Assessing the Veracity of Methods for Extracting Place Semantics from Flickr Tags

This is the author's final version as submitted for publication. The definitive version of this paper was published in Transactions in GIS by Wiley-Blackwell (20130) and is available online. Cite as: Mackaness, W & Chaudhry, O 2013, 'Assessing the Veracity of Methods for Extracting Place Semantics from Flickr Tags' *Transactions in GIS*, vol 17, no. 4, pp. 544-562. DOI: 10.1111/tgis.12043.

Assessing the Veracity of Methods for Extracting Place Semantics from Flickr Tags

William A Mackaness[1] Omair Chaudhry[2]

[1]University of Edinburgh, Drummond Street, Edinburgh, EH8 9XP william.mackaness@ed.ac.uk

[2]Senior GIS Project Scientist, Emergency Response Department, Health Protection Agency Porton, Salisbury, Wiltshire SP4 0JG omair.chaudhry@hpa.org.uk

Keywords: data mining, urban morphology, Flickr, logistic regression, Bayesian inference, vernacular geography, user generated content

Abstract

The volume and potential value of user generated content (UGC) is ever growing. Multiply sourced, its value is greatly increased by the inclusion of metadata that adequately and accurately describes that content - particularly if such data are to be integrated with more formal data sets. Typically digital photography is tagged with location and attribute information that variously describe the location, events or objects in the image. Often inconsistent and incomplete, these attributes reflect concepts at a range of geographic scales. From a spatial data integration perspective, the information relating to 'place' is of primary interest. The challenge therefore is in selecting the most appropriate tags that best describe the geography of the image. This paper presents a methodology based on an information retrieval technique that separates out 'place related tags' from the rest of the tags. Different scales of geography are identified by varying the size of the sampling area within which the imagery falls. This is applied in the context of urban environments, using Flickr imagery. Empirical analysis was then used to assess the correctness of the chosen tags (i.e., whether the tag correctly described the geographic region in which the image was taken). Logistic regression and Bayesian inference were used in order to attach a probability value to each place tag. The high correlation values achieved indicate that this methodology could be used to automatically select place tags and thus hierarchically structure UGC in order that it can be semantically integrated with other data sources.

1 Formal and Informal Contributions to the Geospatial Web

The geospatial web comprises multiply sourced data. Some of it is formal, authoritative, exhaustive and invariably National Mapping Agencies (NMA) are its custodians. Such data reflects an

administrative view of geography. Another increasingly important and complimentary source of data to the geospatial web is 'crowd sourced data'. Referred to as User Generated Content (UGC) or Volunteered Geographic Information (VGI), this is data that are often incomplete, with little quality control or metadata – its variability reflecting the many different motivations for its collection (Ames and Naaman 2007). This source of data reflects more an understanding of 'place' (events and performance) rather than formal and systematic descriptions of 'space' (Goodchild 2007). UGC varies greatly in its detail, is more qualitative in nature and vernacular in its form. Its capture is greatly facilitated by sites such as Open Street Maps, Wikimapia, WikiLocation and Geonames. In some contexts it is even argued that UGC rivals formal ways of capturing geographic information (Howe 2008). When combined, these two sources of data offer highly complementary and synergistic approaches to the mining of geographic data and offer a more intuitive way of understanding place.

1.2 Modelling The Urban

But integrating 'formal' datasets with UGC is much more than simple overlay; much has been written on the need for semantic and ontological modelling in order to automatically conflate the qualitative with the quantitative (Winter 2001, Mustière and van Smaalen 2007). The difficulty of conflation lies in the vagueness omnipresent in the geospatial domain (Bittner and Smith 2001), the problematic notion of space, place and region (Montello et al. 2003) and the granularity inherent in the description of geographic concepts (Sheppard and McMaster 2004).

This is no better illustrated than in the examination of urban morphology, for the urban is forged by complex physical and social processes (Lynch 1960). A shared understanding of its constituent parts and their connections is hard to come by (Fonseca et al. 2000, Laurini 2007). Some elements are crisp in their definition and extent, whilst others have a vernacular form that is sometimes vague in its extent (Smith and Varzi 2000). Despite this, we argue that at a broad scale, the vocabulary used to describe the urban is essentially hierarchical in nature. Thus the city in overview is one that is made up of suburbs, districts and a city centre. At an intermediate scale we might describe it in terms of green spaces, retail parks, and the 'high street' and at finer scales we might think of the city in terms of its streets, buildings and public facilities (Figure 1). In this sense we argue that the city is 'nested' or partonomically constituted (van Smaalen 2003, Chaudhry and Mackaness 2007) and that the terms we use to describe the city carry meaning pertinent to these different scales.

Figure 1: One of many ways in which we might think about 'the urban'.

Various authors have argued that these notions of 'place' are implicit within user generated content, but given its unstructured form, data mining techniques are required in order to extract such 'place' information. The benefits of mining UGC results in a capacity 1) to search for documents and imagery based on references to the geographical (Hill et al. 2000); 2) to model vernacular geographies (Hollenstein and Purves 2010, Jones et al. 2008, Lüscher and Weibel 2010); and 3) to support more intuitive use of web mapping technologies. More broadly it enables us to think differently about how we 'do' geographic information science (Kuhn 2007).

In the presentation of this work, Section 2 discusses the value of Flickr tags as a UGC source, Section 3 reviews information retrieval techniques, Section 4 describes the methodology, with results presented in Section 5. Section 6 demonstrates the use of data mining techniques for 'post selection' of tags that seeks to filter out selected tags that are not geographical in nature. The generalisability of the technique is also explored in this section.

2 Flickr Tags as a source of UGC

The ambition then, is to extract place semantics from unstructured data. In this instance we have chosen the text labels associated with geolocated Flickr images. Flickr is one the biggest sources of images on the web (Winget 2006) with an estimated 153 million geo-tagged images (http://www.Flickr.com/map/). These images have either been geolocated via some mapping software, or increasingly, automatically by the device through which the image was taken (GNSS enabled smartphones for example).

The very nature of this form of UGC means that it is often inconsistent, incomplete and poorly structured (Purves 2011) yet it is its unstructured nature that is both its strength and weakness. An image can be freely tagged with any number of tags, using any words or phrases the user so wishes thus affording maximum flexibility. The weakness lies in the absence of categories and structures that would allow us to readily categorise and integrate such imagery with other geographic data. Some tags are 'non place' tags in that they describe events and emotions or they are tags that are too generic in their description to reflect a shared understanding of that place (e.g., daffodils, church, statue) – unless they can be disambiguated by being combined with other tags (Overell and Rüger 2008). Amongst tags that do describe place, some relate to small and discrete geographic entities (e.g., a monument), whilst others refer to regions big and small (both vernacular or toponym in form). The vernacular reflects a shared description of a place (often with a somewhat vague boundary), whilst toponyms reflect an administrative view – with crisp boundaries (such as street names or public gardens). The aerial extent and crispness of a region also varies considerably (Jones et al. 2008, Campari 1996). For example, a city zoo might be a large region with a precise boundary, whilst the boundary to 'the High Street' is much less distinct. 'The south of the city' has some fuzziness to it, but 'the south of the country' is an even more vague region.

