

Analysing Uncertainty in Home Location Information in a Large Volunteered Geographic Information Database

Nazanin Khalili, Jo Wood, Jason Dykes

giCentre, School of Informatics, City University London EC1V 0HB
Tel.+44 (0)20 7040 0146 Fax +44 (0)20 7040 8584
{ nazanin.khalili-shavarini.1 | jwo | jad7 } @soi.city.ac.uk

KEYWORDS: Vernacular Geography, Vagueness, VGI, Spatio-Social Relation, *Flickr*

1. Introduction

This paper examines the ambiguity and location uncertainty in volunteered geographic information or VGI (Goodchild, 2007). Considering the multi geographies associated with individuals through VGI, the naive geography of confining social entities in a bounding box of a city is not adequate for study and analysis of their complex spatio-social relations. *Flickr* (Yahoo's photo sharing site) is a rich source of spatio-social data among a spatially structured social group (Khalili et al, 2009). The unstructured *flickr* Hometown Location Information (FHLI) associates individuals with places and varies from vernacular geographical terms to precise coordinates. We introduce a method for classifying and disambiguating the uncertainty in FHLI geography to augment bounding box geography and support geographical analysis of FHLI data.

2. Previous Work

There is growing body of work in disambiguating vague, indeterminate location information (Amitay et al. 2004; McCurely, 2001; Jones et al. 2008; Jacquez et al. 2000; Montello et al. 2003). Existing solutions either describe place names according to their structure, the context in which they are examined or refer to gazetteers that are used with GIS. While the spatial terms used in GIS are assigned sharp boundaries, the user generated geographic information on the web is inherently imprecise and fuzzy (Silva et al, 2006). Consequently, the existing disambiguation methods are vulnerable when applied to VGI as:

- They are only applicable to structured databases (Amitay et al, 2004)
- Location information in VGI does not conform to existing geographic boundaries (Purves et al. 2009; Jones et al. 2008)
- Gazetteers exclude vague terms (Popescu et al. 2008).

3. Flickr

Since *flickr* does not apply any restrictions on how users can define their home locations, FHLI varies widely from informal natural language (vernacular geography) to more formal scientific geographical vocabulary and numeric information. The analysis of what Waters and Evans (2003) term "fuzzy psychogeographical" can reflect the people's behaviours in defining geographic places on the web. However, its multifarious nature means that such work is by no means straightforward.

3.1 Sample Dataset

In order to produce an unbiased but indicative dataset that is rich enough for further analysis we generated a structured sample of FHLI data. In consideration of representative and manageable dataset for better analysis and ease of mapping fifteen photos of the highest available accuracy were randomly selected on a daily basis (Table1). Since the sampling periods cover entire *flickr* life from its initial launch there is potential for analysis of changes in geotagging behaviour and friendship network over time.

Intervals	poster	Unique Home Location
19/07/08-18/07/09	1,142	584
19/07/07-18/07/08	991	461
19/07/06-18/07/07	1043	481
19/07/05-18/07/06	947	426
19/07/04-18/07/05	846	381
01/02/04-18/07/04	411	192

Table 1. Number of posters for the randomly selected photos with their unique home locations.

3.2 Specification

Our initial attempts to classify the FHLI revealed six classes of terms that required new disambiguation methods to be developed and new precision measurements to be applied (Table 2). We order these according to different types of vague terms occurred in the dataset.

Vague Classes	Examples
Doesn't exist	<ul style="list-style-type: none"> • never: 'Outer space', 'L???' • current: 'Standel, Kent County'
Multiple Alternatives	<ul style="list-style-type: none"> • same name for different places <ul style="list-style-type: none"> ○ multiple scale: 'Netherland (country, city, town, hamlet) ○ multi-site place: 'Nanyang technological University' • different names for same places: 'Germany, Allemand, Deutshland'
Multiple Entities	<ul style="list-style-type: none"> • 'UK, Paris one day Italy'
Abbreviation	<ul style="list-style-type: none"> • single scale: 'Philly' (Philadelphia) • multiple scale: 'PC, US' (Pacific Coast, Panama City, Park City, Penn Central)
Mis-spelling	<ul style="list-style-type: none"> • 'Toru?, Poland'
Descriptive	<ul style="list-style-type: none"> • 'I live somewhere with lots of sunshine'

Table 2. Classes of vague terms in FHLLI.

