A data driven approach to mapping urban neighbourhoods

Paul Brindley
Horizon Centre for Doctoral Training,
University of Nottingham
Jubilee Campus, NG8 1BB
+44 (0)115 823 2316
psxpb2@nott.ac.uk

James Goulding
Horizon Digital Economy Research,
University of Nottingham
Jubilee Campus, NG8 1BB
+44 (0)115 823 2557
james.goulding@nott.ac.uk

Max L. Wilson
School of Computer Science,
University of Nottingham
Jubilee Campus, NG8 1BB
+44 (0)115 846 6551
max.wilson@nott.ac.uk

ABSTRACT
Neighbourhoods have been described by the UK Secretary of State for Communities and Local Government as the "building blocks of public service society". Despite this, difficulties in data collection combined with the concept's subjective nature have left most countries lacking official neighbourhood definitions. This issue has implications not only for policy, but for the field of computational social science as a whole (with many studies being forced to use administrative units as proxies despite the fact that these bear little connection to resident perceptions of social boundaries). In this paper we illustrate that the mass linguistic datasets now available on the internet need only be combined with relatively simple linguistic computational models to produce definitions that are not only probabilistic and dynamic, but do not require *a priori* knowledge of neighbourhood names.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications - *Spatial databases and GIS, Data mining*.

Keywords
Neighbourhoods; vernacular geography; spatial data mining.

1. INTRODUCTION
The subjective nature of *neighbourhoods* (and ensuing difficulties in collecting data about them) has meant that official definitions of neighbourhoods simply don't exist in many countries. This is despite their utility to a host of social applications, ranging from criminology to epidemiology. Even in the UK, where some gazetteers provide single geospatial points of reference for neighbourhoods, data are often incomplete and rarely agree [2]. In the absence of neighbourhood boundary definitions, many social studies have therefore had to resort to using official administrative units (formal areas used by local and national government for statistical purposes) as proxies for neighbourhoods - despite the fact that these do not correspond to resident perceptions of boundaries, even when they have the same name [9, 12].

SIGSPATIAL '14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

ACM 978-1-4503-3131-9/14/11…$15.00
http://dx.doi.org/10.1145/2666310.2666473

Recent research has demonstrated the potential to extract geographical extents from "Big Data" [4, 6, 9, 10, 11], although previous methods have been highly reliant on gazetteers. In this paper, we show that the large-scale linguistic corpora that are now freely available via the Web need only be combined with relatively simple computational models to automatically generate neighbourhood definitions that are: *probabilistic*, *dynamic* and, uniquely, require *no a priori knowledge* of neighbourhood names.

In order to achieve this we perform harvesting of postal addresses online, allowing us to computationally co-locate *postcodes*[1] with *neighbourhood names* that appear between street and city address elements. Using the geographical locations of postcodes we can then apply kernel density estimates, therefore converting extracted name frequencies into continuous, probabilistic spatial extent surfaces for each neighbourhood. Because such boundaries may be generated dynamically, this opens up the opportunity to examine how neighbourhood makeup changes over time.

2. BACKGROUND
Prior studies of neighbourhood extents and vernacular place names [1, 7] have highlighted the need for geographies which are *fuzzy*, *overlapping*, and embrace the varying perspectives held by different stakeholders. Unfortunately, such features have been broadly unrealised, with proposed methods being unsustainably time consuming, with research being restricted to small case studies and preclude national coverage.

In contrast, the growing availability of mass datasets on the internet has increased research interest in the mining of geospatial features from Web corpora. Prior research has, however, focussed on either specific geographical objects (e.g. landmarks) or acted at larger geographical scales than neighbourhoods (e.g. country districts) [6, 8]. In parallel there is a burgeoning body of research aimed at automatically defining spatial extents by clustering georeferenced social media data, such as that generated by geo-tagged "tweets" or *Foursquare* "check-ins" [3, 13]. Such methods show strength in isolating *functional areas*, but rarely create complete neighbourhood coverage for a city and, understandably, do not generate associated names for the units generated, despite the fact that it is these names that people identify with.

