach is ideal for defining urban locales cores (such as centred around local shopping areas) the focus is more of a representation of the underlying population and on areas likely to produce higher data returns rather than on the actual geographic pattern of the urban locale itself. Consider edge locations (such as on the urban fringe or next to large parks) or in areas of relatively sparse population density, where absolute values will be expected to be low. Lower data volumes in such cases should not necessarily mean that the area is in any way less associated with a particular urban locale name. In fact, it could be that opinion concerning the urban locale name of a shopping area was contested but that in an edge location there is complete agreement that the area belongs within a single urban locale. The absolute number of data returns would not distinguish this, continually giving prominence to areas of high data returns. For such reasoning, the method as illustrated within Chapter 5, also considers the proportion of data assigned to particular urban locales.

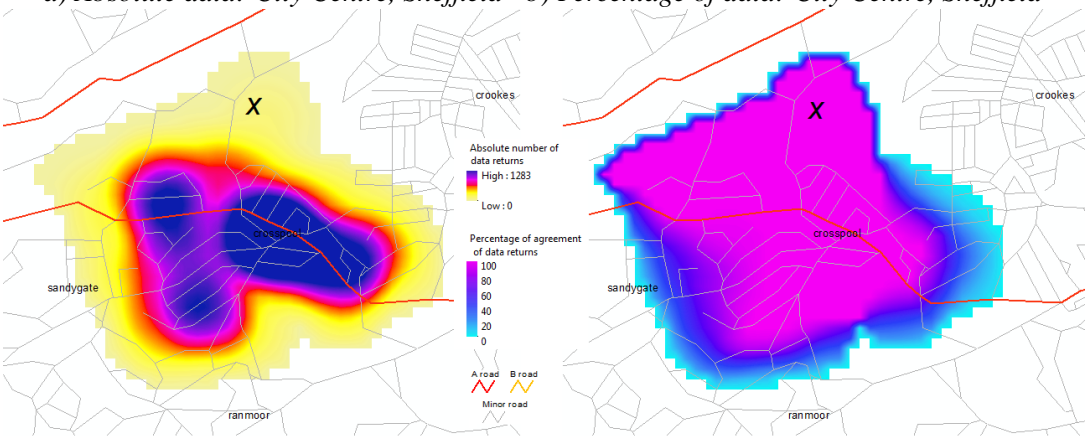
Examples of both absolute data returns and the percentage of records referencing a specific urban locale can be found in Figure 8.1. Whilst the two geographies have similarities, there are also marked differences. For instance, compare the area to the north of Crosspool (Sheffield) labelled with an x in Figure 8.1c and d, where data volumes may be low but there is overall consensus that the area is defined as ‘Crosspool’. Such measures can add to insight where data volumes might be expected to be relatively low (for example on the periphery of towns and cities, near rivers, or in close proximity to large parks).

It is clear that both the absolute and percentage of data returned for a particular locale are important depending on the type of questions being posed. Investigations of where the centre of an urban locale was, or placement of place names on maps would be best implemented through the use of absolute data returns. Queries along the lines of "What do people call *this* space?" are more suited to the percentage of data belonging to a particular locale and are especially important, as in the case of this body of work, when considering vagueness and the differing perspectives of urban locale. This work, hence, advocates the incorporation of both elements - as similar to "two sides of the same coin."

Fig. 8.1 Differences between absolute and percentage returns



a) Absolute data: City Centre, Sheffield *b) Percentage of data: City Centre, Sheffield*



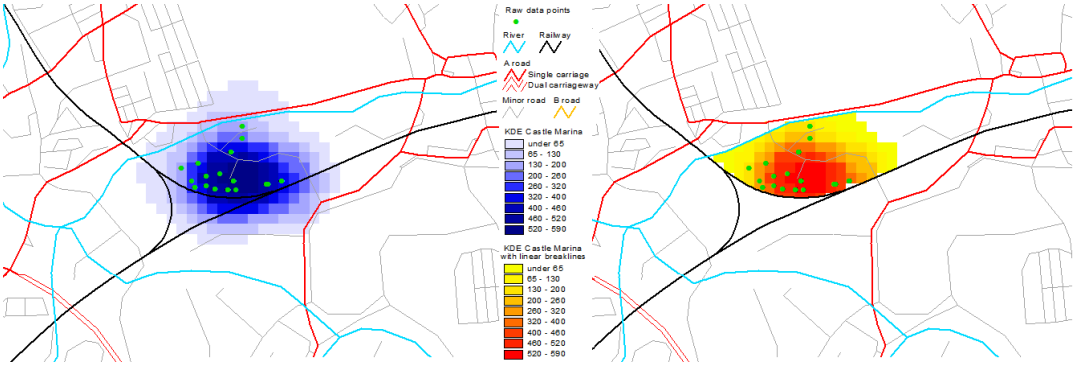
c) Absolute data: Crosspool, Sheffield *d) Percentage of data: Crosspool, Sheffield*

8.1.1.3 Linear features as breakline constrains for smoothing

One criticism of the KDE approach is that it has a tendency to over-smooth (Purves et al., 2005a; Twaroch et al., 2009a). This is particularly problematic if smoothing extends the feature over some unlikely or implausible 'barrier', for example over a river, railway or motorway. The difficultly lies in the fact that at times, urban locales can cross such features. For example, the Victorian urban locale of Broomhall in Sheffield spans a dual carriageway (the A61 ring-road) which was built in the 1960s long *after* the urban locale had formed. Similarly, if the linear feature has many crossing places, it is unlikely to operate as a barrier to movement and thus, may not affect and restrict place naming. This is most likely the explanation why Finsbury Park, London is not constrained in the north by the Crouch Hill - Harringay Green Lanes Overground railway line. The problem of knowing exactly when linear features may or may not act as breaklines appears problematic. This body of work introduces a simple solution to this issue, as described in detail in Chapter 5, based on the premise that crossing specified linear features only occurs *if the underlying data supports the notion*. For example, in Figure 8.2, the raw data points show that the urban locale of Castle Marina, Nottingham is very constrained by the surrounding geography. Even

though this is the case, a standard KDE surface as shown within Figure 8.2a pays no regard for such features and spreads out and beyond accordingly. Constraining by using selected linear features as breaklines, as in Figure 8.2b, can produce more realistic output.

Fig. 8.2 Linear features as breakline constrains for smoothing: Example of Castle Marina, Nottingham

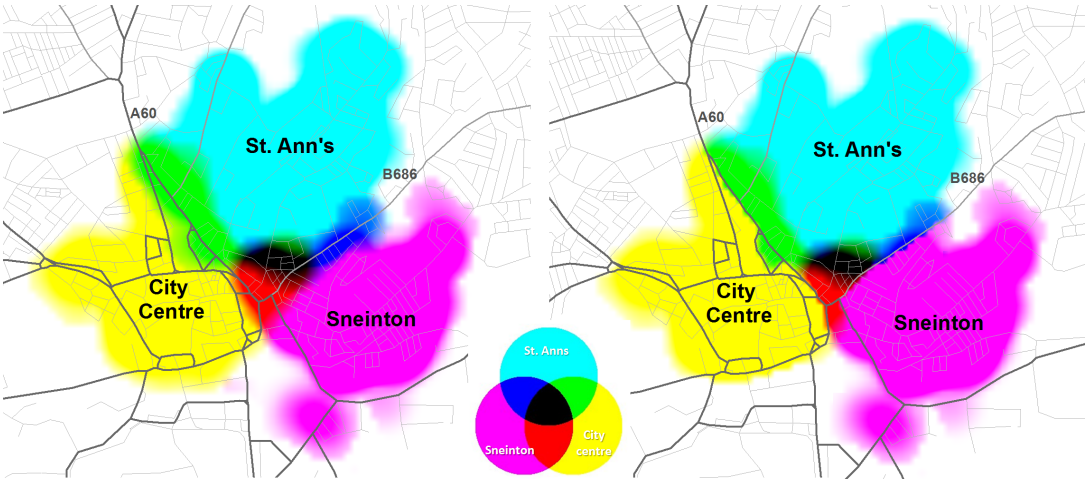


a) Original KDE surface *b) Final, modified KDE surface*

Figure 8.3 demonstrates the potential differences and improvements if the procedure is undertaken, illustrating a colour composite for three urban locales within the centre of Nottingham - both with and without breaking along linear features. The overarching visual interpretation between output with and without using breakines within the KDE process is one of similarity. There are, however, subtle variations. The predominant difference is that with linear features acting as breaklines, the St. Ann's urban locale does not spread either west across the A60 into the City centre, or south traversing the B686 into Sneinton. In contrast, the underlying data demonstrates that the City centre and Sneinton both still permeate into St. Ann's (the green and blue tones on the map respectively). Thus, the effect is a crisping and distinction along certain roads but only when the underlying data are in support. Whilst the method is simple Figure 8.3 clearly demonstrates its worth and visual impact.

The main limitations of the method lie in determining suitable linear features and having enough data points to determine if a linear feature should act as a barrier or not. It is unrealistic to expect all non dual carriageway A-roads to act as barriers to movement in the same way as motorways, rivers and railways. Including them as breaklines creates a number of small 'islands' where data volumes are likely to be low and thus smoothing will not occur simply due to lack of supporting data and not because the road in reality acts as a constraining feature. An example is shown in Figure 8.4. Such problems with voids are not widespread and were only identifiable in 4 out of the total of 102 urban locales for Nottingham (4% ot total locales). For this reason, however, it is recommended that single carriage A-roads are only considered suitable breakpoints if manual checking is incorporated and that if full automation

Fig. 8.3 Cyan-Magenta-Yellow-Key (CMYK) composite to explore the impact of using linear features as breaklines: for the centre of Nottingham

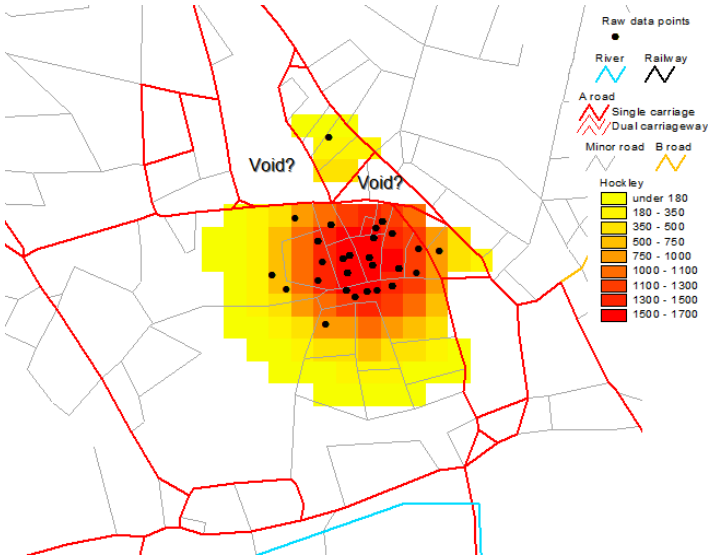


a) Breaklines not used to restrict smoothing *b) Breaklines used to restrict smoothing*

is required only motorways and dual carriageways are used for road type breaklines. Perhaps, in time, data volumes will increase with the growth of information held on the web - which will circumvent the issue.

A further limitation of the method is that dangles (lines that are unconnected at one end - such as a dead end) are not incorporated, as the method is based upon connected features. Such cases, however, are relatively rare when considering major roads, railways and rivers. It is, however, the firm view of the author that the benefit of the procedure far outweighs these minor limitations of the method.

Fig. 8.4 Voids due to sparse data if all A-roads are considered as breaklines: Example of Hockley, Nottingham



8.1.1.4 Other benefits

The prevalence of references explicitly to 'city centre' within the web extraction output was unexpected. Although one might expect businesses to potentially include such information within their postal addresses when posted on the web, it was considered unlikely that residential parts within town and city centres would also incorporate such information. This, however, was not the case and a high number of residential only postcodes were found to mention the area as being within the 'city centre'. In Sheffield 15% of all postcodes consisting of references to the city centre (where there was complete 100% agreement that the area was named as city centre) were from residential only postcodes, whilst in Nottingham the figure was 10% of postcodes. The possibility of producing 'city centre' footprints was unexpected but is of particular interest considering difficulties concerning alternative methods (Haklay, 2007; Thurstain-Goodwin and Unwin, 2000; Vaughan et al., 2010).

It is important to stress that the method is one that is highly scalable, extensible and dynamic. The potential scalability of the processes draws on the fact that searches rest upon postcodes which can easily be processed in batches through parallel cloud computing. This said, the work demonstrated here was all undertaken on a standard PC - with no requirement of high specification computing. Additionally, as mentioned throughout this document:

- output is based upon information from a wide range of source data which should aid robustness;
- the locale names are not required *a priori* (which is important given that Chapter 6 demonstrated the inconsistencies within current urban locale level gazetteers); and
- the tight delimitation of locale names within the address context reduces issues of ambiguity (for instance making it easy to distinguish references to Lenton, Nottingham from those relating to Lenton Abbey, Nottingham).

8.1.2 Vague place names: Identification of misspellings and synonyms

The role of vague place *names* (as outlined as in Chapter 2 as the third element of place) has largely been little considered to this point. This section seeks to address that imbalance. A number of synonyms were identified through the misspelling operation. A sample of these were displayed within Table 6.3. Additionally, it is useful to know

common misspellings that can arise from specific place names. Most gazetteers, however, do not contain such information. Such output arises as a by-product of the proposed methodology.

It is important to explore the extent to which spelling variations are an issue within this work. The procedure identified 181 spelling variants for the 121 urban locales of Sheffield. In total 81 of the urban locales had spelling variants (67% of locales), meaning that on average an urban locale had 1.5 different spellings identified by the method and accounting for 780 data records (1.2%). There is potential for further research to distinguish between true spelling issues (such as 'Chapeltown' and 'Chapletown') and more likely typing mistakes (for example 'Chapeltown' and 'Chapelown') - although such work is outside the scope of this study which simply groups both types together as variants of spelling. Whilst the total data volumes involved may be considered modest, the power of determining the likelihood in which different spelling variants may be likely to be found could be valuable for improving current web searching as the following examples from Sheffield illustrate:

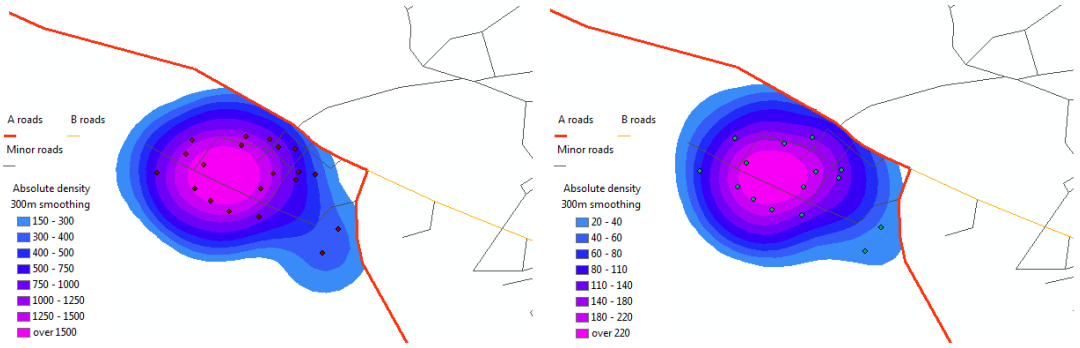
- 36.4% of references to 'Carterknowle' refer to the place as 'Carter Knowle'
- 14.5% of references to 'Frecheville' refer to the place as 'Frechville'
- 3.2% of references to 'Wincobank' refer to the place as 'Winkobank'
- 2.0% of references to 'Attercliffe' refer to the place as 'Attercliff'
- 2.0% of references to 'Crookes' refer to the place as 'Crooks'

Table 6.4 contained the urban locales with highest degree of synonymy. 'Pottington business estate' and 'Pottington industrial estate' urban locales in Barnstable fail the spelling test (the Ratcliff/Obershelp ratio of 0.76 and Levenshtein Distance of 0.78 are both under the 0.8 minimum threshold - refer back to Chapter 5 for details) and so are not considered a spelling issue. They are picked up under the synonym identification with 96% of the variance within a linear grid regression being explained. The similarity of geography between the two is clearly evident within Figure 8.5.

The two locales within the case study areas with the second highest degree of variance explained in the synonym identification procedure are 'Archway' and 'Upper Holloway', London. Here the R^2 falls to 85% and although there are general similarities within the geography, there are also noticeable differences within Figure 8.6 (such as the extension to the south-east in 'Upper Holloway'), indicating that the two names do not share an identical footprint. Importantly, however, such cases give further strong evidence of the need for vague overlapping urban locales. In such cases, it is impossible for both areas to be adequately represented by discrete boundaries. In fact no reference to an areal feature termed 'Archway' can be found in either Quattroshapes, OS MRDP or Wikimapia - instead all three tag the area as 'Upper Holloway'. In

contrast Wikipedia does acknowledge the existence of the area as 'Archway' and states that "The name [Upper Holloway] has fallen out of common use and the area is generally regarded as being a part of Archway or Holloway[2]". If areas can be regarded as being part of multiple urban locales then clearly the models we use to represent them should allow for overlapping areas.

Fig. 8.5 Synonym identification: the case of 'Pottington business estate' and 'Pottington industrial estate' urban locales in Barnstable (R^2 = 96%, p < .001)



a) Pottington business estate, Barnstable *b) Pottington industrial estate, Barnstable*

Fig. 8.6 Synonym identification: the case of 'Archway' and 'Upper Holloway' urban locales in London (R^2 = 85%, p < .001)



a) Archway, London *b) Upper Holloway, London*

The synonym procedure can also be exploited in order to account for issues concerning the theoretical assumptions of the method utilised within this doctoral study. Whilst the approach adopted within this thesis primarily focuses on the level of consensus for the urban locale *name*, there remains the possibility that the locale name could change, but where it would be desirable to conceive the spatial object as persisting. Such a process would be identified through synonym identification.

[2]https://en.wikipedia.org/wiki/Upper_Holloway

In summary, the importance of variants in spelling and synonyms are clear for semantics. When people discuss particular named places, it is important that they are aware of the full potential spatial interpretations that others may perceive with the specified place. The method undertaken within this thesis demonstrates how both spelling and synonyms can be produced which can at the very least provide improvements to existing gazetteer sources that currently lack both.

8.1.3 Fuzzy-babilistic

Similar to the header within Chapter 3 sharing the same name, this section addresses the extent to which people perceive urban locale objects as either fuzzy or probabilistic. When interviewed as part of the boundary validation within Chapter 6, residents in Sheffield were also questioned if they felt that two urban locales could potentially overlap either in their own perception or due to different people considering the same space as separate locales. The responses were used to determine if views were most closely related to discrete (two areas could not co-exist); probabilistic (two areas could co-exist but only as different people perceiving the space differently); or fuzzy boundaries (areas could co-exist within their own perception - as a mix of multiple locales).

The output is shown in Table 8.2, which demonstrates the variation in the different ways that people perceived the spaces, although the probabilistic perspective received the greatest support. The following quotes sum up how the majority felt: "Some might call it Crookes but it's always Broomhill to me" and "I associate with Broomhill but others say it's Crookes". Such views highlight that although there is no one-fit-rule, most people identify with a single named area, although they are aware that other people may hold different views. It should be noted the study area was chosen due to its spatial position between three different urban locales and as such the respondents might be more aware of such issues. People in other locations might be more inclined to perceive more discrete boundaries. The aim of the questioning, however, was to determine if a fuzzy or probabilistic standpoint was more common.

Table 8.2 Comparing approaches to perceiving space

	Count of responses	% of responses
Discrete	4	20
Fuzzy	3	15
Probabilistic	13	65

8.1.4 Limitations of method

Dependency on parameters: The main criticism of the method from conference peer-reviewers was that it is dependant upon a large number of threshold values. Such criticisms form the basis of the rationale behind the parameter testing discussed within Chapter 5 and the detailed validation measures considered in Chapter 6. Due to the issue that such ground-truthing is problematic because of the fact that no suitable gold-standard urban locale dataset exists, a plethora of different approaches were adopted. These included the comparison of *names* with other gazetteers and three different measures of *geographic* validation procedures: (1) the level of agreement with existing geographic urban locale sources; (2) seeking local stakeholders views of specific locale boundaries; and (3) three different resident validation methods (3a: resident views from an on-line questionnaire for places within Nottingham; 3b: information collected from residents for a single locale by a Neighbourhood Forum; and 3c: resident views in targeted streets where web extraction output portrayed evidence of differing views of urban locale name within the same local space). The take-home message from the validation processes was the support of the proposed web extraction method (which generally was closer to resident perceptions than any of the alternative current sources of urban locale information). The strong performance within the validation process supports the various parameter settings used within the method. Whilst such parameters have been carefully chosen through extensive investigation and testing and also validated against a range of other data - they still represent a series of somewhat arbitrary figures. There are, however, no alternatives when quantifying such output. The method incorporates vagueness but itself cannot be created using vague parameters with no prior definition.

Such parameters can have important effects, most notably determining if an area name has enough data points and in a concentrated geographic context to warrant inclusion as an urban locale. Whilst such a process removes spurious text appearing within address forms, it can also exclude very small areas such as Page Hall, High Storrs or Lower Walkley in Sheffield. The merits of including such places is however uncertain given the fact that any KDE surface drawn with such few datapoints (under 40) would not be robust. The issue was debated within the parameter testing of Chapter 5.

The extent that the web extraction output can reflect the units of vernacular geography is dependant on the number of data returns which is largely in itself a reflection on the scale of the geographic units of concern. As such, extremely small vernacular named places (such as Foyes Corner in Southampton) are unlikely to be identified. This does not diminish the value of the work as this research aims to construct robust areas and works best at higher spatial scales (such as neighbourhood and suburb level). As such,

it might not identify the very local Foyes Corner in Southampton but it certainly does identify the Shirley urban locale that it falls within.

As mentioned previously in this chapter, a criticism of the KDE procedure is that it often over-smooths into areas with no data (Purves et al., 2005a; Twaroch et al., 2009a). For these reasons, to help compensate for this a 2% cut-off was applied in order to 'trim the edges'. This was in addition to developing methods (as previously discussed in Chapter 5) to stop KDE surfaces spreading into areas with no supporting point input data through the use of linear features as breaklines.

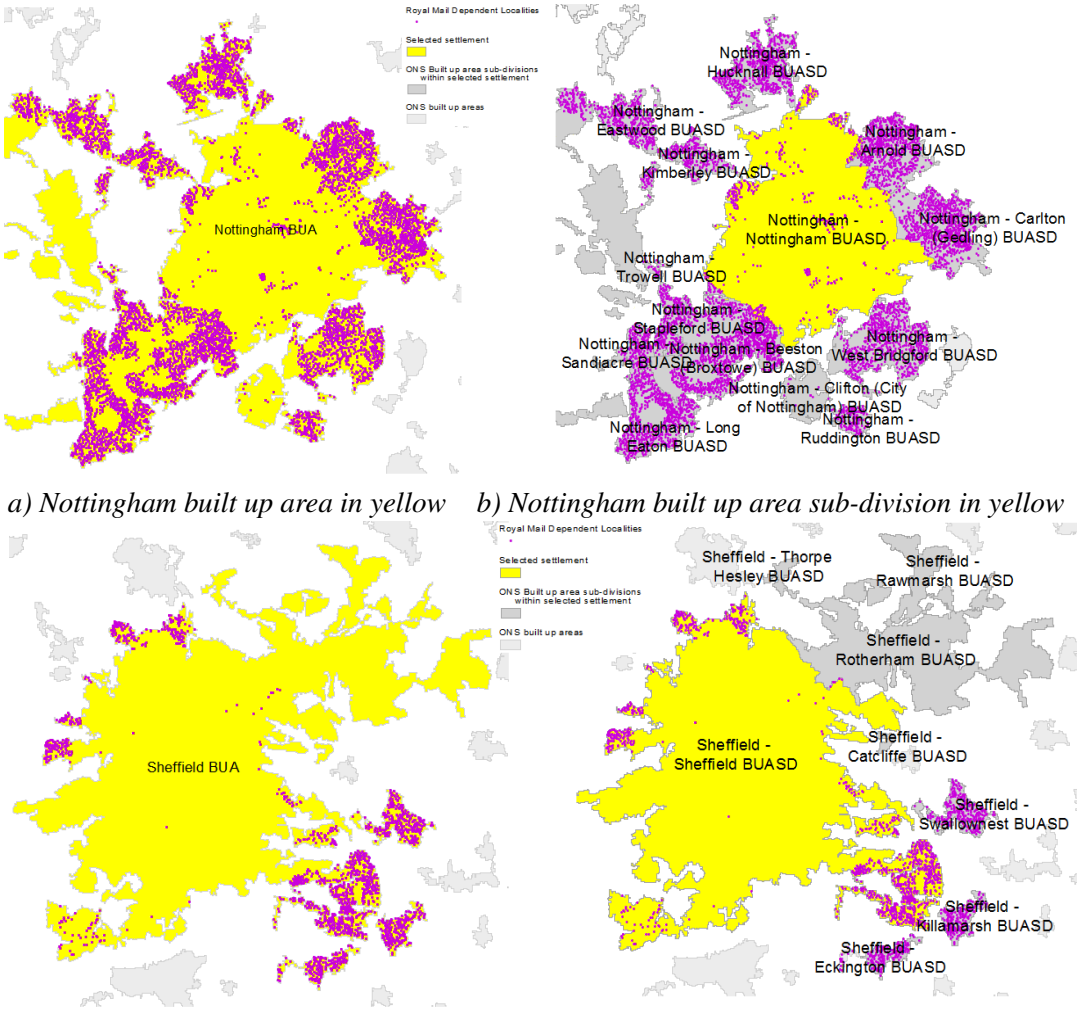
Dependent localities: The degree to which urban locale data are potentially influenced by official Royal Mail designated locality information ("Dependent Localities" as discussed within Chapter 5) is heavily influenced by the areal definition of the settlement study area. Thus, whilst the use of "Dependent Localities" within named town and city areas in a narrow geographic stance ("ONS built-up area sub-divisions") is rare, their use within the 'wider' geographic scale ("ONS built-up areas" which in extent are closely in line with the higher level geography for 2001 urban areas)[3] can be more common. As an example the ONS built-up area of Derby includes the following ONS built-up area sub-divisions: Derby, Borrowash, Breadsall, Duffield and Little Eaton.

Critically, this comes down to the difficult geographical question of "What geographic area is considered *Sheffield*?" or "What is *Nottingham*?". Should adjoining, outlining named areas be considered within or external to the dominant centre? Nottingham is particularly affected by this issue as demonstrated in Figure 8.7a and 8.7b where the different extents of the settlement significantly affect the degree to which Dependent Localities occur within the named area. This is illustrated by the proportion of area of "Nottingham" (coloured in yellow) that is affected by Dependent Locality postcodes (shown in the figure as purple points). The complex nature of sub-divisions of Nottingham is in stark contrast to that for Sheffield shown in Figure 8.7c and 8.7d where the effect of Dependent Localities is far less regardless of the area defined as Sheffield. For this work the resolution rests with the official designation of ONS built up areas. The implications of using this definition for urban extents, however, should not go without discussion, particularly regarding the effect upon the prevalence of Dependent Localities.

Such naming issues for the urban areas of concern, however, are outside the scope of this particular work. The settlement units used here rest upon the official "ONS built-up areas". This was despite the fact that this increased the number of Royal Mail Dependent Localities where web extraction output might simply be a product of the

[3] See http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/built-up-areas—built-up-area-sub-divisions/index.html for further details on built up areas

Fig. 8.7 Defining urban areas



a) Nottingham built up area in yellow *b) Nottingham built up area sub-division in yellow*

c) Sheffield built up area in yellow *d) Sheffield built up area sub-division in yellow*

official addressing conventions. It should be stressed, however, that there is nothing wrong, per se, with web extraction output in such areas designated as Dependent Localities, other than it being influenced by Royal Mail's official view of place naming. In this light, it would be possible to roll out the proposed method to areas beyond the urban setting and include rural towns and villages within the context of this work, with acknowledgement that they would be a reflection on Royal Mail's view of named places. In such a context, the focus of investigations would be upon those data that provided information that were inconsistent from the official Royal Mail stance.

Representativeness of web data: Chapter 2 argued that the sheer size of data held on the web is enormous and rapidly increasing. In 2014, 38 million adults (76%) in Great Britain accessed the web every day and 22 million households (84%) had access to the web in 2014 (a figure that has increased from 57% in 2006) (Office for National Statistics (ONS), 2014). Whilst such numbers represent the majority of the country's population, it must be recognised that an element would not be able to access or post

information to the web. Thus, the urban locales generated here must be seen as a construct from a sub-set of total potential perspectives, however large that section of the population may be. The age demographics and geographic concentrations of those less likely to have access to the web should also be considered[4].

The fact that input data are drawn from a wide range of web data can be seen as both a strength and limitation. The benefit of having data from different sources should increase robustness (where data cannot easily be formed from a single person's perspective) but may also lead to uncertainty concerning where data have been drawn from in order to formulate the urban locale boundaries. The underlying data themselves, however, by definition are not official postal addresses but must constitute some form of user generated content which has subsequently been posted on the web. The formulation of the different web grouping of input URLs also assists in determining the origin of data.

This said, one limitation of the method is that it is possible that if sufficient information were posted on the web, it would potentially be possible for a determined individual to influence how space was described. Although, a relatively high number of data points would be required and other similar methods (for example using Gumtree or Flickr) would also be susceptible to such potential manipulation.

It should also be stressed that the majority of data are drawn from business information and real estate perspectives. It is relatively rare that residents themselves post urban locale information on the web (with the exception of selling items through web based classified advertisements, people finding and dating sites, for example). Within the validation, however, the method was found to generate views that were closer aligned than other comparable approaches, to those obtained from residents. Furthermore, the representativeness is also affected by the degree to which certain websites either allow users to enter manually address elements (for example as in the case of Yellow Pages and 192.com) or prohibit it (such as for business listings within Google Maps where addresses must conform to Royal Mail data and thus not contain urban locale information). Future work could explore the potential for normalising the data in order to balance some of these representative issues (see Section 8.3 Further research).

Populated spaces: The core aim of the urban locale generation was to generate areal boundaries which would facilitate the collection and aggregation of other socio-economic data for the locales (for example from Census of Population data). As such, the principal concern is with populated, rather than unpopulated[5] spaces. This is further demonstrated by the fact that the gridded KDE surfaces are, as a final

[4] see for example the work of The Geography of Digital Inequality - http://oxis.oii.ox.ac.uk/research/geography-of-digital-inequality

[5] such as parks, cemeteries, car parks, derelict land and so forth

process, allocated back to the postcode point locations. For other potential uses of urban locale boundaries information (for example when collating data that are not solely tied to populated places, such as the case for anti-social behaviour or car crime), however, it would be appropriate to model both populated *and* unpopulated areas. It would be possible to combine the current web-based postal address method with other approaches, such as examining Flick photograph tags for urban locale names, in order to identify locales within unpopulated spaces.

Temporal dynamics: Another limitation arises from the short-term temporal instability of data where potentially, due to the way Bing API servers hold search data, there are substantial variations in the input data retrieved during relatively short-term time periods. Such concerns are partially offset by the fact that, as demonstrated in the previous chapter, the geographic patterns of output data appear to be maintained regardless of the short-term temporal variations in input data.

A further limitation is that it is not possible to determine when data including urban locale names were actually created. Some data held on the web is persistent whilst others are more ephemeral and it is impossible to interrogate, with any certain, when exactly data were created. Whilst temporal effects can be handled as particular snapshots in time, it would be useful to eliminate historic perspectives.

8.1.5 Synopsis

Although the method developed within this body of work is relatively simple, and includes a number of limitations, it represents a powerful approach in which to investigate the complex notion of place. Not only does it allow for a vague interpretation, incorporating relatively high data volumes but is also comprehensive in terms of the place names it identifies. Given the requirements of the European INSPIRE Directive (as discussed in Chapter 4), the current method of deriving lists of urban locale (suburb) geographic names using OS 1:50K gazetteer is frequently incomplete. For example, in Nottingham the following urban locales are all not included within OS 1:50K data: Canning Circus, Castle Boulevard, Heronridge, Hockley and Warren Hill, demonstrating the importance of developing more comprehensive solutions. The following section of this chapter further demonstrates the value of the method by drawing on three sample case study applications.

8.2 Case study applications

At the most basic level, this research can improve existing mapping within our urban areas by providing current urban locale names and varying label sizes based on underlying populations. A key driver behind this research is the standpoint that urban locale geographies can provide a robust foundation for a host of social applications. For example, Sampson (2012) discusses how urban locales can influence a wide range of factors from juvenile delinquency, adult crime, teenage sexuality, substance abuse, and mental health among other outcomes. Once urban locales have been defined they can be used as new units for analysis to which other data can be tied to - be this Census of Population, crime, Twitter or just about any other socio-economic data available at a fine spatial resolution. As such it would be impossible to outline all the potential uses of these valuable data. Instead, this chapter will examine three case studies in order to investigate the value and usefulness of the urban locale data generated.

8.2.1 Identification of case study boundaries for research

Many academic, Governmental and other studies are tied to geographic case study areas - frequently based on place names. Defining the extents of those place names for the purposes of the case study research, however, can be problematic due to lack of data and the subjective nature of such information (as discussed previously within Chapter 4). The work here draws inferences from one such case study approach as detailed below.

