Challenges in Geocoding Socially-Generated Data

J. J. Huck[1], J. D. Whyatt[2], P. Coulton[3]

[1] School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA
01524 230854
j.huck2@lancaster.ac.uk

[2] Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ

[3] School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA

Summary: An investigation into the difficulties facing researchers attempting to geocode data derived from social networking sites for analysis is presented. A number of issues are identified including the lack of any inherent scale in either the socially-generated data or the results from a geocoder, and the ambiguous nature of place names. A methodology is therefore presented that may be followed by the researcher in order to address these issues, and as such improve the quality and meaning of spatial analysis that is based upon these data.

KEYWORDS: Geocoding, Social Networking, Twitter, Scale Optimisation, Place Names.

1. Introduction

It has become common practice in academia, the media and beyond to attempt to derive geospatial information from socially-generated data. There are, however, a number of issues with doing so that have yet to be addressed fully in the literature. The purpose of this paper is to address these issues, and suggest a succinct methodology by which the researcher can geocode their data to the greatest effect.

1.1 Twitter

Twitter is an example of a 'micro-blogging' site whereby users can publish short texts of up to 140 characters in length known as 'tweets' in order to share information; described by Twitter as "what's happening?" (Phuvipadawat & Murata, 2010). Over time, however, Twitter has become an important tool for communication and collaboration, the dissemination of news and even marketing; taking the medium far beyond the 'conversational' interaction that it was originally intended for. Tweets are published using both traditional computers and portable platforms such as mobile phones.

1.2 Geocoding data from Twitter

Geocoding refers to the process of attaching spatial information to data that previously did not have it, normally by the comparison of locational identifiers such as place names or postcodes to gazetteer databases in order to determine the most likely location. In recent years it has shifted from being an expensive specialist process relying on skilled operators (Roongpiboonsopit & Karimi, 2010), to being available for free online to the general public (Jung et al. 2011), and has become almost commonplace within academia and the media for tweets to be geolocated on a map in order to allow the identification of spatial patterns relating to a given topic (Field & O'Brien, 2010). As most tweets lack explicit locational information, researchers generally assign coordinates to the textual location that the 'tweeter' has specified within their Twitter profile using an online geocoding service: either commercial, such as the 'Google Geocoding API' (Google, 2011) or 'Yahoo! PlaceFinder' (Yahoo, 2011); or open source, such as 'Nomanitim' (Open Street Map 2012).

2. Background to Study

The sample dataset used within this study is data collected from Twitter regarding the 'Royal Wedding' of Prince William and Kate Middleton which took place on Friday 29th April 2011 (Official Royal Wedding, 2011); with over 1.7 million Tweets collected during the period of a month before and after the event. The locations from which these tweets originated are illustrated on the map in Figure 1.