An alternative to unsupervised clustering-based approaches is to extract boundary data from the Web based on neighbourhood names listed in a gazetteer or equivalent directory (e.g. via business listings in Yahoo Local [9], Gumtree [10, 11], Flickr photos [4, 5] or specific search terms within search engines [6]). Once point data corresponding to a specific gazetteer name has been identified, kernel density smoothing can be applied in order

[1] Postcodes (ZIP codes in the US) are combinations of alphanumeric characters used within postal addresses for the purpose of sorting mail.

to produce a continuous surface area, with cut-off thresholds used to remove noise. Unfortunately, these gazetteers seldom provide a definitive source of neighbourhood names. In the UK, for example, there is only a 6% agreement across Open Street Map (OSM), Yahoo Places, Geonames, Sheffield City Council, and Ordnance Survey (OS - the National Mapping Agency of Great Britain), as to the correct neighbourhoods for the city of Sheffield [2]. Furthermore, this approach 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 [5, 11].

Cross-referencing methods can also suffer from issues of *place ambiguity*. For example, a web query amassing documents for the neighbourhood "Manor, Sheffield, UK" not only returns data pertaining to the place itself, but also information pertaining to other, subjectively-related geographical areas 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).

3. EXPERIMENTAL METHODOLOGY

We now present the methodology underpinning our approach, along with our validation process. To provide focus for our experiments, and to allow validation, we have focussed on UK neighbourhoods. Probabilistic neighbourhood areas are generated via the following steps: 1. *Data collection*; 2. *Data Cleansing / Synonym identification*; 3. *Neighbourhood Derivation*.

The first requirement of the method is to extract postal addresses from mass text corpora that reveal common usage (in our case this aggregated corpus is the Web, but this could easily be extended to include many more forms of data source). Within the UK, an official postal address within an urban area is defined as: *building number and street; city; postcode*[2], and it is this format that is mined for. Our 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 neighbourhood names within that structure.

To determine what to search for, we first obtain a set of postcodes via *Ordnance Survey's Code-Point Open* dataset. These are automatically iterated through the *Bing API*, with relatively simple linguistic pattern matching techniques applied to returned results so as to: (1) extract postal addresses from each document; (2) apply rule based filters to identify street and city names (either via identification in Royal Mail's Postal Address lists, or through detection of common abbreviations (e.g. "*rd"* or "*st"*); (3) illicit the text between these two entities, pairing the resulting *neighbourhood candidate names* with the current postcode to produce a final set of <*candidate name, postcode*> pairs.

3.1 Data cleansing / Synonym identification

All postcodes associated with the same candidate name are then collated and geocoded in order to produce a set of <*candidate name, point-set*> pairs. Eradicating noise from these intermediate results set is key - the nature of the raw data means synonyms can be frequent, with slight variations of spelling referring to the same entity. To address this, a data cleansing process is applied which identifies synonyms via an exhaustive comparison of all name candidates using *Ratcliff/Obershelp* and *Levenshtein Distances* (which return a score between 0-1). Cases where two candidates return a similarity score greater than a threshold, θ, are deemed synonyms, and their associated postcode sets merged. θ is calculated incorporating a distance decay effect that increases the likelihood that synonyms will be merged, the closer the geocoded point-sets are together. This is done via the formula: *0.8 + (0.2 × Euclidian distance)* between the candidates' centroids. For example, if the centroids of *Broomhill* and *Broomhall* were only 10m apart, a threshold of 0.802 would be required (0.8+(0.2×0.010)) to merge them. If however, they were 800m apart this would rise to 0.96 (0.8+(0.2×0.800)). A minimum frequency of 30 data points was required to ensure mapped output would not be dependent upon a small number of sample points.

3.2 Neighbourhood Derivation

In order to construct sets of neighbourhood surfaces, spatial *kernel density estimates* (KDE) [6, 11] were applied to the surviving set of <*candidate name, point-set*> pairs. The quadratic KDE measures used here rely on a cell size of 50m and utilise two different search radii: 300m (as commonly used within the literature [10, 12]) and 1600m (to gauge overall mass at a larger geographic scale). Unlike previous works which at this point produce a single KDE surface of point data for each neighbourhood, we constructed a number of KDE based computational rules to further reduce noise. A KDE with a 300m search radius was produced in those cells that passed one of the following conditions (providing there were >= 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 percentage and sustained over 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 (especially at the edges) and to help compensate for this a 2% cut-off is applied to 'trim the edges' [12].