'Mapping vernacular geographies in places of super-diversity': Katherine Stansfeld[6]*, PhD candidate, Royal Holloway, University of London*

"The research[7]*, takes an in-depth ethnographic approach to explore how people relate to and identify with place in Finsbury Park, London and how this is expressed within everyday processes of conviviality and multiculture through mobilities, sites of sociability and the naming of place. The research aims to investigate how, in the context of 'super-diversity' the 'vernacular geographies' of different people represent both cultural complexity, multiplicity and shared spaces of encounter and civic culture. The project is interested in how contemporary cartographic and geographic information can map places as constellations of trajectories."*

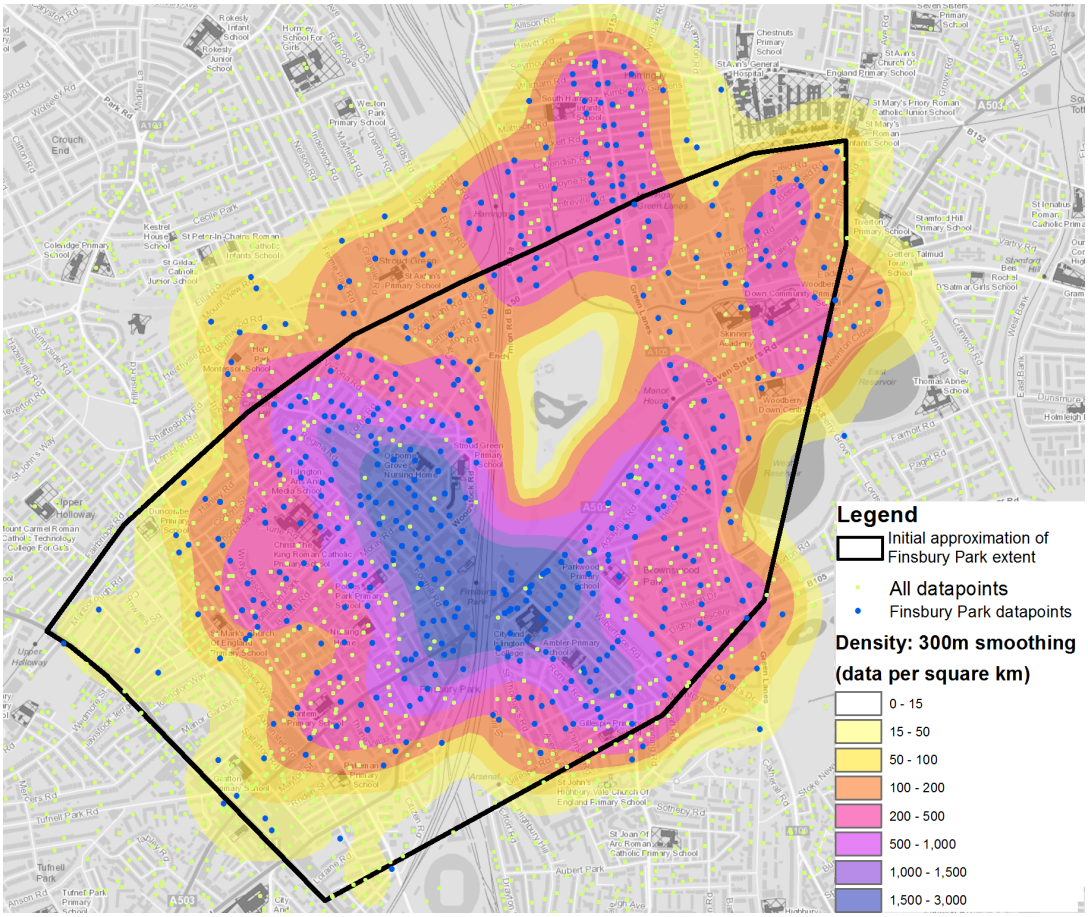
- Katherine Stansfeld

[6] https://pure.royalholloway.ac.uk/portal/en/persons/katherine-stansfeld(e5610901-a709-492b-9c95-4dedd66e34d0).html

[7] funded by a CASE award from ESRC South East Doctoral Training Centre and Ordnance Survey

One of the initial stages of such work would be to define the study area - in this case, Finsbury Park, London. Traditionally, this would be done using a combination of local knowledge, place names appearing on existing maps and potentially the labour-intensive task of discussing area extents with residents and local businesses. An early approximate indication of the area for Katherine's research is illustrated by the thick black line in Figure 8.8. Using the methods developed within this thesis, it is now relatively straightforward to computationally extract the extent of a desired place name from web derived data. This is both less time consuming and hopefully more accurate as it draws on a wide range of perspectives from a variety of differing datasources. Using the methods of this thesis, a map of Finsbury Park was extracted on behalf of Katherine's research and is illustrated by the KDE surface within Figure 8.8.

Fig. 8.8 Defining Finsbury Park for a case study area



The density surface produced by the web extraction output appears to correspond well with the area initially viewed as the extent of Finsbury Park (represented as the black outline within Figure 8.8). The main exception is to the north - where references to the area as Finsbury Park appear to spread further than first anticipated. The line of the expected boundary to the north follows the overground railway (Crouch Hill to Harringay Green Lanes) and it is interesting that this does not act as a barrier. Perhaps this is due to the many bridges and tunnels (12 in total) which mean that the railway

does not restrict movement as it would without them. In contrast, it appears that the reservoirs to the south-east (West and East reservoirs) do act as a barrier.

This map has provided a means for Katherine to identify her case study area in hopefully a more robust and accurate approach and has also opened new areas of research and issues to explore. For example she "found it fascinating that Finsbury Park and Stroud Green are two simultaneously overlapping vernacular areas, definitely given me further opportunities to explore qualitatively". In summary, the generation of urban locale boundaries developed within this work can assist research requiring the identification of neighbourhood level case study areas. Such output can provide efficient and robust extents, facilitating the research process.

8.2.2 Local community groups case study

The feedback from the chair of a local community group was particularly encouraging. The benefits can be split into the following dominant areas of concern:

- **General awareness:**
  There was a recognition that it was important to know the extent of the area a group serves and represents. In this regard, being aware of differences of opinion and not just a singular discrete view of the area's extent is important. Without an extent - it can become difficult to answer and resolve a wide range of issues, for example "Where is the greenspace in the area?", "What is its history?", "Where could walks in the area go?" and so forth.

- **Statistics for the area:**
  Census of Population data (2001 and 2011) were provided to the local group based on extents from the web extraction output (as discussed in Chapter 7: Tying other data to urban locale units). Such data were felt by the chair to be of critical use - without which they would have been reliant on council data which would have been unworkable because they were for the local Electoral Ward and not the urban locale itself. The local Ward includes two urban locales - although it is only named after one (the locale in question). The Ward had a 2011 population of 17,700 compared to 11,100 for the urban locale in question (63% of the total population). The chair felt that due to the affluence of the second urban locale included in the Ward, the official Ward statistics skewed the data and hid the real values. They felt that this had a detrimental effect on the area where the locale was often neglected, not considered to have need, and was viewed as wealthy. This issue should be partially alleviated in the next boundary changes (May 2016) when the Ward name will be changed to include both urban locality names. Such problems are not uncommon and frequently

administrative units are tagged with local names in order to gain well known spatial references. In such an approach, it is not only Electoral Wards but also Council Parking permit zones, speed limit areas, Neighbourhood Watch areas and many other generated administrative areas which get attached local names but whose extents might not be an accurate reflection of the geography of the name they adopt. Over time, such naming conventions can blur distinctions and, in turn, start to influence the naming of urban locale areas.

Even without the ambiguity issues arising through the urban locality name the chair felt that local community groups required such information, for example as that provided by extracting Census data for urban locales, in order to help inform local matters and feedback to local councillors. The chair believed that the statistics estimated for the urban locale (as opposed to official data for the Ward) brought a clarity of information, helped illuminate the issues and concerns of the area and provided evidence to support funding bids and requests for support.

The data also provided the community group with an estimate of the number households within the urban locality (4,500 - compared to the Ward figure of 7,300), which was required in order for the printing and distribution of a community newsletter. Without this estimated information, the chair felt that the newsletter would not have gone ahead as it would have been harder and perhaps impossible to have gained funding to assist with the printing and delivery costs.

- **Physical map:**
  Having a printed map of the area (with or without boundaries) was found to be useful in order to engage with people at community meetings (especially those open to the public). It was felt that the map became a focal point and drew attention - increasing and facilitating engagement.

In summary, the value of defining urban locale boundaries for local community groups lies principally in both the semantics (and the awareness concerning how other people may perceive space) and the aggregation of other data to the locales (in order to generate socio-economic information for the places that people relate to).

8.2.3 Neighbourhood planning case study

Neighbourhood planning was introduced through the Localism Act, 2011. Defining urban locale boundaries is a key element within neighbourhood planning – which seeks to assist local communities to produce plans to shape development within their local

area[8]. The precise boundaries of the neighbourhood area need to be submitted within the Forum's application of designation and require to be agreed by the local authority. It is important to note that the geographical extent of Neighbourhood Forums *cannot* overlap. This is in direct contrast with the evidence provided at the end of Chapter 7, where at a bare minimum 8 and 14% of Sheffield and Nottingham respectively were constituted as multiple, overlapping urban locales.

This work could contribute to two areas within Neighbourhood Planning. Firstly, it could aid local community groups to map their boundaries - as required by the planning process. Secondly, it could assist local council planning officers when evaluating community groups' plans - where they have to determine the accuracy and legitimacy of boundaries supplied.

- **Local community groups:**
  As discussed within Section 4.3, the London Assembly (2014) suggest that the "slow" uptake of neighbourhood planning areas may partially be due to difficulties in defining neighbourhood boundaries within complex urban settings. Maps were sent to a small number of local community groups in order to assist them with defining their boundaries. Feedback included "These maps are interesting and useful" because "different people had different perception of the extent of the boundary" and "Excellent maps. Thank you".

  According to personal communication with CLG, community groups are frequently "unable to draw a boundary because they are unsure how far their neighbourhood extends" and the web extraction output could potentially be "very useful for people who would like to apply as a Neighbourhood Area" because "such mapping assists community groups to be more representative, reach further in to their neighbourhood and help get them on to the first stages of Neighbourhood Planning".

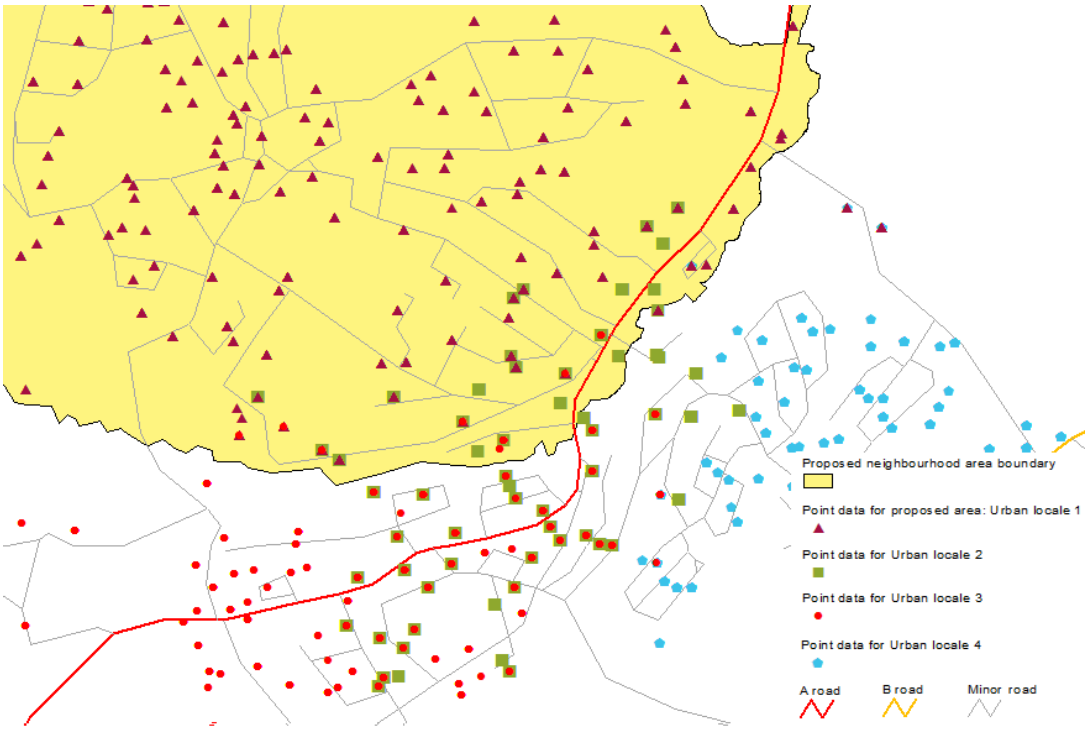
- **Local authority planning officers:**
  When extents of a number of locales were presented to three local planning officers of a local council, the nature of urban locales as vague and overlapping spaces was agreed, with two officers living close to each other but sharing differing names for the area. There was also an awareness of the difficulties in obtaining detailed urban locale information, short of asking residents directly - which would be both time consuming and costly. Data were used by the planning officers for two principal purposes. The first was to generally examine the geography of a proposed neighbourhood area to investigate the legitimacy of boundaries as supplied by the local community group (including if the proposed extent contained other areas with different names). The data were able to help the planning officers make an

[8]https://www.gov.uk/government/policies/giving-communities-more-power-in-planning-local-development/supporting-pages/neighbourhood-planning

informed decision concerning the appropriateness of the proposed boundaries and to inform them of any potential issues that might arise when the plan went for public consultation.

For a different Neighbourhood Plan, the second data requirement related to determining which urban locale name a specified small number of particular roads might identify with, in order to verify if a proposed boundary was representative. The boundary of the proposed area was relatively free-standing and thus unproblematic - except in the south-east corner, where there was more uncertainty, particularity concerning if two roads should be in or out of the proposed area. Figure 8.9 illustrates the general good fit between the proposed boundary and web extraction output for the urban locale name in question (represented in the figure by the triangle symbols) and data addressed the officers concerns about the two roads in question. Also, importantly, it demonstrated the presence of a small urban locale which had not been previously considered (urban locale number 2 within Figure 8.9 and portrayed as square symbols). Whilst this did not affect the boundary itself (the identified locale is small and completely overlaps with other urban locales), it provided the necessary information to ensure planning officers were fully aware of potential issues that could arise (for example objections to the proposal and to ensure that all other relevant community groups in the area were aware of the proposed neighbourhood area designation).

Fig. 8.9 Example of aiding Neighbourhood Planning - designation of Neighbourhood Areas



At the required very fine spatial resolutions essential for this application, the raw point value data were found to be of particular value to the planning officers. One issue was raised by the planning officers in that whilst the data were of high value to them, it was not considered possible for officers to directly work with the output of the web extraction data due to the skills that would be required in handling and querying the raw point data and smoothed surfaces within a GIS. This remark led to the implementation of translating output to kml for use in Google Earth, as demonstrated within Chapter 5.

For neighbourhood planning policy - it is important to recognise the fact that within complex urban settings, urban locales will not form neat discrete areas and there will frequently be differing perspectives on what the same space may be termed. Incorporating such truths within policy can only aid the process. Discrete areas can be formed for neighbourhood planning purposes (where areas represent the majority of support) but one should always consider and embrace the fact that differences of opinion will exist. Being aware of such 'competing' issues should be an important first step within the neighbourhood planning policy.

8.3 Further research

A number of areas of study have been identified in order to further develop the method or in applying similar techniques to other areas of work. These will now be discussed below.

8.3.1 Further development of the method

Short-term future work includes applying the procedures developed to generate urban locales within this thesis to the UK as a whole. As discussed within Chapter 5, this could be achieved with very modest data costs of just over £4,000 for the necessary Bing API requests. The potential to create the most accurate dataset of urban locales at a national scale would be of immense benefit for other research purposes and would have the potential to create significant impact both in academia and in the wider community. Once achieved, it would be possible to tie other data such as Census of Population information to the created units. Thus, it would be possible for people to have statistics for the areas that hold meaning to them as opposed to ad-hoc administrative areas.

Given the relatively ad-hoc, disjoint nature of urban locale names appearing within current gazetteers, there appears merit in combining such datasources in an attempt to

construct a more complete and comprehensive gazetteer list of urban locales. Additionally, there may be particular benefit of combining the web extraction output with Flickr point data. Whilst suitable when considering urban locales within a built-up context (for example when tying Census data to locales), the web extraction method (as discussed previously in this chapter) can not generate output for areas without postal addresses. Flickr data, however, is often available in such non-populated areas and itself is sparsest in quiet suburban settings - which in contrast can be identified comfortably by the web extraction output. The aggregation of the two datasources, thus, appears mutually beneficial. Whilst work would be required to integrate the two sets of data, for example it is important within the Flickr data to be able to disambiguate between features such as 'Crookes Valley Park' and 'Crookes' as they are two separate places. In contrast, such ambiguity is easily identified within the web extraction output due to the tight delimitation between street and settlement elements within postal addresses. Preliminary work has identified that combining the two datasets would be especially useful when parks or cemeteries were on the edge of urban locales.

Whilst methods have been formulated within the context of UK data, there is no reason why the techniques cannot be applied within other countries. There are, however, a number of issues that arise. Notably, the critical aspect is the extent to which locale names are likely to be found in postal addresses held on the web that are not administratively assigned. For example, in many countries, such as Brazil, the postal delivery system is more rigid than found in the UK and only administratively assigned, official names are found in postal addresses - even those found on the web. In some contexts it may be possible to identify certain features but not all urban locales. For example, whilst data in New Zealand generally follows administratively assigned names, it is possible to extract data concerning the CBD (or city centre which is not an official postal description). Similarly in the USA, references to Downtown may be found after the Zip code. In many countries, it is unclear if names found relate to official designation or not and ideally a detailed knowledge of the postal delivery system in the country is required to definitively identify the potential of the method. For example the urban locales are shown in red in the following examples:

- Belgium: "Rue des Canadiens 11, 7022 Hyon - Mons"
- Canada: BC V6R 2B4 Kitsilano
- France: "Avenue de la Motte-picquet Grenelle 75007 Paris"
- Greece: "Kato Patissia, Athens, 111 44"
- Greece: "Votanicos, Athens, 118 55"
- Italy: "Chiaiano 80145 Napoli"
- Malaysia: "Chow Kit 50350 Kuala Lumpur"
- Republic of Ireland: "Stillorgan Road, Blackrock, Dublin 4"
- Republic of Ireland: "Stillorgan Road, Donnybrook, Dublin 4"

- Republic of Ireland: "Unit 3, Northern Cross Business Park, Finglas, Dublin 11"
- Singapore: "168 Jalan Bukit Merah, Surbana One, Singapore 150168"
- Singapore: 26 Greenwood Avenue, Hillcrest Park, Singapore 289195"
- USA: "Atlantic Avenue... Boston, MA, 02210, Waterfront"

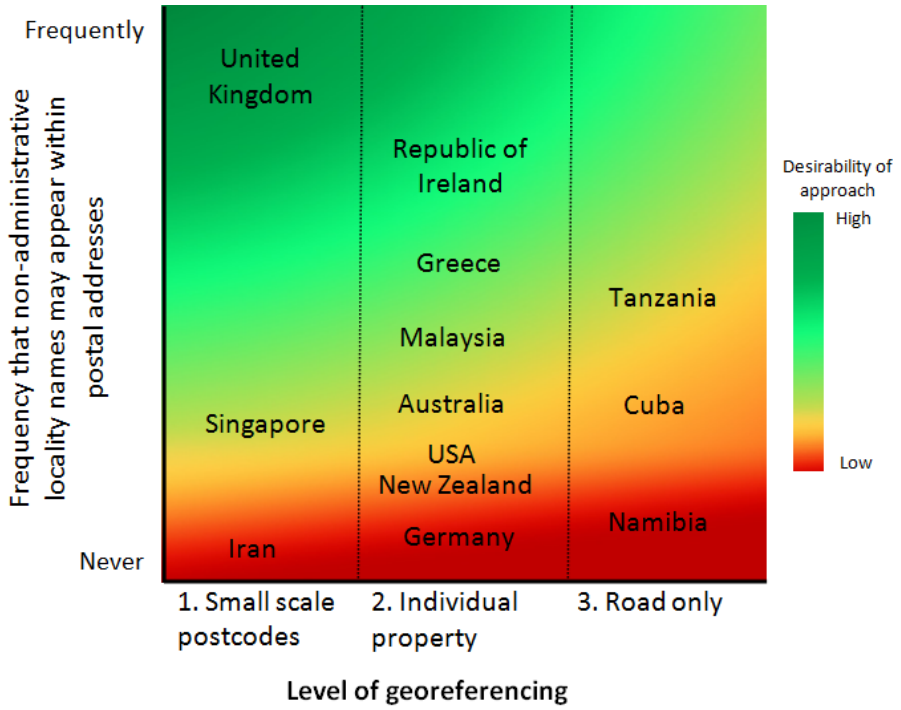
The second main requirement of the method is the need to georeference any locale data identified. This is achieved in the UK context through the postcode but could be achieved using either the individual building or road name. There is, however, a trade-off between spatial resolution and the number of data requests required. Whilst the positional accuracy of searching for *individual* addresses is high (for example house number *and* postcode), such an approach would also require a vastly increased number of search requests. Thus, the most desired outcome is for small scale postcodes whereby the positional accuracy and resolution are low (such as in the UK where each postcode only contains on average 15 houses) which helps reduce the total number of web searches required. In many cases, the only option is to search on road names which significantly reduces the positional accuracy of any identified data. The trade-off between suitability, the positional accuracy and likelihood of finding non-administrative data within postal information on the web is illustrated in Figure 8.10. With the exception of the United Kingdom, the examples portrayed within Figure 8.10 have been derived through experimentation based on manual testing of a small number of postal addresses within selected countries. Whilst a detailed knowledge of the local circumstances would be required, there is potential to adopt similar techniques within different countries.

Constructing more realistic units of analysis may have significant benefits to research concerning how we perceive, use and understand spaces within our complex urban areas. For example it may be possible to explore the degree to which cities can be described by the shapes, hierarchy and overlap of urban locales or further investigate how attachment to place is formed within areas of multiple overlapping locales. Alternatively, the differing locale geographies and intensities produced by the three groups of web classification (business information, real estate and other) could be further explored in order to identify more systematic geographic patterns. It may also be possible to extend the current method by potentially normalising the underlying input data with regard to the nature of the representativeness of the web-site generating population. The possibilities for further research are extensive.

8.3.2 Re-application of the work

Finally, it may be possible to apply a similar set of methods and procedures developed within the thesis to other web based data, as opposed to the postal addresses utilised

Fig. 8.10 A framework for the use of urban locales within postal addresses in differing countries



within this research. Furthermore, there are specific procedures developed within this research that may be applied in a wider context. For example, the noise reduction procedure using differing density scales (see pages 94-95) could easily be replicated for other sources of data where errors are likely to be high. This work demonstrated the importance of considering the use of both absolute data output *and* percentage, demonstrating the differences that these two similar processes produce. Such distinctions are important in many social phenomena where alternative absolute values merely represent the underlying population distribution rather than the desired process itself. The use of linear breaklines developed in order to determine if KDE surfaces should smooth beyond specified linear features can also be applied to other contexts using 'hotspot mapping' techniques (for example high volume crime data).

8.4 Summary and urban locale conclusions

This section firstly summarises this chapter, before concluding on the topic of vague urban locales (drawing on content from Chapters 4 to 8). In summary, this chapter has outlined advantages and limitations of the method not discussed in the previous chapter and has further demonstrated the value to the work through the use of three case study applications, before turning attention to examining the potential related further work. Urban locales are considered a key element of UK Government social

policy and yet no national set of boundaries exist that accurately reflect these valuable social units. Instead administrative areas are used as proxies even though current academic literature has demonstrated that these are not the same as the areas that people intuitively conceive and identify with (Flemmings, 2010; Schockaert and De Cock, 2007; Twaroch et al., 2009a).

The urban locale research (discussed in Chapter 4 through to Chapter 8) has demonstrated the feasibility of mapping urban locale boundaries by combining some simple linguistic models with web based data generated by individuals. This is achieved through the passive mining of postal addresses held online. Such data represent current perspectives, especially when compared to present mapping alternatives where urban locale names may persist from nineteenth century surveying of opinions from a small select number of local privileged occupations. Through the use of linear breaklines, the method can account for the fact that features, such as major roads, rivers and railways, at times act as a strict barrier to urban locale extension but this is not always the case (as the example of Broomhall, Sheffield crossing the dual carriageway, A61, demonstrates). The various settlement sizes for which data have been generated (ranging from small urban areas such as Ashby-de-la-Zouch and Ludlow through to extensive built up areas of Nottingham and Sheffield and the densely populated districts of Camden and Westminster) demonstrate that the method can be applied equally to urban areas of any size and highlight the potential for generating a national coverage. Importantly, the validation process has demonstrated that urban locales identified through such an approach (using postal addresses held online) are representative of locales generally and do not simply reflect a particular 'kind' or sub-set of urban locales.

The unique contribution of the proposed methodology includes the fact that the procedure does not require the names of the units of interest to be known *a priori*. This is of particular importance given that urban locale level gazetteers have been shown to rarely be comprehensive. The tight delimitation between street and city elements also reduces issues of ambiguity, facilitating for example the clear disjuncture of references to 'Birley' (Sheffield) from those of 'Birley Carr' (Sheffield). The proposed method is one that is highly scalable, extensible and dynamic. It is then possible to aggregate other socio-economic data (such as Census information) to these important social units. With the growth of digital data availability – this work is timely and can provide a framework for just about any other socio-economic data to be aggregated to. Thus, data can be generated, analysed and presented using the *real* units that people understand and emphasise with. The work has demonstrated that it is possible to construct dynamic, probabilistic and actionable definitions of urban locale that can be used as the substrate for important business, policy and social decisions.

Output is based upon information from a wide range of source data which will aid robustness. Thus, it is possible to construct different geographies based on the source URL of the data and compare Real estate views of urban locale boundaries to those obtained from Business information. Using such an approach it was possible to demonstrate that urban locale geographies from Real estate sources were generally similar to output from Business information data (97% of cells in Nottingham having the same majority urban locale when smoothed to 100m), although noticeable differences for specific urban locales were also evident.

The work also highlights the potential utility of leveraging the percentage of search returns that relate to a named neighbourhood (in addition to the usual absolute number of data returns). Whilst absolute data maps tend to identify concentrations focused around the core (commonly shopping areas) of the locale, the percentage maps exhibit a rather different geography, identifying where the majority of people may associate with the urban locale names. It should be stressed that the method to identify urban locales only generates perspectives for populated spaces and makes no explicit attempt to model unpopulated areas.

It was also possible to investigate the temporal change in urban locales in Sheffield by running the analysis for two separate time periods (2012 and 2014). Comparison of the majority urban locale function revealed only a 5% change in cells within the two year period, demonstrating the high degree of temporal stability. Tying other data to the urban locale units (for example Twitter tweets, house prices or Census of Population data) demonstrate the potential of urban locales to add value to both sets of data. The aggregation of Census of Population data to the urban locales also allowed for further analysis to be undertaken concerning the urban locale data units, including testing homogeneity. Results demonstrated that whilst urban locales in Nottingham did not appear homogeneous in terms of socio-economic conditions, they were more homogeneous than other alternative administrative data units at a similar spatial scale. The extent of heterogeneity within urban locales appears to give support to the definition of urban locales based on identification of areas sharing the same *name*, as opposed to those based on some measurement of homogeneity of units. (in terms of socio-economic characteristics). For example, the work of Martin (1998) seeks to identify neighbourhoods in terms of homogeneous Census data, whilst other work has been undertaken to construct homogeneous geographic spaces in terms of similar “tweets” (Arribas-Bel, 2015) or Foursquare “check-ins” (Noulas et al., 2011). Whilst such work identifies functional areas, the generated units lack an association with *named areas* to which people are more likely to identify with and relate to. As such, this work places the *name* of urban locales at the centre of the process.

Comparisons were also made between the level of data returns and the underlying population - in order to identify geographic areas with higher or lower than expected urban locale data output. Finally, the degree of overlap between urban locales was investigated and found to be *at least* 8% of area in Sheffield and 14% in Nottingham. Critically, this illustrates the degree of error that any discrete formation of urban locales could not seek to avoid and demonstrates the importance of the principle of vague urban locales.

The benefits of the proposed work are pervasive and far beyond the small number of case study applications discussed within this chapter. There is considerable potential to inform and contribute to existing academic and governmental research concerning neighbourhood level effects. Sampson (2012) discusses how neighbourhoods can influence a host of factors from juvenile delinquency, adult crime, teenage sexuality, substance abuse, and mental health among other outcomes. The debate concerning precisely how neighbourhood level effects transpire, however, is ongoing (see Galster (2012)). The generation in the UK of the vague urban locale social units by which we live our lives by would open up a number of research areas within academia.

The work holds particular importance not only for semantics and web search services more generally but also to governmental research where the ability for residents to understand and empathise with various governmental data may in part be linked to the appropriateness of the statistical units used to disseminate that data. There are also a great number of advantages of this work to industry – especially current mapping agencies but also to all the vast number of industries that provide public data. Instead of producing statistical data for administrative areas, it would be possible to generate output for the named areas that the public understand and relate to. Benefits to wider society would extend from this, enabling a better working understanding of how social processes are formed and reinforced. Current mapping only contain urban locales as point references but this body of work demonstrates how digital maps could contain full area extents. It would also be able to identify historic named places within existing maps and scale place names by population, making maps easier to read and interpret. The use of the method for assisting completeness for the European INSPIRE Directive and within neighbourhood planning is also clear.

This research has demonstrated that the mass linguistic data now available online can be combined with computational models to automatically produce urban locale definitions that are both probabilistic and dynamic. Currently this has been achieved through passive mining of postal addresses from the web and, uniquely, without a priori knowledge of neighbourhood names. Even with the relatively simple methods used to produce viable geospatial intelligence, the potential of combining large scale web data with linguistic models is evident.

Chapter 9

Conclusions: Generating Vague Places from web data

9.1 Introduction

This final chapter summarises the findings of the research, the aims of which were outlined in Chapter 1. The following section discusses how the aims were achieved though the implementation of the research objectives. Section 9.3 outlines the main contributions made by this doctoral work, whilst the last section highlights some final key remarks.

9.2 Summary of research findings

The opening paragraph of Chapter 1 included Davies' (2009) statement that vagueness is both a blessing and a curse for GISc. A blessing for it provides the opportunity to devise innovative ways in order to help overcome the issue, but a curse - for systems are not designed to handle it (Davies, 2009). This body of work has demonstrated that using very simple approaches, vagueness as a concept, can easily be incorporated within geographic information, be this in terms of vague attributes (as demonstrated through the use of vague settlement classification), or within vague place names or boundaries. By incorporating vagueness, GISc models will represent a truer reflection of the ways in which we perceive the social constructs of our world.

The aim of this thesis was to (1) investigate an approach to automatically generate vague, probabilistic geographical information concerning place through extracting

differing perspectives from the web, whilst (2) assuring the quality of the vague information produced and (3) demonstrating the value of the vague information with examples of its use. This was achieved through applying the concept of vagueness to all three elements of place as identified by Goodchild and Hill (2008): place types, place names and spatial boundaries of places.

Chapter 1 established that in order to achieve these research aims, a number of ancillary objectives needed to be accomplished. Table 1.1 presented each chapter and their related research objectives. The work of each chapter will be discussed in terms of its success at meeting its stated research objective.

The **first objective** was to investigate the concept of vagueness within geographic information concerning place. This was addressed predominately in Chapter 2 which explored vagueness through the notions of natural vagueness (as demonstrated by sorites paradox) and multiperson disagreement. The chapter outlined different methods of modelling vagueness, focusing on fuzzy set theory and probabilistic approaches. The use of passive crowdsourcing of the web for differing perspectives on place was introduced. Such work built on that of Jones et al. (2008a) but extended the method so that data were extracted from a range of different sources (as opposed to data extraction from only accommodation sources for example). Furthermore, output was not limited to a set number of search returns and critically the research applied the concept of vagueness, not only to spatial boundaries but also to the other place elements: place types (or attributes more generally) and place names (through spelling issues and synonym identification). The work concerning this objective was complemented through further objectives which sought to construct vague place type / settlement classification (Chapter 3) and investigate the concept of vague 'neighbourhood' level spatial boundaries and vague place names (Chapter 4).

The **second objective** was to explore *how* a vague settlement classification could be formulated from data held on the web. This was achieved by developing a methodology in Chapter 3 that searched for online data containing differing viewpoints on how people might define specific settlements. This was realized by using the search term "* is a ..." where the asterisk would represent the settlement name and some form of categorisation may occur following the search term. Scale was achieved through the use of searching through the Bing API. A parent term was also required within the returned output in order to ensure that the settlement in the search returns actually referred to the place of concern - reducing issues of ambiguity.

The findings of the second objective were implemented in order to construct vague ***place types*** from web data - which was the **third research objective** of this thesis and can also be found in Chapter 3. Online material containing references to settlement

terms could be extracted from web sources in order to construct a vague and probabilistic settlement classification. Chapter 3 also demonstrated the high level of variations in the manner in which people classify places and as such proved the importance of constructing vague settlement classification information. On average, questionnaire responses classified a settlement in 2.8 different classes of the 15 available with at least a level of 10% agreement amongst participants - demonstrating the highly variable degree to which different people perceive and classify place. When size distinctions were ignored, the figure was still high, standing on average at 2.1 of the 9 classes with a minimum of 10% agreement. In addition, the potential of extracting the language and terms of how place was generally described was established (for example for the identification of 'market towns' or 'picturesque villages'). When population numbers were tied to the different categories, the large variations both within and between specific classes was evident. For example, a 'small town' could be used for a description of a settlement which could contain anything from 3,000 households to 17,000. The substantial overlaps between the differing categorises (as shown within Figure 3.4) demonstrates that population size alone is not an accurate measure on which to base a settlement classification when embracing how people perceive place.