Additionally we must acknowledge that the 'free' choice of those tags will be governed by the context, the user's geographical knowledge, and their inclination (Ames and Naaman 2007). So in the example of Figure 1, we might ask: how might we extract 'place' information among the tags used to describe this image (the tags being: Edinburgh, Royal Mile, Mound, high rise, tenement, PTlens, Scotland, CT3A), and how might it be structured so as to facilitates its integration and future retrieval?

Figure 2: A geolocated image with tags: Edinburgh, Royal Mile, Mound, high rise, tenement, PTlens, Scotland, CT3A

In seeking to extract meaning from Flickr tags, we assume that the image tags reflect a particular geography of place and space. This is akin to Rattenbury and Naamam's idea of 'place semantics'(2009). More particularly, we make the following assumptions (drawn from Ahern et al. (2007)):

- The total number of photographs taken in a location is an indication of the relative importance of that location;
- The importance of a location increases with the number of *individual* photographs taken of that location;
- Users are likely to use a common set of tags in the description of a place;
- The more users that used a particular tag in an area, the more representative the tag is of that area.

To this list we make the additional assumption that where tags occur in a concentrated area, they are likely to represent a particular place, whereas diffuse homogenous patterns of the same tag are likely to be representative of a region.

3 Information Retrieval Techniques

In essence, the task of identifying 'place tags' is one of data mining using ideas developed in Information retrieval (IR) (Buttcher, Clarke and Cormack 2010). IR techniques were originally applied to text (Cunningham 2002), the assumption being that frequency of occurrence is an indication of a word's significance, relative to the size of the document being searched. Any IR technique (whether its applied to text or spatial data) must cope with data for which there is no schema, the data are unstructured, and there is no clear semantic correlation between the types of queries and the data. IR methodologies must therefore attempt to infer semantics using a variety of metrics (Manning, Raghavan and Schutze 2008). In a geographical context, such techniques must

also deal with the inherently fuzzy nature of geographic phenomena (Purves, Clough and Joho 2005) as well as their hierarchical and partonomic nature (Mennis, Peuquet and Qian 2000, Bittner and Smith 2001). Various researchers have applied information retrieval techniques to user generated content (Girardin et al. 2008), in the analysis of spatio-temporal data (Dubinko et al. 2006) and the analysis of geo-referenced imagery. Variously the ambition of such research has been to link content to visualisations tools (Ahern et al. 2007), to generate meta-gazetteers (Smart, Jones and Twaroch 2010), for reverse geocoding (attaching meta data to other (untagged) imagery deemed similar to imagery for which meta data does exist) (Sarin et al. 2007), or to facilitate the integration of user generated content with other sources of data (Jain et al. 2009).

Various authors have presented techniques for extracting structured information from tagged imagery (Jaffe et al. 2006, Ahern et al. 2007, Girardin et al. 2008, Rattenbury and Naaman 2009), and unstructured lists (Purves 2011, McCurley 2001). Some techniques exist for measuring the intensity of tags (Toyama et al. 2003), such as naive and spatial scans - (Kulldorff 1999) and a number of named entity recognition techniques have been developed to identify place names and regions from unstructured data (Mikheev, Grover and Moens 1999, Grothe and Schaab 2009, Ahern et al. 2007, Kessler et al. 2009).

A commonly used IR technique is to measure a term's importance by examining the frequency of occurrence in a given document together with an examination of how low its frequency is, across a whole collection of documents. This 'term frequency–inverse document frequency' (TF-IDF) approach attempts to identify semantically significant terms that enable one document to be differentiated amongst a collection of documents (Jaffe et al. 2006, Dubinko et al. 2006). In a geographical context, the idea of TF-IDF translates to one of counting the number of images with a particular tag, that fall within a specified region or grid cell. TF-IDF assigns a high score to tags that have a larger frequency within a grid cell, as compared with the region outside that cell. A threshold can then be used to determine which tags are deemed 'place tags' and those that are 'a-spatial' and don't describe any geography. This is much more productive than merely identifying frequently used terms that may do little to help differentiate one region from another.

Various authors have observed a scale dependency in the uniqueness and frequency of tags associated with such imagery (O'Hare and Murdock 2013). This scale dependency reflects 'part of' hierarchies inherent among the tags – something that Bittner and Smith (2001) refer to as 'ontological zooming'. Serdyukov et al. (2009) used a coarse set of spatial granularities (1, 5, 10, 50 and 100 km^2) to examine these dependencies, and Gschwend and Purves (2012) examined geomorphometric landscape measures at two levels of resolution. Crandall et al. (2009) explored issues of scale and place semantics by also looking at two levels of granularity – the city level (100 km^2) and the individual placename level (100m^2), though this was limited to a specific set of landmarks for a fixed set of cities. In essence all these approaches seek to analyse the distribution of occurrences over varying spatial domains (and in the process seek to address the issue of the modifiable areal unit problem (Openshaw 1984).

In order to explore this scale dependency we used a TF-IDF approach that is similar to Ahern et al. (2007) and Rattenbury and Naaman (2009), in that we applied this technique at a number of different levels of detail in order to reveal a geographical hierarchy to the tags with the ambition of

understanding the partonomic relationships that exist among the tags, and identify the limits of reliability of any given tag in describing place semantics. We did this by dividing the region into a set of regular grids. Searching using small grid cells identifies tags that are locally important, whilst increasing the 'document size' (larger cell size) tends to identify tags important at the level of the city. Additionally our research extended these ideas by exploring the generalizability of our approach, by using Logistic regression and Bayesian inference to measure the correctness of identified place tags.

4 Methodology

The ambition then, is to extract place semantics from unstructured text labels associated with geolocated Flickr images. We take as input two elements: the geolocation of a photo (latitude and longitude), and its associated collection of tag descriptors. Flickr provides a non-commercial API that enables us to query these images by date, by tags, or geographic location. In addition we utilised a free Flickr API programming kit – Flickrj (http://www.Flickr.com/services/api/) which allows API queries to be embedded within our own code. In our case study, images for the city of Edinburgh, Scotland were accessed. A total of 134,986 images with their id, user tags, URL, user id, latitude, and longitude values were thus obtained. As can be seen from Figure 3, the pattern of locations is far from even, focusing across the central region of the city, tourist attractions and places with a high visual amenity.