4. Methods

In the light of the above cases, in order to study how people define their home locations on the web a method is required that can successfully complete the following three steps:

1. Disambiguation
2. Precision Measurement
3. Uncertainty Classification

The majority of the existing algorithms for disambiguating the vague terms are based on very strong discourse effects between words in a single document. Therefore, they apply the discourse constraints through probability (Smith and Mann, 2003) or one sense per discourse analysis (Gale et al. 1992; Li et al. 2003). As the nature of the FHLLI is not that of a well edited single document, these methods cannot independently disambiguate the home locations successfully. We have therefore, applied and adopted both methods according to FHLLI specifications (section 3.2).

The method we propose considers:

$$P(\text{London, London UK}) > P(\text{London, London Ontario}) \text{ If} \\ \text{Occurrence}(\text{London UK}) > \text{Occurrence}(\text{London Ontario})$$

(P stands for probability)

In cases in which the occurrences of the alternative places are of equal value or there is no occurrence of the alternative names in the dataset, the disambiguation is achieved by selecting the case with higher population. This approach, adopted from Rauch et al. (2003) itself relies upon an uncertain entity that may be associated with an uncertain extent.

The next step is to measure the precision of the disambiguated names. In the first step attempts were made to apply the *flickr* geo photos' classification model:

- World level = 1
- Country ~3
- Region ~6
- City ~11
- Street ~16

Considering the fuzzy vernacular geographic terms that are frequently found in FHLI and in order to achieve a more spatially precise classification, we have extended the *flickr* model to include more detailed hierarchical spatial units. Accordingly, fourteen distinct precision levels were identified (Table 3).

<i>Precision Level</i>	<i>Spatial Unit</i>	<i>Precision Level</i>	<i>Spatial Unit</i>	<i>Precision Level</i>	<i>Spatial Unit</i>
0	Blank	5	Region	10	Village
1	Unknown	6	State	11	Street
2	World	7	City	12	Postcode
3	Continent	8	Town	13	House No.
4	Country	9	Borough	14	Coordinates

Table 3. Precision classification for FHLI.

Classifying the home locations (section 3.1) according to spatial units identified above have resulted in some inconsistencies which were due to the facts summarized in Table 4.

Description	Example
Different internal administrative names for land units in each country	' <i>Parroquia</i> ' in Spain, ' <i>Ward</i> ' in Japan.
Different internal organizations (land divisions) exclusive to each country	' <i>Province</i> ' in China and Canada.
Inconsistency between size and population and the hierarchy of administrative divisions.	' <i>Ipswich</i> ' (town) larger than ' <i>St Davids</i> ' (city) within a single country – Britain.

Table 4. Inconsistencies in measuring precision for the spatial units.

Accordingly, in order to minimize the mentioned inconsistencies in spatial units across nations the population of each alternative spatial unit is also considered for the precision measurements of FHLI. A suitable uncertainty number (from 1 to 5) reflects our confidence in this classification in each case (Table 5).

Uncertainty Classification	Description
1	Less uncertain than the following uncertainties ('London, UK')
2	Nested spatial units e.g. city and county ('Denver', 'New York')
3	Different places in one country ('Portland, US', 'Cangas, Spain')
4	Different places in different countries ('Netherland') or several places for a single user ('Anchorage, Los Angeles, Someday New York, may be Paris').
5	Blank or information that cannot be associated with any place in the world ('Outer Space, L????').

Table 5. Uncertainty classification for FHLI.

5. Results

Classifying the sampled FHLI according to the proposed method can demonstrate how precisely people refer to their home locations on the web. As Figure (1) demonstrates there are remarkably consistent patterns in all the examined time periods with the most significant number for unknown and city level.

Figure (2) illustrates the percentage of FHLI in each category associated with each of the five types of examined uncertainties. Comparing the UC3 to UC4 (Table 5 and Figure 2) indicates that uncertainty and ambiguity in place names are more considerable in national level than across nations.