Undertaking the **fourth objective** was a critical component of the vague settlement classification research. This objective sought to assure the quality of the vague geographic information extracted from the web and ensure that it was representative of that found when interviewing people directly. Without such a comparison, the output of the vague place types would be of severely limited value as it might not necessary reflect general public opinion concerning settlement classification. The validation for the vague settlement classification (Chapter 3) centred on the comparison of output from the web extraction content and that from 281 questionnaire responses which classified 60 different settlements. It found that initially only 46% ($y = 1.1073x - 0.6709, p < .001$) of the variation in questionnaire responses was explained by the web extraction output through linear regression. This figure, however, rose to 72% ($y = 1.0654x - 0.6132, p < .001$) when size distinctions were removed (ie not distinguishing between 'small village' and 'village' categorises). This supported the notion that the questionnaire responses were more likely to include a size distinction within their categorisation (probably due to having pre-defined lists containing size distinction data). The difficulty of categorising places that were within or part of a larger settlement (urban locales) was also highlighted.

The **fifth objective** was to examine the extent of vagueness within ***place names*** extracted from web data. Chapter 5 established a suitable method to identify the vagueness in place names of urban locales through misspelling and synonym identification. Spelling variants were revealed through a combination of both the Ratcliff/Obershelp and Levenshtein Distance algorithms. Synonym identification was undertaken during

post-processing using linear grid regression on the final output grids for each urban locale. Chapter 8 specifically discussed the extent to which specified places were recognised by different names/spellings. Whilst the total data volumes involved may be considered modest (1.2% of data records for the Sheffield case study involved spelling variants identified by the method), most urban locales exhibited some degree of spelling variations (67% of locales). Furthermore, the power of determining the likelihood in which different spelling variants may be likely to be found could be valuable for improving current web searching (for example 36% of references to 'Carterknowle' within the urban locale data refer to the place as 'Carter Knowle' and 15% of data for 'Frecheville' spell the place as 'Frechville').

The investigation of *how* bounded "urban locales" could be formulated (**objective six**) was considered in Chapters 4 and 5. The theory and background to this objective was discussed in Chapter 4 which also explored the existing literature concerning the generation of such units of analysis, whilst the proposed method was outlined in Chapter 5. Urban locales were defined simply as any named area, or sub-part, of a town or city (combining concepts such as suburbs, neighbourhoods, housing estates, non-residential industrial estates and business parks). The construction of such urban locale information was achieved through extracting differing perspectives from web data - specifically from 'neighbourhood' level places appearing within postal addresses held on the web. Such instances, by definition, must be generated by individuals, as official postal addresses do not contain such 'neighbourhood' level references[1].

Point data relating to each urban locale were then cleaned, with erroneous data removed, before KDE surfaces could be constructed. Such grids allow for vague and overlapping geographies to exist. Both the absolute number of datapoints and the percentage of data were estimated. Each measure provides different insights of the underlying data - with the absolute data highlighting urban locale centres, whilst the percentage data illustrate areas of high agreement concerning the place name. A method was also developed that only permitted KDE surfaces to expand across significant linear features (such as rivers and railways) where the underlying point data supported the notion but where the lines acted as hard break-lines when no support was found. Such an approach allowed the necessary flexibility, where due to the frequently historical nature of urban locale creation, there are no fixed rules of whether locales cross such linear features or not. Additionally, the initial method was expanded within Chapter 7 to include three additional elements:

- distinguishing between three different sets of source information (business information, real estate data and other data) in order to define urban locale areas based upon different web groups,

[1] with the exception of Dependent Localities discussed in Chapter 5

- the ability to define temporal change in urban locales through time-slices, and
- the process of how to tie other (point or polygon attribute information) to the urban locale bounded data.

This was followed-up by the **seventh objective** that implemented the previous findings in order to construct vague ***spatial boundaries*** of urban locales using data mined from the web. The method for this objective was outlined within Chapter 5, whilst Chapter 6 contained sample output and demonstrated that the method was highly scalable, extensible and dynamic. Critically, the web extraction output, supported by the validation (specifically the divergent views of asking residents the name of their area in three Sheffield streets) demonstrated the importance of vague urban locales. The degree to which urban locales overlapped was *at least* 8% of the area in Sheffield and 14% in Nottingham. Such figures highlight that by de-facto, discrete methods of mapping urban locales must be wrong in this bare minimum of cases, stressing the value of vague mapping. The various settlement sizes for which data were generated demonstrated that the method can be applied equally to urban areas of any size and highlighted the potential for generating a national coverage. The tight delimitation between street and city elements also reduced issues of ambiguity. Furthermore, the strength of the method lies in the fact that the procedure does not require the names of the units of interest to be known *a priori*. This is of particular importance given that urban locale level gazetteers were shown to rarely be comprehensive.

The **eighth objective** underpins the importance of the research through quality assuring the vague urban locale data using a mixture of statistical techniques and surveys. For this research (concerning the generation of vague geographic information from the web) to hold value it must be demonstrated that the outputs are comparable to that found when asking people directly. The validation of the urban locale names and boundaries (Chapter 6) was conducted using a range of different comparisons. Firstly, *names* in the case study areas of Sheffield and Nottingham were compared with various centre-point gazetteers. Although a great deal of disparity was found within the different gazetteer sources, the web extraction data was found to have a high degree of agreement between the other datasources (outperforming OSM, Geonames and Yahoo! GeoPlanet data and achieving similar levels of agreement to OS data). For the urban locale names where at least three other datasources were in agreement, the web extraction data identified the joint highest proportion in Nottingham (98.3%) and the second highest proportion (behind the combined OS data) in Sheffield (93.0%). Given that data are drawn from postal address information harvested from the web, the high degree of similarity of names with the other urban locale datasources demonstrates that the data are representative of urban locales as a whole and are not merely a specific 'kind' or sub-set of urban locales.

The validation of urban locale *geographies* was undertaken using three approaches. Firstly, comparisons were made against the limited number of datasources containing urban locale boundary information. Whilst the web extraction output achieved good levels of agreement with the other sources, the limitation of such an approach is that it is unclear which data represent a truer reflection, as the level of agreement between the sources may not necessarily reflect a more accurate viewpoint. Secondly, a number of stakeholder views were collected (from local council planning officers and members of a local community group). Whilst again the web extraction output performed well, due to limits of detailed local knowledge such validation is limited to a few case study areas and are merely the reflections of a limited number of stakeholders. The most suitable validation is through the comparison against resident views. This was achieved using three different approaches:

1 68 responses of urban locale names and postcodes from Nottingham residents were compared with the web extraction output and different locale geographies. It was found that there was an 98% agreement between resident views and the web extraction output. The datasources with the next highest levels of accord were Wikimapia and OS MRDP which both had only 70% agreement.

2 the detailed geography of a single locale was compared with over 400 resident views and again resident output was closest to the web extraction data (72% agreement of cells), compared to OS MRDP (32% agreement), Wikimapia (26%) and Quattroshapes (19%).

3 residents of particular streets in Sheffield were asked what they thought their area was called, targeting three streets where the web extraction output indicated a highly variable dis-juncture of opinion. No statistical difference was found between the web extraction data and resident views for the frequency of residents identifying with different locales in the streets ($X^2 = 0.09; df = 2, p < 0.05$). This not only supports the validation of the web extraction output (being similar to that when asking residents directly) but the output also more fundamentally proves the notion and requirement of *vague* urban locale boundaries, where residents within close proximately may hold very different views on what the area should be called.

Finally, the validation of the three different web grouping classification was discussed in Chapter 7 and showed a successful categorisation rate of 87.87% compared to a manual assignment of 121,146 URLs for the Nottingham case study. In order to check for bias in the manual categorisation, two independent members classified 100 URLs, with high levels of agreement (94% and 93%).

A key driver behind the research was the standpoint that urban locale geographies provide a robust foundation for a host of social applications ranging from criminology to epidemiology. Once urban locales have been defined they can be used as new units for analysis to which other data can be tied to. The **ninth objective** was to show evidence of the added value of the vague data when linked to other socio-economical data from other sources. The method of linking data to urban locale boundaries was demonstrated in Chapter 7 for both point (Twitter tweets and house price information) and polygon data (Census of Population information). The chapter showed that such data linkages were relatively straight-forward, using existing GIS techniques and simple statistical weighting approaches. The added value of such data was discussed in Chapter 8 through the use of a case study whereby a local forum were able to calculate appropriate socio-economic statistical output for their area of concern to a degree that was not previously possible. Census of Population data were estimated for their urban locale which contrasted to the administrative (Ward level) data that was previously used and whose boundaries approximately contained two different locales. It was also possible to identify longer-term trends by comparing change in Census data (2001 to 2011).

The final objective (**ten**) was to evaluate the success of the work, examine potential uses for such information and discuss possibilities for further work. This was principally achieved in Chapters 3, 8 and 9. The work has been able to successfully demonstrate that vague geographic information are more likely to be a truer reflection of how people perceive the social process within the real world, as opposed to forming discrete objects to model features. This was not only the case for vague settlement classifications but also for vague urban locale names and boundaries and where value was demonstrated through the validation procedures. Natural vagueness is omnipresent in the classification of space - there cannot be a universal solution to discrete categorisation of a variable measured on a continuous scale. As such, there will always be discussion over when a 'village' becomes a 'town' or if a town is a 'market town' or not. In a similar vein, natural vagueness also persists within the spatial definition of vague objects, as such constructs rarely tie neatly to the physical world with complete agreement and through the problem of definition it is impossible to draw truly discrete boundaries within GISc. Such phenomena are further compounded by the fact that different people will view space on an individual basis, leading to differences of interpretation through multiperson disagreement. This chapter has argued the extent to which the work has met its research objectives and has demonstrated the high level of success for each objective.

In order to examine potential uses of the vague data, Chapter 8 outlined three example case study application areas. The case studies included: (1) the designation and delimiting of case-study areas for research purposes; (2) the use of specific locale

boundaries and associated Census data to assist local forum groups; and (3) to support neighbourhood planning within local councils. Although such examples highlight example uses, they do not demonstrate the extensive potential utility that the development of vague urban locales could hold at a national coverage. The ways in which neighbourhood effects transpire are as yet still unclear (Galster, 2012), although it is widely recognised that areal units at the urban locale scale have important ramifications for society - influencing a wide range of factors including juvenile delinquency, adult crime, teenage sexuality, substance abuse, and mental health among other outcomes (Sampson, 2012). Given such issues, it is clear that a set of national urban locales for the UK could only assist this neighbourhood effect research. Furthermore, once urban locale boundaries have been constructed the value of linking other socio-economic data to these important social units has been demonstrated.

It is also important to note that given the requirements of the European INSPIRE Directive to record lists of geographic names (including suburbs) and the fact that this thesis has demonstrated the incompleteness of such data including the OS 1:50K gazetteer used to fulfil the INSPIRE obligation; the proposed method could support work to construct more comprehensive gazetteers for urban locale level places. A further potential use of the work would be to provide alternative spellings within urban locale gazetteers. The variants in spelling and synonyms have been shown to be important for semantics - ensuring that people are not only aware when apparently different names may actually refer to identical spatial units but also being informed of the potential spatial interpretations that others may associate with the specified place name (both in terms of name and boundaries).

The potential avenues for developing and enhancing this work were discussed within the future work section of Chapter 8. It argued that there were two bodies of work that would (1) continue to develop and expand the method or (2) would re-apply the work in different ways/contexts. The first, included the potential of applying the method to create a national set of urban locales. Whilst such a process was outside the financial constraints of this research, only a very modest sum would be required to purchase the necessary Bing API search requests (just over £4,000) to undertake such an endeavour. Other work listed the potential of aggregating current gazetteers and web extraction output in order to create a more comprehensive list of urban locales (in line with INSPIRE directive requirements); combining web extraction data with Flickr photograph point data; and a framework for the use of urban locales within postal addresses in differing countries. Finally, the creation of urban locale data would enable a plethora of other academic research to be undertaken utilising these important social units. Such analysis would be a similar vein to that undertaken within Chapter 7: assessing the homogeneity and extent of overlap of units and the relationship between urban locale data returns and the underlying population density.

The second area of future work concerned re-applying the method to different sets of data. For example, the procedure of using differing density scales for noise reduction could easily be replicated for data where errors are likely to be high. The work highlighted the importance of considering both the absolute values *and* percentage of data output when analysing *any* spatial data, otherwise results may simply be reflective of the underlying population distribution. Finally, the use of linear breaklines within KDE surfaces could be applied in many other social contexts (rather than current uses limited to terrain operations).

9.3 Contributions of this research

This doctoral work makes a number of contributions to knowledge through demonstrating the general need for vague geographic information - specifically within settlement classification (vague attributions) and urban locale (vague names and boundaries). Each will be discussed in turn.

9.3.1 Vague attributes for settlement classification

Settlement classifications usually consist of discrete categorisations of place, whereby each settlement can only be identified as falling within a single class or categorisation. This work demonstrated that different people will view and interpret place in a variety of ways, and hence, a vague classification would represent a more realistic approach to how people generally categorise place. For example, the validation established that whilst 87% of respondents classified Nottingham as a 'city', a further 8% categorised it as a 'town'[2]. A novel method for deriving vague settlement classification was constructed using search requests sent through the Bing API (as briefly outlined previously in this chapter and discussed in more detail in Chapter 3). Using such an approach, the web extraction output was found to report comparable figures and in the case of Nottingham returned a response of 82% of data classifying Nottingham as a 'city'. Similarly, the web extraction output defined Thurmaston (near Leicester) as both a village (62%) *and* a suburb (38%). Such findings give strong weight to the importance of defining vague, rather than discrete, settlement classifications if one desires to capture the real views and perceptions of place that people hold. Critically, the validation of the output demonstrated that vague settlement classifications generated from web data were representative of those obtained from interviewing people.

[2]irrespective of size qualifiers

9.3.2 Vague place names for urban locales

Although the need for different spelling variants within place gazetteers is discussed in the literature (for example Jones et al. (2001)), the difficulties of collecting such information means that few gazetteers currently offer such attribute data. Spelling variants are identified as part of the data cleansing process of urban locale generation (see Chapter 5) and opens up the potential of further research concerning the extent to which spelling variants and typing errors manifest themselves within place names posted on the web. The synonym identification method (see Chapter 5) uses a simple linear grid regression comparison in order to find synonyms, not in terms of names but in space - through their spatial boundaries.

9.3.3 Vague spatial boundaries for urban locales

The most significant contributions to knowledge made by this work arise from the construction of vague spatial boundaries for urban locales. Firstly, the concept of urban locales is developed, encompassing all named 'neighbourhood' level, sub-parts of urban areas (such as neighbourhoods, suburbs, industrial estates and so forth). Such a construction is useful as it draws together the separate literature from neighbourhood and suburb sources. The identification of 'neighbourhood' level places through the aggregation of administrative units based on measures of homogeneity (with some attempt of post process naming) was shown to be questionable due to the levels of heterogeneity found within the units. In contrast, the concept of urban locale developed within this body of research rests on the assumption that it is the '*name*' that critically binds the space together. The generated method, thus, draws differing perspectives of urban locale point location from the passive crowdsourcing of web based data and subsequently uses KDE to generate continuous surfaces as vague extents. The method, thus, maintains the *name* at the centre of the process. Interestingly, the method was able to identify boundaries for 'city centre' urban locales whose identification has proved troublesome (Thurstain-Goodwin and Unwin, 2000). Importantly, the method was shown to hold general applicability to urban settlements with a wide spectrum of population sizes. The work contributes an automated method that allows for the national collection of robust vague urban locales with high data volumes from diverse source data, without the need to know names *a priori* and with reduced levels of ambiguity.

Secondly, and most fundamentally, the validation supported the notion that urban locales (in terms of both names and boundaries) generated from web data were representative of those obtained through asking people. Refer back to the previous

section (objective eight of the Summary of Research Findings) for a detailed account. The third contribution to knowledge was the demonstration of the extent to which urban locales nest and overlap. This was discussed elsewhere in this chapter where the degree to which urban locales overlapped was *at least* 8% of area in Sheffield and 14% in Nottingham. This illustrates the essential need for vague urban locales, as opposed to discrete boundaries. Such findings have important implications, for example to neighbourhood planning which assumes definitive extents.

The fourth contribution rests with applying the principle of linear breaklines within the KDE process (usually restricted to terrain functions) to a new context - within urban locale identification. The procedure could equally be applied to many other social processes where linear features (such as rivers and railways) may, but not necessarily always, act as hard breaklines. Assigning the URLs of the urban locale input point data into three different groupings (business information sources; real estate sources and other sources) demonstrated how different boundaries could be generated for varying groups and represented the fifth research contribution. Importantly, the work demonstrated that, although exceptions existed, generally real estate views were consistent with boundaries formed from the other classifications.

The sixth knowledge contribution relates to the analysis undertaken, demonstrating that urban locales in Nottingham were not necessarily homogeneous in terms of socio-economic data but were generally more homogeneous than other alternative administrative data of similar spatial scales. This is important given that frequently 'neighbourhood' level places are identified through the process of aggregating smaller statistical units based on measures of homogeneity of characteristics. The final novel contribution was the investigation of the spatial relationships between the frequency of address information held on the web and the underlying population density. Exploration included the overall web visibility (the total number of addresses regardless of urban locale references) and specifically the extent to which people included urban locale level information within postal addresses held on the web and the underlying population.

9.4 Final remarks

Vagueness, through both natural vagueness and multiperson disaggregate, is an inherent property of geographical data (Fisher, 2000). GIS, however, are more commonly used to generate and process discrete geographic objects. This thesis has been able to demonstrate that vague information can relatively easily be incorporated within GIS. The methods developed within the thesis generated vague geographic information for

places, based on differing opinions and perceptions extracted from web derived data. Vagueness was applied to all three elements of place as defined by Goodchild and Hill (2008): place types, place names and spatial boundaries of places.

The aim of the research has been to (1) investigate an approach to automatically generate vague, probabilistic geographical information concerning place through extracting differing perspectives from the web, whilst (2) assuring the quality of the vague information produced and (3) demonstrating the value of the vague information with examples of its use. In order to achieve the first aim, the thesis developed methods to extract differing perspectives of place from the passive crowdsourcing of web data - constructing vague representations for:

- place type settlement classification (attribute data),
- place names (through misspelling and synonym identification), and
- spatial boundaries for 'neighbourhood' level places - termed urban locales.

Data were compared against similar information obtained from asking public opinion, in order to validate the quality of the web extraction output. Critically, the output generated were found to be representative of views obtained through asking people. Whilst data drawn from directly interviewing participants will be limited in terms of sample sizes, data extracted from the web can be achieved at scale through automated and dynamic procedures at national coverages. Finally, case studies of vague urban locale data were introduced in order to demonstrate the value of such data.

The novel method developed to generate vague urban locale boundaries did not require the knowledge of locale names *a priori*. Furthermore, it was able to use the underlying data to assess if KDE smoothing was required beyond specific linear feature physical barriers (such as rivers and railways) and due to the tight identification process (between street and settlement names with postal addresses held on the web) had reduced issues of ambiguity. The method was also able to generate vague place names through the identification of spelling variants and potential synonyms.

Importantly, the validation data, drawn from public opinion, supported the notion that vagueness was omnipresent within geographic information concerning place. The incorporation of vagueness within place models can, thus, improve the accuracy of real-world representations within GIS.

References

Abdelhaq, H., Gertz, M., and Sengstock, C. (2013). Spatio-temporal characteristics of bursty words in twitter streams. In *ACM International Conference on Advances in GIS (SIGSPATIAL)*, pages 194–203, Orlando, Florida, USA.

Ahlers, D. (2012). Local web search examined. In Lewandowski, D., editor, *Web Search Engine Research (Library and Information Science)*, pages 47–78. Emerald Group Publishing Limited: Bingley, UK.

Ahlqvist, O., Bibby, P., Duckham, M., Fisher, P., Harvey, F., and Schuurman, N. (2005). Not just objects: Reconstructing objects. In Fisher, P., editor, *Re-Presenting GIS*, pages 17–25. John Wiley: New York.

Alani, H., Jones, C. B., and Tudhope, D. (2001). Voronoi-based region approximation for geographical information. *International Journal of Geographical Information Science*, 15(4):287–306.

Alazzawi, A. N. and Abdelmoty, A. I. (2012). What can I do there? Towards the automatic discovery of place-related services and activities. *International Journal of Geographical Information Science*, 26(2):345 – 364.

Almadani, F., Fisher, P., and Jarvis, C. (2013). Modelling spatial extents of village territories from postal address records. In *GIS RESEARCH UK (GISRUK), Proc. Of the GIS Research UK 21st Annual Conference*, Liverpool, UK.

Ames, D. and Anselmo, A. (2008). Bayesian network integration with GIS. In Shekhar, S. and Xiong, H., editors, *Encyclopedia of GIS*, pages 39–45. Springer: New York.

Angelaccio, M. and Buttarazzi, B. (2000). A visualisation system for web local search. In *IEEE International Conference on Information Visualisation*, page 474 – 478, London, UK.

Arampatzis, A., Kreveld, M. V., Reinbacher, I., Jones, C. B., Vaid, S., Clough, P., Joho, H., and Sanderson, M. (2006). Web-based Delineation of Imprecise Regions. *Computers Environment and Urban Systems*, 30(4):436–459.

Arribas-Bel, D. (2015). The spoken postcodes. *Regional Studies, Regional Science*, 2(1):457–460.

Bailey, N. and Pill, M. (2014). The potential for neighbourhood regeneration in a period of austerity: changing forms of neighbourhood governance in two cities. *Journal of Urban Regeneration and Renewal*, 7(2):150–163.

Bar-Ilan, J. (2004). The use of web search engines in information science research. *Annual Review of Information Science and Technology*, 38(1):231–288.

Batty, M., Crooks, A., Hudson-Smith, A., Milton, R., Anand, S., Jackson, M., and Morley, J. (2010). Data mash-ups and the future of mapping. A report by the Centre for Advanced Spatial Analysis (CASA), University College London and the Centre for Geospatial Science (CGS), University of Nottingham for JISC - Technology and Standards Watch (TechWatch).

Beckett, J. (2005). *City Status in the British Isles, 1830–2002*. Aldershot, Ashgate.

Bennett, B. (2001). What is a forest? On the vagueness of certain geographic concepts. *Topoi*, 20:189–201.

Bergman, M. K. (2001). White paper: the deep web: surfacing hidden value. *Journal of electronic publishing*, 7(1).

Bibby, P. (2005). GIS, worldmaking and natural language. In Fisher, P., editor, *Re-Presenting GIS*, pages 55–83. John Wiley: New York.

Bibby, P. R. (2009). Land use change in Britain. *Land Use Policy*, 26S:S2–S13.

Bibby, P. R. and Brindley, P. (2006). Residential development since 2000. Report for Communities and Local Government (CLG).

Bibby, P. R. and Brindley, P. (2007). Historic farm complexes in current socio-economic context: High Weald. Report for High Weald AONB.

Bibby, P. R. and Brindley, P. (2010). Current use of historic farm property in the West Midlands: Introduction and overview. Report for Advantage West Midlands and English Heritage.

Bibby, P. R. and Brindley, P. (2013). Urban and Rural Area Definitions for Policy Purposes in England and Wales: Methodology. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/239477/RUC11methodologypaperaug_28_Aug.pdf.

Bibby, P. R. and Brindley, P. (2014). 2011 Rural-Urban Classification of Local Authority Districts in England: Methodology. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/389779/RUCLAD2011_Methodology.pdf.

Bibby, P. R. and Shepherd, J. (2004). Developing a new classification of urban and rural areas for policy purposes – the methodology. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/239084/2001-rural-urban-definition-methodology-technical.pdf.

Bittner, T. and Stell, J. (2003). Stratified rough sets and vagueness. In W. Kuhn, M. W. and Timpf, S., editors, *COSIT 2003: spatial information theory - foundations of Geographic Information Science*, page 270–286. Springer-Verlag: Berlin.

Black, M. (1937). Vagueness. an exercise in logical analysis. *Philosophy of Science*, 4(4):427–455.

Blockland, T. (2001). Bricks, mortar, memories: neighbourhoods and networks in collective acts of remembering. *Journal of Urban and Regional Research*, 25(2):268–283.

Bowman, A. and Azzalini, A. (1997). *The Kernel Approach with S-Plus Illustrations*. Clarendon Press: Oxford.

Brindley, P., Goulding, J., and Wilson, M. L. (2014a). A data driven approach to mapping urban neighbourhoods. In *ACM International Conference on Advances in GIS (SIGSPATIAL)*, pages 437–440, Dallas, Texas, USA.

Brindley, P., Goulding, J., and Wilson, M. L. (2014b). Mapping urban neighbourhoods from internet derived data. In *GIS RESEARCH UK (GISRUK), Proc. Of the GIS Research UK 22nd Annual Conference*, pages 355–364, Glasgow, UK.

Brunsdon, C. (1995). Estimating probability surfaces for geographic point data: an adaptive kernel algorithm. *Computers and Geosciences*, 21(7):877–894.

Burrough, P. (1989). Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Science*, 40(3):477–492.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special (Letters on Geographic Information Retrieval)*, 3(2):16–19.

Chen, Y., Suel, T., and Markowetz, A. (2006). Efficient query processing in geographic web search engines. In *SIGMOD*, pages 277–288, Chicago, USA.

Chisholm, R. (1973). Parts as essential to their wholes. *Review of Metaphysics*, 26:581–603.

Christoforaki, M., He, J., Dimopoulos, C., Markowetz, A., and Suel, T. (2011). Text vs. space: efficient geo-search query processing. In *CIKM 11*, pages 423–432, Glasgow, UK.

Clarke, C., Cormack, G., and Lynarn, T. (2001). Exploiting redundancy in question answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 358–365, New York, NY, USA. ACM.

Clementini, E. and Di Felice, P. (1996). An algebraic model for spatial objects with indeterminate boundaries. In Burrough, P. and Frank, A., editors, *Geographic objects with indeterminate boundaries*, page 155–169. Taylor and Francis: London.

Clementini, E. and Di Felice, P. (2001). A spatial model for complex objects with a broad boundary supporting queries on uncertain data. *Data and Knowledge Engineering*, 37:285–305.

Clough, P. and Pasley, R. (2010). Images and perceptions of neighbourhood extents. In *Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10, Zurich 18-19 February 2010*, New York, New York, USA. ACM Press.

Cohn, A. and Gotts, N. (1996). The 'egg-yolk' representation of regions with indeterminate boundaries. In Burrough, P. and Frank, A., editors, *Geographic objects with indeterminate boundaries*, pages 171–188. Taylor and Francis: London.

Coombes, M. (2004). Multiple dimensions of settlement systems: coping with complexity. In Champion, T. and Hugo, G., editors, *New forms of urbanization: beyond the urban-rural Dichotomy*, pages 307–324. Aldershot: Ashgate.

Coombes, M. and Raybould, S. (2001). Public policy and population distribution: Developing appropriate indicators of settlement patterns. *Environment and Planning C: Government and Policy*, 19:223–248.

Coulton, C., Chan, T., and Mikelbank, K. (2010). Finding Place in Making Connections: Applying GIS to Residents' Perceptions of Their Neighborhoods. Availble from: http://www.urban.org/research/publication/finding-place-making-connections-communities.

Coulton, C., Chan, T., and Mikelbank, K. (2011). Finding Place in Community Change Initiatives: Using GIS to Uncover Resident Perceptions of their Neighborhoods. *Journal of Community Practice*, 19:10–28.

Crang, M., Graham, S., and Crosbie, T. (2006). Variable geometries of connection: urban digital divides and the uses of information technology. *Urban Studies*, 43:2551–2570.

Cranshaw, J., Schwartz, R., Hong, J., and Sadeh, N. (2012). The livehoods project : Utilizing social media to understand the dynamics of a city. In *Sixth International AAAI Conference on Weblogs an Social Media (ICWSM 2012)*, pages 58–65, Trinity College in Dublin, Ireland.

Cresswell, T. (2009). Place. In Thrift, N. and Kitchen, R., editors, *International Encyclopedia of Human Geography*, pages 169–177. Elsevier: Oxford.

Crumley, C. L. (1976). Toward a Locational Definition of State Systems of Settlement. *American Anthropologist*, 78:59–73.

Cunha, E. and Martins, B. (2014). Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. *International Journal of Geographical Information Science*, 28(11):2220–2241.

Dalton, N. (2007). Is neighbourhood measurable? *Proceedings, 6th International Space Syntax Symposium, Istanbul, 2007*.

Dalton, N. (2011). Synergy, inteligibility and revelation in neighbourhood places. *PhD thesis: University College London*.

Davies, C. (2009). Are places concepts? familarity and expertise effects in neighborhood cognition. In Hornsby, K., Claramunt, C., Denis, M., and Ligozat, G., editors, *Spatial Information Theory*, volume 5756 of *Lecture Notes in Computer Science*, pages 36–50. Springer Berlin Heidelberg.

Davies, C., Holt, I., Green, J., Harding, J., and Diamond, L. (2009). User Needs and Implications for Modelling Vague Named Places. *Spatial Cognition Computation*, 9(914458409):174–194.

Davis, B. and Miller, C. (2004). Modeling wildfire probability using a gis. In *Proceedings of the American Society for Photogrammetry and Remote Sensing (ASPRS) 2004 Annual Conference*, Denver, USA.

de Bruin, S. (2000). Querying probablistic land cover data using fuzzy set theory. *International Journal of Geographical Information Science*, 14(4):359–372.

de Certeau, M. (2000). *The Everyday Practice of Life*. University of California Press: Los Angeles.

de Finetti, B. (1974). *Theory of Probability*. Wiley: New York.

de Smith, M., Goodchild, M., and Longley, P. (2007). *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*. Troubador Publishing Ltd: Leicester.

Department of Trade and Industry (2000). Closing the Digital Divide: Information and Communication Technologies in Deprived Areas, a report by Social Exclusion Unit Policy Action Team 15.

Derungs, D. and Purves, R. (2014). Where's near? Using web n-grams to explore spatial relations. In *8th International Conference on Geographic Information Science (GIScience 2014)*, Vienna, Austria.

Devillers, R., Stein, A., Bedard, Y., Chrisman, N., Fisher, P., and Wenzhong, S. (2010). Thirty years of research on spatial data quality: Achievements, failures, and opportunities. *Transactions in GIS*, 14(4):387–400.

Dobson, M., Cowen, D., and Guptill, S. (2011). Researching Address and Spatial Data Digital Exchange and Data Integration. A report to the Geography Division, U.S. Bureau of the Census.

Du, H., Alechina, N., Stock, K., and Jackson, M. (2013). The logic of near and far. In Tenbrink, T., Stell, J., Galton, A., and Wood, Z., editors, *Spatial Information Theory*, Lecture Notes in Computer Science, pages 475–494. Springer International Publishing: Cham Switzerland.

Einstein, A. (1954). Foreword in Max Jammer, Concepts of space. Harvard University Press: Cambridge, Massachusetts, USA.

English Heritage (2007). Suburbs and the historic environment.

Evans, A. J. and Waters, T. (2007). Mapping vernacular geography: web-based GIS tools for capturing "fuzzy" or "vague" entities. *International Journal of Technology, Policy and Management*, 7(2):134–150.

Fairclough, N. (2000). *New Labour: new language?* Routledge: London.

Feick, R. and Robertson, C. (forthcoming). A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems*.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage Publications Limited: London.

Fisher, P. (1991). First experiments in viewshed uncertainty: the accuracy of the viewable area. *Photogrammetric Engineering and Remote Sensing*, 57:1321–1327.

Fisher, P. (1999). Models of uncertainty in spatial data. In P.A. Longley, M.F. Goodchild, D. M., editor, *Geographical Information Systems: Principles and Technical Issues 2nd edition*, pages 191–205. John Wiley: New York.

Fisher, P. (2000). Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, 113:7–18.

Fisher, P. (2003). Fuzzy modelling. In Abrahart, R. J. and Openshaw, S., editors, *GeoComputation*, pages 165–191. CRC Press: London.

Fisher, P., Cheng, T., and Wood, J. (2007). Higher order vagueness in geographical information: Empirical geographical population of type n fuzzy sets. *GeoInformatica*, 11:311–330.

Fisher, P., Comber, A., and Wadsworth, R. (2006). Approaches to uncertainty in spatial data. In Devillers, R. and Jeansoulin, R., editors, *Fundamentals of Spatial Data Quality*, pages 43–59. ISTE: London.

Fisher, P. and Orf, T. (1991). An investigation of the meaning of near and close on a university campus. *Computers, Environment and Urban Systems*, 15(1):23–35.

Fisher, P. and Unwin, D. (2005). Re-presenting Geographical Information Systems. *Representing GIS Wiley Sons London*, pages 1–14.

Fisher, P., Wood, J., and Cheng, T. (2004). Where is Helvellyn? Fuzziness of multi-scale landscape morphometry. *Transactions of the Institute of British Geographers*, 29:106–128.