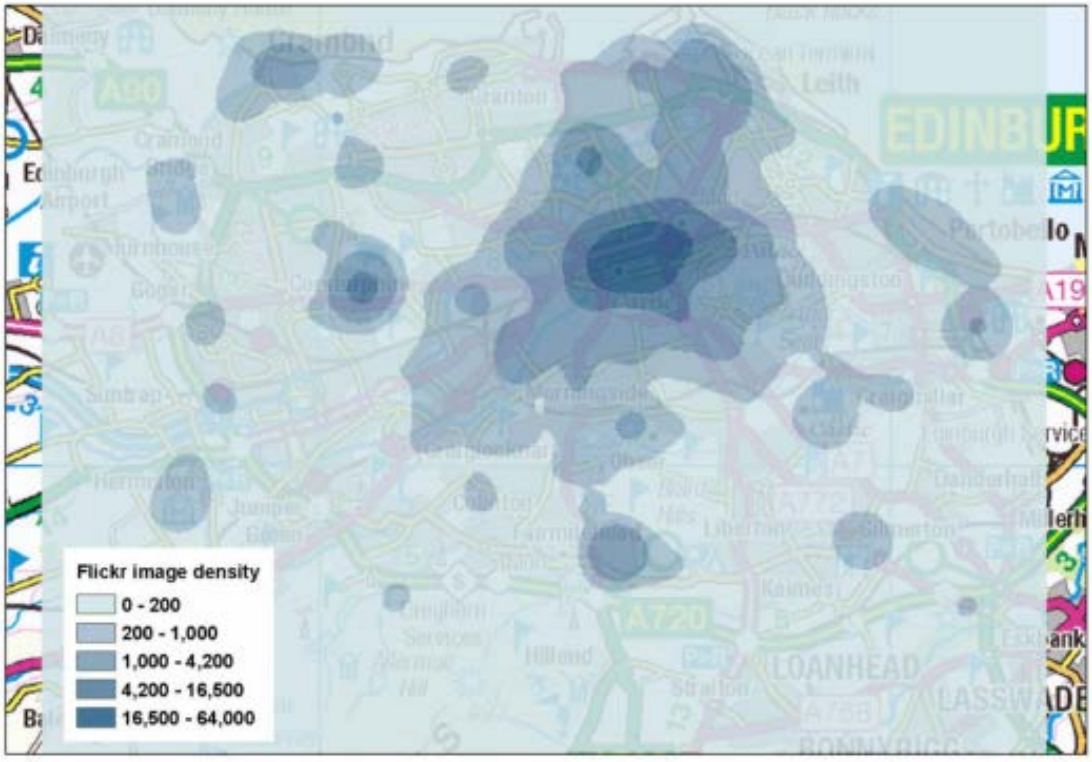


Figure 3: A kernel density map (using a bandwidth of 500m) of Edinburgh summarising the distribution pattern of Flickr images across the city.

We begin by diving the region into a set of regular grids. We chose a range of granularities at relatively fine scales (100, 500, 1000, 2000, and 4000 m^2). At the $100m^2$, there were a total of

20,400 100m^2 cells covering the city. Of these, 4,404 cells contained one or more Flickr images and out of those 4,404 cells, 3,993 cells contained images that had been variously tagged. The tags were lowercased, with white space and commas removed. So 'Royal Mile, Edinburgh Castle' becomes 'royalmile edinburghcastle' which gives us, for each cell, a 'bag of tags' rather than a bag of words. A bag of tags approach has been shown to be a little more accurate than a bag of words approach. We rank the importance of a tag's usage pattern falling within each grid cell at each of the different scales. From this we formed a hierarchical structure that reflects, to some degree, the containment of placenames within more regional descriptions of urban space.

4.1 Modelling the Local Context

In our study, for each tag contained within a cell we computed its 'term frequency' (TF) by dividing the number of times the tag occurred within the cell by the total number of all tags within that cell. Inverse document frequency (IDF) was computed by taking the logarithm of the total number of cells that contain one or more tagged images (in our case 3,993) and dividing it by the total number of cells that contain that particular tag. TF-IDF is the product of TF and IDF. This product (TF-IDF) is the 'weight' that is assigned to any given tag. The highest weighted tag within a cell is then selected as the place tag for that cell. This approach ensures that those tags which are frequent within one cell but occur rarely in other cells are given a high weight. So 'Princes Street' (frequently occurring within a particular cell) will have a high weight, whereas 'Edinburgh' or 'Scotland' will have a low weight since they occur frequently both within the cell as well as across the whole collection of cells.

4.2 Combining TF-IDF with an Object's View

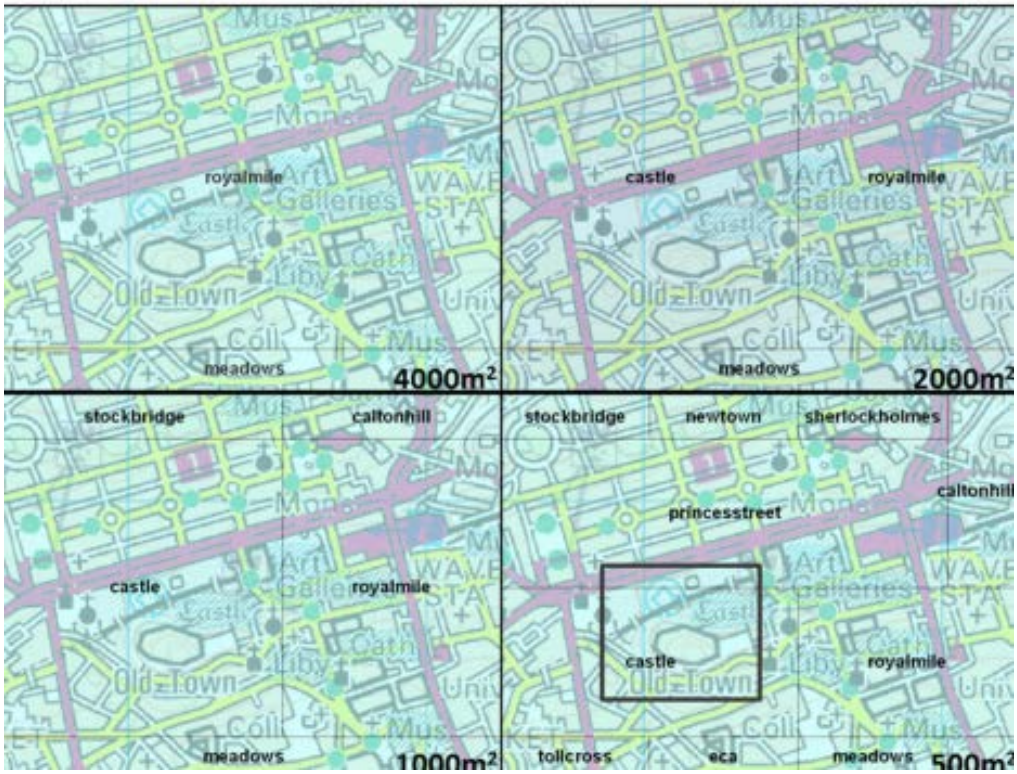
The TF-IDF approach identifies tags 'local' to a region, but it does not remedy the problem of 'tag distortion'. Tag distortion arises where a single individual records a relatively large number of images, and uses the same tag to describe an event (rather than a place). For example 'Elaine's wedding' was one such tag common to a selected cell – one in which 20 separate images were spatially contained within a particular cell. We can resolve this problem if we take an *object*'s perspective of the tag, rather than a *subject*'s perspective. We might realistically expect *different* people to use the *same* tag, thus corroborating the validity of that tag. In the example of 'Elaine's Wedding', it is very unlikely that other people would use this tag within the same cell. So by attaching importance to the number of different users who use a particular tag, we overcome the distorting effect of a single user attaching the same tag to multiple images falling in the same cell. So instead of using tag occurrences we use user frequencies associated with each tag when calculating TF-IDF weights. Using the unique user count reduces the TF for tags such as 'Elaine's wedding' and IDF ensures that tags with a high user count, such as 'Edinburgh' and 'Scotland', will have low IDF values. This results in low weights (TF-IDF) for both of these types of tags. All tags contained within a particular cell are then sorted in descending order and the tag with the highest weight is selected.