3.3 Validation

Validation is not only problematic due to the subjective nature of neighbourhood geographies, but also due to the fact that their efficacy can only be accurately assessed in reference to some specific end application. In aiming to derive neighbourhood extents that have generalized utility we must therefore attempt to compare outputs with some form of "ground truth" representation - and as noted in §2, these are extremely difficult to come by.

We have been able to isolate two urban areas (Nottingham and Sheffield) where it was possible to source detailed quantitative as well as qualitative knowledge of neighbourhood extents. This was notable for the city of Sheffield, UK, for which it was possible to obtain a unique resource of *bounded* neighbourhood areas derived by Sheffield City Council through an exhaustive surveying process. This enables us to perform comparisons between our results and 100 council defined neighbourhoods. In addition to this, derived neighbourhood *names* were assessed against a composite picture derived from multiple sources (City Council definitions, OSM, Yahoo Places, Geonames and OS data). Given that these differing sources are known to be incomplete and

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

derived in a relatively ad-hoc fashion, the degree to which our method can derive a superset of their contents is also of interest.

4. RESULTS

We performed detailed examinations of the neighbourhoods generated for the two case study cities - Nottingham and Sheffield - containing 18,946 and 18,244 postcodes respectively (including historic postcodes). This produced 312,760 document returns for Nottingham and 288,071 for Sheffield. Of these records, 139,472 and 84,245 respectively contained additional information between the street name and settlement, which were used to identify 126 neighbourhoods in Nottingham and 121 in Sheffield. For purposes of illustration, four examples of the grids produced by the automated procedures are shown in Figure 1a-d.

Importantly, our method also allows for investigation of areas where perceptions vary and different people or organisations refer to the area by different names. Figure 2 illustrates an example of differing views of three Nottingham neighbourhoods. Preliminary manual investigation of domain names from Sheffield data shows that the majority of information harvested was derived 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.

4.1 Validation of Generated Extents

Due to the boolean nature of the ground-truth data (defined by local councils) it was impossible to draw direct comparisons with the probabilistic values from postal web defined neighbourhoods. Therefore, boundary data produced from both methods were converted to 50m boolean grids and agreement represented the number of grid cells where both sources carried the same name. The similarity between our technique and those from ground-truth areas were very encouraging. At a 50m grid level, we produced an overall 73% agreement, examples of which are shown Figure 3. It is worth noting that complete agreement is impossible due to council defined areas precluding overlapping of neighbourhoods, and being constrained topographically to Census Output Areas.

When comparing neighbourhood *names* (rather than boundaries) with the portfolio of existing gazetteers - it was found that of the 121 neighbourhoods defined by our method, 106 (88%) were also found in at least three of the existing sources of Sheffield neighbourhoods (Sheffield City Council defined, OSM, Yahoo Places, Geonames and OS data). Given the general poor agreement between the differing sources of neighbourhood names (see §2), such a positive fit is promising. Some of the neighbourhood names that our method identified that were not found in the other existing gazetteer sources of neighbourhoods contained well known Sheffield neighbourhoods such as “Hunters Bar”, “Shalesmoor” or “Kelham Island” - further demonstrating the incompleteness of existing sources of neighbourhoods.

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. Whilst all of these were to be found within the data output they had low levels of returns and failed the rules of inclusion. The rationale was that any density map produced using such small numbers may not be meaningful.

Of the fifteen neighbourhoods identified by our method but not supported by at least three other gazetteers, only four do not appear in either OS data or OSM. There are, however, appropriate references to three of the four (“Newhall”, “Parkway Industrial Estate” and “Holbrook Industrial Estate”) within old OS maps. Whilst the remaining neighbourhood - “Wadsley Park Village” (built on the former site of Middlewood Hospital, 2001-06) does not appear in any of the existing neighbourhood gazetteers, it does have its own community website (http://www.wpvonline.co.uk).