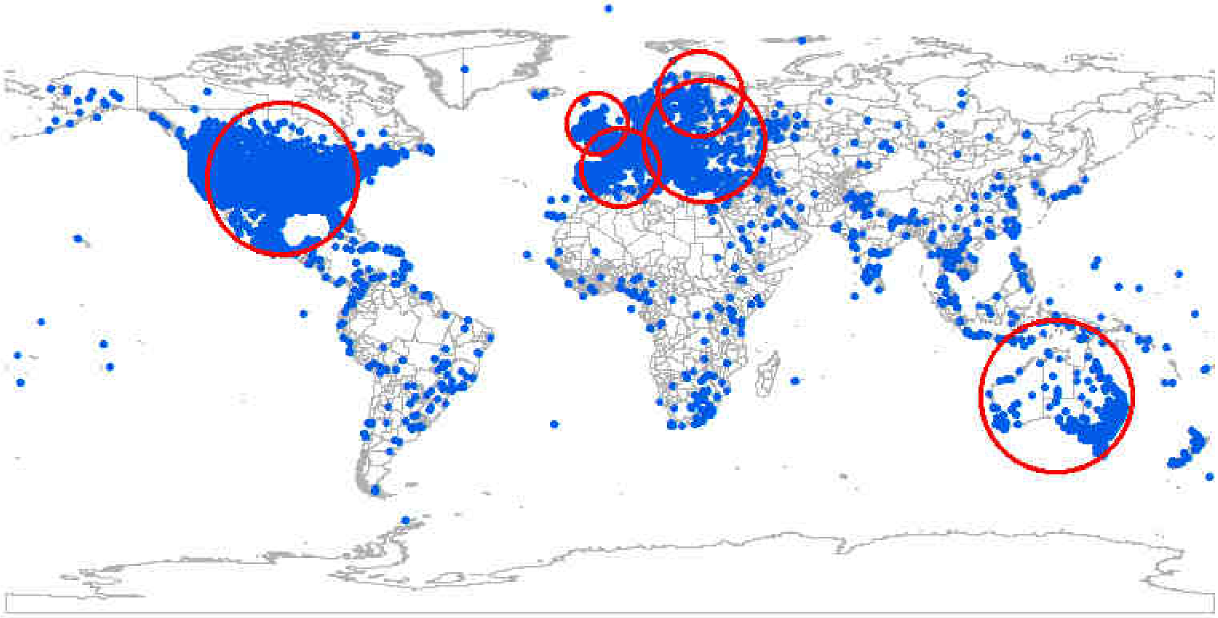

Figure 1. **'**First pass' geocoded locations for the tweets collected within this investigation. The areas upon which the data collection focused are illustrated in red.

The spatial distribution of the data in Figure 1 is purely indicative, as the geocoding is a 'first pass' attempt using the Google Maps Geocoding API (Google, 2011) that does not address any of the issues in this paper. There are obvious concentrations in the USA and Europe, and a smaller concentration in Australia; though it should be noted that these areas represent the areas of search that were used to capture Tweets (illustrated by red circles in Figure 1), and so may not represent the complete global distribution of Twitter activity relating to the Royal Wedding. Additionally, as the US-based Google Maps Geocoder (Google, 2011) was used to geocode the data displayed here, there is likely to be a positive bias towards the USA.

2.1 False hotspots

One of the major issues associated with geocoding socially-generated data is that of scale; whereby there is no implicit scale associated with either the data returned from a geocoder (Whitsel, 2008), or the textual representation of location given in a Twitter users profile. 'Scale' in this sense refers to a general indication of the 'level of detail' attained by the data returned from the geocoder, and not a specific numeric scale as would be found on a map. In the event that no normalisation work is performed upon the data returned, it is likely that false 'hotspots' will tend to form at the centroid of administrative areas; appearing as a dense cluster of data-points on the map, but in reality being nothing more than an artefact caused by data being viewed at the wrong scale (e.g. a cluster of Twitter users who list their location as "UK" should not be compared as like-for-like with a cluster of Twitter users who list their location as "LANCASTER").