As a refinement, the tag is selected only if it has a 'user count' of at least two. This extra condition ensures that at least two distinct users have used the same tag. If this condition is not met then the next tag in the sorted list is checked and so on until both conditions are met. By applying this technique 3,951 cells were assigned a tag at the 100m^2 spatial resolution for the city of Edinburgh.

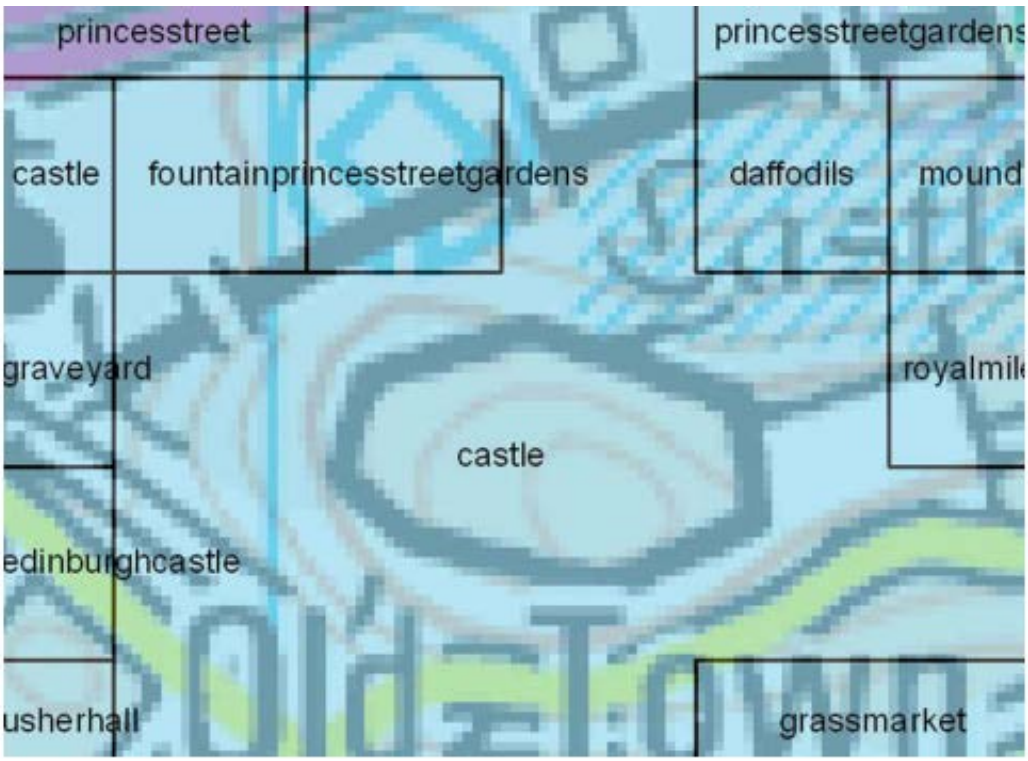
Given our interest in exploring the hierarchical nature of place tags, we applied the same methodology but to increasingly larger cell sizes, but covering the same region. We applied the technique to grid cells with resolution: $500m^2$, $1000m^2$, $2000m^2$, $4000m^2$ covering the city of Edinburgh. Once the label was selected for each cell at a specific level ($100m^2$ to $4000m^2$) we aggregated adjacent cells if they had the same label. This created continuous regions that shared the same tag.

5 Results

Figure 4 shows the result from around Edinburgh city centre (around Edinburgh Castle and along a street called 'The Royal Mile'). Figure 4 shows the selected tags as labels for each cell at the different levels of detail. At each higher level the most dominant tag (highest TF-IDF weight) is selected as the label. The output is a surface of place semantics, reflecting a socially defined process of people's understanding of place that more readily conforms to people's experience of place ((Davies et al. 2009). The fact that it reflects a more intuitive sense of place is the very reason why it is relevant to the design of intuitive interfaces. For example, output from this research been incorporated into a mobile, dialogue based interactive system that supports tourists exploration of the city (Mackaness et al. 2013).



(a)



(b)

Figure 1: (a) Selected labels from tags at different levels (4000m^2 to 500m^2) of spatial detail. (b) Selected tags at 100m^2 for the region highlighted in (a) at 500m^2

From the information so derived, we created an interactive tree view (Figure 5)[1] in order to explore these hierarchal relationships in more detail. The cells are connected hierarchically via their spatial relationship – 'contained by'. The numbers next to each tag name in Figure 5 is the user frequency of that tag at a particular level of detail. For instance tag 'royalmile' at 4000m^2 has a user frequency of 413 meaning there are 413 different users that have used this tag. This visualisation was created by building upon the freely available prefuse java library (Heer, Card and Landay 2005).

5.1 Evaluation

The question now becomes: 'at these various scales, how correct are these place tags in describing these vernacular regions?'. In other words, we want to evaluate the outputs and assess the veracity (or correctness) of this approach in selecting tags representative of place. We wanted to know how generalisable the approach was, and whether we could predict the likelihood of the correctness of a place tag for any urban environment, not just Edinburgh. Initially we compared our results against Geonames (www.geonames.org) but rather than act as a yardstick, it highlighted instead the incompleteness of Geonames! Instead the evaluation took the form of manual inspection of the output. In other words the place tags selected by the above approach were compared with results from web based searches of the place tag, comparing textual and pictorial locations with the relevant grid cell locations. The authors also used their local knowledge of the area to classify the tags. Whilst this might introduce bias, in reality it was a simple task to identify non-place tags.

[1] Visualisation applet at: http://www.omairchaudhry.net84.net/City_Viz/TreeView_UserFreq.html

Review of these visualisations revealed that there were still ‘non place’ tags present among selected tags, most notably at the finest level of detail ($100m^2$). Date tags were a typical example of this. This happens because a tag will have less weight either because there are very few distinct users that have used that tag or because it is common to a broader collection of cells. Among tags at the $100m^2$ scale, it was found that out of a total of 3,951 cells ($100m^2$) assigned a tag, only 34% contained a place tags; the remainder were unrelated to place (non-place tags). Similar manual inspections were carried out for the selected tags at all other scales. Figure 6 illustrates how ‘non-place tags’ selected by this approach are greatest at the most detailed level. But at the broader scale (larger grids), the TF-IDF values for non place tags were less as compared to place tags. This is because the spatial extent is larger and there are more images that have tags describing place.