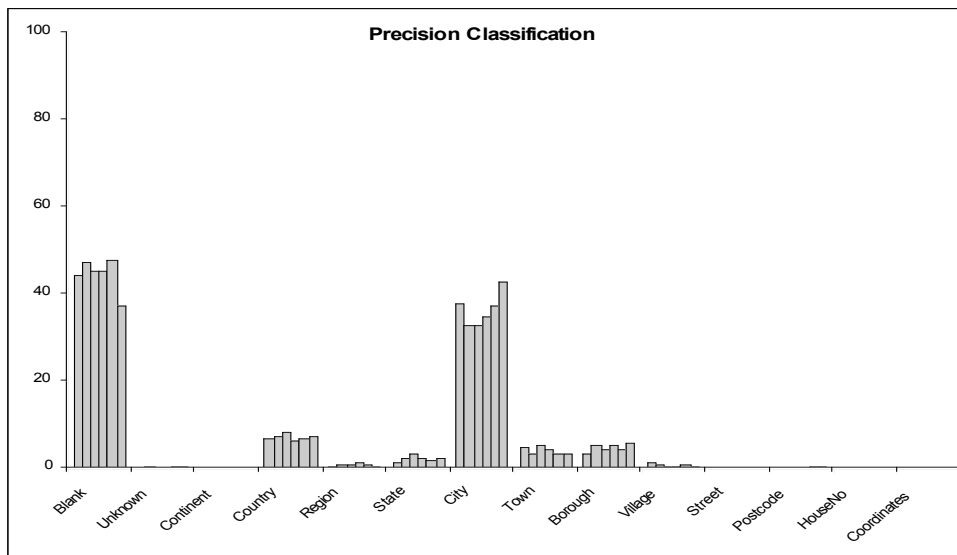


Figure1. Precision classification for the sample data

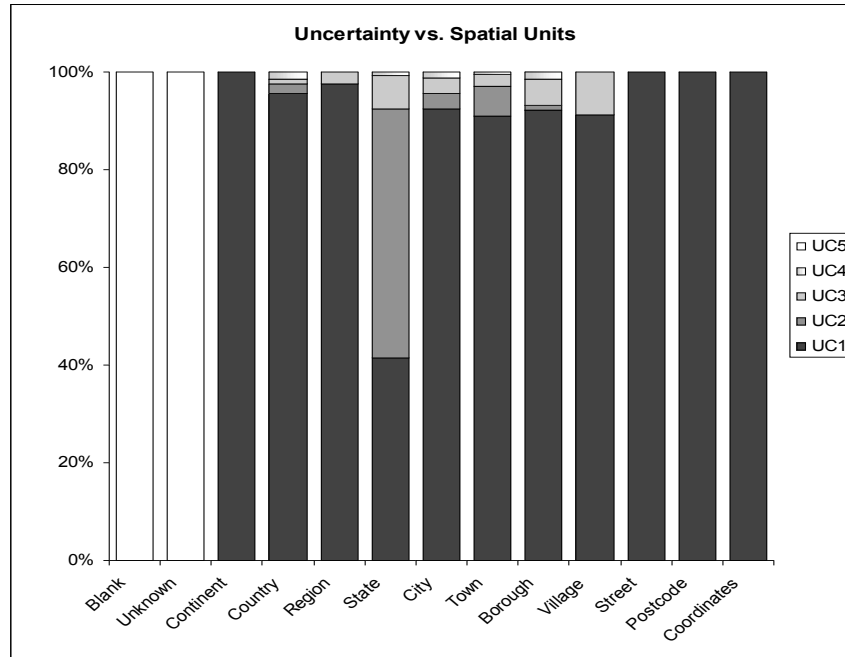


Figure2. Uncertainty assigned to spatial units.

6. Conclusion

This paper examines the complexity in classification and disambiguation of the vague geographic terms in VGI. The above preliminary analysis is conducted in order to assess and evaluate the specifications of the FHLI. The initial results are expected to contribute towards selecting a rich, unbiased and indicative sample FHLI. According to the introduced classes of vague terms in FHLI (Table2) and the fuzzy vernacular nature of the dataset, it is not feasible to fully automate the proposed approach. However, there is potential to automate the classification process for some of the FHLI by referring to the location classifications used in gazetteers (Smith and Mann, 2003; Hill, 2000). Overall, this paper concludes that:

- Focusing on the national scope of FHLIs can increase the ambiguity and uncertainty in FHLI and
- Analysis of the distribution of geotagged photo collections in line with analysis of FHLI might improve our confidence in classification and disambiguation process.