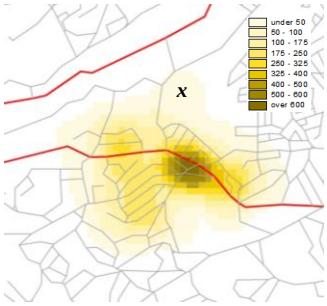

Figure 1a. Absolute 300m search radius: Crosspool, Sheffield

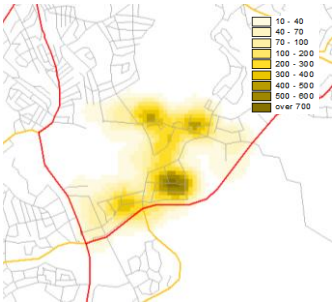

Figure 1b. Absolute 300m search radius: Norton, Sheffield

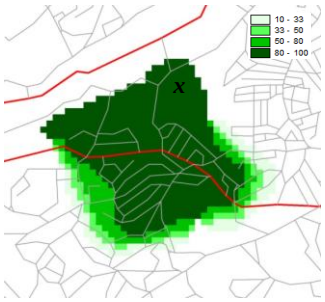

Figure 1c. Percentage 300m search radius: Crosspool, Sheffield

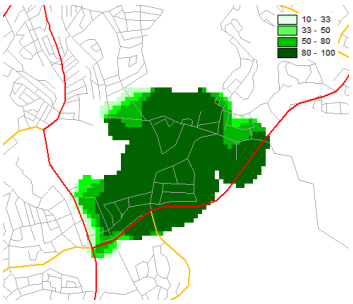

Figure 1d. Percentage 300m search radius: Norton, Sheffield

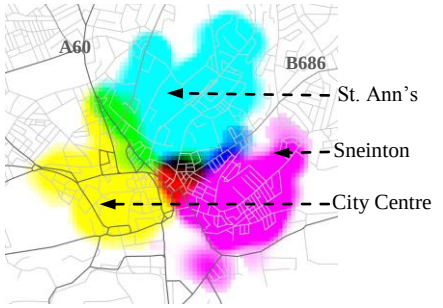

Figure 2. CMYK composite of Nottingham. A mix of colours represents conflicting views (e.g. *green* = “City Centre” and “St. Ann’s”, *black* = all 3 neighbourhoods).

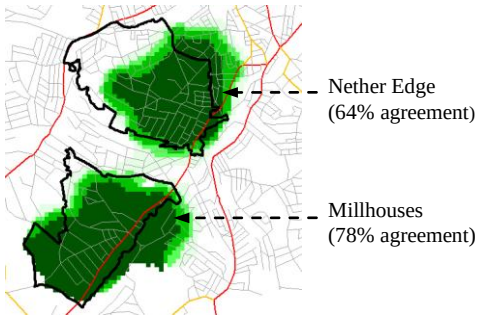

Figure 3. Comparison of selected internet derived neighbourhoods and Sheffield council boundaries

5. DISCUSSION

Although this work has been applied within the UK, such an approach could be utilised elsewhere providing that the postal delivery service in the country facilitates the use of additional information within the address between street and city elements. Whilst the scope of this paper is currently limited to neighbourhoods as found within a fixed postal address structure, our results indicate that the proposed methodology not only holds utility for adding to the comprehensiveness of neighbourhood level gazetteers, but that it can do so without any requirement for the names of the units of interest to be known *a priori*. The tight delimitation between street and city elements reduces issues of ambiguity, facilitating for example the clear disjuncture of references to "Birley" (Sheffield) from those of "Birley Carr" (Sheffield). 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 the absolute data maps (Figures 1a and 1b) identify concentrations focused around the core (commonly shopping areas) of the neighbourhoods, the percentage maps (Figures 1c and 1d) exhibit a rather different geography, identifying where the majority of people may associate with the neighbourhood names. For instance, compare the differences for the area to the north of Crosspool (Sheffield) labelled with an *x* in Figures 1a and 1c 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).