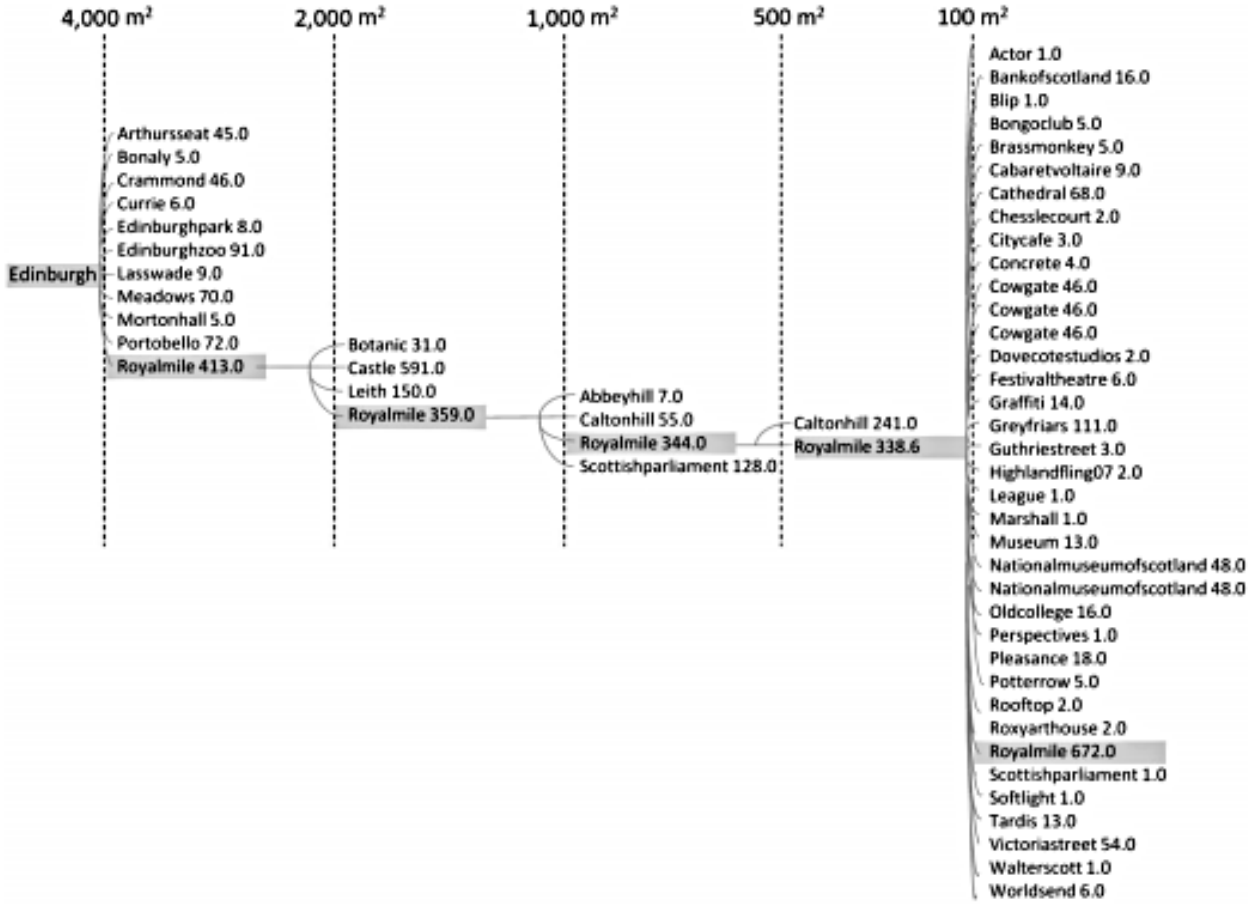


Figure 2: A ‘treeview’ visualisation of selected tags and their unique user frequency at different levels of detail

Figure 3: Result of manual inspection revealing increased presence of non-place tags at smaller cell sizes for Edinburgh, UK.

6 Post Selection Refinement using Data mining Techniques

Techniques such as 'stop words' and 'controlled vocabularies' have been proposed to minimise the chances of selecting non place tags (Pasley et al. 2008, Croft, Metzler and Strohman 2009). As an alternative to these approaches, we used data mining techniques as a way of attaching a confidence value to selected tags at the various granularities. The aim was to further refine the above approach such that a confidence value, representing the probability that it is indeed a place tag rather than anything else can be attached to each selected tag. This value can then be used as a basis for further refining the selection process. We first explored the use of logistic regression and subsequently experimented with a second approach that used Bayesian inference. In both approaches we undertook a manual assessment to build and test the accuracy of the two approaches. For each approach we randomly selected 70 % of the manually classified cases at $100m^2$ to build the model (logistic regression). We also used it as the training data for the Bayesian approach. The remaining 30% of the manually classified data at $100m^2$ and 100% of the data at all other levels of detail (500 m^2 to $4000m^2$) were used to assess the validity of the approach.

6.1 (Binary) Logistic Regression

Logistic regression is similar to multiple regression except that the dependent variable in the logistic regression is sampled as a binary variable i.e. non-place (y=0) or place tag (y=1). Logistic regression therefore models the probability of presence and absence for a given observed value among the predictor variables. The probability function can be written as (Allison 2001):

$$P(Y=1) = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+....+\beta_n x_n)}} \qquad \text{Eq1}$$

In Equation 1, y is the dependent variable, α is the intercept, β is the coefficient(s) of the independent variable(s), x. Equation 1 can be used to calculate the probability that the outcome (dependent variable) will be 1. In this research y is 1 if the selected tag is considered to be the correct place tag for a given cell (using manual inspection), otherwise it is 0 (non-place tag).

Table 1: variables used in the logistic regression model

x1	the user frequency of a selected tag within a cell;
x2	the user frequency of the selected tag in the whole collection (all cells);
x3	the selected tag frequency within a cell;
x4	the selected tag frequency in the whole collection;
x5	the total number of images contained by a cell;
x6	the total user frequency for all the tags contained by a cell;
x7	the total raw frequency of all the tags contained by a cell.

For each cell and its selected tag, we calculated a number of variables, (xn).These are listed in Table 1. Stepwise binary logistic regression was carried out in SPSS (Kleinbaum, Klein and Pryor 2010).We randomly selected 70% of the manually classified cases at $100m^2$ to build the model. The remaining 30% at $100m^2$ and all the selected tags at the rest of the scales were used to test the accuracy of the model.