Flemmings, R. (2010). Revealing the fuzzy geography of an urban locality. In *GIS RESEARCH UK (GISRUK), Proc. Of the GIS Research UK 18th Annual Conference*, pages 345–351, London, UK.

Foody, G. (2003). Uncertainty, knowledge discovery and data mining in gis. *Progress in Physical Geography*, 27(1):113–121.

Forsyth, A. (2012). Defining suburbs. *Journal of Planning Literature*, 27(3):270–281.

Frey, W. (2004). The fading of city-suburb and metro-nometro distinctions in the United States. In *New forms of Urbanization*, pages 67–88. Aldershot: Ashgate.

Fumia, D. (2010). Divides, high rise and boundaries: A study of toronto's downtown east side neighbourhood. *Ethnologies*, 32(2):257–289.

Galster, G. (2012). The mechanism(s) of neighbourhood effects: Theory, evidence, and policy implications. In van Ham, M., Manley, D., Bailey, N., Simpson, L., and Maclennan, D., editors, *Neighbourhood Effects Research: New Perspectives*, pages 23–56. Springer Netherlands.

Gardner, J. and Oswald, A. (2001). Internet use: the digital divide. In Park, A., Curtice, J., Thompson, K., Jarvis, L., and Bottomley, C., editors, *In British Social Attitudes: the 18th Report: Public Policy, Social Ties*. Sage, London.

Garrioch, D. (2006). Introduction: The Social History of Urban Neighborhoods. *Journal of Urban History*, 32:663–676.

Gittings, B. (2014). Creating a definitive place name gazetteer for scotland. In *GIS RESEARCH UK (GISRUK), Proc. Of the GIS Research UK 22nd Annual Conference*, pages 117–121, Glasgow, UK.

Giunchiglia, F., Dutta, B., Maltese, V., and Farazi, F. (2012). A facet-based methodology for the construction of a large-scale geo-spatial ontology. *Journal on Data Semantics*, 1(1):57–73.

Glass, R. (1948). *Cliché of urban doom and other essays*. Basil Blackwell: Oxford.

Goldberg, D., Wilson, J. P., and Knoblock, C. (2009). Extracting geographic features from the Internet to automatically build detailed regional gazetteers. *International Journal of Geographical Information Science*, 23:93–128.

Goodchild, M. (2008). Imprecision and spatial uncertainty. In Shekhar, S. and Xiong, H., editors, *Encyclopedia of GIS*, pages 480–483. Springer: New York.

Goodchild, M. F. and Hill, L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22:1039–1044.

Goovaerts, P. (2008). Spatial uncertainty in medical geography: A geostatistical perspective. In Shekhar, S. and Xiong, H., editors, *Encyclopedia of GIS*, pages 1106–1112. Springer: New York.

Griffiths, S., Vaughan, L., Haklay, M., and Jones, C. (2009). The sustainable suburban high street: a review of themes and approaches. *Geography Compass*, 2(4):1155–1188.

Grothe, C. and Schaab, J. (2008). An evaluation of kernel density estimation and support vector machines for automated generation of footprints for imprecise regions from geotags. In *International Workshop on Computational Models of Place (PLACE'08)*, pages 15–28, Park City, Utah, USA.

Grothe, C. and Schaab, J. (2009). Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition and Computation*, 9(3):195–211.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Hahmann, S. and Burghardt, D. (2010). Connecting linkedgeodata and geonames in the spatial semantic web. In *6th International GIScience Conference*, page 28–34, Zurich, Switzerland.

Hahmann, S. and Burghardt, D. (2013). How much information is geospatially referenced? networks and cognition. *International Journal of Geographical Information Science*, 27(6):1171–1189.

Haklay, M. (2007). Successful Suburban Town Centres - a study of the relationship between morphology, sociability, economics and accessibility. EPSRC reference: EP/D06595X/1.

Halfacree, K. (2004). Rethinking 'rurality'. In Champion, T. and Hugo, G., editors, *New forms of urbanization: beyond the urban-rural dichotomy*, pages 285–304. Aldershot: Ashgate.

Hall, M. M. and Jones, C. B. (2008). A field based representation for vague areas defined by spatial prepositions. In *Proceedings of Methodologies and Resources for Processing Spatial Language, Workshop at LREC'2008*.

Hall, M. M., Jones, C. B., and Smart, P. D. (2015). Spatial natural language generation for location description in photo captions. In *Proceedings of Spatial Information Theory COSIT 2015*, Santa Fe, New Mexico, USA.

Hall, M. M., Smart, P. D., and Jones, C. B. (2011). Interpreting Spatial Language in Image Captions. *Cognitive Processing*, 12(1):67–94.

Harris, R. (2010). Meaningful types in a world of suburbs. In Clapson, M. and Hutchinson, R., editors, *Suburbanization in Global Society*, pages 15–50. Emerald: Bingley, England.

Heuvelink, G. and Brown, J. (2008). Uncertain environmental variables in GIS. In Shekhar, S. and Xiong, H., editors, *Encyclopedia of GIS*, pages 1184–1189. Springer: New York.

Hillier, B. and Vaughan, L. (2007). The spatial syntax of urban segregation: The city as one thing. *Progress in Planning*, 67(4):205–294.

Hollenstein, L. (2008). Capturing Vernacular Geography from Georeferenced Tags. *MSc thesis: University of Zurich*.

Hollenstein, L. and Purves, R. S. (2010). Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1:21–48.

Holloway, L. and Hubbard, P. (2001). *People and place: the extraordinary geographies of everyday life*. Harlow: Pearson Education.

Humayun, M. and Schwering, A. (2013). Selecting a representation for spatial vagueness: A decision making approach. In Vandenbroucke, D., Bucher, B., and Crompvoets, J., editors, *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography, pages 115–131. Springer International Publishing: Cham Switzerland.

Humayun, M. I. and Schwering, A. (2012). Representing vague places: Determining a suitable method. *CEUR Workshop Proceedings*, 881:19–25.

Independent Transport Commission (2004). The future of suburbs and exurbs. Available from: http://www.theitc.org.uk/docs/12.pdf.

Janowicz, K. and Keßler, C. (2008). The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 22(10):1129–1157.

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620–630.

Jenks, M. and Dempsey, N. (2007). Defining the neighbourhood: Challenges for empirical research. *Town Planning Review*, 78(2):153–177.

Johar, I. and Maguire, C. (2007). Sustaining our suburbs. Available from: http://www.rudi.net/files/paper/optional_file/40149_Sustain_Suburbs_5_0.pdf.

Jones, C. B., Alani, H., and Tudhope, D. (2001). Geographical Information Retrieval with Ontologies of Place. In *D. Montello (ed), Spatial Information Theory Foundations of Geographic Information Science, COSIT 2001, Lecture Notes in Computer Science*, pages 323–335. Springer: Berlin.

Jones, C. B. and Purves, R. S. (2008). Geographical Information Retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Jones, C. B., Purves, R. S., Clough, P., and Joho, H. (2008a). Modelling Vague Places with Knowledge from the Web. *International Journal of Geographical Information Science*, 22(10):1045–1065.

Jones, C. E., Haklay, M., Griffiths, S., and Vaughan, L. (2008b). Visualising London's suburbs. In *GIS RESEARCH UK (GISRUK), Proc. Of the GIS Research UK 16th Annual Conference*, Manchester Metropolitan University.

Kearns, A. and Parkinson, M. (2001). The significance of neighbourhood. *Urban Studies*, 38(12):2103–2110.

Keen, A. (2015). *The Internet is not the answer*. Atlantic Monthly Press: New York.

Keßler, C., Janowicz, K., and Bishr, M. (2009a). An agenda for the next generation gazetteer: Geographic information contribution and retrieval. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS09*, pages 91–100.

Keßler, C., Maué, P., Heuer, J., and Bartoschek, T. (2009b). Bottom-up gazetteers: Learning from the implicit semantics of geotags. In Janowicz, K., Raubal, M., and Levashkin, S., editors, *GeoSpatial Semantics*, volume 5892 of *Lecture Notes in Computer Science*, pages 83–102. Springer: Berlin.

Kilgarriff, A. and Grefenstette, G. (2003). Web as corpus. *Computational Linguistics*, 29:1–15.

Knights, C. (2008). Urban Regeneration: A Theological Perspective from the West End of Newcastle-upon-Tyne. *The Expository Times*, 119:217–225.

Knox, P. and Pinch, S. (2009). *Urban Social Geography: An Introduction*. Routledge: London.

Kulik, L. (2001). A geometric theory of vague boundaries based on supervaluation. In Montello, D., editor, *Spatial Information Theory: Foundations of Geographic Information Science*, pages 44–59. Springer: Berlin.

Kwok, C., Etzioni, O., and Weld, D. (2001). Scaling question answering to the web. In *Proceedings of the 10th International World Wide Web Conference*, pages 150–161, New York, NY, USA. ACM.

Laere, O. V., Schockaert, S., Tanasescu, V., Dhoedt, B., and Jones, C. B. (2014). Georeferencing wikipedia documents using data from social media sources. *ACM Transactions on Information Systems (TOIS)*, 32(3):12:1–12:32.

Lam, C. S., Wilson, J. P., and Holmes-Wong, D. (2002). Building a Neighborhood-Specific Gazetteer for a Digital Archive. *Annual ESRI International User Conference, San Diego, California*.

Li, L. and Goodchild, M. F. (2012). Constructing places from spatial footprints. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '12*, pages 15–21.

Li, W., Yang, P., and Zhou, B. (2008). Internet-based spatial information retrieval. In Shekhar, S. and Xiong, H., editors, *Encyclopedia of GIS*, pages 596–599. Springer: New York.

Liu, Y., Wang, F., Kang, C., Gao, Y., and Lu, Y. (2014). Analyzing relatedness by toponym co-occurrences on web pages. *Transactions in GIS*, 18(1):89–107.

Local Futures Group (2007). The state of the suburbs. Available from: http://www.localleadership.gov.uk/docs/suburbs.pdf.

London Assembly (2012). Beyond consultation: The role of neighbourhood plans in supporting local involvement in planning.

London Assembly (2014). Localism in London. What's the story? Available from: http://www.london.gov.uk/sites/default/files/14-11-20-Localism-Report_0.pdf.

Lupton, R., Fenton, A., Tunstall, R., and Harris, R. (2011). Place typologies and their policy applications. A report prepared for the Department of Communities and Local Government.

Lüscher, P., Weibel, R., and Burghardt, D. (2009). Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data. *Computers, Environment and Urban Systems*, 33(5):363–374.

Lynch, K. (1960). *Image of the City*. The MIT Press: Cambridge MA, USA.

Mackaness, W. A. and Chaudhry, O. Z. (2011). Automatic Classification of Retail Spaces from a Large Scale Topographic Database. *Transactions in GIS*, 15(3):291–307.

Martin, D. (1998). Automatic neighbourhood identification from population surfaces. *Computers, Environment and Urban Systems*, 22(2):107–120.

Martin, D. (2003). Developing the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, 17(2):181–196.

Massey, D. (1991). A global sense of place. *Marxism Today*, page 24–9. Reprinted in D. Massey, Space, Place and Gender, pages 146–156, Cambridge: Polity Press.

Mayhew, S. (2009). *A Dictionary of Geography (4 ed.)*. Oxford University Press Print Publication: Oxford.

McCurley, K. S. (2001). Geospatial mapping and navigation of the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 221–229, New York, NY, USA. ACM.

McKenzie, R. (1921). The neighbourhood: a study of local life in Columbus, Ohio. *American Journal of Sociology*, 27:780–799.

Miller, V. (2006). The unmappable: Vagueness and spatial experience. *Space and Culture*, 9(4):453–467.

Montello, D., Friedman, A., and Phillips, D. (2014). Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, 28(9):1802–1820.

Montello, D. R., Goodchild, M. F., Gottsegen, J., and Fohl, P. (2003). Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition & Computation*, 3:185–204.

Moulaert, F., Martinelli, F., Gonzalez, S., and Swyngedouw, E. (2007). *Can Neighbourhoods Save the City?: Community Development and Social Innovation (Regions and Cities)*. Routledge: London.

Nagel, S. (2005). An ontology of German place names. *Corela(on-line), HS-2*, pages 1–7.

NISRA (2005). Northern Ireland Statistics and Research Agency. Statistical Classification and Delineation of Settlements. Report of the Inter-Departmental Urban-Rural Definition Group.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011). Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Fifth International AAAI Conference on Weblogs and Social Media*, pages 570–573, Barcelona, Catalonia, Spain.

OED Online (2015). June 2015. Oxford University Press: Oxford.

Office for National Statistics (ONS) (2006). Neighbourhood Statistics Service: annual report to Ministers 2005/06. Available from: http://www.neighbourhood.statistics.gov.uk/HTMLDocs/images/NeSS-Report_2005-06_tcm97-51089.pdf.

Office for National Statistics (ONS) (2014). Internet access – households and individuals 2014. Available from: http://www.ons.gov.uk/ons/dcp171778_373584.pdf.

Office of the Deputy Prime Minister (ODPM) (2005). Sustainable communities: People, places and prosperity. London: HMSO. Available from: http://www.eukn.org/dsresource?objectid=143362.

Orford, S. and Leigh, C. (2013). The Relationship between Self-reported Definitions of Urban Neighbourhood and Respondent Characteristics: A Study of Cardiff, UK. *Urban Studies*, 51:1891–1908.

O'Sullivan, D. and Unwin, D. (2003). *Geographic Information Analysis*. Wiley: London.

Overell, S. (2011). The problem of place name ambiguity. *SIGSPATIAL Special (Letters on Geographic Information Retrieval)*, 3(2):12–15.

Pallant, J. (2013). *SPSS survival manual: a step by step guide to data analysis using IBM SPSS*. Open University Press: Maidenhead.

Parker, M. (2015). Interview on the BBC program - Timeshift (Series 15: 2), A Very British Map: The Ordnance Survey Story. First shown: 9 Sep 2015.

Pasley, R., Clough, P., Purves, R. S., and Twaroch, F. A. (2008). Mapping geographic coverage of the web. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS '08*, New York, New York, USA. ACM Press.

Pasley, R., Clough, P., and Sanderson, M. (2007). Geo-Tagging for Imprecise Regions of Different Sizes. In *Proceedings of Workshop on Geographic Information Retrieval GIR'07*.

Pateman, T. (2004). Rural and urban areas: comparing lives using rural / urban classifications. *Regional Trends*, 43:1–77.

Perry, C. (1992). The neighbourhood unit. In Lewis, H., editor, *Neighbourhood and community planning*. Regional Plan of New York and the Environs, New York.

Phithakkitnukoon, S. and Olivier, P. (2011). Sensing Urban Social Geography Using Online Social Networking Data. *The Social Mobile Web*, pages 36–39.

Pickles, E. (2010). Eric Pickles' speech to the Local Government Association annual conference - 7 July 2010. Available from: https://www.gov.uk/government/speeches/local-government-association-conference-6-july-2010.

Pontikakis, E. and Navratil, G. (2006). Physical vs. Web Space – Similarities and Differences. In Riedl, A., Kainz, W., and Elmes, G., editors, *Progress in Spatial Data Handling*, pages 17–26. Berlin: Springer.

Pumain, D. (2000). Settlement systems in the evolution. *Geografiska Annaler: Series B, Human Geography*, 82(2):73–87.

Purves, R., Edwardes, A., and Wood, J. (2011). Describing place through user generated content. *First Monday*, 16(9).

Purves, R. S., Clough, P., and Joho, H. (2005a). Identifying imprecise regions for geographic information retrieval using the web. In *In Proceedings of GIS RESEARCH UK 13th Annual Conference, Glasgow, UK*, pages 313–318.

Purves, R. S., Clough, P., Joho, H., and Jones, C. B. (2005b). Spatially-Aware Information Retrieval on the Internet Modelling Vague Places with Knowledge from the Web. In *Deliverable number: D24 3301, Spatially-Aware Information Retrieval on the Internet (SPIRIT) project*.

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A. K., Vaid, S., and Yang, B. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7):717–745.

Radev, D., Libner, K., and Fan, W. (2002). Getting answers to natural language questions on the web. *Journal of the American Society for Information Science and Technology*, 53(5):359–364.

Rattenbury, T. and Naaman, M. (2009). Methods for extracting place semantics from flickr tags. *ACM Trans. Web*, 3(1):1–30.

RERC (2005). Rural Evidence Research Centre. Defra Classification of Local Authority Districts and Unitary Authorities in England. http://archive.defra.gov.uk/evidence/statistics/rural/documents/rural-defn/LAClassifications_technicalguide.pdf.

Ridge, M. (2014). Crowdsourcing our cultural heritage: introduction. In Ridge, M., editor, *Crowdsourcing our cultural heritage*. Aldershot: Ashgate.

Roberts, B. (1996). *Landscapes of settlement: prehistory to the present*. London: Psychology Press.

Roberts, B. (1997). *Geographical Information System and Computer Cartography*. London: Longman.

Rogerson, P. (2014). *Statistical Methods for Geography: A Student's Guide*. Sage Publications Limited: London.

Rosser, J. and Morley, J. (2010). Rate-my-place: a social network application for crowd-sourcing vernacular geographic areas. In *GIS RESEARCH UK (GISRUK), Proc. Of the GIS Research UK 18th Annual Conference*, pages 155–160, University College London, UK.

Sainsbury, R. (1989). What is a vague object? *Analysis*, 49(3):99–103.

Sampson, R. J. (2012). Moving and the Neighborhood Glass Ceiling. *Science*, 337(2012):1464–1465.

Savage, V. (2009). Place names. In Thrift, R. K., editor, *International Encyclopedia of Human Geography*, pages 178 – 184. Elsevier, Oxford.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471.

Schockaert, S. (2011). Vague regions in geographic information retrieval. *SIGSPATIAL Special (Letters on Geographic Information Retrieval)*, 3(2):24–28.

Schockaert, S. and De Cock, M. (2007). Neighborhood restrictions in geographic IR. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 167.

Schuth, A., Marx, M., and de Rijke, M. (2007). Extracting the discussion structure in comments on news-articles. In *9th annual ACM international workshop on Web information and data management (WIDM)*, pages 97–104, Lisbon, Portugal.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: New York.

Smailes, A. (1944). The urban hierarchy in england and wales. *Geography*, 29:41–51.

Smith, B. and Varzi, A. (2000). Fiat and bona fide boundaries. *Philosophy and Phenomenological Research*, 60(2):401–420.

Smith, I., Lepine, E., and Taylor, M. (2007). *Disadvantaged by Where You Live?: Neighbourhood Governance in Contemporary Urban Policy*. Policy Press: Bristol.

Stanojević, M. (2009). Cognitive synonymy: a general overview. *Facta Universitatis, Linguistics and Literature series*, 7(2):193–200.

Stefanidis, A., Crooks, A., and Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78:319–338.

Stock, K., Pasley, R., Gardner, Z., Brindley, P., Morley, J., and Cialone, C. (2013). Creating a corpus of geospatial natural language. In Tenbrink, T., Stell, J., Galton, A., and Wood, Z., editors, *Spatial Information Theory*, volume 8116 of *Lecture Notes in Computer Science*, pages 279–298. Springer International Publishing: Cham Switzerland.

Talen, E. (1999). Constructing neighborhoods from the bottom up: The case for resident-generated GIS. *Environment and Planning B: Planning and Design*, 26:533–554.

Taylor, J. (1997). *An Introduction to Error Analysis: the study of uncertainties in physical measurements*. University Science Books: Sausalito, California.

Taylor, M. and Wilson, M. (2006). The importance of the neighbourhood. Report by Joseph Rowntree Foundation: York.

Tayyebi, A., Perry, P. C., and Tayyebi, A. H. (2014). Predicting the expansion of an urban boundary using spatial logistic regression and hybrid raster-vector routines with remote sensing and GIS. *International Journal of Geographical Information Science*, 28(4):639–659.

The Economist (2014a). A planet of suburbs. Available from: http://www.economist.com/suburbs.

The Economist (2014b). The transformation of cities: A suburban world. Available from: http://www.economist.com/news/leaders/21635486-emerging-world-becoming-suburban-its-leaders-should-welcome-avoid-wests.

The Telegraph (2014). Save our suburbs. Available from: http://www.telegraph.co.uk/news/uknews/1562531/Save-our-suburbs.html.

Thomas, B., Pritchard, J., Ballas, D., Vickers, D., and Dorling, D. (2009). A tale of two cities: the Sheffield project. A report by the Social and Spatial Inequalities Research Group University of Sheffield, Sheffield.

Thurstain-Goodwin, M. and Unwin, D. (2000). Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations. CASA Working paper series, UCL: London.

Tuan, Y.-F. (1975). Place: An experiential perspective. *Geographical Review*, 65(2):151–165.

Tuan, Y.-F. (1977). *Space and Place: The Perspective of Experience*. Minneapolis: University of Minnesota Press.

Twaroch, F. A. and Jones, C. B. (2010). A web platform for the evaluation of vernacular place names in automatically constructed gazetteers. *Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10*.

Twaroch, F. A., Jones, C. B., and Abdelmoty, A. I. (2008). Acquisition of a vernacular gazetteer from web sources. *Proceedings of the first international workshop on Location and the web - LOCWEB '08, Beijing, China*, pages 61–64.

Twaroch, F. A., Jones, C. B., and Abdelmoty, A. I. (2009a). Acquisition of Vernacular Place Names from Web Sources. In *Ricardo Baeza-Yates and Irwin King (Eds.): "Weaving Services and People on the World Wide Web"*, pages 195 – 214. Springer Berlin Heidelberg.

Twaroch, F. A., Purves, R. S., and Jones, C. B. (2009b). Stability of Qualitative Spatial Relations between Vernacular Regions Mined from Web Data. In *Proceedings of Workshop on Geographic Information on the Internet, Toulouse, France, April 6th, 2009*.

Uryupina, O. (2003). Semi-supervised learning of geographical gazetteers from the internet. *Proceedings of the HLTNAACL 2003 workshop on Analysis of geographic references*, 1:18–25.

Vallee, J., Le Roux, G., Chaix, B., Kestens, Y., and Chauvin, P. (2014). The 'constant size neighbourhood trap' in accessibility and health studies. *Urban Studies*, pages 1–20.

Valls, J. and Ontañón, S. (2012). Natural language generation through case-based text modification. In *20th International Conference on Case-Based Reasoning (ICCBR)*, pages 433–457, Lyon, France.

van Ham, M. (2012). Neighbourhood effects research at a crossroads. Ten challenges for future research. *Environment and Planning A*, 44:2787–2793.

Varzi, A. (1998). Basic problems of mereotopology. In Guarino, N., editor, *Formal Ontology in Information Systems*, pages 29–38. IOS Press, Amsterdam.

Vasardani, M., Winter, S., and Richter, K.-f. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27:2509–2532.

Vaughan, L., Griffiths, S., Haklay, M., and Jones, C. (2009). Do the suburbs exist? discovering complexity and specificity in suburban built form. *Transactions on the Institute of British Geographers*, 34(4):475–88.

Vaughan, L., Jones, C., Griffiths, S., and Haklay, M. (2010). The spatial signature of suburban town centres. *The Journal of Space Syntax*, 1(1):77–91.

Vining, R. (1955). A description of certain spatial aspects of an economic system. *Economic Development and Cultural Change*, 3:147–195.

Wacholder, N., Ravin, Y., and Choi, M. (1997). Disambiguation of proper names in text. in proceedings of the fifth conference on applied natural language processing. In *Proceedings of the fifth conference on Applied natural language processing*, Association for Computational Linguistics, pages 202–208, Washington DC, USA.

Wadsworth, R., Comber, A., and P., F. (2006). Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. In A. Riedl, W. Kainz, W. and Elmes, G., editors, *Progress in Spatial Data Handling*, pages 197–213. Berlin: Springer.

Wallgrün, J. O., Klippel, A., and Baldwin, T. (2014). Building a corpus of spatial relational expressions extracted from web documents. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*, GIR '14, New York, NY, USA. ACM.

Waters, T. and Evans, A. J. (2003). Tools for web-based GIS mapping of a "fuzzy" vernacular geography. *Proceedings of the 7th International Conference on GeoComputation*.

Williamson, T. (1994). *Vagueness*. Routledge: London.

Wilson, J. P., Lam, C. S., and Holmes-Wong, D. (2004). A New Method for the Specification of Geographic Footprints in Digital Gazetteers. *Cartography and Geographic Information Science*, 31(4):195–207.

Wilson, M. W. (2012). Location-based services, conspicuous mobility, and the location-aware future. *Geoforum*, 43:1266–1275.

Wilson, R. (2009). Why neighborhoods matter: the importance of geographic composition. *Geography and Public Safety*, 2(2):1–2.

Wilson, R. and Filbert, K. (2008). Crime mapping and analysis. In Shekhar, S. and Xiong, H., editors, *Encyclopedia of GIS*, pages 180–186. Springer: New York.

Wittgenstein, L. (1965). *Philosophical investigations*. The Macmillan Company: New York.

Worboys, M. (1998). Imprecision in finite resolution spatial data. *Geoinformatica*, 2:57–79.

Worboys, M. (2001). Nearness relations in environmental space. *International Journal of Geographical Information Science*, 15(7):633–651.

Worboys, M. (2005). Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19(1):1–28.

Worboys, M. and Duckham, M. (2004). *GIS: A Computing Perspective*. CRC Press: Boca Raton, Florida.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Zhang, A. X., Noulas, A., Scellato, S., and Mascolo, C. (2013). Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. *Proceedings - SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013*, pages 69–74.

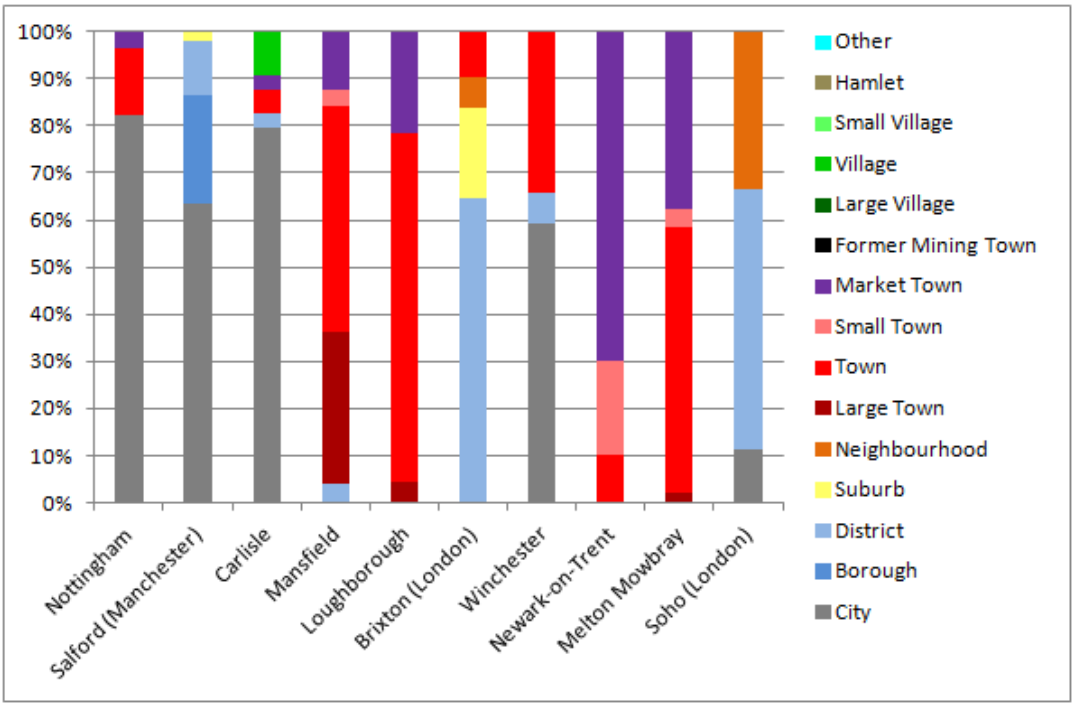
Zhang, J. and Foody, G. (1998). A fuzzy classification of sub-urban land cover from remotely sensed imagery. *International Journal of Remote Sensing*, 19(4):2721–2738.

Appendix A

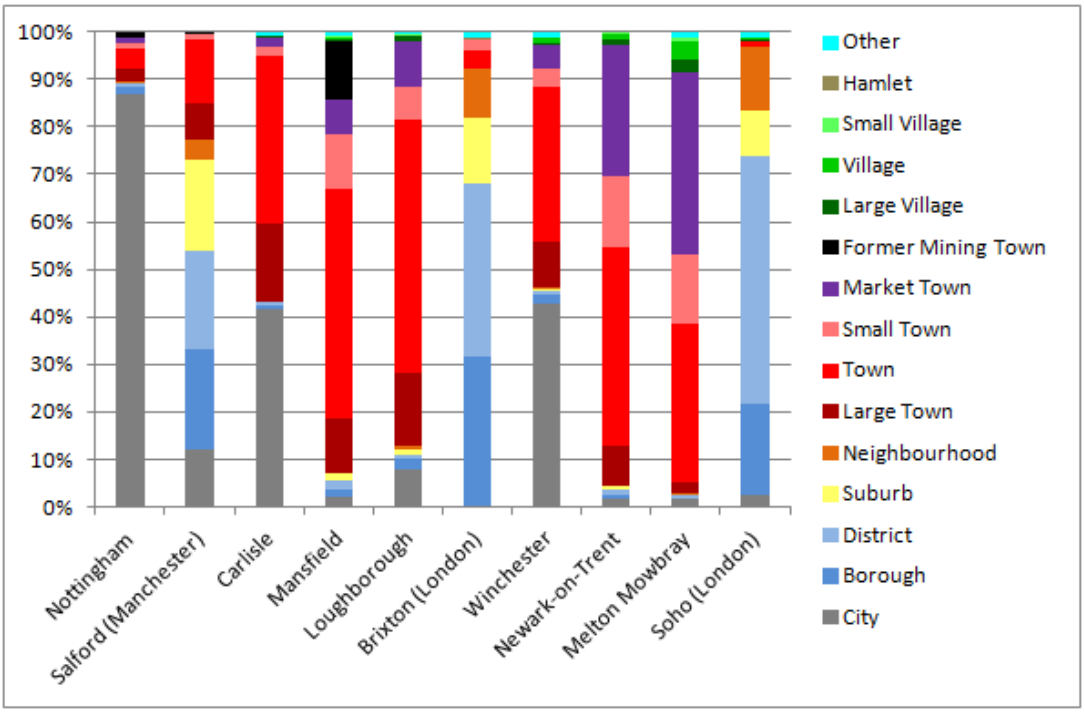
Comparison of settlement classification data between web extraction method and questionnaire responses

The data within this appendix represents that which was used in the validation process of the vague settlement classification formed from web data. Figure 3.6 within Chapter 3 contained output from the web extraction method for 10 of the total 60 settlements. This appendix contains comparisons for all the case study settlements. Whilst the overall impression from the figures below is one of similarity - there are also noticeable disparities. The underlying data of the figures in this appendix were used in regression modelling for the validation of the settlement classification - as shown within Table 3.11 for individual settlements and collectively within Figure 3.7.

