Mapping Urban Neighbourhoods from Internet Derived Data

Paul Brindley[1], James Goulding[1], Max L. Wilson[1]

[1]The University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB
Tel. +44 (0)115 823 2316 Fax +44 (0)115 8232518
psxpb2@nottingham.ac.uk
http://www.nottingham.ac.uk

1. Introduction

Neighbourhoods are geographical units to which people connect and identify with. They represent a key element within government agenda (in 2010 the Secretary of State for Communities and Local Government described them as "the building blocks of public services society") as well as holding wide social utility, with Sampson (2012) arguing that the ways that neighbourhoods affect our lives often go unrecognized. Whilst much research emphasises the merits of neighbourhood units, the subjective nature of the concept and the resulting difficulties in collecting data, means that there are, as of yet, no officially defined neighbourhoods in the UK either in terms of names or boundaries. In this paper we seek to address this situation via large-scale web mining.

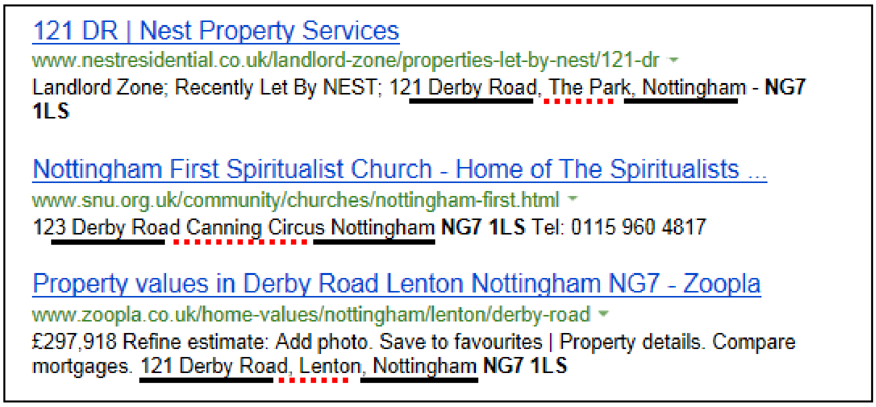
In recent years, research in vernacular geography has grown steadily, along with interest in the mapping of geographic objects via internet data. However, research has not yet focussed on 1. the scalability of such processes up to a national coverage; 2. the identification of concepts such as neighbourhoods when names are unknown *a priori*. In this light, we present a new approach that is able to automatically map neighbourhood extents without prior knowledge via the passive mining of postal address information published on the internet.

The picture of neighbourhoods is especially problematic within our urban areas, where boundaries are frequently less clear and more open to subjective interpretation. From a bureaucratic perspective, it is easiest to perceive urban neighbourhoods as discrete, distinct objects that nest within current administrative boundaries. However, administrative boundaries do not necessarily fit with neighbourhood extents even with the same name (Twaroch *et al*, 2009a; Flemmings, 2010). There has been a long tradition of geographers mapping neighbourhood areas, for example the work in 1960s in America by Lynch (1960). Such research is a strong influence on more recent studies for deriving vernacular place names (Bentley *et al*, 2012; Montello *et al*, 2003; Orford and Leigh, 2011). Although these works have great merits, they usually utilise a small number of selected city case studies due to the nature of data collection (for example through asking residents to draw perceived neighbourhoods). Hence, methods like this are usually both time-consuming and not viable for a wider scale such as a national coverage. We, therefore, present an automated solution to this issue.

The notion of mining the web for geospatial features is not new, although prior research has been concerned with larger spatial entities appearing in existing gazetteers (Jones *et al*, 2008; Purves *et al*, 2007; Purves *et al*, 2005). Research to explicitly identify neighbourhoods has traditionally relied on knowing the required neighbourhood names *a priori* – either through searching resources such as *Gumtree* (Twaroch *et al*, 2009b), using *Flickr* photos (Grothe and Schaab, 2008; Hollenstein and Purves, 2010) or using specific search terms within search engines (Flemmings, 2010; Jones *et al*, 2008). This is a potential weakness in method due to the fact that there is no single comprehensive dataset that contains neighbourhood names, and there are many inconsistencies between the various existing gazetteers[1]. A benefit of the method we propose is that such ambiguities can be reduced due to the tight delimitation of the neighbourhood terms it produces between an expected road and settlement name. This is not the case when undertaking a simple trigger string approach (for example by entering "Clifton in Bristol" or "* is in Clifton") such as that employed by Jones *et al* (2008) or Flemmings (2010). For example, within the work of Flemmings (2010) it was not possible to separate those references to Bedminister Down, Bristol from those relating to Bedminister, Bristol – skewing the geographical representations. We therefore extend existing work by using a technique that is dependent upon extracting neighbourhood terms from within structured postal addresses held on the internet.