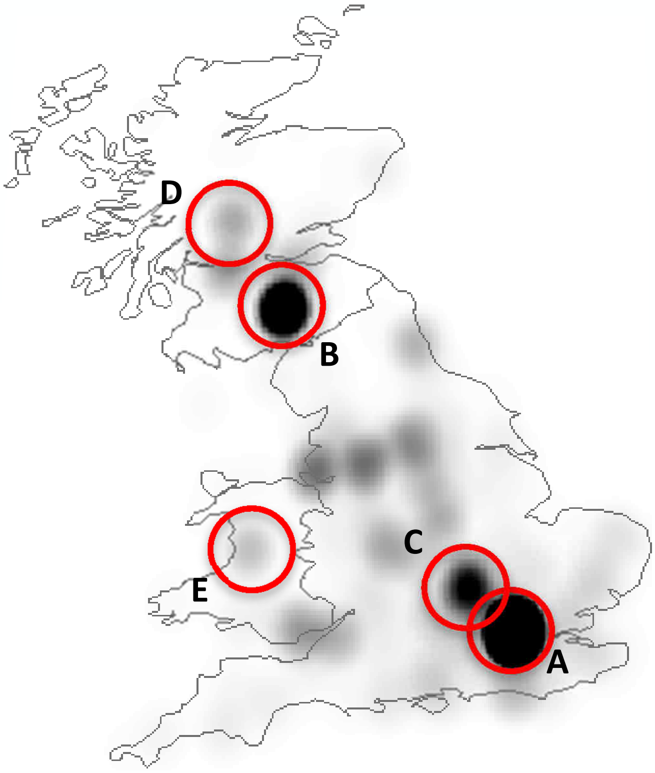

Figure 2. A density map of 'first pass' geocoded tweet locations in the UK. Hotspots are all illustrated in red. A clear hotspot is evident over London (A), as well as 'False hotspots' at the centroid of the UK (B) and each individual country (C-E).

For example, the distribution of tweets collected during the Royal Wedding study across the UK exhibits two significant 'hotspots'. One of these is located in London, a major population centre and the location of the Royal Wedding itself, the other is located in the Scottish Borders, and does not represent a population base of corresponding size. In fact, the reason for this second cluster of data is that the geocoding service returns this location as the centroid for the location "UK" OR "UNITED KINGDOM". As such, any Twitter users who list their location as such will be placed in the Scottish Borders by the geocoding service, when in reality this is most likely not the case. This is illustrated in Figure 2, with the two hotspots clearly visible, along with smaller hotspots at the centres of England, Wales and Scotland.

2.2 Place name ambiguity

Geocoding is not a process that will absolutely return a single correct set of coordinates for each textual location that it is passed. It is likely that, in many instances, a list of possible location 'matches' will be returned; and merely accepting the first result in the list (although this is usually the location that the geocoding service deems the most likely) is not sufficient to prevent bias in the data. This problem is particularly prevalent with the use of place names, which are intrinsically ambiguous (Longley et al. 2011) (e.g. there are 9 places called 'WHITCHURCH' in the Ordnance Survey 1:50,000 gazetteer), and this is compounded in informal data such as that found in social networking profiles, with 'vernacular' or non-official place names often causing problems for geocoding services, once again leading to the misleading coordinates being attached to data.

3. Suggested Methodology

In order to address the issues raised, a methodology has been developed which is illustrated in Figure 3 below.