Fig. A.1 Comparison between web extraction output and questionnaire responses for selected settlements

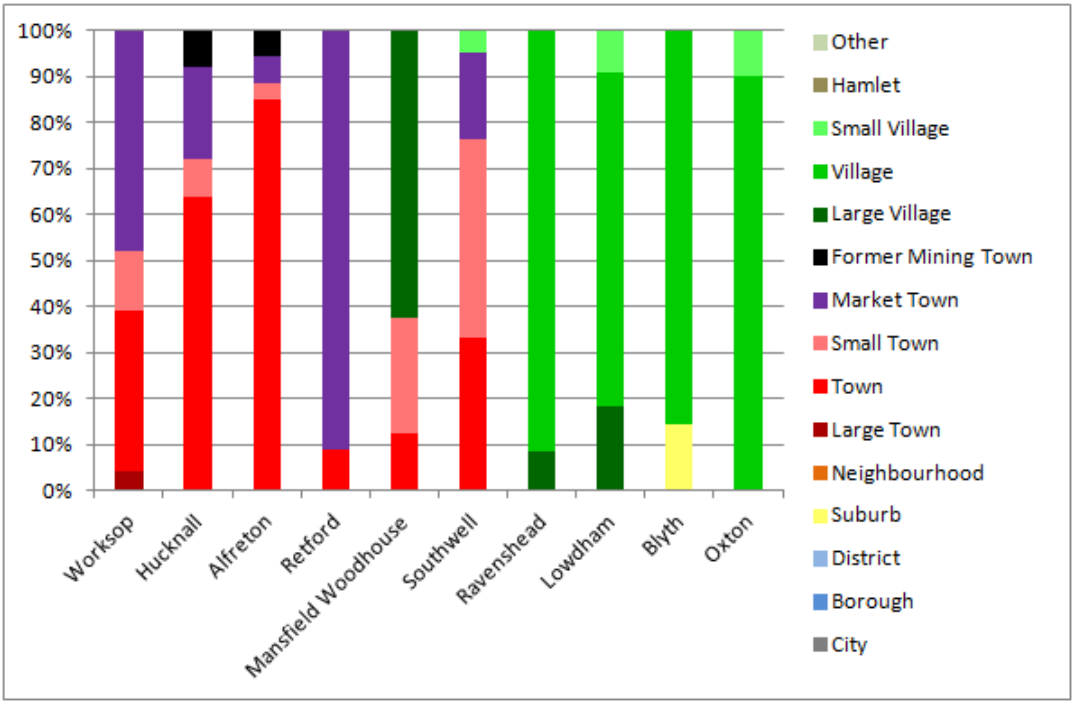


a) From web extraction output

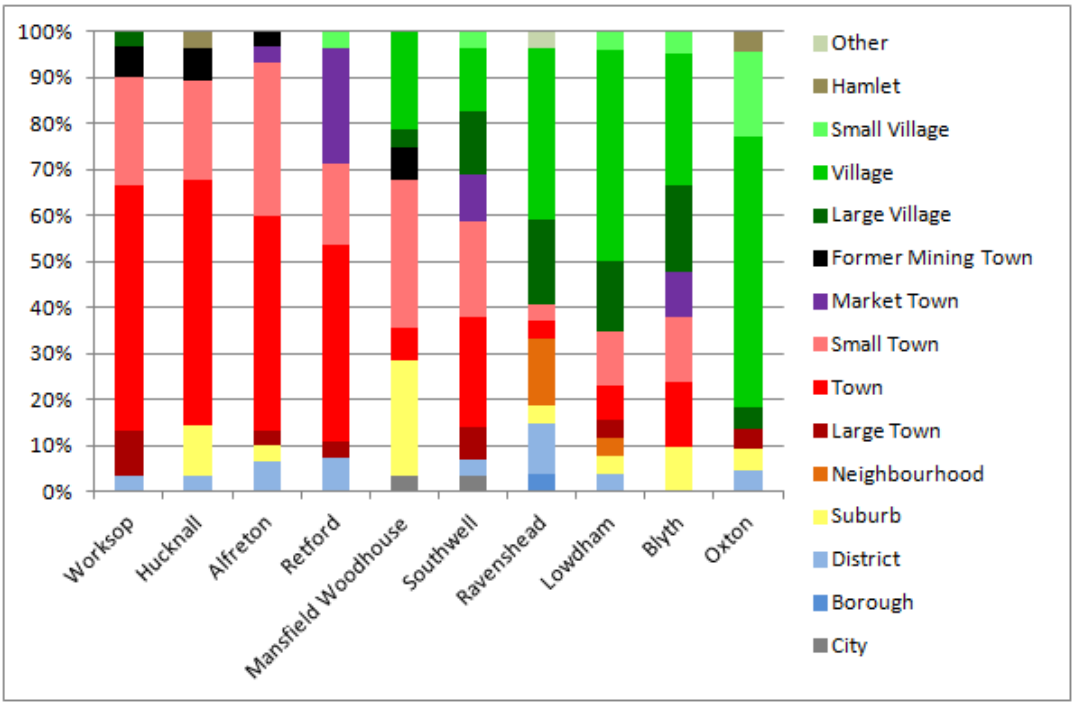


b) From questionnaire responses

Fig. A.2 Comparison between web extraction output and questionnaire responses for selected settlements: 'North Nottinghamshire'

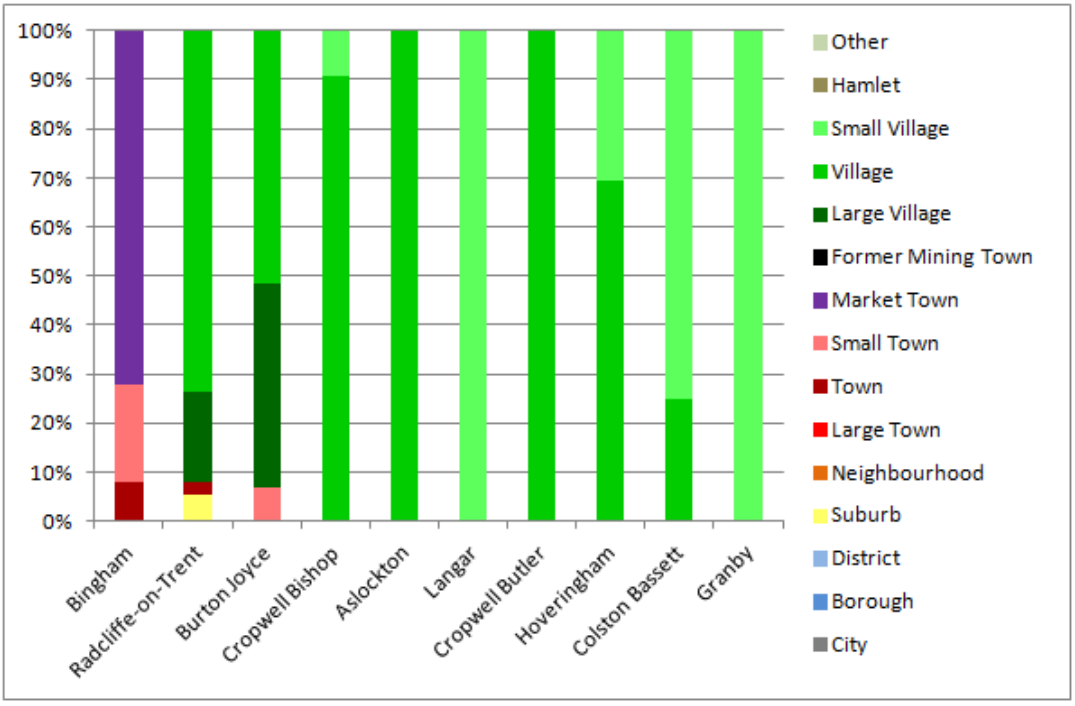


a) From web extraction output

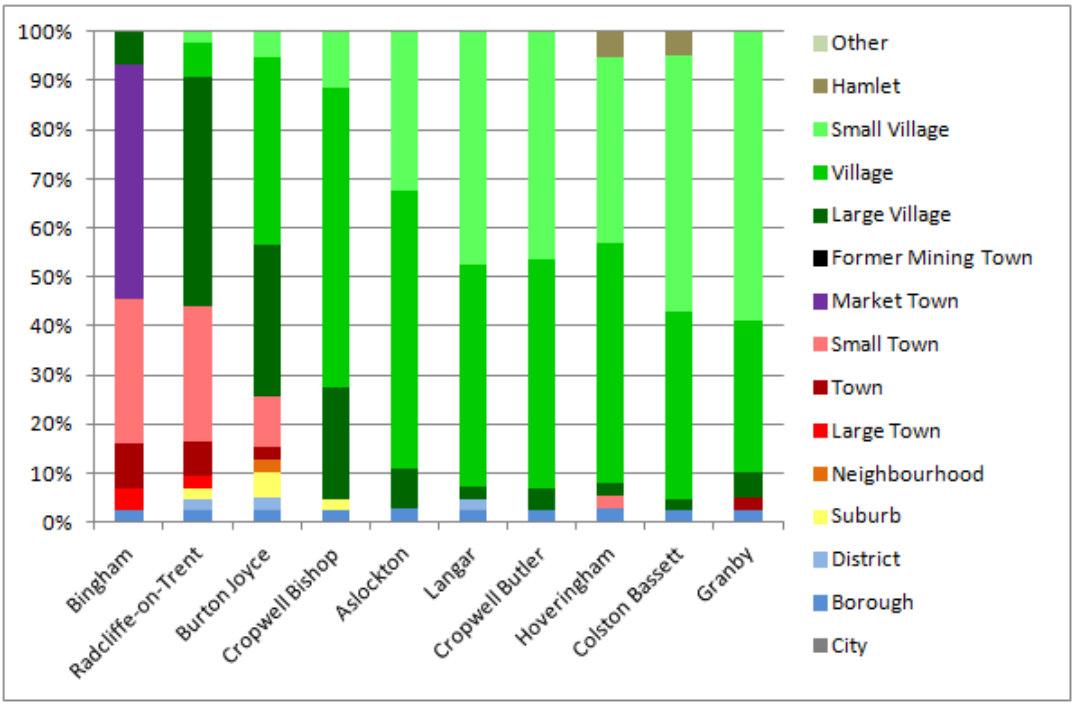


b) From questionnaire responses

Fig. A.3 Comparison between web extraction output and questionnaire responses for selected settlements: 'East Nottinghamshire'

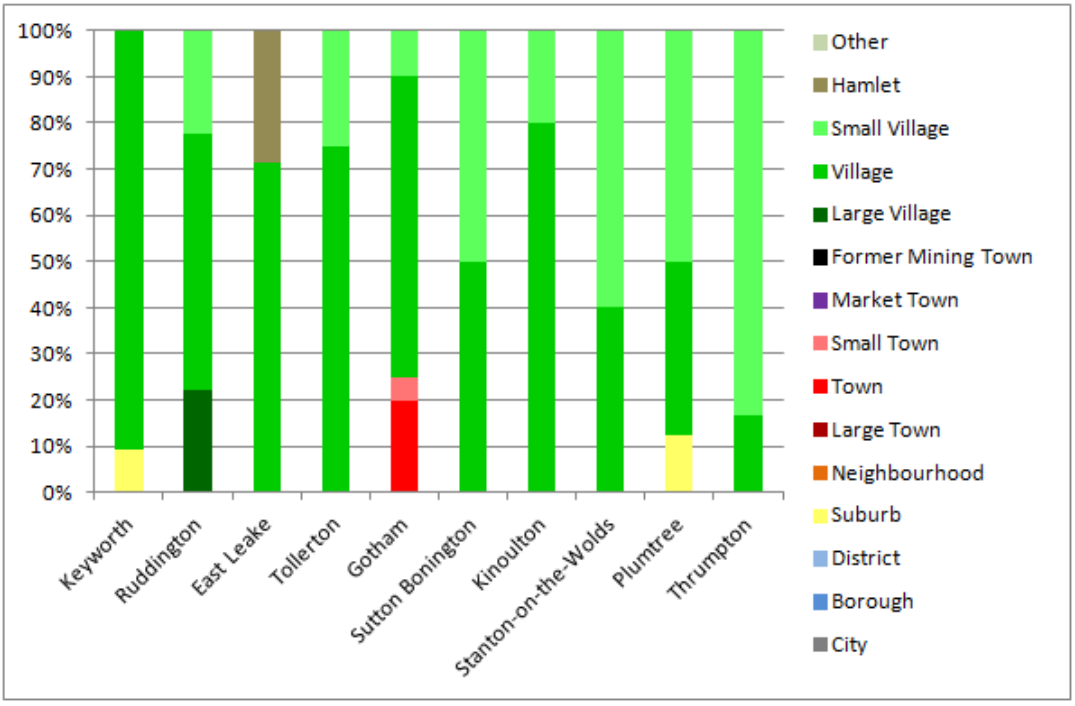


a) From web extraction output

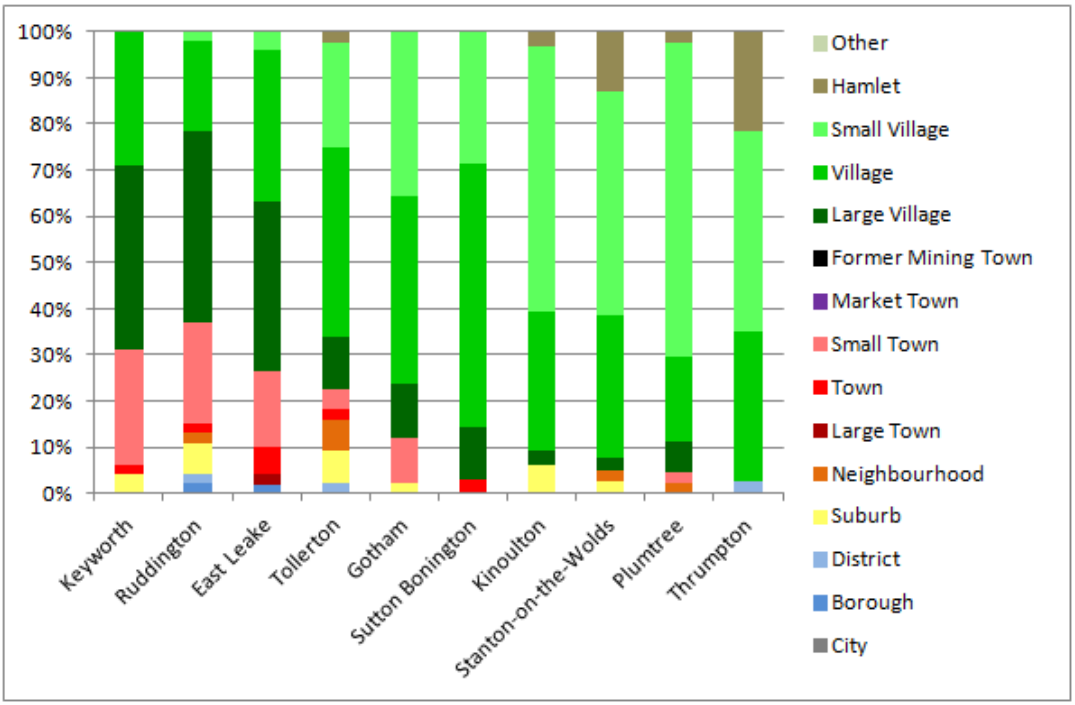


b) From questionnaire responses

Fig. A.4 Comparison between web extraction output and questionnaire responses for selected settlements: 'South Nottinghamshire'

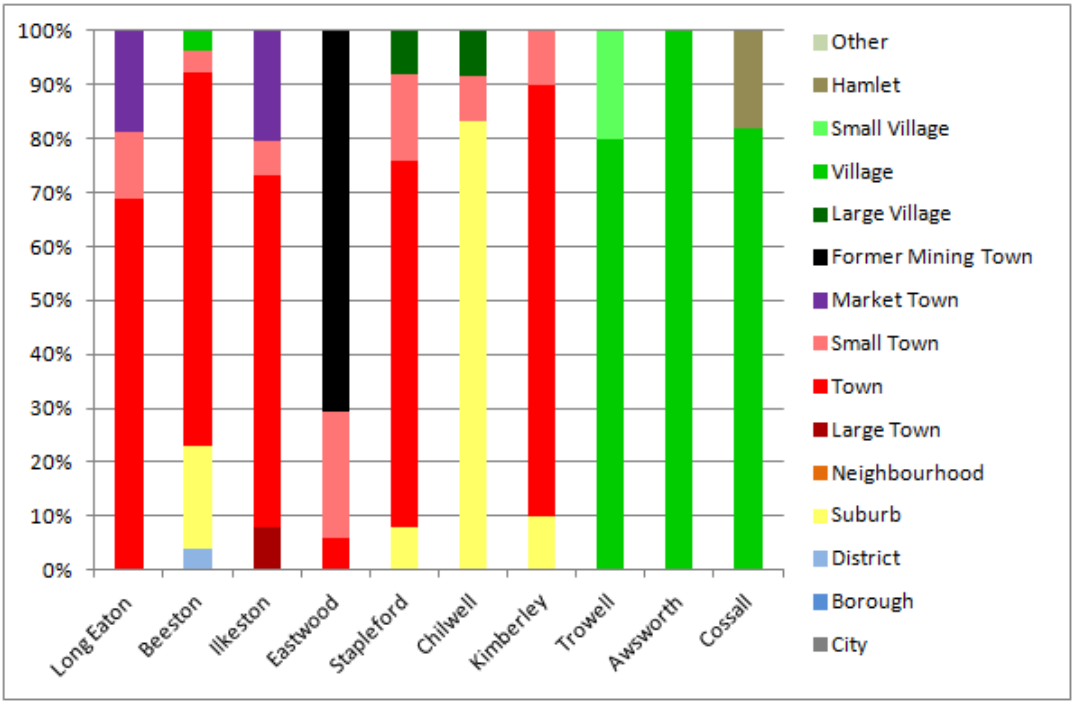


a) From web extraction output

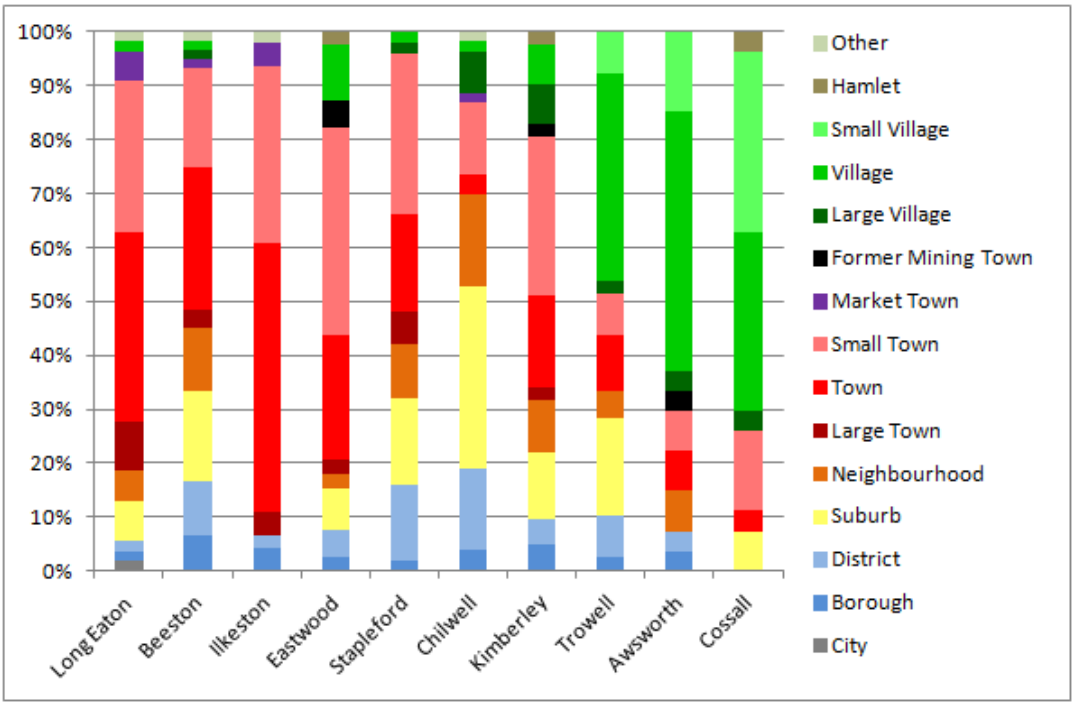


b) From questionnaire responses

Fig. A.5 Comparison between web extraction output and questionnaire responses for selected settlements: 'West Nottinghamshire'

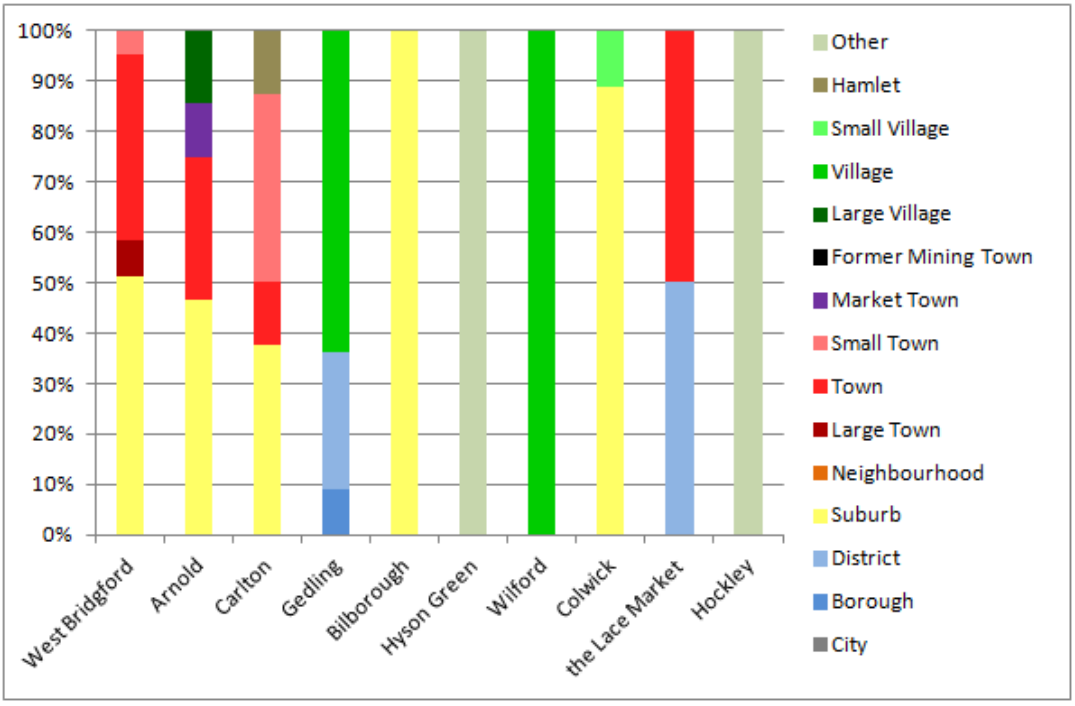


a) From web extraction output

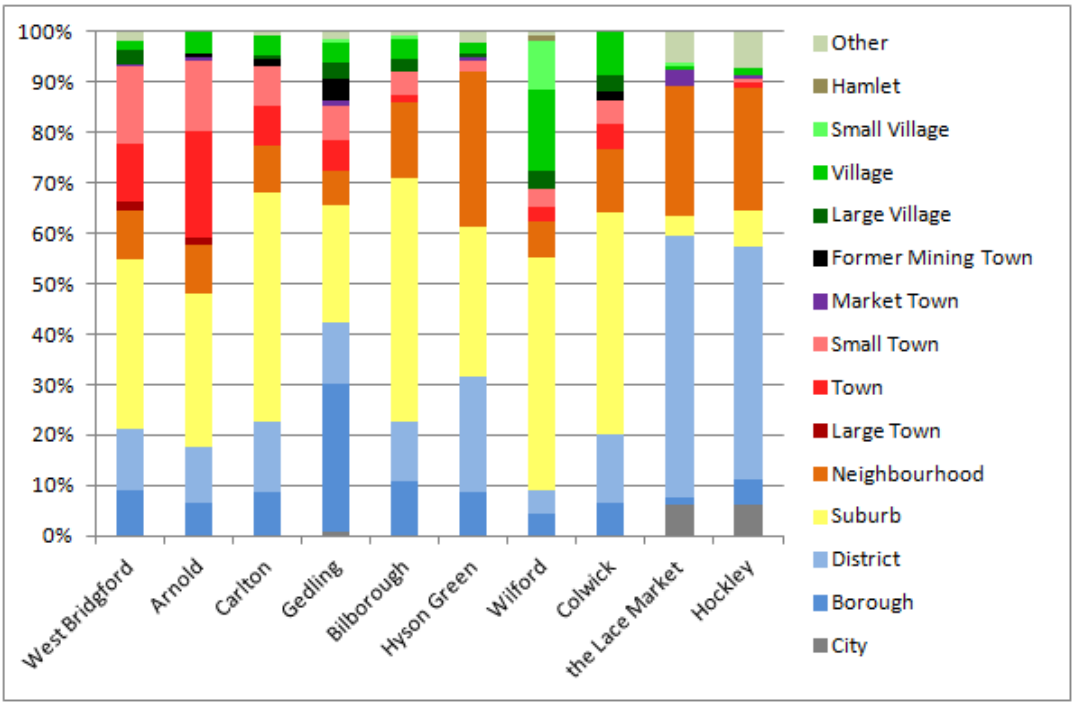


b) From questionnaire responses

Fig. A.6 Comparison between web extraction output and questionnaire responses for selected settlements: 'Central Nottingham'



a) From web extraction output



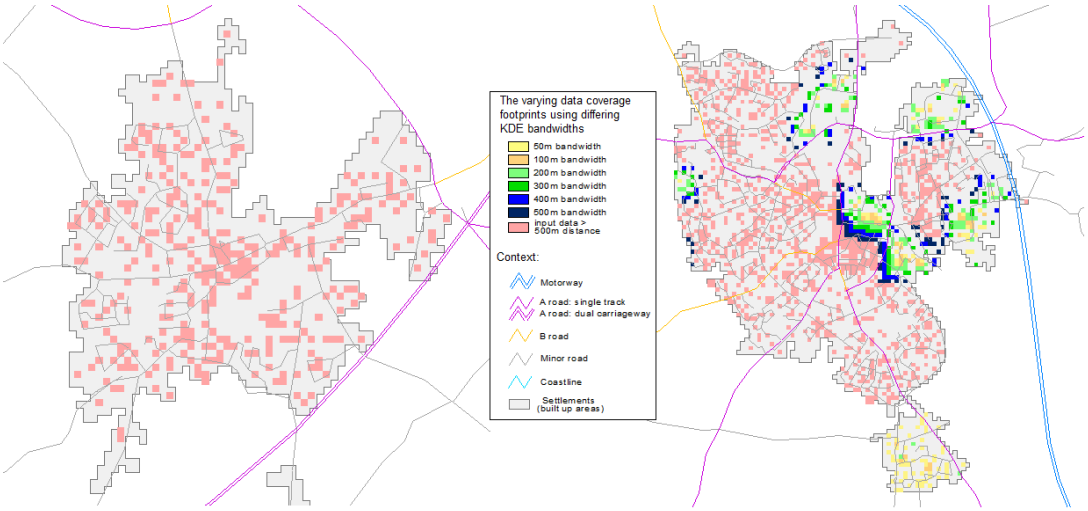
b) From questionnaire responses

Appendix B

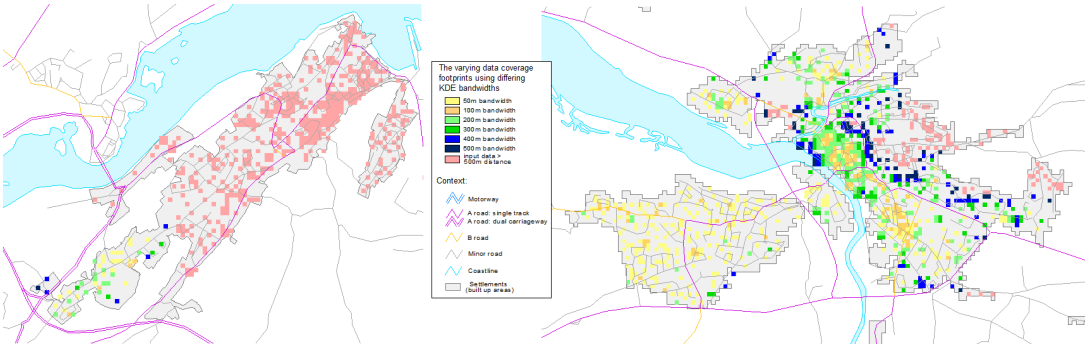
Data coverage using different KDE bandwidths

This appendix illustrates the changing footprint of areas with urban locale names when using differing bandwidths in the KDE process. It builds on the information within Figure 5.9 which contained output for just two of the case study areas, whilst this appendix contains output for all case study settlements.

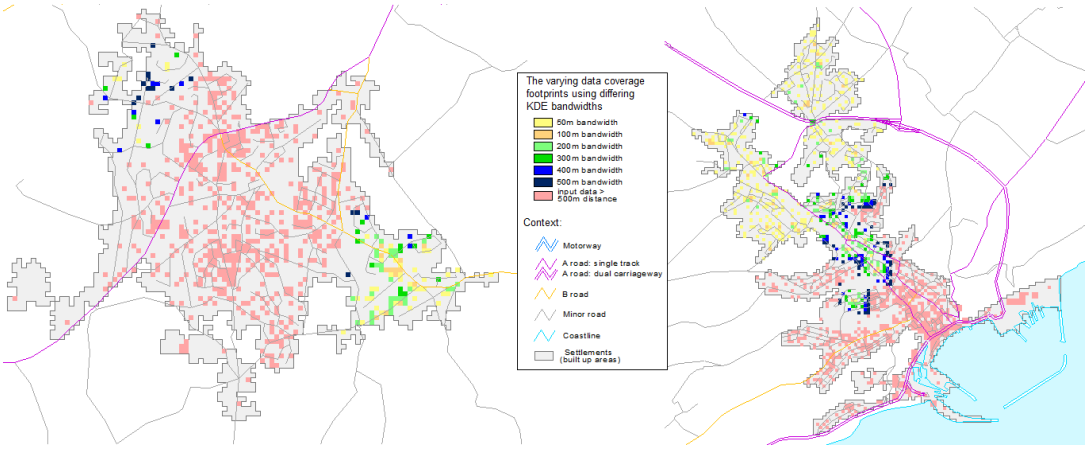
Fig. B.1 Spread of data coverage using different KDE bandwidths