2. Methodology

Our methodology is based on the fact that even though official Royal Mail addresses do not require such information to be entered between street and city elements, in web content, publishers often provide neighbourhood names alongside corresponding postcodes (as shown in Figure 1). Thus, an exhaustive automated web search[2] followed by linguistic parsing of content allows us to generate neighbourhood names. Subsequent mapping of data is possible through geo-referencing the postcodes.



121 DR | Nest Property Services
www.nestresidential.co.uk/landlord-zone/properties-let-by-nest/121-dr
Landlord Zone; Recently Let By NEST; 121 Derby Road, The Park, Nottingham - **NG7 1LS**

Nottingham First Spiritualist Church - Home of The Spiritualists ...
www.snu.org.uk/community/churches/nottingham-first.html
123 Derby Road Canning Circus Nottingham **NG7 1LS** Tel: 0115 960 4817

Property values in Derby Road Lenton Nottingham NG7 - Zoopla
www.zoopla.co.uk/home-values/nottingham/lenton/derby-road
£297,918 Refine estimate: Add photo. Save to favourites | Property details. Compare mortgages. 121 Derby Road, Lenton, Nottingham **NG7 1LS**

Figure 1. Identifying neighbourhood terms in between street and settlement content

Regular expression matching was used to identify street information by first looking for an official street (as recorded by Royal Mail's Postal Address File) – including common abbreviations (such as "rd" for "road"), before identifying non-official road names (a term followed by common street suffixes). It was then possible to extract out the text between the road and settlement name within the address. A distance decay effect was additionally

[1] As discovered in the report *Investigating the feasibility of Natural Language Spatial Querying for Sub-Settlement areas* (http://www.nottingham.ac.uk/~psxpb2/phd.html), which found that existing neighbourhood level gazetteers (including Open Street Map, Yahoo Places, Geonames and Ordnance Survey data) were incomplete (with several substantial towns and cities containing no neighbourhood level data) and inconsistent with each other (Open Street Map and Yahoo Places only agreed on neighbourhood names for 14% of data).

[2] Using the Bing Search API, with automation achieved via Python.

exploited for the matching likelihood, so that near things were more likely to match with a close spelling than more distant objects. This resulted in a set of geospatial points (identified via postcode geo-referencing) labelled with neighbourhood name occurrences.

Unlike existing work in this area which at this point present resulting neighbourhoods via a *kernel density estimate* (kde) for each neighbourhood (such as Jones *et al*, 2008; Twaroch *et al*, 2009b; Flemmings 2010), our work went on to construct a series of further computational rules. The rules use a combination of absolute and percentage of data at a variety of density scales in order to identify neighbourhoods that were either substantial (in terms of the number of data returns: *[density smoothed to 300m]>75 and [density smoothed to 1600m]>= 15*) or locally significant (a high percentage but also sustained over a specific area: *[cells with 70%+ agreement] >= 40 and [density smoothed to 300m] >75*).

3. Experimental Results

In order to test the efficacy of our approach the urban area of Sheffield was selected, an area covering 18,244 postcodes (including historic postcodes). A case study area approach of the city of Sheffield was undertaken due to the unique resource of neighbourhood area identification undertaken by Sheffield Council which would enable comparison between the internet derived output and 100 council defined neighbourhood named areas.

288,071 data returns were obtained from the 18,244 postcodes for the urban area of Sheffield. 84,245 of these records contained additional information between the street name and settlement, which were used to identify 121 different neighbourhoods (see Table 1). A small number of examples of the grids produced by the automated procedures can be found in Figure 2. The method also allows for the investigation of those areas where perceptions vary and different people or organisations may call the area by different names. Figure 3 illustrates an example of differing views of three neighbourhoods within Sheffield.