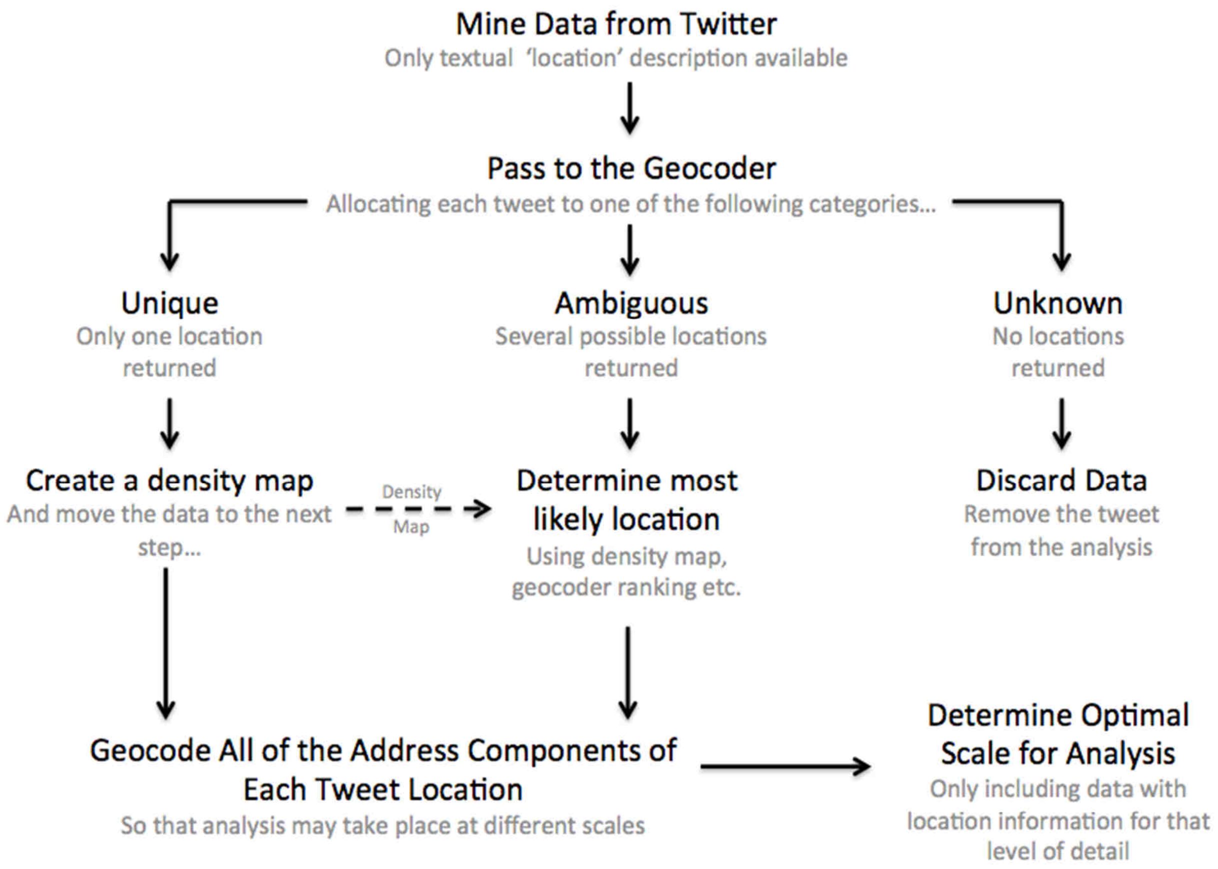
Mine Data from Twitter
Only textual 'location' description available
Pass to the Geocoder
Allocating each tweet to one of the following categories...
Unique
Only one location returned
Ambiguous
Several possible locations returned
Unknown
No locations returned
Create a density map
And move the data to the next step...
Density Map
Determine most likely location
Using density map, geocoder ranking etc.
Discard Data
Remove the tweet from the analysis
Geocode All of the Address Components of Each Tweet Location
So that analysis may take place at different scales
Determine Optimal Scale for Analysis
Only including data with location information for that level of detail

Figure 3. Flow chart illustrating the proposed methodology to be followed in order to minimise the impact of unknown scale and place name ambiguity in analysis of Twitter data.

Upon the collection of the data, it should be submitted to a geocoding service, allowing the data to be separated into three groups: ‘unique’ (whereby the geocoder returns a single location); ‘ambiguous’ (where the geocoder returns several possible locations); and ‘unknown’ (where the geocoder is unable to return any locations). Unknown data can be discarded at this stage, whilst unique data will be accepted.

It is then necessary to determine the most likely location for the ambiguous data as it is not sufficient to rely on the ranking given by the geocoder, which will generally exhibit geographical bias (e.g. whereby locations in the US will receive a higher ranking). Tobler’s ‘First Law of Geography’ states that; "*Everything is related to everything else, but near things are more related than distant things."* (Tobler, 1970). If this law is applied to the phenomenon of tweeting on a specific topic, one can assume that a tweet location is likely to be close to other known tweet locations. A density surface can, therefore, be generated based upon the unique locations (Figure 4). Every potential location for each of the ambiguous tweets can then be assigned a value representing the density of unique tweets in that area, which can be used in order to assess the most likely location. Although it is not possible to define a definite ‘correct’ value, increases confidence in the data compared to simply relying upon the ranking value assigned by the geocoder.