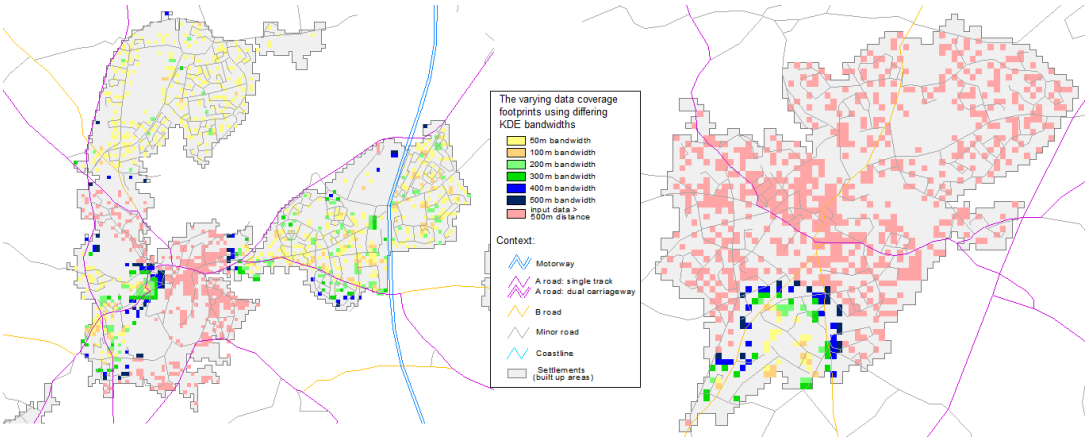
a) Ashby-de-la-zouche *b) Banbury*



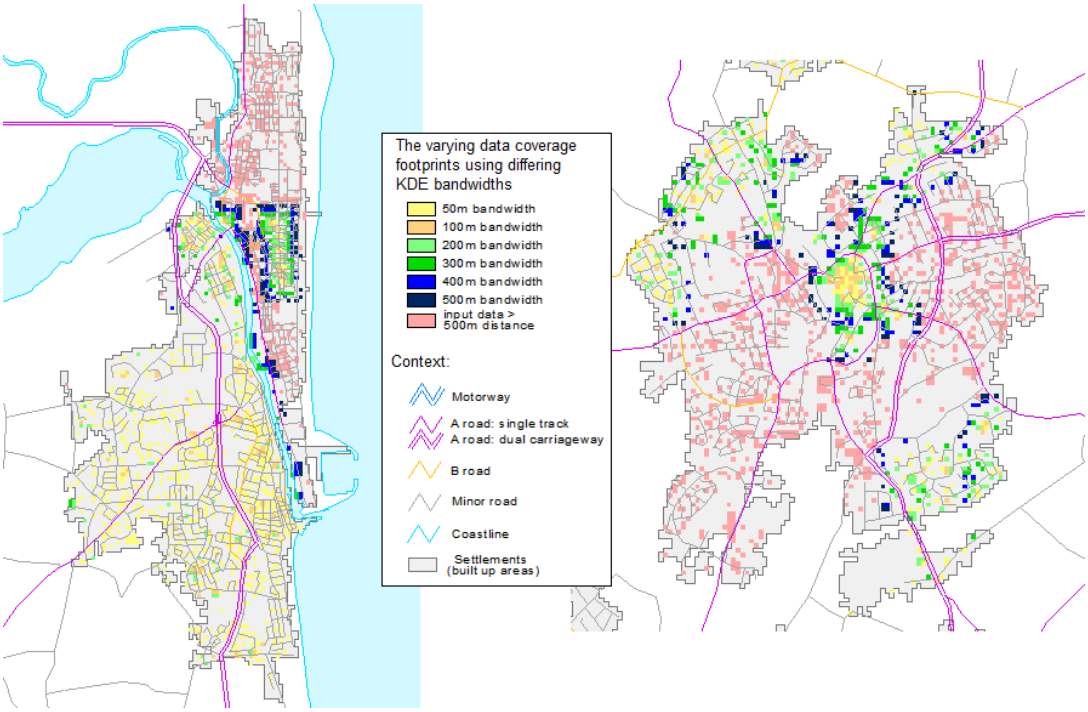
c) Bangor *d) Barnstaple*



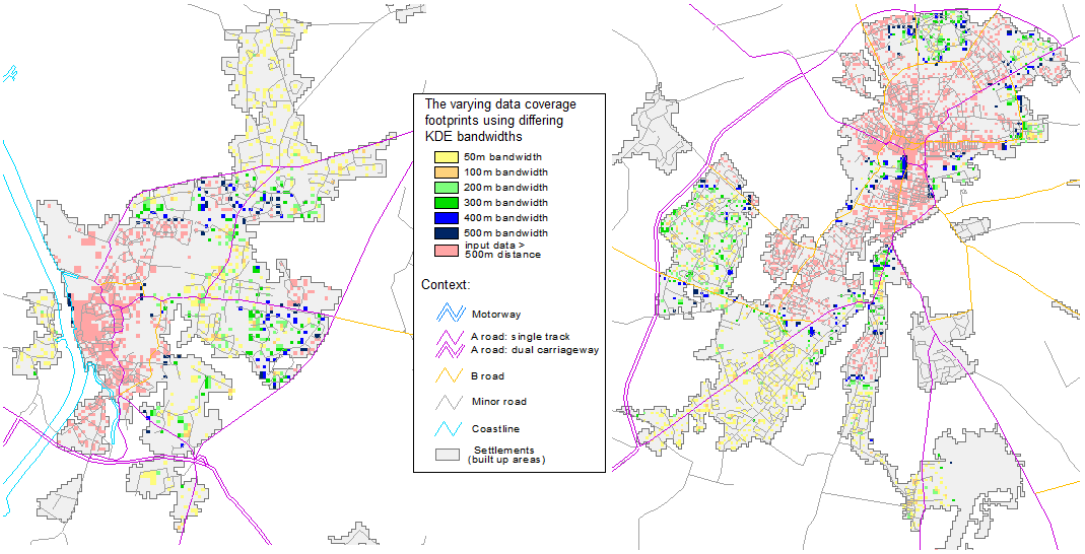
e) Crowborough *f) Dover*



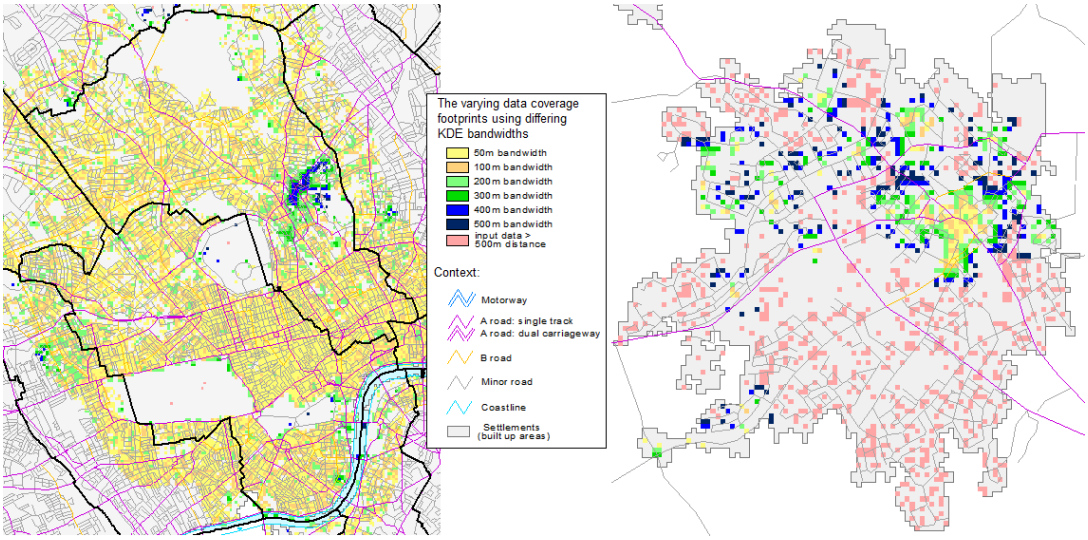
g) Durham *h) Frome*



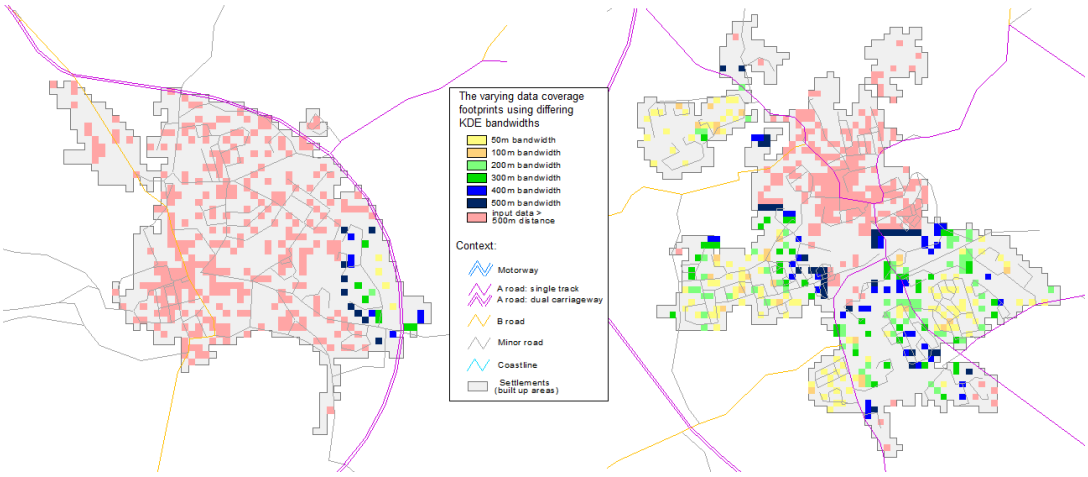
i) Great Yarmouth *j) Kiddeminister*



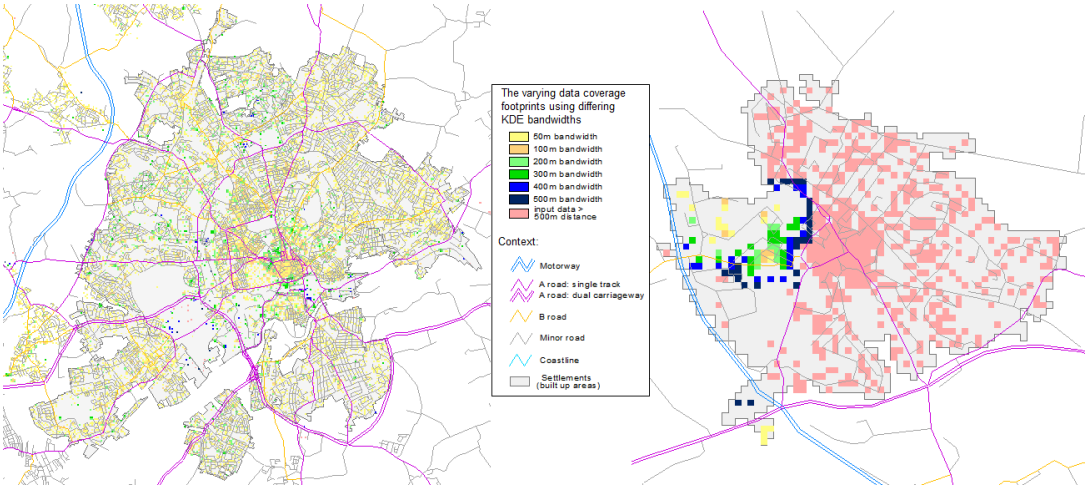
k) King's Lynn *l) Lincoln*



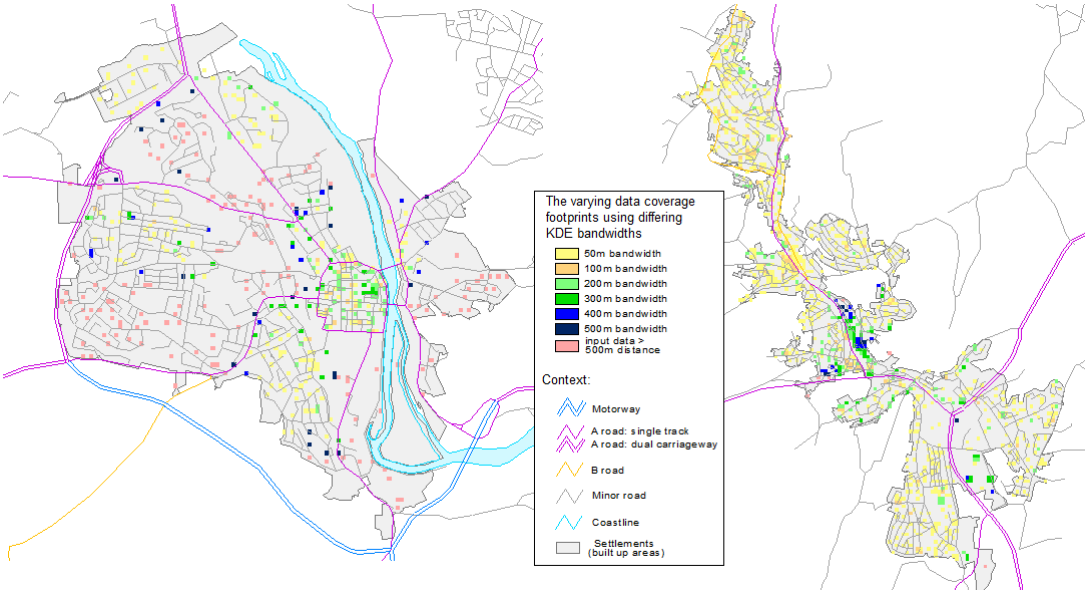
m) Cambden and Westminister, London *n) Loughborough*



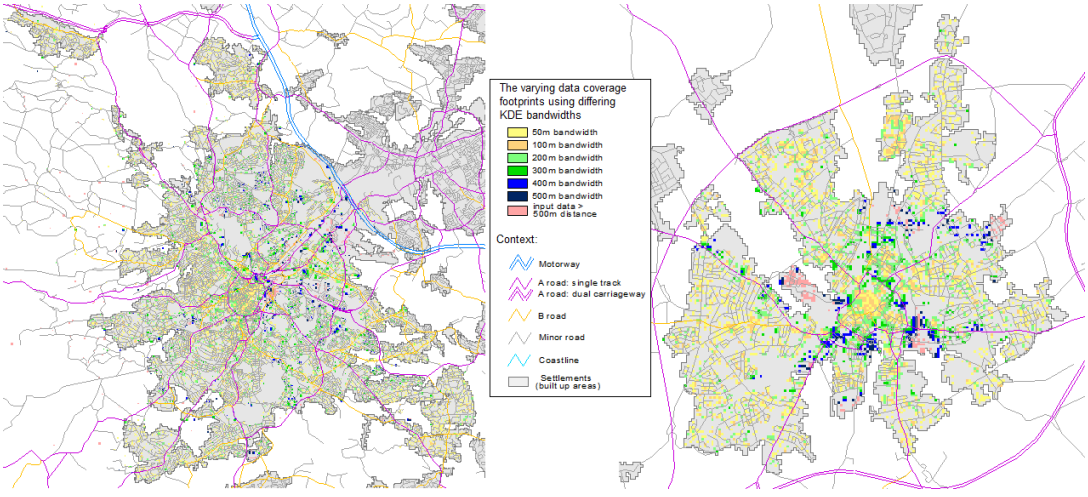
o) Ludlow *p) Morpeth*



q) Nottingham *r) Penrith*



s) Perth *t) Pontypool*



s) Sheffield *t) York*

Appendix C

Exploring domain name frequencies of urban local data

This appendix investigates the origin of data used to generate urban locale data and specifically explores the domain name frequencies of data output. It supplements data found in Chapter 6, concerning domain name frequencies, and Chapter 7 relating to the web classification of URLs. The frequency distributions of sources returning differing numbers of data returns varied considerable within the case study areas. As shown within Table C.1 below, the proportion of data that came from domain names that returned only a single data point varied from 46% and 42% in Crowborough and Bangor to 5.2% in Nottingham and Sheffield case areas.

The domain names that generated the largest number of data points (for all case study areas combined) can be found in Table C.2. The table demonstrates that even the Zoopla domain, which produced the largest number of data returns, only contributed less than 4% of total data - supporting the notion that data are drawn from a wide range of different sources. In contrast, over 16% of total data came from domains that contained 5 or fewer data points (with nearly 7% coming from domains with only a single data return). Such frequency distributions reinforce the importance of websites returning small amounts of data and downplays any assumption that data are biased by a select number of domains supplying substitutional numbers of data points. Critically, it does not appear that data are as heavily skewed, in terms of the proportion of data originating from a small number of sources, as other web based data such as photograph contributors. For instance, Hall et al. (2015) found that roughly 90% of the image captions from the Geograph site[1] were produced by only approximately 2% of the contributors.

[1] http://www.geograph.org.uk

Table C.1 Frequency of data with unique domain names by case study area

	Banbury		**Bangor**		**Barnstable**		**Crowborough**		**Dover**		**Durham**		**Frome**	
Frequency	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data
1	66.3	29.7	68.9	42.2	63.7	17.1	69.8	46.0	60.0	14.4	61.2	18.3	61.9	34.9
2	17.2	15.4	23.3	28.6	18.9	10.1	18.0	23.7	18.5	8.9	19.8	11.8	25.0	28.2
3	5.0	6.7	4.4	8.2	3.6	2.9	5.0	10.0	5.8	4.2	5.1	4.6	4.7	7.9
4	3.2	5.8	1.1	2.7	2.6	2.8	5.8	15.2	2.8	2.7	4.0	4.7	1.7	3.8
5 to 9	4.5	13.9	1.1	6.1	5.8	10.2	1.4	5.2	5.8	9.0	4.9	9.6	6.4	25.1
10 to 14	2.0	11.1	0.0	0.7	0.9	2.9	0.0	0.0	2.5	7.4	1.6	5.3	0.4	0.0
15 to 19	0.5	3.8	1.1	11.6	0.8	3.7	0.0	0.0	0.8	3.2	1.1	5.7	0.0	0.0
20 to 29	1.2	13.6	0.0	0.0	1.5	9.9	0.0	0.0	0.9	5.4	0.7	5.3	0.0	0.0
30 to 49	0.0	0.0	0.0	0.0	1.3	12.9	0.0	0.0	1.3	11.7	1.0	12.0	0.0	0.0
50 to 99	0.0	0.0	0.0	0.0	0.6	10.9	0.0	0.0	1.2	20.3	0.6	10.6	0.0	0.0
100 to 199	0.0	0.0	0.0	0.0	0.2	8.5	0.0	0.0	0.1	3.7	0.0	0.0	0.0	0.0
200 to 499	0.0	0.0	0.0	0.0	0.1	8.1	0.0	0.0	0.1	9.4	0.1	12.0	0.0	0.0
500 to 1999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2000 and over	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

	Great Yarmouth		**King's Lynn**		**Kidderminister**		**Lincoln**		**Camden and Westminster**		**Ludlow**		**Loughborough**	
Frequency	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data
1	58.9	11.0	59.0	14.0	56.7	16.6	57.7	12.1	61.7	10.8	65.9	42.0	61.4	23.4
2	19.7	7.4	22.6	10.7	20.8	12.2	22.1	9.3	21.2	7.4	24.2	30.8	16.5	12.6
3	5.3	3.0	4.3	3.1	5.5	4.8	6.3	4.0	5.0	2.7	2.2	4.2	4.7	5.4
4	3.2	2.4	3.4	3.3	4.1	4.8	3.2	2.7	2.7	1.9	3.3	8.4	4.7	7.2
5 to 9	6.0	7.0	5.1	7.9	6.1	11.4	4.9	6.4	4.6	5.3	4.4	14.7	9.4	21.3
10 to 14	1.8	4.2	1.3	3.5	2.7	9.7	1.7	4.1	1.5	3.1	0.0	0.0	2.4	10.2
15 to 19	1.3	4.0	0.4	1.5	1.4	6.6	1.0	3.4	0.8	2.3	0.0	0.0	0.0	0.0
20 to 29	1.1	5.1	1.4	8.1	1.0	7.3	0.9	4.7	0.9	3.7	0.0	0.0	0.0	0.0
30 to 49	0.6	4.4	1.0	10.0	0.7	7.4	0.5	4.4	0.6	4.1	0.0	0.0	0.0	0.0
50 to 99	0.9	11.2	0.9	13.5	1.0	19.4	0.9	13.8	0.5	5.4	0.0	0.0	0.8	19.8
100 to 199	0.6	14.3	0.5	16.5	0.0	0.0	0.5	14.4	0.2	6.1	0.0	0.0	0.0	0.0
200 to 499	0.4	18.6	0.1	7.8	0.0	0.0	0.3	12.4	0.2	10.6	0.0	0.0	0.0	0.0
500 to 1999	0.1	7.4	0.0	0.0	0.0	0.0	0.1	8.3	0.1	18.1	0.0	0.0	0.0	0.0
2000 and over	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	18.6	0.0	0.0	0.0	0.0

Table C.1 continued - Frequency of data with unique domain names by case study area

	Nottingham		**Morpeth**		**Penrith**		**Perth**		**Pontypool**		**Sheffield**		**Loughborough**	
Frequency	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data	% of sites	% of data
1	52.4	5.2	65.4	27.2	63.2	32.0	58.8	24.1	53.7	6.7	49.8	5.2	61.0	13.0
2	24.9	4.9	16.8	13.9	20.9	21.2	24.4	20.0	19.0	4.7	24.8	5.2	19.8	8.5
3	5.7	1.7	4.3	5.4	5.5	8.4	5.3	6.5	6.2	2.3	6.4	2.0	5.1	3.3
4	3.8	1.5	3.2	5.4	3.3	6.7	1.5	2.4	2.4	1.2	5.1	2.1	2.9	2.5
5 to 9	5.5	3.5	6.5	17.8	4.9	18.9	5.3	14.5	8.4	6.8	6.2	4.1	5.7	7.8
10 to 14	2.6	3.0	1.1	4.9	2.2	12.8	1.8	8.4	3.4	4.9	2.1	2.6	1.9	4.7
15 to 19	1.1	1.8	1.1	7.9	0.0	0.0	2.1	14.8	1.2	2.6	1.3	2.4	0.6	2.1
20 to 29	1.0	2.4	1.1	10.8	0.0	0.0	0.6	5.5	0.9	2.5	1.1	2.9	0.9	4.8
30 to 49	1.0	3.9	0.5	6.7	0.0	0.0	0.3	3.7	1.6	7.3	1.0	3.9	0.6	4.9
50 to 99	0.9	6.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	11.6	0.9	6.2	0.6	8.2
100 to 199	0.4	6.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	19.2	0.4	6.2	0.6	18.3
200 to 499	0.4	10.8	0.0	0.0	0.0	0.0	0.0	0.0	0.6	20.2	0.4	13.8	0.2	13.8
500 to 1999	0.2	25.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	9.9	0.3	32.0	0.1	8.4
2000 and over	0.1	24.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.5	0.0	0.0

Table C.2 Most common 50 domain names (combined for all case study areas)

Domain name	Frequency	Percentage of total data returns
www.zoopla.co.uk	13,451	3.67
www.rightmove.co.uk	6,648	1.81
bizzy.co.uk	5,333	1.46
www.checksure.biz	5,163	1.41
www.yell.com	5,149	1.41
www.smilelocal.com	5,089	1.39
www.zettai.net	4,934	1.35
www.ufindus.com	4,714	1.29
www.192.com	4,687	1.28
www.londontown.com	4,583	1.25
homes.trovit.co.uk	4,393	1.20
www.viewlondon.co.uk	4,386	1.20
www.mandis.co.uk	4,133	1.13
www.cylex-uk.co.uk	4,103	1.12
www.allinlondon.co.uk	4,073	1.11
www.freeindex.co.uk	3,428	0.94
nottingham.enquira.co.uk	3,162	0.86
www.ilondon.co.uk	3,161	0.86
www.sugarvine.com	3,108	0.85
www.locallife.co.uk	3,070	0.84
www.thomsonlocal.com	3,062	0.84
www.moreuk.com	3,052	0.83
www.yelp.co.uk	2,962	0.81
directory.independent.co.uk	2,875	0.78
www.ukcities.co.uk	2,828	0.77
www.touchlocal.com	2,571	0.70
www.bigreddirectory.com	2,337	0.64
www.london-eating.co.uk	2,314	0.63
www.yorkshirebd.co.uk	2,299	0.63
www.london-gazette.co.uk	2,277	0.62
www.118.com	2,070	0.57
www.companies-uk.co.uk	2,013	0.55
london.enquira.co.uk	1,996	0.54
www.scoot.co.uk	1,966	0.54
company-director-check.co.uk	1,833	0.50
www.primelocation.com	1,827	0.50
www.visitlondon.com	1,741	0.48
www.nottinghamonline.co.uk	1,737	0.47
www.localdatasearch.com	1,730	0.47
companycheck.co.uk	1,713	0.47
www.londonnet.co.uk	1,649	0.45
www.bizwiki.co.uk	1,607	0.44
www.hotfrog.co.uk	1,604	0.44
www.misterwhat.co.uk	1,596	0.44
sheffield.enquira.co.uk	1,503	0.41
www.wheresbest.co.uk	1,502	0.41
www.companieslist.co.uk	1,366	0.37
www.myhouseprice.com	1,342	0.37
www.shiftbusiness.co.uk	1,338	0.37
www.gbet.com	1,242	0.34
TOTAL	366,316	100.00

Appendix D

Data output for all case-studies

This appendix contains illustrative output for all case-studies not contained within the main body of work within Chapter 6.