Figure 2 denotes an example whereby people within a synonymous space may have varying points of view of what the area may be named. This ties with the need to identify probabilistic perceptions of neighbourhoods, whereby, although we all have individual opinions of the extents of neighbourhood areas, it is only when examined collectively that sense can be deduced (for example this is where the majority of people believe *x* might be). Therefore, there are rarely easily defined boundaries but we see the emergence of fuzzy, probabilistic views containing differing perspectives. Nevertheless, the disjuncture along specific roads also needs to be recognised. Within Figure 2, there is some spreading of the City Centre north of the A60 (Huntingdon Street) into St Ann's but St Ann's does not appear to extend significantly south of the road. Ongoing work is developing methods to allow for the non-spreading of KDE surfaces across specified barriers (such as rivers) where data do not support the traversing of such features. Further evaluation to justify the many parameters and rules used within our method is also ongoing.

6. CONCLUSION

This research has demonstrated that the mass linguistic data now available online can be combined with computational models to automatically produce neighbourhood definitions that are both probabilistic and dynamic. Currently this has been achieved through passive mining of postal addresses online and, uniquely, without *a priori* knowledge of neighbourhood names. Even with the relatively unprepossessing initial method, the potential of combining big data with linguistic models, to produce viable geospatial intelligence, is evident. Expanding the sophistication of linguistic models used can only serve to expand its utility.

There is, additionally, further scope for fruitful research into the validation of the geography of such areas by eliciting crowd-sourced data from residents concerning their perceptions of neighbourhood boundaries. There is substantial potential in fully automated and scalable procedures such as this in the field of urban geography.

7. ACKNOWLEDGEMENTS

This work is supported by Horizon Centre for Doctoral Training (RCUK Grant No. EP/G037574/1) and by RCUK's Horizon Digital Economy Research Institute (RCUK Grant No. EP/G065802/1).

8. REFERENCES

[1] Abdelhaq, H., Gertz, M. & Sengstock, C. (2013) Spatio-temporal characteristics of bursty words in Twitter streams. In *Proceedings of* SIGSPATIAL'13. ACM, New York, NY, USA, pp. 194-203

[2] Brindley, P., Goulding, J. & Wilson, M.L. (2014) Mapping Urban Neighbourhoods from Internet Derived Data. In *Proceedings of* GISRUK 2014, pp. 355-364.

[3] Cranshaw, J., Schwartz, R., Hong, J. & Sadeh, N. (2012) The livehoods project: utilising social media to understand the dynamics of a city. In *Proceedings of* ICWSM'12, pp. 58-65.

[4] Grothe, C. & Schaab, J. (2009) Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition and Computation* **9** (3): pp. 195-211.

[5] Hollenstein, L. & Purves, P. (2010) Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science* **1**: pp. 21-48.

[6] Jones, C.B., Purves, R.S., Clough, P.D. & Joho, H. (2008) Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science* **22** (10): pp. 1045-1065.

[7] Montello, D.R., Goodchild, M.F., Gottsegen, J. & Fohl, P. (2003) Where's downtown? Behavioural methods for determining referents of vague spatial queries. *Spatial Cognition and Computing* **3**: pp. 185-204.

[8] Pasley, R., Clough, P., Purves, R. S. & Twaroch, F. A. (2008). Mapping geographic coverage of the web. In *Proceedings of* SIGSPATIAL'08. ACM, NY, USA pp. 1-9.

[9] Schockaert, S. & De Cock, M. (2007). Neighborhood restrictions in geographic IR. In *Proceedings of* SIGIR '07, pp. 167-174.

[10] Twaroch, F. A., Jones, C. B. & Abdelmoty, A. I. (2008). Acquisition of a vernacular gazetteer from web sources. In *Proceedings of* LOCWEB '08 pp. 61-64.

[11] Twaroch, F. A., Purves, R. S. & Jones, C. B. (2009). Stability of Qualitative Spatial Relations between Vernacular Regions Mined from Web Data. In *Proceedings of Workshop on Geographic Information on the Internet*.

[12] Twaroch, F., Jones, C.B. & Abdelmoty, A.I. (2009) *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.

[13] Zang, A., Noulas, A., Scellato, S. & Mascolo, C. (2013) Hoodsquare: Modeling and Recommending Neighborhoods in Location-Based Social Networks. Proceedings *of the 2013 International Conference on Social Computing*, pp. 69-74.