Preliminary investigation of the domain names from data records shows that the majority of information derives from business directory sources (56% of records). Other classifications of data include estate agents (18% of records), company reports (such as Companies House; 14%) and restaurant/pub guides (6%). Ongoing work is investigating the geographies of neighbourhoods that these different groupings may produce.

3.1 Validation

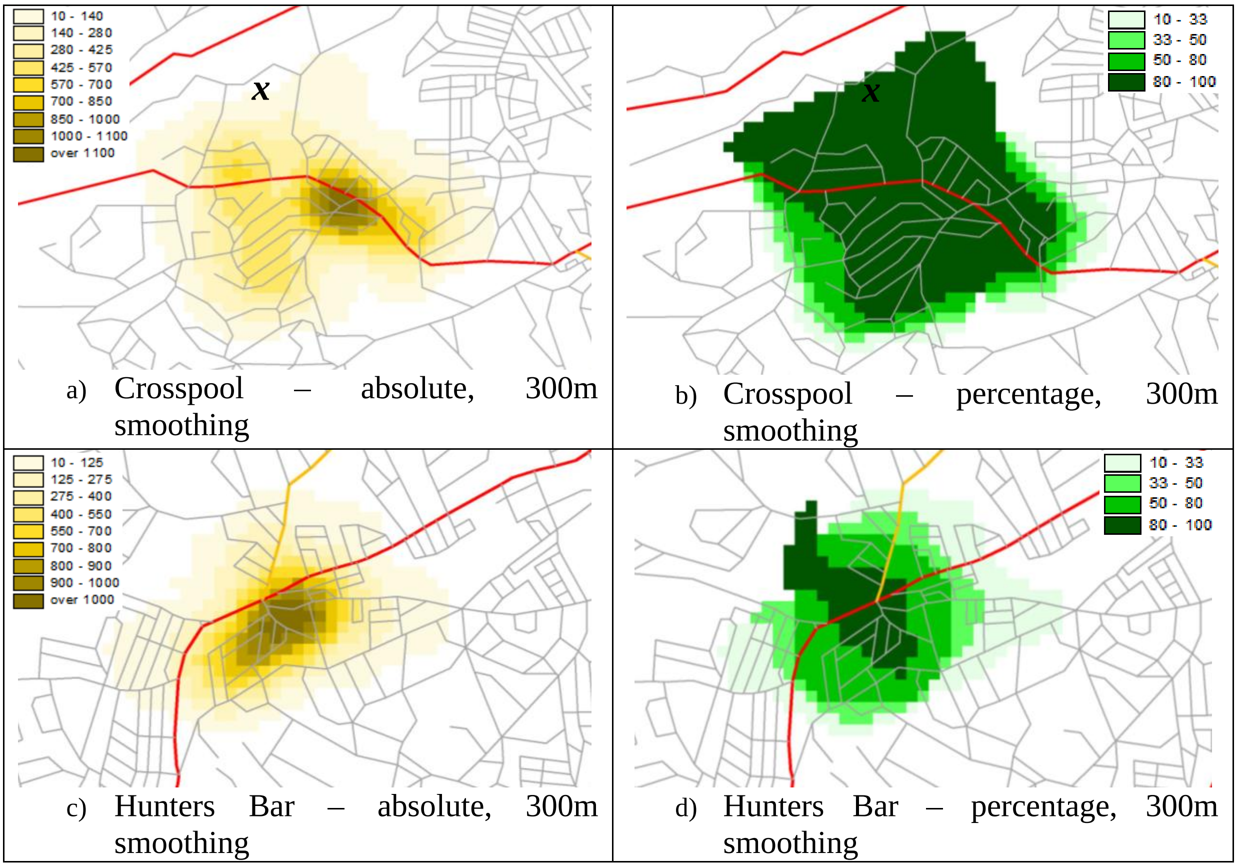
Validation of the *geography* of the neighbourhoods is problematic as there is no true representation to test against (existing gazetteers such as Geonames record the neighbourhoods as single centre points). In order to address this, work is therefore ongoing to survey residents' actual definitions of the neighbourhoods in which they live (which will necessarily be a sample, as to provide full coverage via data collection in this manner is not logistically tractable). The similarity between outputs from our technique and local council defined neighbourhoods is encouraging (an example of which is shown Figure 4), with 73% agreement in geography where neighbourhood names were the same. Complete agreement would be unlikely as the council defined areas do not allow overlap and are constrained by the geography of administrative zones (Output Areas).