Nagelkerke is a measure of model fit for logistic regression. It is a modified version of Cox and Snell's measure of significance with a value that lies between 1 and 0 (Nagelkerke 1991). In this analysis, Nagelkerke's R^2 value for the model was 0.423. Table 2 lists the selected variables (x1, x2 and x4) from the last stage of the stepwise logistic regression together with their coefficient values - the standard output from SPSS (Kleinbaum et al. 2010). The significance value of 0.000, being less than 0.05 in fact indicating that the variable is highly significant in accounting for the likely correctness of a place tag. Table 3 summarises the results in the final classification stage. The cut off value used in Table 3 to separate between cases classified as 0 or 1, is 0.5. This simply means that if the resultant probability for a selected tag is 0.5 or more it will belong to class 1, i.e., it is a place tag. If it is 0.49 or less it will be deemed to belong to class 0 (a non-place tag). This approach revealed that in about 81% of the cases, the tag was correctly identified among the selected classes using these three variables (Table 3). The same model was subsequently applied to the remaining 30% of the selected tags at $100m^2$, where just over 82% of the tags were correctly identified.

Table 2: Selected variables in the model

Variables	B	S.E.	Wald	df	Sig.	Exp(B)
x 1	1.413	.090	246.095	1	.000	4.107
x 2	.013	.002	46.500	1	.000	1.013
x 4	-.001	.000	35.958	1	.000	.999
Intercept (α)	-2.738	.121	512.187	1	.000	.065

Table 3: Classification table with 0.5 as the cutoff value at the 100m^2 scale

Observed		Predicted					
		Selected Cases (70%)			Remaining Cases (30%)		
		Class		Percentage Correct	Class		Percentage Correct
		0	1		0	1	
Class	0	1697	112	93.8	751	37	95.3
	1	410	549	57.2	172	223	56.5
Overall Percentage				81.1			82.3

Once the model was built (Table 3) at the most detailed scale we applied the same model to the remaining larger cells (coarser scales). The results from the model were evaluated against the manual inspection carried out in the previous section. The results of observed against predicted are presented in Table 4. For example, at the 1000m squared grid cell, the logistic model correctly predicted 111 place tags out of 132 tags, and identified 39 tags as indeed being non place tags out of 50 tags. The table illustrates that, the ability of the model to predict correct result (in terms of sensitivity and specificity) increases significantly (especially for true positive) as the scale reduces.

Table 4: Evaluation of the logistic model (Table 2) as compared against manual classification at various levels of detail. The probability cut off value is 0.5 – the same as in Table 3.

			predicted		
		Class	**0**	**1**	**% Correct**
observed	**Scale: 500m^2**	**0**	220	29	88.35
		1	99	216	68.57
	Scale: 1000m^2	**0**	39	11	78.00
		1	21	111	84.09
	Scale: 2000m^2	**0**	1	2	33.33
		1	0	47	100.00
	Scale: 4000m^2	**0**	0	0	
		1	0	10	100.00

6.2 Bayesian Inference

The overall average for correct results using logistic regression was above 80% (the overall percentage from Table 3). However the percentage of true positives at the finest scale of 100m^2 was not as good, at about 57% (Table 3). We therefore decided to test another data mining technique, namely Bayesian inference, to assess whether this approach could improve upon the reliability in such cases. Bayes' Rule is a simple way of calculating conditional probabilities (Hacking 2001). Conditional probabilities are those probabilities whose value depends upon the value of another

probability value (Duda, Hart and Stork 2001). The Bayesian decision rule tries to minimize the probability of error in a decision by deciding the most probable outcome. Probabilistic inference in Bayesian networks is a well understood approach (Russel and Norvig 2003).

Using a Bayesian approach, we can answer questions of the following form: 'For a given selected tag for a cell with a specific set of characteristics (as listed in Table 2), what is the likelihood that it belongs to the list of place tags, or non-place tags, given these specific set of characteristics?'. What is returned is a probability value reflecting the likelihood that the tag should indeed be selected as a meaningful tag for that cell. Here we used an approach similar to that proposed by Luscher et al. (2009). They used normal kernel density estimation techniques to determine joint probability density values for different building types in the Topography Layer of Ordnance Survey MasterMap (Ordnance Survey 2007). The value was then used to classify buildings into terraced or non-terraced houses. In essence this involved comparing an unknown with a sample of 'knowns' (training sample) and classifying the unknown according to how similar it was to the 'knowns'.

The joint conditional probability for a classification of an unknown is given by Equation 2 (Lüscher et al. 2009):

$$P_c(\vec{f} \mid C = c) = \frac{1}{N \parallel \vec{h} \parallel} \sum_{i=1}^{N} K\left(\frac{\vec{f} - \vec{f}_i}{\vec{h}} \right) \quad \text{eq 2}$$

Where P_c is the conditional probability of the unknown for the predicted class C, N is the number of samples in the training dataset, $\vec{h}$ is a smoothing parameter called the bandwidth, K is the standard normal distribution function, $\vec{f}$ is the vector of properties of unknown, $\vec{f}_i$ is the vector of the same properties of training dataset.

The same 70% of randomly selected manually classified tags, as used in the logistic regression, were used as the training dataset. The remaining 30% at 100m^2 and the rest of the dataset at lower levels of detail (500-4000m^2) were 'withheld' in order to assess the accuracy of this approach. Table 5 presents the result of the Bayesian approach for all the unselected cases i.e. the non-training dataset for Edinburgh at all levels of detail. It is important to point out that the attributes (or properties) used for Bayesian classification are the same as the ones selected for logistic regression i.e. x1 (the user frequency of a selected tag within the cell); x2 (user frequency of the selected tag in the whole collection); and x4 (selected tag frequency in the whole collection). The overall accuracy in classification of the non-training dataset at 100m^2 using this approach is 85% - marginally better than the overall accuracy using logistic regression (Table 3). A major difference is in the improvement of the correct selection of place tags (y=1) - namely 71% (Table 5) using this approach as compared to 56% (Table 3) using logistic regression at 100m^2. Similarly at other scales there is an improvement in the correct prediction of classification of both non-place and place tags (Table 4 and 5).

Table 5: Classification result using Bayesian inference for Edinburgh at different levels of detail

	predicted				
observed	**Scale**	**Class**	**0**	**1**	**% Correct**
	100m^2	0	727	61	**92**
		1	113	282	**71**
	500m^2	**0**	239	8	**97**
		1	72	245	**77**
	1000m^2	**0**	48	3	**94**
		1	30	101	**77**
	2000m^2	**0**	2	2	**50**
		1	7	39	**84**
	4000m^2	**0**			
		1		10	**100.00**

By examining a large number of images at a range of scales, our approach is able to predict the reliability with which place semantics can be inferred from image tags. The analysis has shown how scale is an inherent property of place semantics – that groupings of images reveal different geographical extents (from statues to plazas, from parks to cities). We mapped these probability values against the selected tag (visualised in Figure 7)[2]. In effect these value show how confident the approach is in deciding that a tag is indeed a place tag.