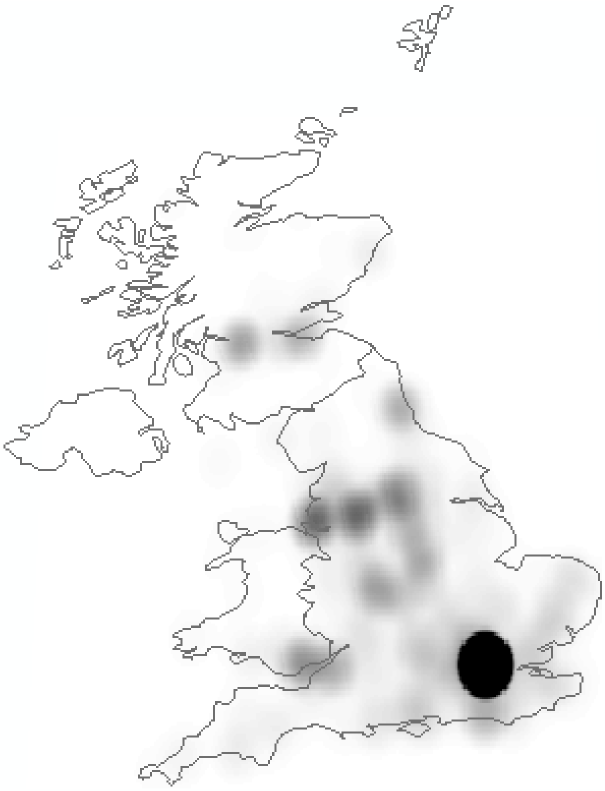

Figure 4: Density map of non-ambiguous tweets in the UK (with false hotspots removed).

The next step in the process is to determine a suitable scale at which analysis should take place in order to avoid the issue of 'false hotspots' of data forming at the centroid of administrative areas. The process of identifying the scale (*level of detail*) of each geocoded location is trivial, and the specifics will depend upon the format in which the data is returned from the researcher's chosen geocoder, but the principle involves simply identifying each of the 'address components' that make up each location, and geocoding all of them. A distribution of the greatest level of detail at each of the locations can then be created, and used in order to determine the appropriate scale for analysis, with the location found for the address component at the selected scale used to locate each tweet. This is a trade-off, as any data at a lower level of detail will need to be discarded from the analysis, and any at a greater level of detail will need to use a lower level of detail than the best available. A coarser scale will therefore sacrifice detail and maximise the amount of data used, whereas a finer scale will sacrifice more data, but allow for a more detailed analysis.

Once this process has been completed, the researcher will be left with a 'normalised' dataset, which is at a specific scale and has a minimised level of ambiguity arising from the use of free-text place names. The 'normalisation' process of a tweet with a greater level of detail than the selected 'optimal' level is illustrated in Figure 5.

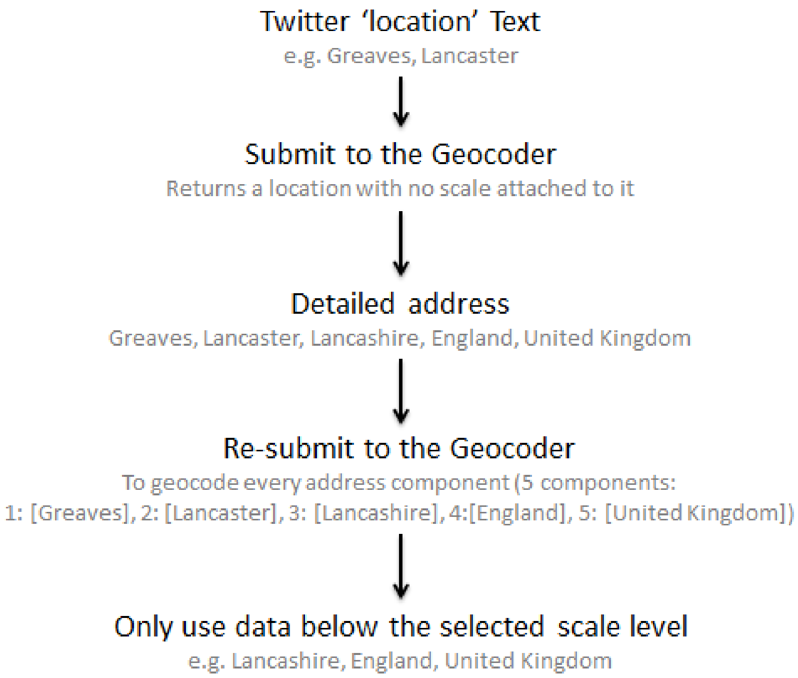
Twitter 'location' Text
e.g. Greaves, Lancaster
Submit to the Geocoder
Returns a location with no scale attached to it
Detailed address
Greaves, Lancaster, Lancashire, England, United Kingdom
Re-submit to the Geocoder
To geocode every address component (5 components:
1: [Greaves], 2: [Lancaster], 3: [Lancashire], 4:[England], 5: [United Kingdom])
Only use data below the selected scale level
e.g. Lancashire, England, United Kingdom