Validation of the names of the neighbourhoods is more straightforward. However, given the poor level of agreement between existing neighbourhood gazetteers[3] – perfect agreement is

[3] The five sources (Sheffield Council defined, Open Street Map, Yahoo Places, Geonames; Ordnance Survey

unlikely. However, of the 121 neighbourhoods defined by the proposed method – 106 (88%) were also found in at least three of the existing sources of Sheffield neighbourhoods (Sheffield Council defined, Open Street Map, Yahoo Places, Geonames and Ordnance Survey (OS) data (a combination of the 50k gazetteer, VectorMap Local and data extracted from the Multi Resolution Data Programme). Some of the areas that were not found in the existing gazetteer sources of neighbourhoods were well known Sheffield neighbourhoods such as Hunters Bar, Shalesmoor or Kelham Island – further demonstrating the incompleteness of existing sources of neighbourhoods within current gazetteers.

A full list of the names of neighbourhoods not supported by evidence from the existing sources of neighbourhoods can be found in Table 2. There were 8 neighbourhoods that were found in at least three of the existing sources of neighbourhoods but that were not identified by extracting terms from postal addresses held on the internet – these can be found in Table 3.



a) Crosspool – absolute, 300m smoothing

b) Crosspool – percentage, 300m smoothing

c) Hunters Bar – absolute, 300m smoothing

d) Hunters Bar – percentage, 300m smoothing

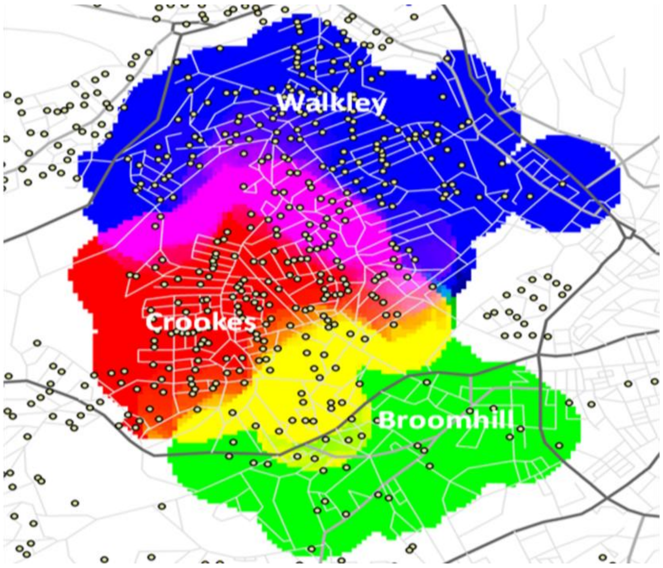
Figure 2. Examples of neighbourhoods from postal address information on the internet *(Contains Ordnance Survey data © Crown copyright and database right 2014)*

data) produced a total of 276 different neighbourhoods. There was agreement between all five sources in only 6% of neighbourhoods. A further 8% of neighbourhoods were found in at least four of the sources. In 52% of cases – the neighbourhood name was only found in a single source.

Table 1. Sheffield neighbourhoods extracted from postal address information on the internet