Fig. D.1 Example output: (a) Banbury and (b) Bangor

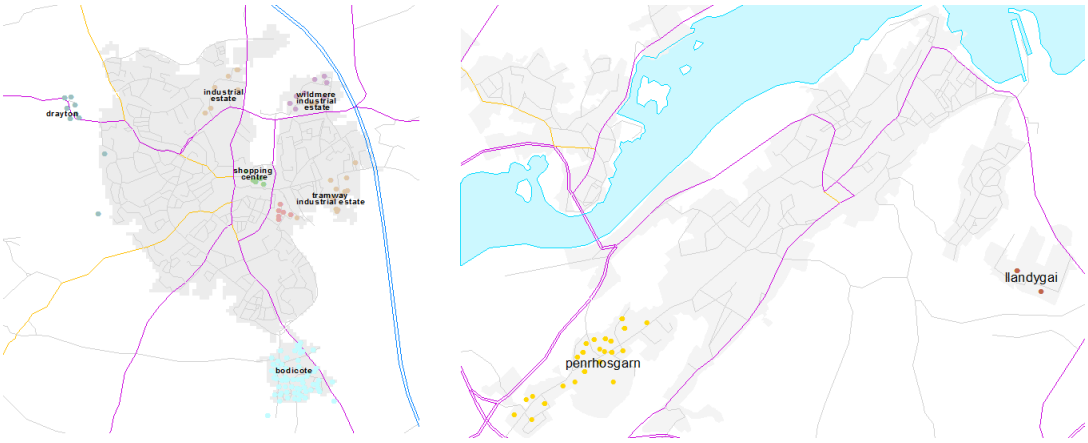


Fig. D.2 Example output: (a) Barnstaple and (b) Crowborough

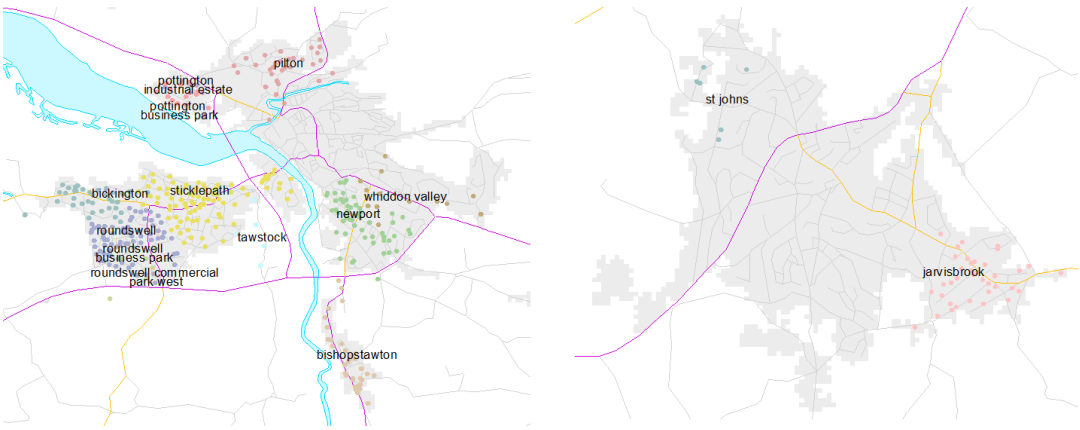


Fig. D.3 Example output: (a) Dover and (b) Durham

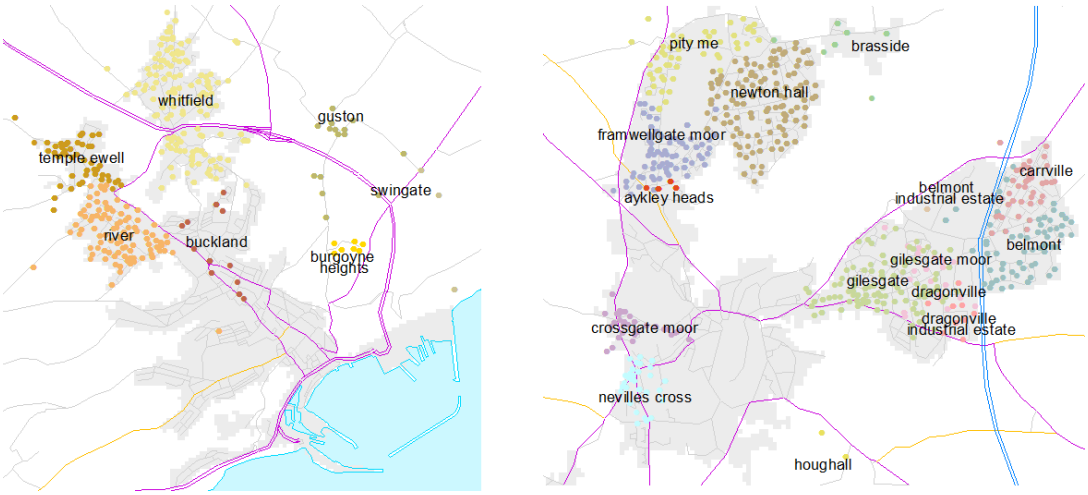


Fig. D.4 Example output: (a) Frome and (b) Great Yarmouth

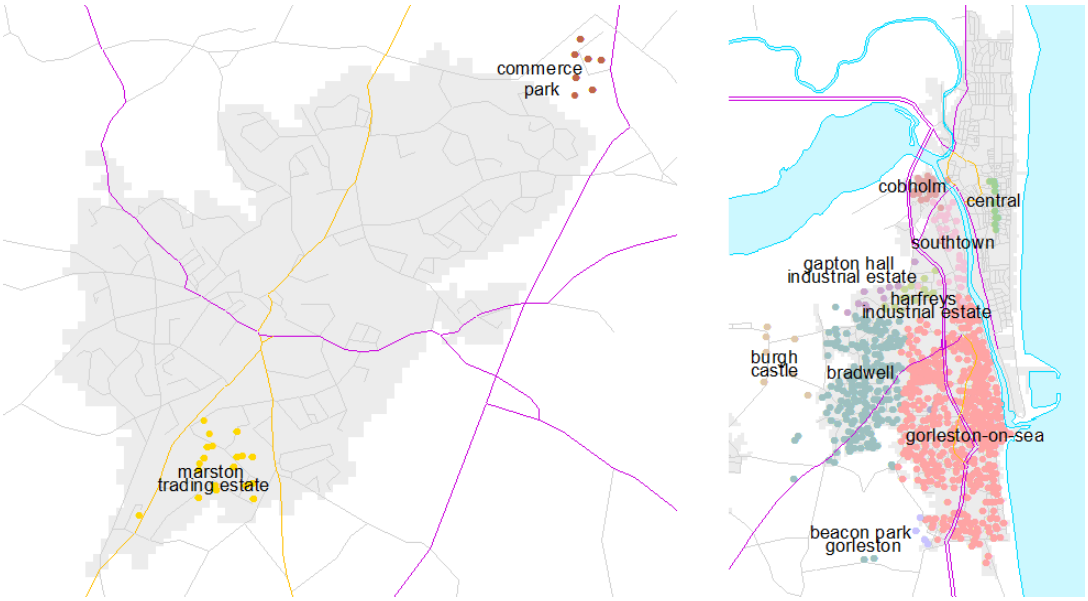


Fig. D.5 Example output: (a) Kidderminister and (b) King's Lynn

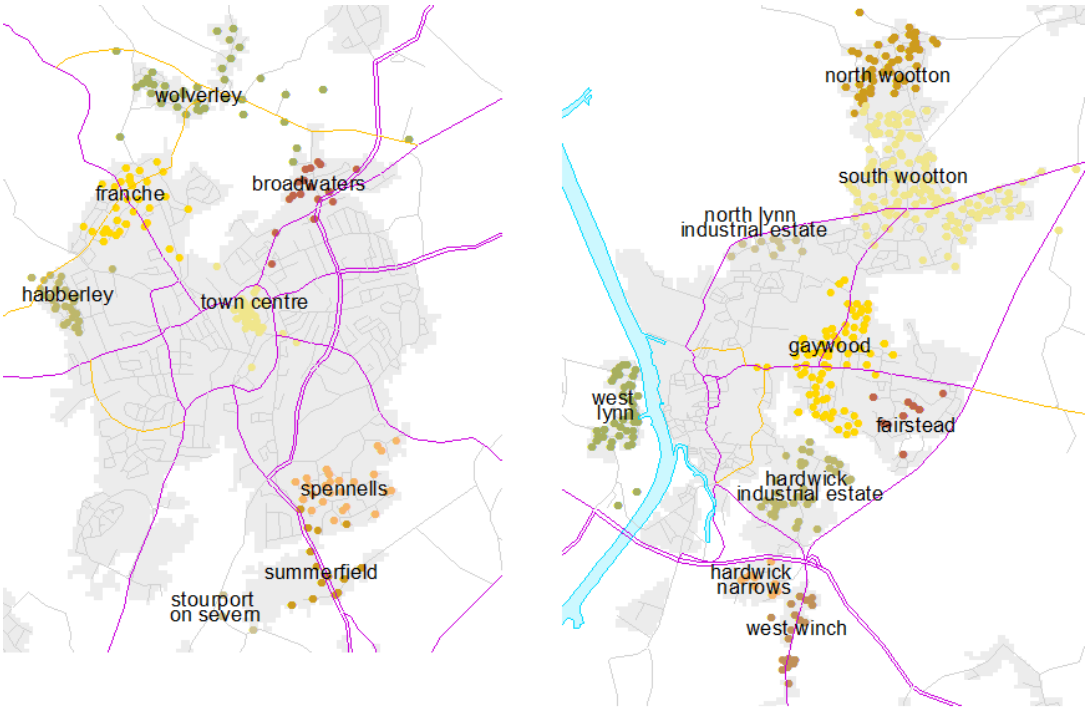


Fig. D.6 Example output: (a) Lincoln and (b) Loughborough

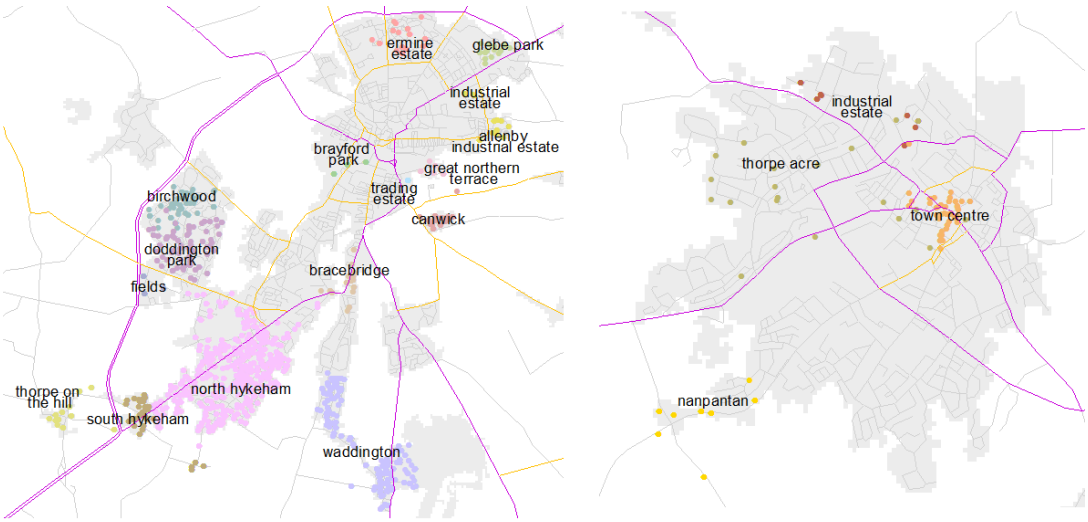


Fig. D.7 Example output: (a) Ludlow and (b) Morpeth

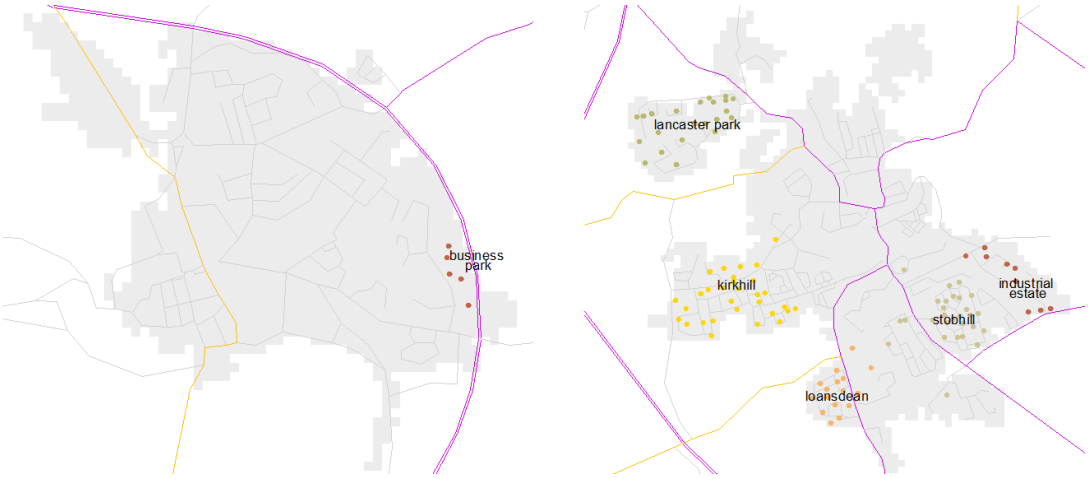


Fig. D.8 Example output: Nottingham

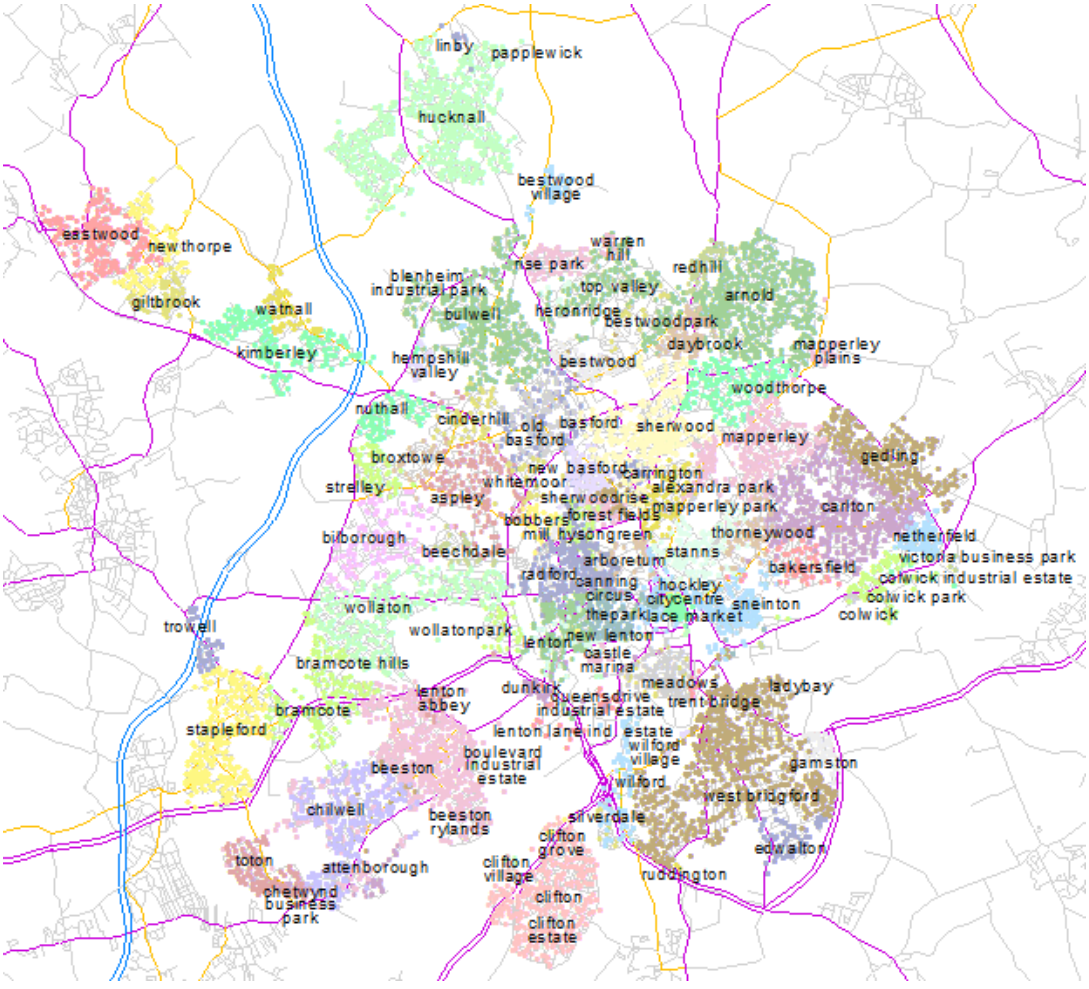


Fig. D.9 Example output: (a) Penrith and (b) Perth

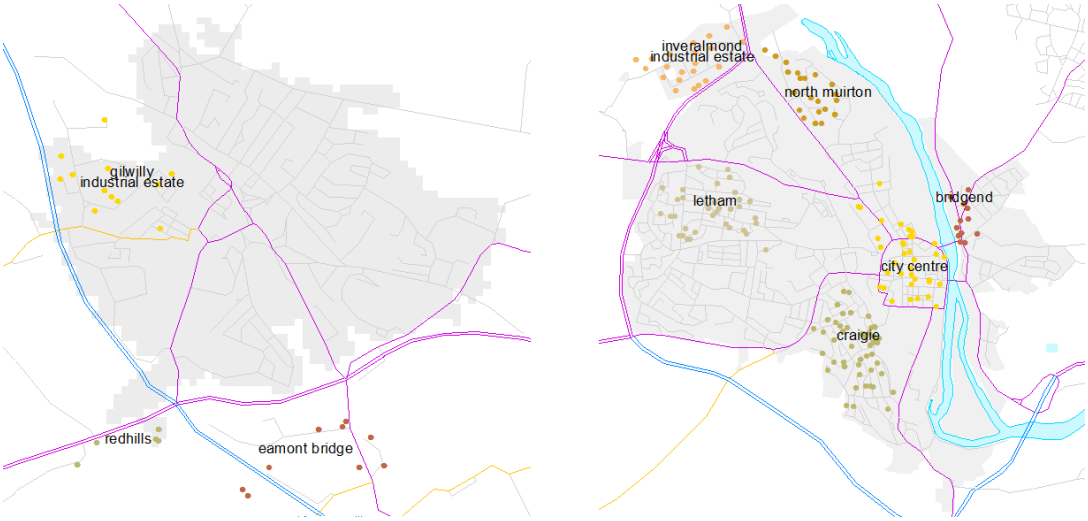
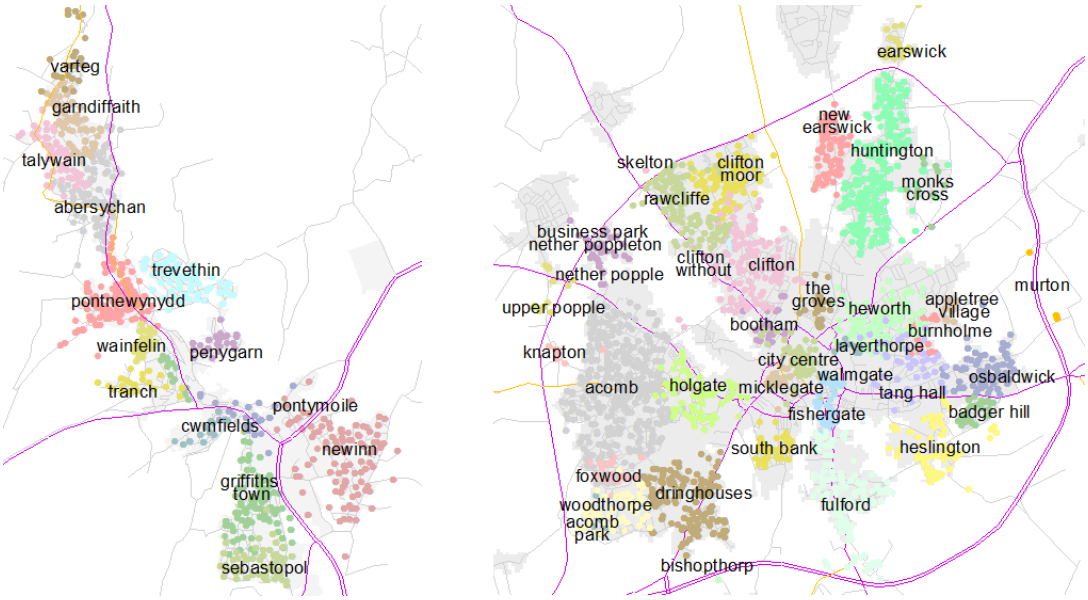


Fig. D.10 Example output: (a) Pontypool and (b) York



Appendix E

Levels of urban locale name agreement between different datasources

This appendix contains supplementary information to Chapter 6 specifically concerning the level of urban locale name agreement between output from different datasources for the two study areas of Sheffield and Nottingham.

Table E.1 Level of agreement between the different urban locale datasources in Sheffield

No. of sources	Source(s)	Freq.	Urban Locales
6	Council defined, Geonames, OS combined, OSM, Yahoo, web extraction output	17	Wadsley, Hillsborough, Dore, Wincobank, Walkley, Greenhill, Stannington, Hackenthorpe, Nether Edge, Darnall, Totley, Brightside, Handsworth, Heeley, Tinsley, Woodhouse, Grenoside
5	Council defined, OS combined, OSM, Yahoo, web extraction output	10	Highfield, Halfway, Beauchief, Waterthorpe, Arbourthorne, Norton, Fulwood, Loxley, Gleadless, Bradway
5	Council defined, Geonames, OS combined, OSM, web extraction output	8	Beighton, Middlewood, Ecclesfield, Brincliffe, Ecclesall, Woodseats, Longley, Manor
5	Geonames, OS combined, OSM, Yahoo, web extraction output	4	Frecheville, Grimesthorpe, Attercliffe, Neepsend

Continued on next page

Table E.1 – continued from previous page

No. of sources	Source(s)	Freq.	Urban Locales
4	Council defined, OS combined, OSM, web extraction output	35	Crosspool, Birley, Greystones, Sharrow, Sothall, Meersbrook, Jordanthorpe, Norfolk Park, Burncross, Westfield, Lodge Moor, Shirecliffe, Upperthorpe, Crookes, Firvale, Ranmoor, Richmond, Whirlow, Lowedges, Wybourn, Owlthorpe, Burngreave, Basegreen, Wisewood, Woodthorpe, Mosborough, Bents Green, Millhouses, Crookesmoor, Shiregreen, Broomhill, Southey Green, Netherthorpe, Firth Park, High Green
4	OS combined, OSM, Yahoo, web extraction output	1	Totley Rise
4	Geonames, OS combined, OSM, web extraction output	1	Parson Cross
4	Council defined, Geonames, OS combined, web extraction output	1	Chapeltown
4	Geonames, OS combined, OSM, Yahoo	1	Carbrook
4	Council defined, OS combined, OSM, Yahoo	1	Hemsworth
4	Council defined, Geonames, OS combined, OSM	1	Hollins End
3	OS combined, OSM, web extraction output	13	Banner Cross, Norton Lees, Wadsley Bridge, Sharrow Vale, Carterknowle, Manor Park, Sandygate, Lane Top, Nethergreen, Meadowhead, Intake, Pitsmoor, Malin Bridge
3	Council defined, OS combined, web extraction output	7	Stocksbridge, Worrall, Oughtibridge, Abbeydale, Wharncliffe Side, Deepcar, Gleadless Valley
3	Council defined, OSM, web extraction output	4	Foxhill, Broomhall, Charnock, Citycentre
3	Geonames, OSM, web extraction output	2	Manor Top, Crystal Peaks
3	Geonames, OS combined, web extraction output	2	Bradfield, Meadowhall
3	OS combined, OSM, Yahoo	2	Lowfield, Greenland
3	Council defined, OS combined, OSM	2	Park Hill, Batemoor

Continued on next page

Table E.1 – continued from previous page

No. of sources	Source(s)	Freq.	Urban Locales
3	OS combined, Yahoo, web extraction output	1	Woodhouse Mill
3	Geonames, OS combined, OSM	1	Herdings
2	OS combined, OSM	8	Newfield Green, Common Side, Birley Edge, Parkhead, Philadelphia, Normanton Spring, Lower Walkley, Broomfield
2	OS combined, web extraction output	7	Dungworth, Birley Carr, Bolsterstone, Rotherham, Parkwood Springs, Holbrook, Kelham Island
2	Council defined, OS combined	4	Stubbin, Woodland View, Woodside, Walkley Bank
2	OS combined, Yahoo	2	Tapton Hill, New Totley
2	Geonames, OS combined	2	Gleadless Townend, Elm Tree Hill
2	Geonames, web extraction output	1	Shalesmoor
2	Geonames, Yahoo	1	High Hazels
2	Council defined, OSM	1	Endcliffe
2	Council defined, Geonames	1	Langsett
1	OSM	11	Hinde House, Page Hall, Stradbroke, Wicker, Saint Philip's, Littledale, Atlas, Saint George's, Saint Vincent's, Osgathorpe, Devonshire Quarter
1	Council defined	11	Abbeyfield, Granville, New Parson Cross, Acres Hill, Housteads, Firshill, Rural Area, Old Parson Cross, Brushes, Flower, Colley
1	web extraction output	7	Hunters Bar, Meadowhall Centre, Newhall, Wadsley Park Village, Thorncliffe Park Chapeltown, Holbrook Industrial Estate, Parkway Industrial Estate
1	Geonames	7	Castle Square, Park Grange, Park Grange Croft, Copthorne, Fitzalan Sq - Ponds Forge, Valley Centertainment, University of Sheffield
1	Yahoo	2	Crookesmoon, Sheffield Technology Park

Continued on next page

Table E.1 – continued from previous page

No. of sources	Source(s)	Freq.	Urban Locales
1	OS combined	104	Bell Hagg, Plumbley, Meersbrook Bank, Fairleigh, Pismire Hill, Summerfield, Lane End, Town End, Rivelin Glen, Little Common, Arbourthorne Estate, Hallowmoor, Carr Head, Lydgate, Brick Houses, Ecclesfield Common, Hollin Busk, Hallam Head, Little Norton, Charltonbrook, Spink Hall, Little Sheffield, Den Bank, Deep Pit, The Wheel, Lower Crabtree, Fourlane Ends, Rollestone, Pot House, Wood Willows, Stephen Hill, Howard Hill, Old Haywoods, Bradway Bank, Townend House, East Whitwell, Abbeydale Park, Liberty Hill, Normandale, Greystones Cliffe, Birkendale, Drakehouse,Coisley Hill, Owlerton, Endcliffe Crescent, Upper Crabtree, Knoll Top, Sheffield Park, Endcliffe Edge, Charnock Hall, Upper Bradway, Causeway Head, Totley Brook, Backmoor, Steel Bank, Hurl Field, Knowle Top, Bolehill, Grange Lane, Potter Hill, Killamarsh, Clough Field, Lower Bradway, Cricket Houses, Sheffield City Airport, Wadsley Grove, Low Wincobank, Shay House, Templeborough, Park Head, Handsworth Hill, Townhead, Oxclose, Manor Oaks, Goole Green, Hillfoot, Port Mahon, Attercliffe Hill Top, Norton Hammer, Hill Top, Studfield Hill, Park Side, Sharrow Head, The Cross, Garden Village, Mortomley, Nethergate, Loxley Bottom, Carsick, Boulder Hill, Hanging Water, Mount Pleasant, Wood Royd, Orgreave, Warren, Royd, Norton Woodseats, Stumperlowe, Carsick Hill, High Wincobank, Four Lane Ends, Bracken Moor, Smithy Moor, Horner House

Table E.2 Level of agreement between the different urban locale datasources in Nottingham

No. of sources	Datasource(s)	Frequency	Urban Locales
5	Geonames, OS combined, OSM, Yahoo, web extraction output	4	Chilwell, Toton, Sneinton, Bulwell
4	OS combined, OSM, Yahoo, web extraction output	14	Sherwood, Aspley, Mapperley Park, The Park, Whitemoor, Radford, Broxtowe, Lenton Abbey, Top Valley, Lenton, Old Basford, Cinderhill, Dunkirk, Rise Park
4	Geonames, OS combined, OSM, web extraction output	4	Gedling, Carlton, Edwalton, Clifton
4	Geonames, OS combined, Yahoo, web extraction output	3	Meadows, Woodthorpe, Bramcote
3	OS combined, Yahoo, web extraction output	14	Bestwood, New Basford, Sherwood Rise, Thorneywood, Carrington, Wollaton, St Anns, Bakersfield, Bestwood Village, Mapperley, Daybrook, Beechdale, Forest Fields, Bilborough
3	Geonames, OS combined, web extraction output	10	Stapleford, Redhill, West Bridgford, Hucknall, Arnold, Watnall, Wilford, Beeston, Attenborough, Colwick
3	OS combined, OSM, web extraction output	7	Bramcote Hills, Bestwood Park, Hempshill Vale, Gamston, Wollaton Park, Beeston Rylands, Hyson Green
3	Geonames, Yahoo, web extraction output	1	Basford
3	OS combined, OSM, Yahoo	1	Highbury Vale
2	OS combined, web extraction output	22	Castle Marina, Bobbers Mill, Greasley, Strelley, Arboretum, New Lenton, Giltbrook, Newthorpe, Trowell, Eastwood, Kimberley, Colwick Industrial Estate, Trent Bridge, Clifton Grove, Linby, The Lace Market, Awsworth, Nuthall, Alexandra Park,Netherfield, Mapperley Plains, Ruddington
2	OSM, web extraction output	5	Lady Bay, Bestwood Park Estate, The Meadows, Silverdale, Bestwood Estate
2	OS combined, Yahoo	3	Bulwell Forest, Standard Hill, Porchester
1	web extraction output	18	Lenton Lane Industrial Estate, Papplewick, Clifton Village, Wilford Village, Victoria Business Park Netherfield, Citycentre, Colwick Park, Queens Drive Industrial Estate, Chetwynd Business Park, Canning Circus, Clifton Estate, Boulevard Industrial Park Beeston, Blenheim Industrial Estate, Old Market Square, Hockley, Warren Hill, Castle Boulevard, Heronridge
1	OS combined	11	Old Lenton, Wilford Hill, Adbolton, Wighay, Broomhill, Hazelgrove, Ruffs, Butler's Hill, Westville, Beauvale, Chettles Industrial Estate
1	Yahoo	3	Highfields Science Park, Stockhill, Sneiton
1	OSM	3	Compton Acres, Assarts Farm Estate, Abbey Park
1	Geonames	3	Watnall Cantelupe, Watnall Chaworth, Hucknall Torkard

Appendix F

Web grouping: ancillary information

This appendix contains ancillary information relating to the method for the classification of data into differing 'web groupings' (as discussed within Chapter 7), whereby data from real estate sources can be separated from that obtained relating to business sources or other data.

1. URL keywords

Real Estate: 'estateagent', 'estate-agent', 'accommodation', 'letting', 'houseprice', 'house-price', 'estates'

Business information: 'directory', 'findlocal', 'www.finda', '.mylocal', '.ufindus.', 'classified', 'companies', 'companysearch', 'localbusiness', 'businessregister', 'businesspages', 'ufindus', 'companydirector', 'company-director'

2. Metadata words

Real Estate: ' property ', ' estate agent ', ' estate agents ', ' real estate ', ' rent ', ' renting ', ' sale ', ' sales ', ' auction ' ' auctions '

Business information: 'business directory', 'company directory', 'business listing', 'company listing', 'business Search', 'yellow pages', 'company report', 'credit report', 'company accounts', 'company documents', 'company credit', 'company search', 'credit rating', 'company house', 'companies house', 'company information', 'business information', 'local directory', 'tradesmen directory', 'find your local '

3. Web-page terms

Real Estate: ' property ', ' estate agent ', ' estate agents ', ' real estate ', ' rent ', ' renting ', ' sale ', ' sales ', ' auction ', ' auctions ', ' bedroom', 'detached ', ' flat ', '

flats ’, ‘ homes ’, ‘ house ’, ‘ houses ’, ‘ price ’, ‘ properties ’, ‘ terrace ’, ‘ terraced ’, ‘ apartment’, ‘ landlords ’, ‘ tennants ’

Business information: ‘business directory’, ‘company directory’, ‘business listing’, ‘company listing’, ‘business Search’, ‘yellow pages’, ‘company report’, ‘credit report’, ‘company accounts’, ‘company documents’, ‘company credit’, ‘company search’, ‘credit rating’, ‘company house’, ‘companies house’, ‘company information’, ‘business information’, ‘classified listing’, ‘local directory’

Table F.1 Examples of manually assigned URLs

Domain URL	Frequency	Class
www.zoopla.co.uk	6,371	Real estate
www.smilelocal.com	3,036	Business info.
www.ufindus.com	2,738	Business info.
www.rightmove.co.uk	2,737	Real estate
www.yell.com	2,200	Business info.
www.zettai.net	2,129	Business info.
www.cylex-uk.co.uk	2,066	Business info.
www.ukcities.co.uk	2,060	Business info.
www.yelp.co.uk	1,856	Business info.
www.moreuk.com	1,799	Business info.
www.192.com	1,682	Business info.
www.mandis.co.uk	1,528	Business info.
www.freeindex.co.uk	1,394	Business info.
www.checksure.biz	1,357	Business info.
www.touchlocal.com	1,335	Business info.
www.thomsonlocal.com	1,318	Business info.
www.london-gazette.co.uk	1,272	Other
www.118.com	1,137	Business info.
www.scoot.co.uk	1,066	Business info.
www.gbet.com	976	Business info.
www.bigreddirectory.com	942	Business info.
www.misterwhat.co.uk	669	Business info.
www.ib2b.co.uk	661	Business info.
www.myhouseprice.com	647	Real estate
www.companies-uk.co.uk	647	Business info.
www.wheresbest.co.uk	619	Business info.
www.thetimenow.com	594	Business info.
www.bizwiki.co.uk	535	Business info.
www.docstoc.com	495	Other
www.touchnottingham.com	433	Business info.
www.primelocation.com	425	Real estate
www.nhs.uk	393	Business info.
www.accommodationforstudents.com	379	Real estate
www.theconstructioncentre.co.uk	365	Business info.
www.ofsted.gov.uk	364	Other

Continued on next page

Table F.1 – continued from previous page

Domain URL	Frequency	Class
www.nethouseprices.com	352	Real estate
www.applegate.co.uk	349	Business info.
www.ordertakeaways.co.uk	345	Other
www.schools.org.uk	336	Business info.
www.autoinsider.co.uk	321	Business info.
www.britishpostcodes.info	317	Real estate
www.beerintheevening.com	316	Other
www.hotfrog.co.uk	312	Business info.
www.locallife.co.uk	310	Business info.
www.uk-web-index.co.uk	299	Business info.
www.shoplocally.co.uk	283	Business info.
www.cdrex.com	265	Business info.
www.housingcare.org	260	Business info.
www.dft.gov.uk	244	Other
www.takeawaysulike.co.uk	244	Other
www.zoominfo.com	240	Business info.
www.businessmagnet.co.uk	230	Business info.
www.localdatasearch.com	230	Business info.
www.uk-local-search.co.uk	222	Business info.
www.footballtradedirectory.com	217	Other
www.thediscdirectory.co.uk	215	Business info.
www.carehome.co.uk	200	Business info.
www.bdexcel.com	173	Business info.
www.ukbusinessconnect.com	167	Business info.
www.cityvisitor.co.uk	161	Business info.
www.creditgate.com	161	Business info.
www.linkedin.com	160	Other
www.eigroup.co.uk	159	Real estate
www.sugarvine.com	145	Other
www.whatdotheyknow.com	142	Other
www.findmyschool.co.uk	142	Business info.
www.ebay.com	141	Other
www.nuroa.co.uk	134	Real estate
www.123people.co.uk	131	Other
www.ebay.co.uk	131	Other
www.home.co.uk	130	Real estate
www.near.co.uk	127	Business info.
www.touchdoncaster.com	105	Business info.
www.finder.co.uk	100	Business info.
...	...	...
Grand Total	59,003	

Appendix G

Homogeneity testing: ancillary information

This appendix contains additional box-plots exploring the degree of urban locale homogeneity for a number of different socio-economic variables from the 2011 Census of Population. Output within the appendix are for selected Nottingham locales.

Using a single socio-economic variable from the 2011 UK Census of Population, it is possible to illustrate the degree to which data vary at a smaller spatial scale (for example Census Output Areas - OAs) within the larger urban locale units. Examination of the box-plots within this appendix demonstrates that some variables exhibit greater homogeneity within urban locale units than others (for example social class measures and ethnicity). It should be noted that smaller locales are more likely to exhibit less variation (greater homogeneity) due to probability and fewer input Output Area (OA) data. Even in cases of greater homogeneity, however, the variations within the boundary units (urban locales) are highly pronounced, as shown by the interquartile ranges (the spread of the red boxes within the figures below) and demonstrating that the units themselves are rarely completely homogeneous.

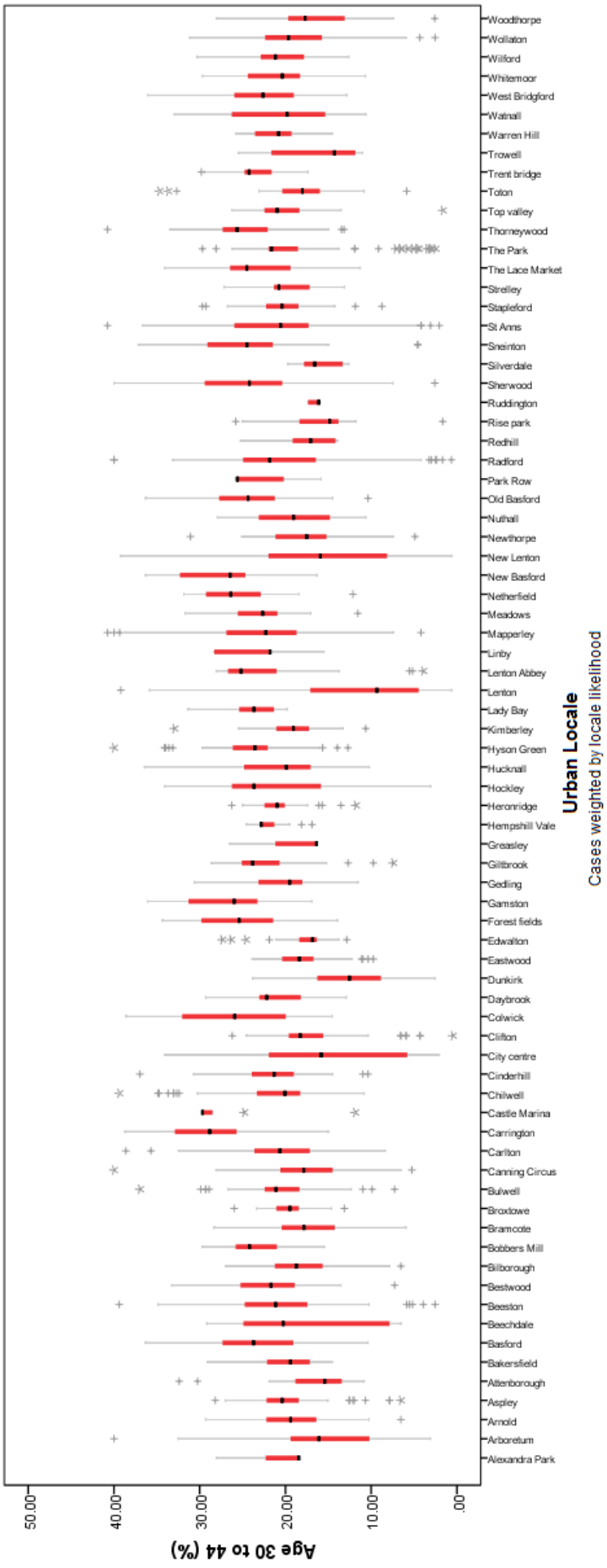


Fig. G.1 Variations within the percentage of people aged 30-44 in selected Nottingham urban locales

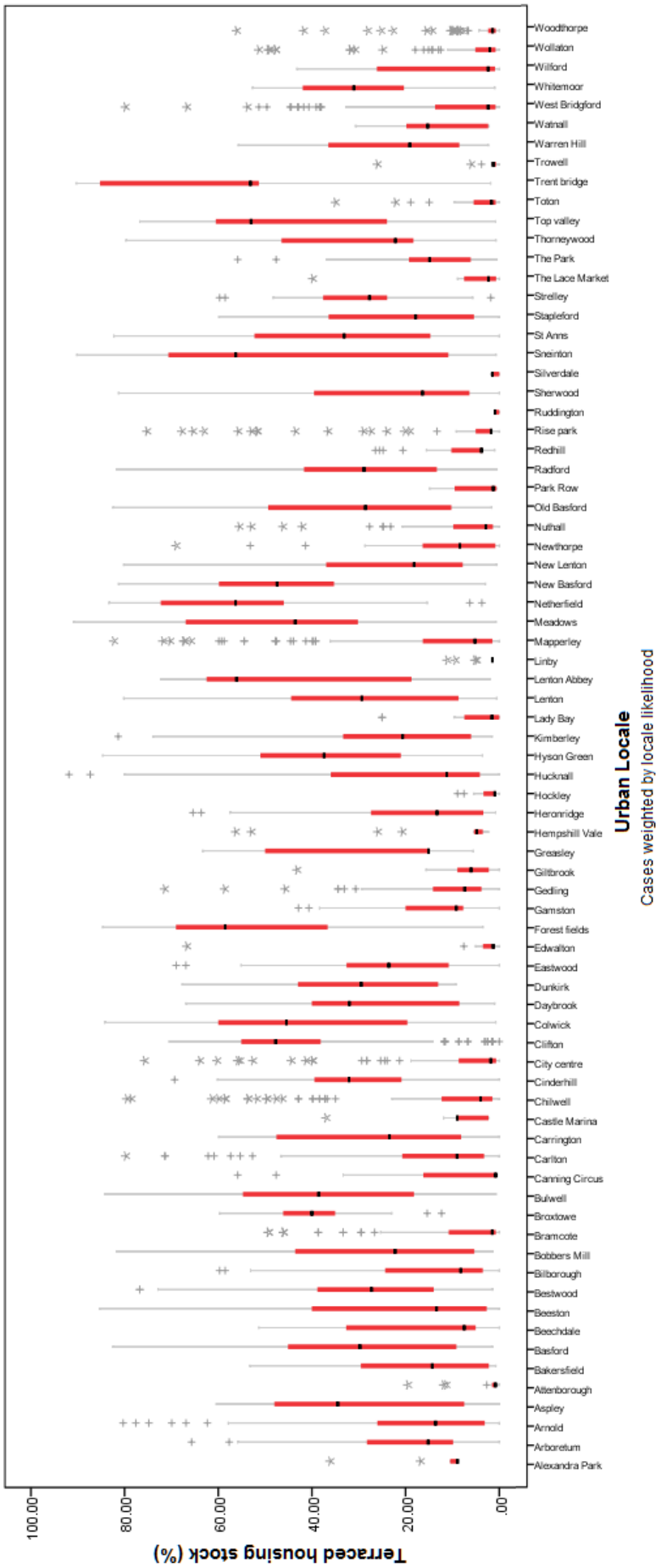


Fig. G.2 Variations within the percentage of terraced housing in selected Nottingham urban locales

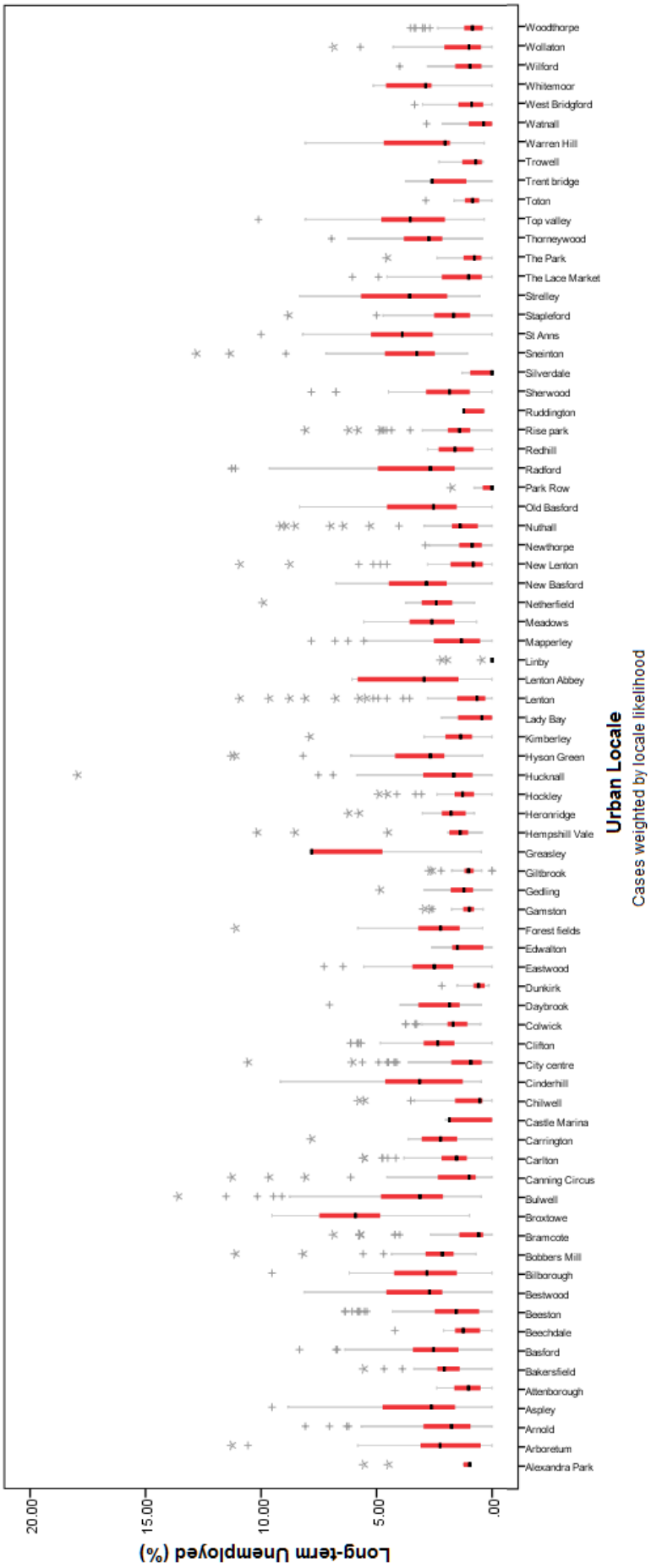


Fig. G.3 Variations within the percentage of people who are long-term unemployed in selected Nottingham urban locales