Figure 5. Flow diagram illustrating the 'normalisation' process of a tweet which has a level of detail greater than the selected 'optimal' scale.

4. Discussion and Conclusions

As the use of geocoding services and socially generated data increases in both academia and the media, the value of these data as a resource for gauging public interest and opinion will be increasingly recognised and exploited, allowing it to influence decision making. The spatial analysis of such data is an inevitable and already prevalent extension to this and, as such, standardising data to maximise the quality of analysis is vital to ensuring that conclusions are meaningful and representative of true spatial patterns. This paper identifies two significant issues that need to be given consideration during this process, and suggests a methodology by which the quality of geocoded socially-generated data can be increased: both in terms of the removal of bias (by the eradication of scale-related 'false hotspots'); and of a reduction in the ambiguity arising from the use of free-text place names for locating tweet origin.

5. Acknowledgements

The authors thank Gemma Davies for her invaluable contribution to this project.

6. References

Field, K. and O'Brien, J. (2010) 'Cartoblography: Experiments in using and organising the spatial context of micro-blogging', *Transactions in GIS*, vol. 14, no. s1, pp. 5-23.

Google (2011) *Google Maps API Web Services*, [Online], Available: http://code.google.com/apis/maps/documentation/geocoding/ [23 August 2011].

Jung, C., Knopp, S., Luxen, D. and Sanders, P. (2011) 'Efficient error-correcting geocoding'.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2011) *Geographic Information Systems and Science*, Third Edition edition, John Wiley and Sons.

Official Royal Wedding (2011), [Online], Available: http://www.officialroyalwedding2011.org/ [23 August 2011].

Open Street Map (2012) *Nomanitim - OpenStreetMap Wiki* [Online], Available: http://wiki.openstreetmap.org/wiki/Nominatim [7th February 2012].

Phuvipadawat, S. and Murata, T. (2010) 'Breaking news detection and tracing in Twitter', International Conference on Web Intelligence and Intelligent Agent Technology.

Roongpiboonsopit, D. and Karimi, H.A. (2010) 'Comparative evaluation and analysis of online geocoding services', *International Journal of Geographical Information Science*, vol. 24, pp. 1081-1100.

Tobler, W. (1970) 'A computer movie simulating urban growth in the Detroit region', *Economic Geography*, vol. 46, no. 2, pp. 234-240.

Twitter (2011) *Twitter Developers Documantation*, [Online], Available: https://dev.twitter.com/docs [23 August 2011].

Whitsel, E.A. (2008) Error and bias in geocoding school and students' home addresses*.* *Environmental Health Perspectives*, vol 116 no. 8 [Correspondance].

Yahoo (2011) *Yahoo! Placefinder*, [Online], Available: http://developer.yahoo.com/geo/placefinder/ [30 August 2011].

7. Biography

Jonny Huck is a 2nd year part-time PhD student researching geospatial computer science jointly within the School of Computing and Communications, and the Lancaster Environment Centre at Lancaster University. During the day he is the Technical Manager for an industrial wind energy developer.

Duncan Whyatt is a Senior Lecturer in GIS within the Lancaster Environment Centre, Lancaster University.

Paul Coulton is a Senior Lecturer in the School of Computing and Communications at Lancaster University, and founder of the Mobile Radicals (http://www.mobileradicals.com/).