Name	Frequency	Name cont.	Freq. cont.	Name cont.	Freq. cont.
Chapeltown	4,498	Rotherham	457	Neepsend	126
Stannington	3,390	Wharncliffe Side	423	Banner Cross	118
Stocksbridge	2,942	Wadsley	404	Meadowhall	118
Ecclesfield	2,498	Greenhill	403	Whirlow	115
Hillsborough	2,323	Wincobank	398	Brincliffe	114
High Green	2,295	Beauchief	380	Dungworth	111
Deepcar	2,176	Norton Lees	358	Woodhouse Mill	110
Beighton	2,028	Firth Park	351	Grimesthorpe	109
Woodhouse	1,888	Parson Cross	349	Firvale	108
Grenoside	1,887	Holbrook Ind. Estate	347	Shalesmoor	101
Dore	1,773			Southey Green	99
Mosborough	1,689	Intake	343	Greystones	90
Handsworth	1,662	Meersbrook	329	Abbeydale	86
Halfway	1,532	Shiregreen	324	Manor Top	81
Oughtibridge	1,518	Hunters Bar	323	Parkwood Springs	80
Norton	1,490	Ranmoor	283	Bents Green	78
Walkley	1,428	Frecheville	278	Foxhill	78
Darnall	1,051	Crystal Peaks	277	Lane Top	76
Loxley	1,045	Wadsley Bridge	277	Newhall	75
Totley	829	Bradfield	275	Longley	74
Waterthorpe	828	Totley Rise	266	Brightside	71
Crookes	827	Meadowhall Centre	228	Gleadless Valley	70
Broomhill	820	Lodge Moor	224	Jordanthorpe	65
Nether Edge	807	Malin Bridge	222	Birley	63
Woodseats	798	Pitsmoor	210	Meadowhead	62
Sothall	757	Crookesmoor	201	Manor	61
City Centre	750	Holbrook	185	Parkway Ind. Estate	61
Ecclesall	708	Upperthorpe	177	Sharrow Vale	61
Owlthorpe	683	Wisewood	174	Wadsley Park Village	61
Hackenthorpe	677	Broomhall	168	Woodthorpe	61
Bradway	674	Sharrow	166	Birley Carr	60
Fulwood	648	Lowedges	161	Carterknowle	55
Tinsley	629	Norfolk Park	156	Netherthorpe	55
Gleadless	623	Richmond	156	Middlewood	52
Millhouses	596	Nethergreen	149	Wybourn	52
Westfield	585	Burngreave	143	Kelham Island	51
Heeley	575	Arbourthorne	141	Basegreen	48
Crosspool	542	Bolsterstone	141	Manor Park	44
Attercliffe	533	Highfield	140	Thorncliffe Park, Chapeltown	44
Worrall	533	Shirecliffe	137		
Burncross	531	Sandygate	135	Charnock	41

Contains Ordnance Survey data © Crown copyright and database right 2014



RGB COMPOSITE:

Blue: Walkley 'neighbourhood'
Red: Crookes 'neighbourhood'
Green: Broomhill 'neighbourhood'

A mix of colours represents contested spaces eg purple is both Walkely & Crookes.

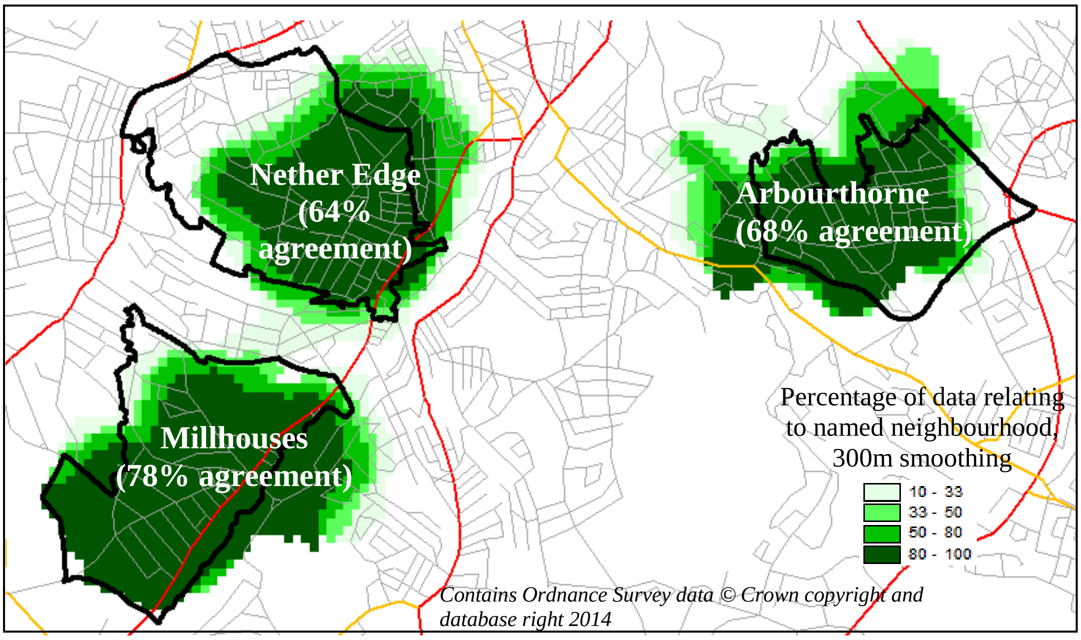
Figure 3. Differing perceptions of neighbourhood