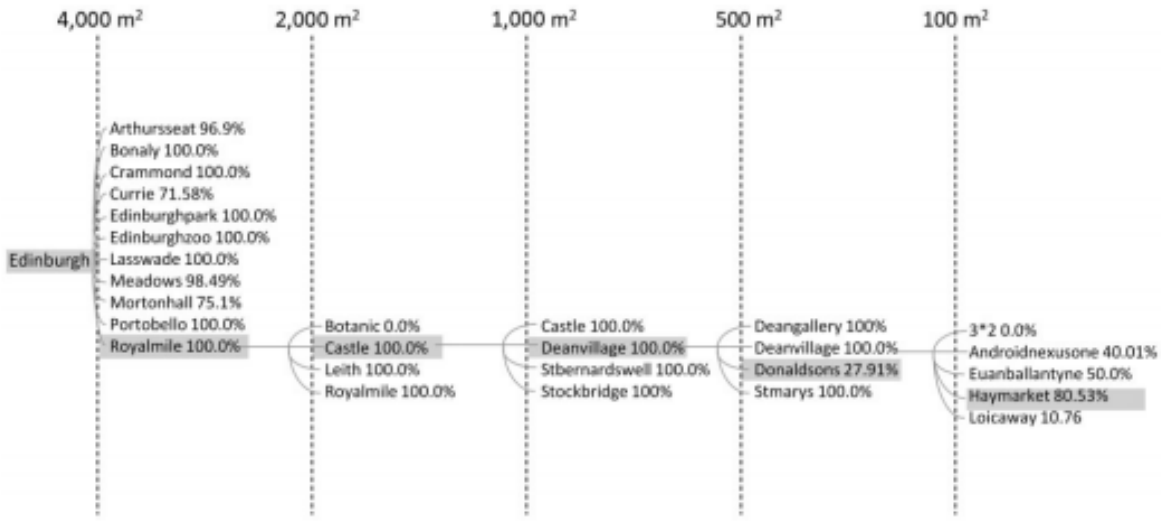


Figure 4: Tree view of selected tags with associated probability value that indicates the likelihood of it being a place tag as predicated by the Bayesian logic

[2] Visualisation applet available at: http://omairchaudhry.net84.net/City_Viz/Tree_View_Tags_Confidence_Bayes.html

6.3 Observations and Discussion

The broader motivation for mining geographical user generated content is to enable data integration between structured and unstructured data, to support reverse geocoding, to create meta gazeteers, make more intuitive interfaces, and ultimately to enable a richer modelling of 'place'. Whilst data mining techniques have had to deal with the fuzzy and multi scaled nature of geographic information, they have not examined the reliability of these approaches, nor explored their generalisability. Thus the motivation for this research has been to 1) explore the multi scaled nature of user generated content, and 2) apply statistical techniques to a large number of images at a range of scales in order to be able to predict the reliability with which place semantics can be inferred from image tags. Thus our first research question was to explore how varying the document size (cell size) changed the prominence of each tag. We observed that place specific names were replaced by more regional ones at more granular scales, but that typical tag lists could not be expected to represent all granularities. 'Place' is not homogenous; the intensity and extent of place is governed by cultural, geographical and historical drivers, and its mapping will depend on visual amenity and popularity – the very reason why it is so important to be able to assess the reliability of a tag. Our research shows that as you move away from popular attractions and the city centre, that tags become less reliable. The ambition of this approach has been to filter out non place semantics, but this is not to ignore the value of 'non place' tags. For example, such tags can be used to model emotions or saliency characteristics of a place (Schmitz 2006).

Our results show that for a given set of geo located images the reliability of place tags at the fine scale are less reliable that at more granular scales. The confidence values of intermediate geographies are highest and at very granular levels (i.e., at the 'city' scale) the confidence values are also high. At the scales in between, tags are less effective at describing the partonomic structure of the city. This is because tags are rarely 'even' in their step change of description. So you tend not to get a 'nice' set of tags such as: 'panda, zoo, Corstophine district, North West Edinburgh, Edinburgh Scotland'. More likely would be the case of: 'Panda, Edinburgh Zoo' (i.e., an absence of tags describing various granularities of location). This may well be sufficient from the user's point of view since they have no further interest (or knowledge) with which to further enrich the tags.

We acknowledge limitations to this work. As with all research exploring the utility of UGC, errors can arise in both the assignment of coordinates to an image, and in the choice of tags associated with an image. Purves (2011) also warns of a circularity that can arise in the choice of tag as to whether it reflects actual local knowledge or is selected based on a consensus derived from the geospatial web itself. Furthermore, a tag may not so much represent its location as the object in the field of view. This is more problematic in rural contexts than urban ones where the distance between the observer and the object of interest in a rural context can be large. Evidence suggests however, that generally users seek to record their location rather than the location of the object in the image (Hollenstein and Purves 2010). The positional accuracy of the device may affect the results. Typical positional accuracy of GNSS enabled cameras is 10-30m. Given the nature of the task, and the size of the grid cells used, this was not deemed to affect the precision of the results. We also acknowledge that the boundary of a cell may intersect part way across a region and therefore fail to capture its true extent, though this problem is partially addressed by varying the cell sizes and merging cells that have the same place tag, thus creating continuous regions.

Future work will explore different 1) grid shapes, and the use of KDE techniques (Hollenstein and Purves 2010, Grothe and Schaab 2009), and 2) its utility in rural contexts, and the use of 'leveraging strategies' such as smoothing values using neighbouring cells (O'Hare and Murdock 2013), though Serdyukov et al. (2009) observed limited improvement in performance using this approach. It

would be interesting to use viewshed modelling as a way of helping us differentiate tags that refer to the image taker's location rather than what is in the field of view, however in an urban context this has been found to be a manageable problem, particularly where a high density of images exist (Jaffe et al. 2006).

7 Conclusion

There is considerable potential in the exploitation of the geospatial web - the assumption being that tags reflect concepts that humans readily understand and use in their day to day language (despite their vague and imprecise nature). We argue that the benefits of combining user generated content with 'formal' sources can best be achieved by systematically attaching metadata to UGC. The challenge is in reliably extracting descriptions of space among a bag of tags associated with any given image. In this paper we have deepened our understanding of the relationship between scale and the reliability of a place tag in describing place. By using logistic regression and Bayesian inference we have identified three characteristics that most govern the likely correctness of a place tag selected by the TF-IDF approach. The use of these data mining techniques enables us to predict the confidence with which we can identify place tags for any given region (not just Edinburgh). Although the Bayesian approach proved to be better at all scales the reliability of both models increases as the scale of observation decreases. The use of data mining techniques (logistic regression and Bayesian inference) has enabled us to explore the veracity of this approach and thus assess the reliability of this technique for any given urban region.

Acknowledgement

We wish to acknowledge the thoughtful, thorough and helpful suggestions of anonymous reviewers and for comments from researchers at the GISRUK conference in 2012.

References

Ahern, S., M. Naaman, R. Nair & J. Yang. 2007. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections In *Seventh ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) ACM*, 1-10. New York.

Allison, P. D. 2001. *Logistic Regression Using the SAS System: Theory and Application*. New York: Wiley Interscience.

Ames, M. & M. Naaman. 2007. Why we tag: motivations for annotation in mobile and online media. In *CHI '07 - Proceedings of the SIGCHI conference on human factors in computing systems*, 971-980.