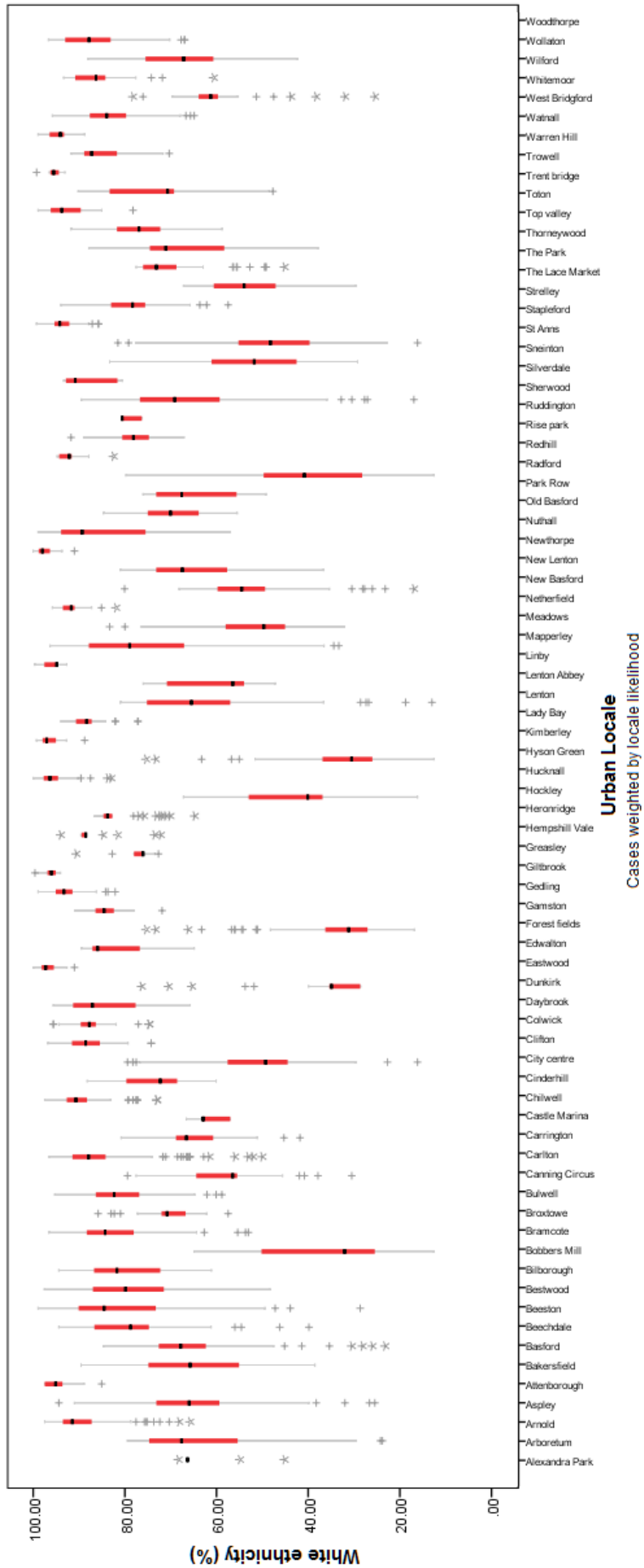


Fig. G.4 Variations within the percentage of people of white ethnicity in selected Nottingham urban locales

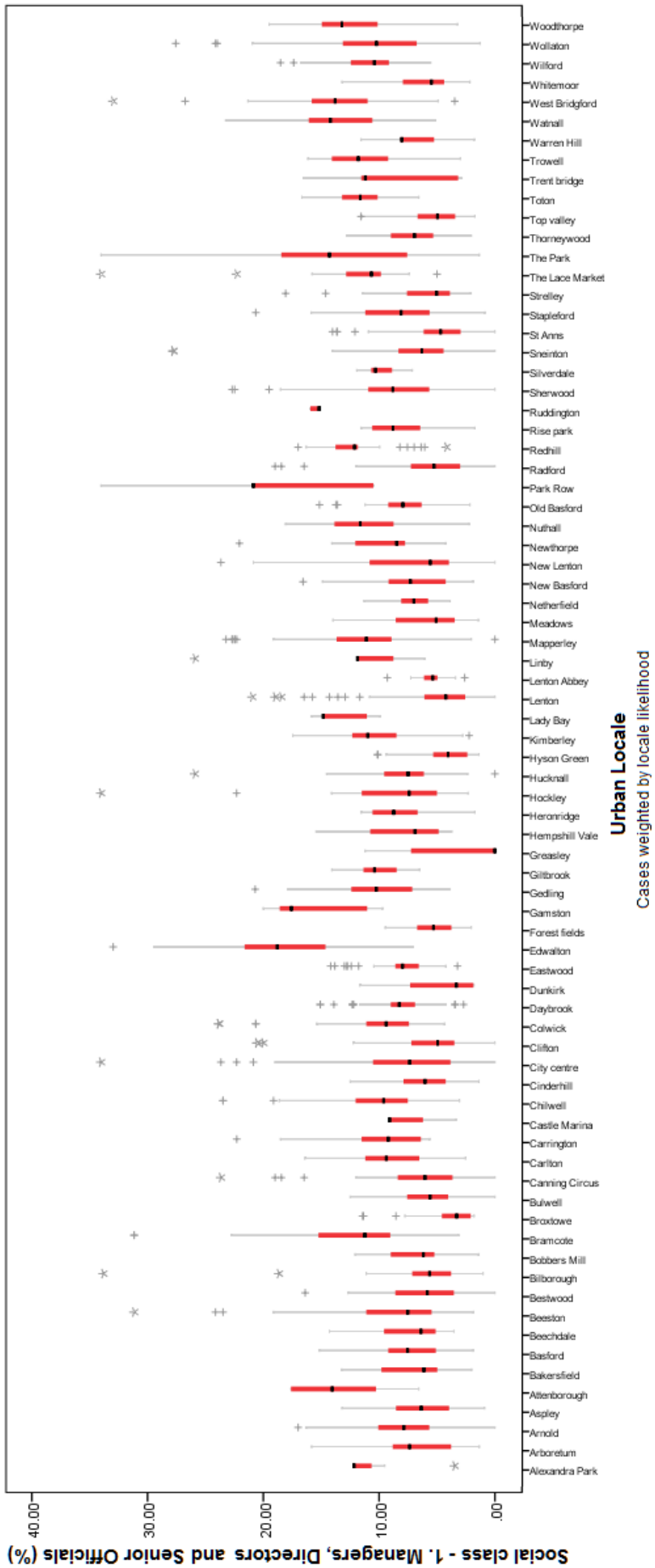


Fig. G.5 Variations within the percentage of people who are managers, directors or senior officials Nottingham urban locales

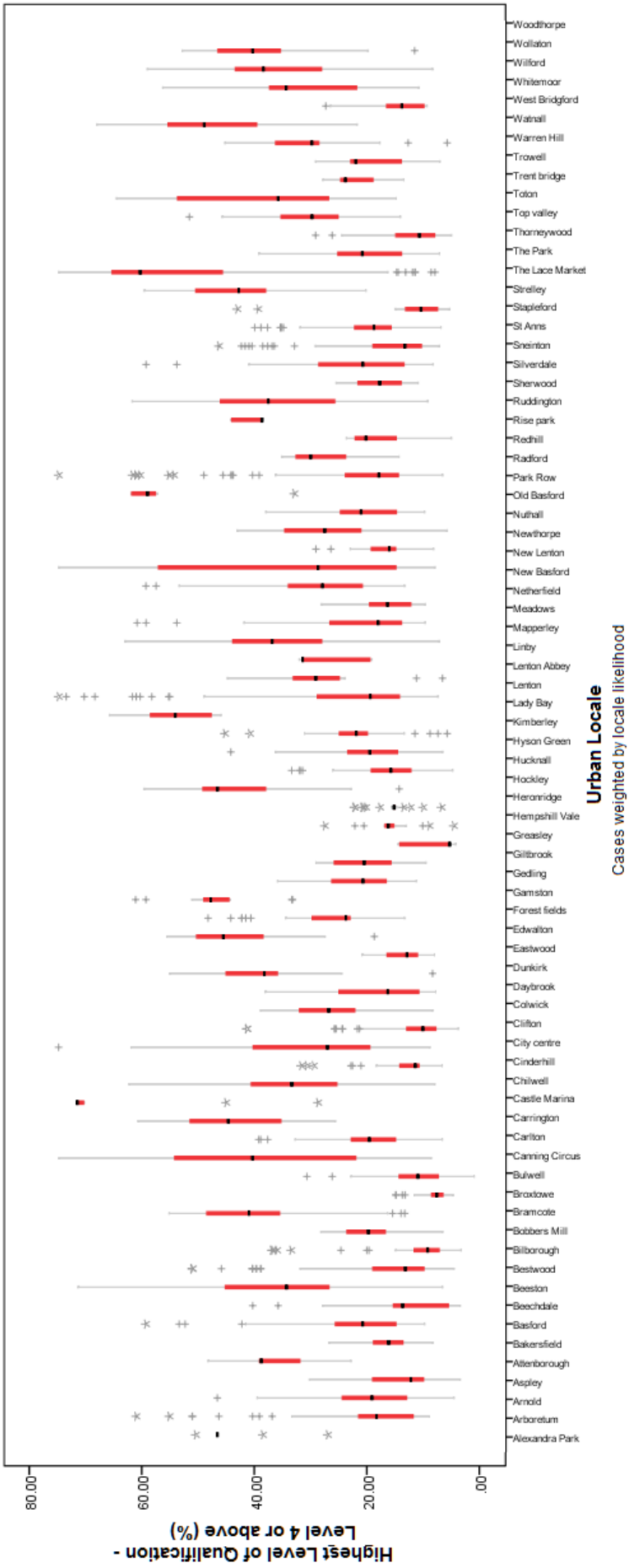


Fig. G.6 Variations within the percentage of people with their highest qualification at level 4 or above in selected Nottingham urban locales

Appendix H

Relationship with underlying population

This appendix provides supplementary information regarding the investigation of data returns (using the method as described in Chapter 5) with the underlying population (as discussed within Chapter 7: Comparison with underlying population geography). One might, for example, expect higher data returns from areas with greater populations.

1. Nottingham: Relationship between underlying population and all data returned from Bing API requests

This first relationship compares the underlying population (in terms of numbers of residential and non-residential properties as supplied via the OS data Codepoint) with the numbers of data returns from the Bing API irrespective of references to urban locales (essentially a measure of web-visibility).

Adjusted R-Squared = 0.9482

$$D = 0.0777 + 0.6277 * [res300] + 11.9527 * [nres300]$$

where :

D = the number of total data returns (regardless of urban locale content within them)

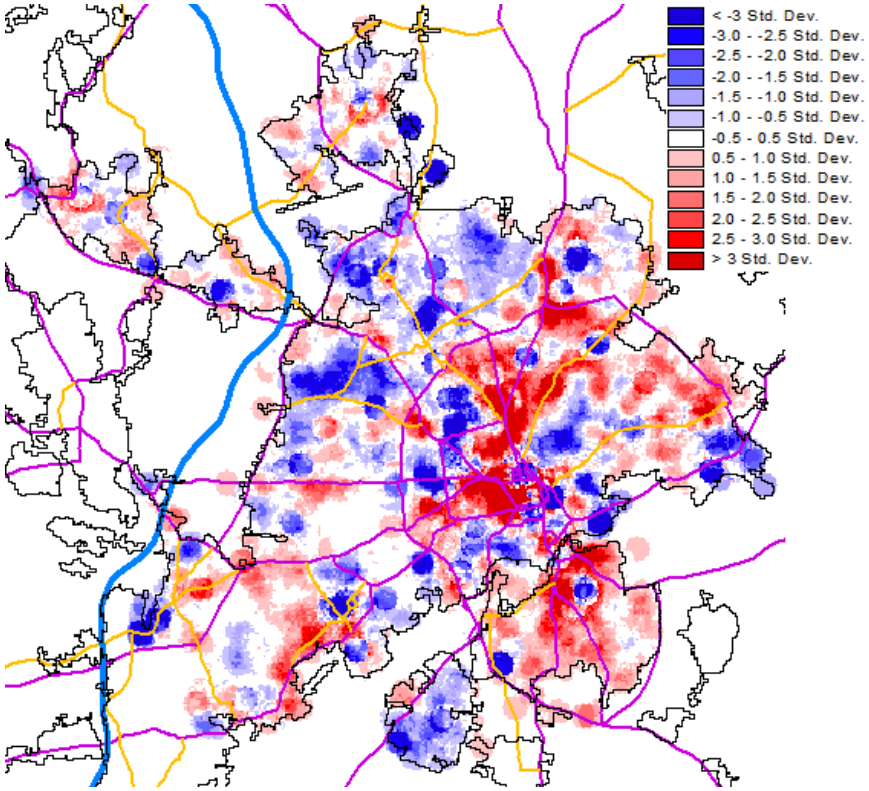
res300 = average number of residential delivery points within 300m

nres300 = average number of non-residential delivery points within 300m

Table H.1 Regression output: relationship between all data and delivery points: Nottingham

Variable	Coefficient	Std. Error	t-Value	Significance
[Intercept]	0.0777	0.0031	25.40	<0.00001
[res300]	0.6277	0.0011	560.65	<0.00001
[nres300]	11.9527	0.0103	1163.39	<0.00001

Fig. H.1 Residuals for the relationship between the total number of data returns and numbers of residential and non-residential properties



2. Sheffield: Relationship with urban locale data and underlying population

This relationship does not take Royal Mail designated Dependent localities into consideration, and hence, the relationship is weak and Dependent localities appear as positive outliers to the model.

Adjusted R-Squared = 0.3683

$$UD = 0.0712 + 0.1816 * [res300] + 0.4562 * [nres300]$$

where :

UD = the number of data returns with an urban locale within it

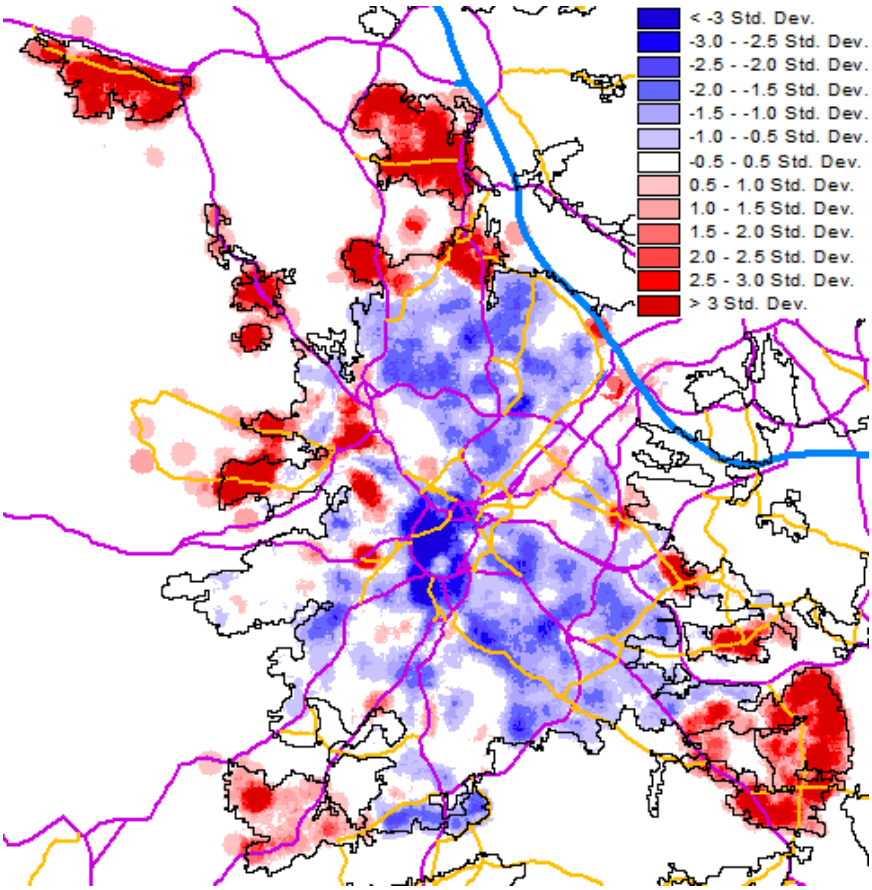
res300 = average number of residential delivery points within 300m

nres300 = average number of non-residential delivery points within 300m

Table H.2 Regression output: relationship between named urban locale data and delivery points: Sheffield

Variable	Coefficient	Std. Error	t-Value	Significance
[Intercept]	0.0712	0.0014	51.90	<0.00001
[res300]	0.1816	0.0006	293.94	<0.00001
[nres300]	0.4562	0.0069	65.66	<0.00001

Fig. H.2 Residuals for the relationship between the total number of named urban locale data and numbers of residential and non-residential properties



3. Nottingham: Relationship with urban locale data and underlying population

The information below repeats that for the relationship in model 2 of this appendix (above) but is for Nottingham.

Adjusted R-Squared =0.524422

$$UD = 0.0962 + 0.3646 * [res300] + 1.1931 * [nres300]$$

where :

UD = the number of data returns with an urban locale within it

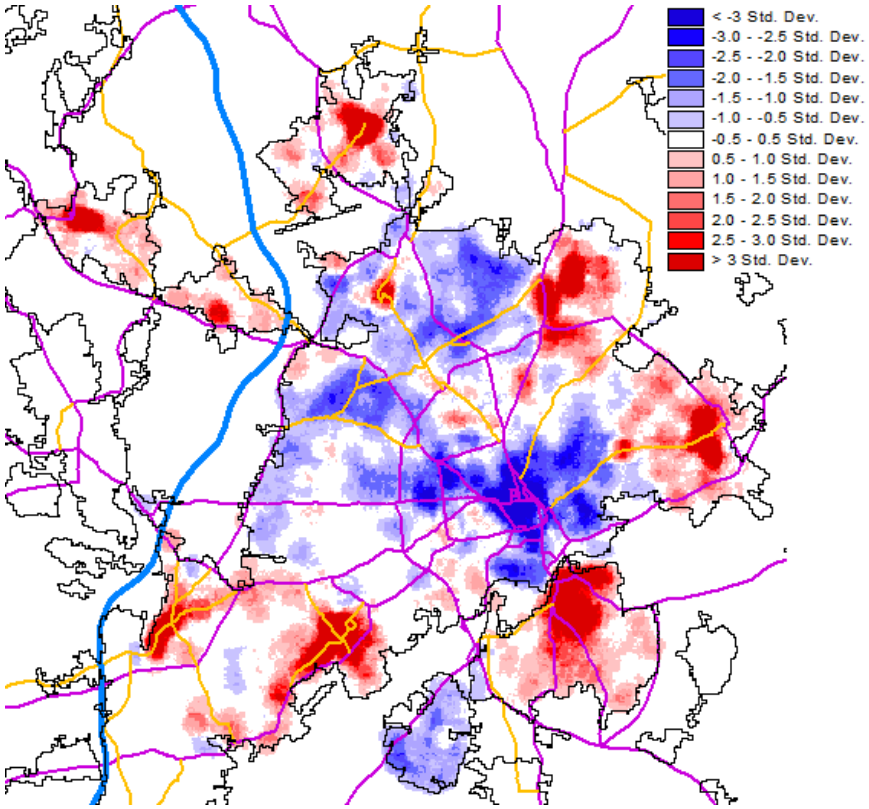
res300 = average number of residential delivery points within 300m

nres300 = average number of non-residential delivery points within 300m

Table H.3 Regression output: relationship between named urban locale data and delivery points: Nottingham

Variable	Coefficient	Std. Error	t-Value	Significance
[Intercept]	0.0962	0.0032	29.62	<0.00001
[res300]	0.3646	0.0012	306.55	<0.00001
[nres300]	1.1931	0.0109	109.32	<0.00001

Fig. H.3 Residuals for the relationship between the total number of named urban locale data and numbers of residential and non-residential properties



4. Nottingham: Relationship with urban locale data and underlying population, when accounting for Dependent localities

The model below includes populations within Dependent localities as a separate variable - allowing for different slopes within the regression analysis.

Adjusted R-Squared = 0.791800

$$UD = 0.0644 + 0.1886 * [res300] + 0.4728 * [nres300] + 0.2407 * [dlres300] + 4.4913 * [dlnres300]$$

where :

UD = the number of data returns with an urban locale within it

res300 = average number of residential delivery points within 300m

nres300 = average number of non-residential delivery points within 300m

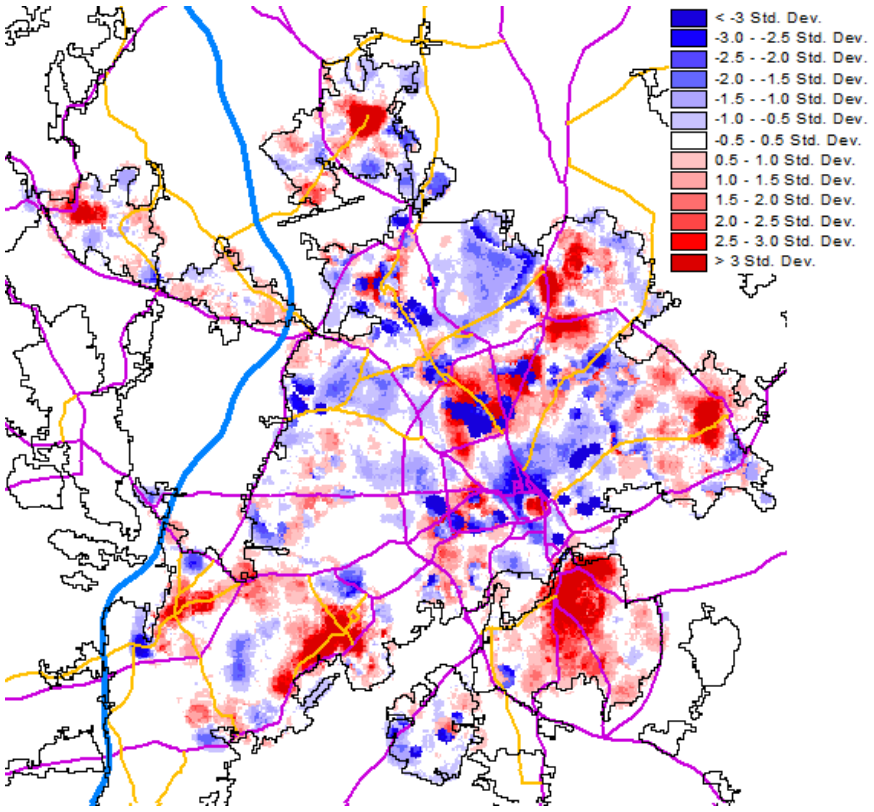
dlres300 = average number of residential delivery points within 300m in Dependent locality areas

dlnres300 = average number of non-residential delivery points within 300m in Dependent locality areas

Table H.4 Regression output: relationship between named urban locale data and delivery points: Nottingham

Variable	Coefficient	Std. Error	t-Value	Significance
[Intercept]	0.0644	0.0022	29.93	<0.00001
[res300]	0.1886	0.0010	183.14	<0.00001
[nres300]	0.4728	0.0082	57.52	<0.00001
[dlres300]	0.2407	0.0014	167.54	<0.00001
[dlnres300]	0.4728	0.0082	57.52	<0.00001

Fig. H.4 Residuals for the relationship between the total number of named urban locale data and numbers of residential and non-residential properties



While the residuals from the model shown in Table H.5 at first appear to prescribe to strong tendencies relating to affluence and peripherality, there are exceptions such as the highly positive residuals for Hyson Green. It should be noted that Hyson Green exhibited negative residuals within model 1 in this appendix - meaning that it has

Table H.5 The fifteen highest and lowest average regression residuals when aggregated to urban locales: Nottingham

Urban locale	Average residual score
Sneinton	-2.64
City centre	-2.08
St Anns	-1.85
Forest Fields	-1.00
Strelley	-0.87
Bulwell	-0.71
Clifton	-0.65
Bestwood Village	-0.64
Thorneywood	-0.60
Top Valley	-0.42
Broxtowe	-0.41
Bestwood Park	-0.40
The Meadows	-0.38
Arboretum	-0.36
Carrington	-0.36
...	...
Trowell	0.30
Gamston	0.31
Alexandra Park	0.37
Beeston	0.45
Edwalton	0.45
Hyson Green	0.54
Woodthorpe	0.58
Stapleford	0.64
Daybrook	0.71
Gedling	0.76
The Lace Market	0.77
Eastwood	0.80
Carlton	0.90
Hucknall	1.25
West Bridgford	1.32

generally lower levels of postal addresses appearing on the web generally - irrespective of the number of references to urban locales. Although it is outside the scope of this work to further explore this, it is especially interesting that it appears that addresses appearing on the web referring to Hyson Green are highly likely to include a reference to the urban locale within them, whilst in contrast, similar areas (in terms of geography and deprivation) tend not to.

Appendix I

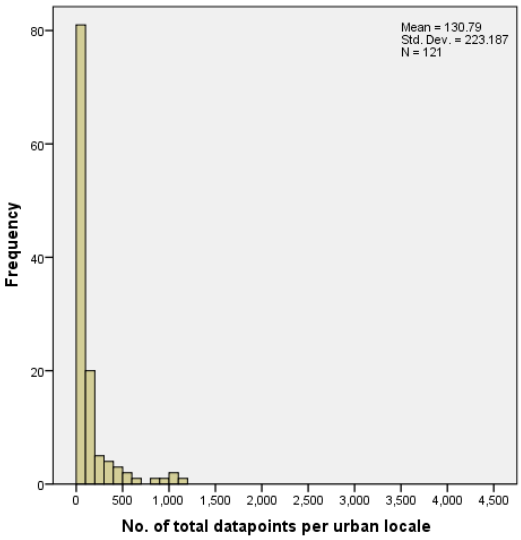
Relationship with Flickr

This appendix compares data volumes of urban locales extracted from the photograph site Flickr with the web extraction output and builds upon the 'Data volumes' section within Chapter 8. It uses a case example of the city of Sheffield to compare the number of data records extracted using place names found within Flickr tags of geotagged Flickr photos in September 2014. The aim of the comparison is to demonstrate if the web extraction method can provide sufficient data volumes concerning urban locale locations.

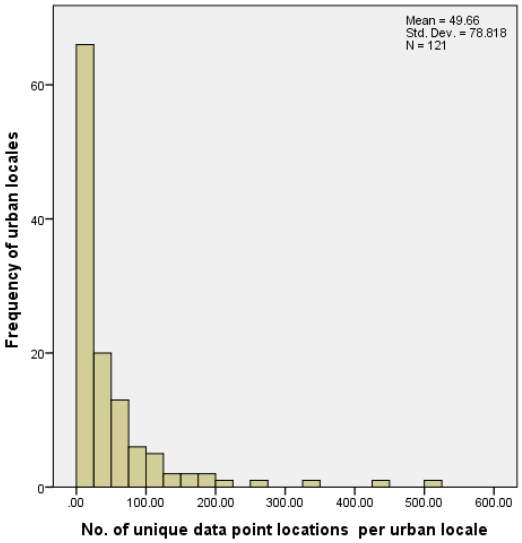
The web extraction method on average produced a higher number of data returns than from Flickr photos for the case study area of Sheffield. The median number of data returns from Flickr photos for Sheffield urban locales was 41 (mean of 131), compared to 213 from the web extraction output (mean of 520). It is, however, also important to recognise the potential spatial footprint of data returns. A good spread of data points within the extent of the urban locales is desirable to ensure complete coverage. The web extraction output are tied to postcodes and so data points will by default consist of fewer unique sets of coordinates than Flickr photographs where georeferencing is not constrained in such a manner. This said, whilst the web extraction output contained a higher degree of spatial repeats, it continued to produce a greater number of unique data points for mapping with a median of 43 different data point locations per urban locale (mean of 59) compared to only 19 from Flickr (mean of 50). Figure I.1 demonstrates the skewed nature of total data records from the Flickr data - with a high number of urban locales containing relatively few data points.

Figure I.2 further explores the differences in data volumes between the two methods. In relation to total data records (Figure I.2a), the web extraction method produces higher than expected numbers of data points compared to Flickr in locales on the edge of the city. These principally include areas of Dependent Localities (as first

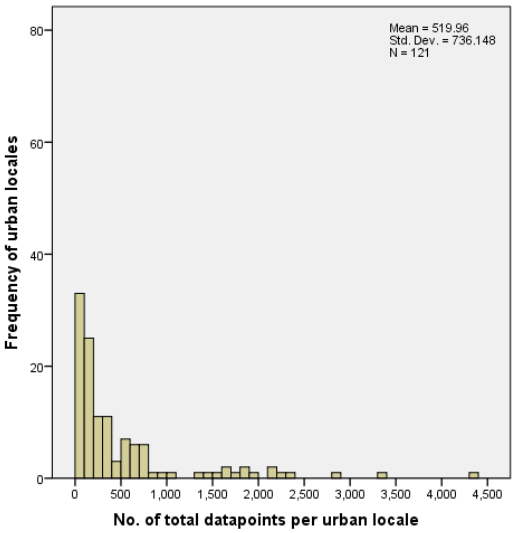
Fig. I.1 Data distributions for urban locale names using Flickr API data and web extraction output



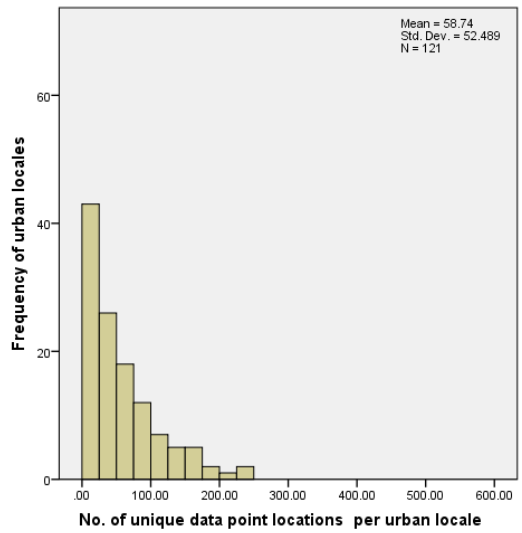
a) Flickr: Total data points



b) Flickr: Unique data locations



c) Web extraction output: Total data points



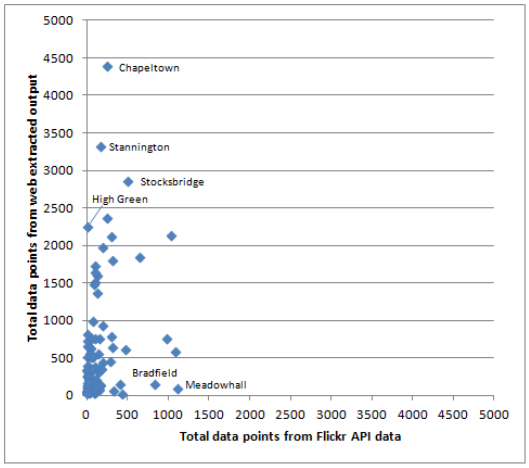
d) Web extraction output: Unique data locations

discussed in Chapter 4 and analysed in Chapter 7 and Appendix H. In contrast, Flickr contained higher data numbers in areas containing tourism and leisure activities - such as Meadowhall (Meadowhall shopping centre) and Bradfield (a toursity village in the Peak District[1]). The main difference when comparing the number of unique spatial data locations (Figure I.2b) is that the data volumes concerned are more even. A similar pattern, however, remains - with the web extraction output offering high data quantities for more peripheral locales, whilst Flickr data is more prominent in areas of tourism and leisure activities. The difference in the number of data records for the urban locale of Tinsley is pronounced, with its many photos of trains at the train

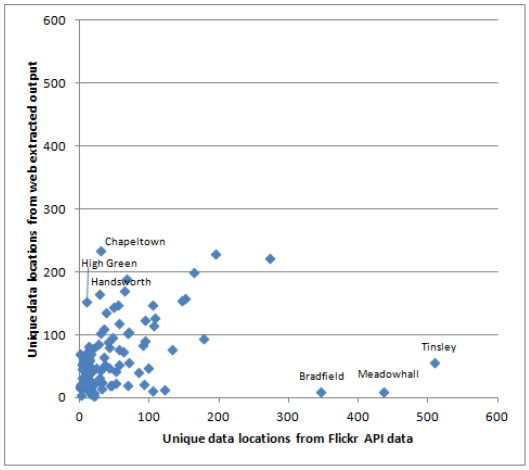
[1] Additionally, 16% of photos (n=134) made reference to the Tour de France in either the title or tags. The Tour de France passed through Bradfield earlier in the summer

depot and abandoned buildings awaiting demolition - both sources that would not be expected to appear within the web extraction output. The potential of combing data from the two sources is discussed within Chapter 8.

Fig. I.2 Comparison of data returns - Flickr and web extraction output



a) Total data points



b) Unique data locations

Whilst it would be possible to compare *geographies* it is beyond the scope of this work which is specifically concerned with data volumes produced by both the web extraction method and Flickr data.