Table 2. Neighbourhoods identified by our method which were not included in at least three of the other neighbourhood gazetteers

Name	Frequency	Found in OS data	Found in OSM data
Birley Carr	60	Yes	No
Bolsterstone	141	Yes	Yes but as 'village' class
Dungworth	111	Yes	Yes but as 'village' class
Holbrook	185	Yes	No
Holbrook Industrial Estate	347	No	No
Hunters Bar	323	No	Yes but as 'locality' class
Kelham Island	51	Yes	Yes but as 'island' class
Meadowhall Centre	228	Yes	Yes but as 'mall' class
Newhall	75	No	No
Parkway Industrial Estate	61	No	No
Parkwood Springs	80	Yes	No
Rotherham	457	Yes	Yes
Shalesmoor	101	Yes	No
Thorncliffe Park, Chapeltown	44	No	Yes but as 'commercial' class
Wadsley Park Village	61	No	No

Table 3. Neighbourhoods in at least three of the other gazetteers but not identified by our method

Name	Frequency
Batemoor	21
Carbrook	36
Greenland	20
Hemsworth	16
Herdings	29
Hollins End	10
Lowfield	11
Park Hill	25



Contains Ordnance Survey data © Crown copyright and database right 2014

Figure 4. Comparison of selected internet derived neighbourhoods and Sheffield Council neighbourhood boundaries

4. Discussion and Conclusions

This work demonstrates the feasibility of using postal addresses held on the internet in order to identify neighbourhood names and subsequently map them. Not only is the proposed methodology useful for aiding the comprehensiveness of neighbourhood level gazetteers but the approach does not require that the names of the units of interest are known *a priori*. The use of the *percentage of data returns* that relate to a named neighbourhood (in addition to the usual absolute number of data returns) is a useful indicator of neighbourhood locations when data volumes may be expected to be relatively low (on the edge of cities, by rivers or next to large parks for example). Whilst the absolute data maps (Figures 2a and 2c) show concentrations focused around the shopping areas of the two neighbourhoods, the percentage maps (Figures 2b and 2d) for the corresponding areas demonstrate quite a different geography, identifying where the majority of people may identify with the neighbourhood names. For example, compare the differences for the area to the north of Crosspool labelled with an *x* in Figures 2a and 2b where data volumes may be low but there is total agreement that the area is defined as ‘Crosspool’.

Of those neighbourhoods not identified by the proposed method (Table 2), only four names do not appear in either OS data or OSM as a different class of object. These include: Newhall, Wadsley Park Village, Holbrook Industrial Estate and Parkway Industrial Estate. There are references within old OS maps in the area outlined as 'Newhall' by our proposed web extraction method (and not on Newhall Road itself) that identify with the name 'Newhall', for example 'Newhall County School' in 1950s and 1960s maps. 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, they do have their own community website and forum (http://www.wpvonline.co.uk).

It is unclear to what extent industrial areas should be identified as 'neighbourhoods'. Our web extraction method takes no account of differences between residential and non-residential areas, merely of named areas. Reference to an 'industrial estate' can be found within Holbrook in OS maps from the 1980s onwards, however not since the 1970s was it explicitly mentioned as 'Holbrook Industrial Estate'. Similarity, there is no reference 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.

There were eight neighbourhoods that were found in at least three of the existing sources of neighbourhoods but that were not identified by extracting terms from postal addresses held on the internet – these can be found in Table 3. Whilst all of these were to be found within the data output (the frequency column within Table 3) they had low levels of returns and were therefore not considered robust enough to be included in our further analysis. The rationale was that any density map produced using such small numbers may not be meaningful.

Figure 3 demonstrates the large areas where people within the same space may have different opinions of what the area may be called. This fits well with the need to identify probabilistic perceptions of neighbourhood, where we all have our own different views of the areal extent of neighbourhood areas and it is only when taken collectively that sense can be extracted (for example this is where the majority of people believe *x* might be). Thus, there are rarely clear cut boundaries but fuzzy, probabilistic views containing differing perspectives. Whilst further work on the validation of the geography of such areas is required, there is a substantial potential of undertaking such a fully automated procedure to the field of urban geography.