Therefore, we plan to apply this method to a broader subset of geotagged photos relating to the UK during the same time periods. Geotagged photo collections will be used as extra location information for estimating home locations. The assigned uncertainties to the disambiguated terms will be used in analysis of the FHLI and visualization of the multi geographies associated with social entities. The final models and techniques are expected to be applicable to variety of VGI available on the web.

References

- Amitay, E., Har'El, N., Sivan, R. and Soffer, A. (2004) Web-a-Where: Geotagging Web Content. 27th annual international conference Special Interest Group on Information Retrieval (ACM SIGIR), Sheffield, UK, 273-280.
- Gale, W.A., Church, K.W. and Yarowsky, D. (1992). One Sense Per Discourse. Proceedings of the 4th Defence Advanced Research Projects Agency (DARPA) Speech and Natural Language Workshop, 233-237.
- Goodchild, M.F. (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, 211-221.
- Hill, L.L., (2000) Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, J.L.Borbinha and T.Baker, Eds. Lecture Notes in Computer Science, 1923, 280-290.
- Jones, C., Purves, R.S., Clugh, P.D. and Joho, H. (2008) Modelling Vague Places with Knowledge from the Web. *International Journal of Geographical Information Science*, 22(10), 1045-1065.
- Jacquez, G., Maruca, S. and Fortin, M. (2000) From Fields to Objects: A Review of Geographic Boundary Analysis. *Journal of Geographical System*, 2(3), 221-241.
- Khalili, N., Wood, J. and Dykes, J. (2009) Mapping geography of social networks, Proceedings of the GIS Research UK 17th Annual Conference, (Fairbairn, D., Eds.), pp. 311-315, University of Durham, Durham, UK.
- Li, H., Srihari, R., Niu, C. and Li, W. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In Workshop on the Analysis of Geographic References, Edmonton, Canada.
- 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 Computation*, 3(2&3), 185-204.
- McCurley, S. (2001) Geospatial mapping and navigation of the web. In Proceedings of the 10th International WWW Conference Hong Kong, 221-229.
- Purves, R. Dykes, J., Edwards, A., Hollenstein, L., Mueller, D., and Wood, J. (2009) Describing the space and place of digital cities through volunteered geographic information. *GeoViz Workshop on Contribution of Geovisualization to the concept of the Digital City*, Hamburg, Germany.
- Popescu, A., Grefenstette, G. and Moëllic, P. (2008). Gazetiki: Automatic Creation of a Geographical Gazetteer. *International Conference on Digital Libraries*, 85-93.
- Rauch, E., Bukatin, M. and Baker, K. (2003) A confidence-based framework for disambiguating geographic terms. In Workshop on the Analysis of Geographic References, Edmonton, Alberta, Canada.
- Silva, M.J., Martins, B., Chaves, M., Cardoso, N. and Afonso, A. (2006) Adding Geographic Scopes to Web Resources. *Computers Environment and Urban Systems*, 378-399.
- Smith, D.A. and Crane, G. (2002) Disambiguating Geographic Names in a Historical Digital Library. Proceedings of the 5th European conference on research and advanced technology for digital libraries, London, UK. Springer-Verlag, 127-136.

Smith, D.A. and Mann, G. (2003) Bootstrapping toponym classifiers. In Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL), Workshop on analysis of geographic references, Morristown, NJ, USA, 45-49.

Waters, T. and Evans, A. (2003) Tools for web-based GIS mapping of a “fuzzy” vernacular geography. GIS Research UK (GISRUK), 9-11.

Biographies

Nazanin Khalili Shavarini is a PhD Candidate at the giCentre, City University London with research interests in visualization and geo social networks. She graduated from Tehran Azad University in 2005 in Computer Hardware Engineering and has an MSc in Information Systems and Technology from City University London.

Dr. Jo Wood is a Reader in Geographic Information at the giCentre at City University London with research interests in geovisualization, terrain modelling and object oriented programming for spatial sciences.

Dr. Jason Dykes is a Senior Lecturer at the giCentre, City University London undertaking applied and theoretical research in, around and between information visualization, interactive analytical cartography and human-centred design.