A number of improvements to the proposed method have been identified. These include:

- the use of hardlines to stop the smoothing from crossing specific line features (such as rivers, railways, duel carriageways, motorways and so on) if no evidence of the neighbourhood can be found on both sides of the linear feature;
- the use of existing gazetteers to provide evidence of neighbourhood names so that areas with lower data volumes of returns are not excluded (such as from Table 3);
- a further pass of data to identify neighbourhood references within business names to further increase data volumes of returns.

This research has demonstrated the feasibility of extracting urban neighbourhood names from structured postal addresses appearing within internet data and their subsequent mapping. It has shown that passive mining of postal address information published on the internet can be used to automatically map neighbourhood extents without prior knowledge of the neighbourhood names themselves.

5. References

BENTLEY, F., CRAMER, H., HAMILTON, W., BASAPUR, S. (2012) Drawing the city: differing perceptions of the urban environment. Presentation at CHI 2012, May 9th 2012.

FLEMMINGS, R. (2010) Revealing the Fuzzy Geography of an Urban Locality. GISRUK 2010 proceedings, pp 345-351.

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. International Workshop on Computational Models of Place (PLACE'08). Park City: Utah, USA.

HOLLENSTEIN, L. AND PURVES, P. (2010) Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science* **1**: 21-48.

JONES, C.B., PURVES, R.S., CLOUGH, P.D. AND JOHO, H. (2008) Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science* **22** (10): 1045-1065.

LYNCH, K. (1960). *The Image of the City*. The M.I.T. Press: Boston.

MONTELLO, D.R., GOODCHILD, M.F., GOTTSEGEN, J. AND FOHL, P. (2003) Where's downtown? Behavioural methods for determining referents of vague spatial queries. *Spatial Cognition and Computing* **3**: 185-204.

ORFORD, S. AND LEIGH, C. (2011) Where to draw the line? Mapping perceived neighbourhoods onto Lower Super Output Areas. GISRUK 2011 proceedings, pp 350-357.

PICKLES, E. (2010) Eric Pickles' speech to the Local Government Association annual conference - 7 July 2010.

PURVES, R., CLOUGH, P. D., AND JOHO, H. (2005) Identifying imprecise regions for geographic information retrieval using the Web. GISRUK 2005 proceedings pp. 313–318

PURVES R.S., CLOUGH P., JONES C.B., ARAMPATZIS A., BUCHER B., FINCH D., FU G., JOHO H., KHIRINI A.S., 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 Geographic Information Systems* **21**(7): 717-745.

SAMPSON, R.J. (2012) *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press: Chicago.

TWAROCH, F., JONES, C.B. AND ABDELMOTY, A.I. (2009a) Acquisition of vernacular place names from web sources in R. Baeza-Yates and I. King (eds) *Weaving services and people on the world-wide-web*. Springer: Berlin.

TWAROCH, F. A., PURVES, R. S., & JONES, C. B. (2009b). Stability of Qualitative Spatial Relations between Vernacular Regions Mined from Web Data. Proceedings of Workshop on Geographic Information on the Internet, Toulouse, France, April 6th, 2009.

6. Acknowledgements

The primary author is supported by the Horizon Doctoral Training Centre 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).

We would also like to thank Ordnance Survey Research Department for the supply of data to aid the validation process undertaken.

Biography

Paul Brindley *is a second year PhD student within the Horizon Doctoral Training Centre at The University of Nottingham. His research interests include probabilistic classification and mapping of geographic objects derived using the internet as a datasource.*

Dr James Goulding *is a research fellow at Horizon Digital Economy Research. His research focuses on novel analyses of behavioural data via periodicity analysis, motif extraction and subspace clustering and he has won the ACM Englebart prize for work in data theory and the Centre for DE prize for Data Visualization.*

Dr Max L. Wilson *is a lecturer at the University of Nottingham. Max's research focuses on the interplay between Human-Computer Interaction and Information Retrieval (HCIR), with a recent interest in Social Media. His publications include a book on Search User Interface Design.*