Bittner, T. & B. Smith. 2001. A taxonomy of granular partitions. In *Proceedings of the Conference on Spatial Information Theory - COSIT'2001, Lecture Notes in Computer Science*, 28-43. Berlin-Heidelberg: Springer-Verlag.

Buttcher, S., C. L. A. Clarke & G. V. Cormack. 2010. *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge: MIT Press.

Campari, I. 1996. Uncertain boundaries in Urban Space. In *Geographic Objects with Indeterminate Boundaries,* eds. P. Burrough & A. U. Frank, 57–69. London: Taylor and Francis.

Chaudhry, O. Z. & W. A. Mackaness. 2007. Utilising Partonomic Information in the Creation of Hierarchical Geographies. In *10th ICA Workshop on Generalisation and Multiple Representation*. Moscow, Russia.

Crandall, D., L. Backstrom, H. D. & J. Kleinberg. 2009. Mapping the World's Photos. In *Proceedings 18th International World Wide Web Conference*. Madrid, Spain.

Croft, B., D. Metzler & T. Strohman. 2009. *Search Engines: Information Retrieval in Practice*. Boston: Addison-Wesley.

Cunningham, H. (2002) GATE, A General Architecture for Text Engineering. *Computing and the Humanities,* 36**,** 223-254.

Davies, C., I. Holta, J. Greena, J. Hardinga & L. Diamonda (2009) User Needs and Implications for Modelling Vague Named Places. *Spatial Cognition & Computation: An Interdisciplinary Journal,* 9**,** 174-194.

Dubinko, M., R. Kumar, J. Magnani, J. Novak, P. Raghavan & A. Tomkins. 2006. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, 193-202. New York, USA: ACM Press.

Duda, R. O., P. E. Hart & D. G. Stork. 2001. *Pattern classification*. New York: John Wiley & Sons.

Fonseca, F., M. Egenhofer, C. Davis & K. Borges (2000) Ontologies and knowledge sharing in Urban GIS. *Computers Environment Urban Systems,* 24**,** 232-251.

Girardin, F., F. Calabrese, F. D. Fiore, C. Ratti & J. Blat (2008) Digital Footprinting: Uncovering Tourists with User-Generated Cotent. *Pervasive Computing,* 7**,** 36-43.

Goodchild, M. F. (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal,* 69**,** 211-221.

Grothe, C. & J. Schaab (2009) Automated Footprint Generation from Geotags with Kernel Density Estimation and Support Vector Machines *Spatial Cognition and Computation,* 9**,** 195-211.

Gschwend, C. & R. Purves (2012) Exploring Geomorphometry through User Generated Content: Comparing an Unsupervised Geomorphometric Classification with Terms Attached to Georeferenced Images in Great Britain. *Transactions in GIS,* 16**,** 499-522.

Hacking, I. 2001. *An introduction to probability and deductive logic*. Cambridge: Cambridge University Press.

Heer, J., S. K. Card & J. A. Landay. 2005. prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Portland, Oregon, USA: ACM.

Hill, L., L. Carver, Larsgaard M, Dolin R, S. T.R., J. Frew & M.-A. Rae (2000) Alexandria Digital Library: User Evaluation Studies and System Design. *Journal of the American Society for Information Science,* 51**,** 246-259.

Hollenstein, L. & R. S. Purves (2010) Exploring place through user-generated content: Using Flickr tags to describe city cores. *JOURNAL OF SPATIAL INFORMATION SCIENCE,* 1**,** 21-48.

Howe, J. 2008. *Crowdsourcing: How the Power of the Crowd is Driving Business*. New York: Crown Business.

Jaffe, A., M. Naaman, T. Tassa & M. Davis. 2006. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. Santa Barbara, California, USA: ACM.

Jain, P., P. Z. Yeh, K. Verma, C. A. Henson & A. P. Sheth. 2009. SPARQL Query Re-writing Using Partonomy Based Transformation Rules. In *Third International Conference on Geospatial Semantics,* eds. K. Janowicz, M. Raubal & S. Levashkin, 140-158. Mexico City.

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

Kessler, C., P. Mau, J. T. Heuer & T. Bartoschek. 2009. Bottom-up gazetteers: Learning from the implicit semantics of geotags. In *GeoS '09: Proceedings Third International Conference on GeoSpatial Semantics* 83-102. Berlin: Springer.

Kleinbaum, D. G., M. Klein & E. R. Pryor. 2010. *Logistic Regression: A Self-Learning Text*. NY: Springer.

Kuhn, W. 2007. Volunteered Geographic Information and Geographic Information Science. In *NCGIA and Vespucci Specialist Meeting on Volunteered Geographic Information,* ed. M. F. Goodchild. Santa Barbara, CA.

Kulldorff, M. 1999. Spatial scan statistics: models, calculations, and applications. In *Scan Statistics and Applications,* eds. J. Glaz & M. Balakrishnan, 303-322. Birkhauser.

Laurini, R. 2007. Pre-consensus Ontologies and Urban Databases. In *Ontologies for Urban Development,* eds. J. Teller, J. R. Lee & C. Roussey. Berlin Heidelberg: Springer Verlag.

Lüscher, P. & R. Weibel. 2010. Semantics Matters: Cognitively Plausible Delineation of City Centres from Point of Interest Data. In *13th workshop of the ICA commission on Generalisation and Multiple Representation*. Zurich, Switzerland.

Lüscher, P., R. Weibel & D. Burghardt (2009) Integrating ontological modelling and Bayesian inference for pattern classification in topographic vector data. *Computers, Environment and Urban Systems,* 33**,** 363-374.

Lynch, K. 1960. *The Image of the City*. Cambridge Massachussettes: MIT Press.

Mackaness, W. A., P. Bartie, T. Dalmas, S. Janarthanam, O. Lemon, X. Liu & B. Webber. 2013. SpaceBook: Designing and Evaluating a Spoken Dialogue Based System for Urban Exploration. In *GISRUK 2013*. Liverpool, UK.

Manning, C. D., P. Raghavan & H. Schutze. 2008. *Introduction to Information Retrieval Cambridge*. Cambridge University Press.

McCurley, K. S. 2001. Geospatial Mapping and Navigation of the Web. In *10th international conference on World Wide Web*. Hong Kong: ACM.

Mennis, J. L., D. J. Peuquet & L. Qian (2000) A conceptual framework for incorporating cognitive principles into geographical database representation. *International Journal of Geographical Information Science,* 14**,** 501-520.

Mikheev, A., C. Grover & M. Moens (1999) XML tools and architecture for named entity recognition. *Journal of Markup Languages: Theory and Practice,* 1**,** 89-113.

Montello, D. R., M. F. Goodchild, J. Gottsegen & P. Fohl (2003) Where's downtown? Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition and Computation,* 3**,** 185–204.

Mustière, S. & J. van Smaalen. 2007. Database Requirements for Generalisation and Multiple Representations. In *Generalisation of Geographic Information: Cartographic Modelling and Applications,* eds. W. A. Mackaness, A. Ruas & L. T. Sarjakoski, 113-136. Oxford: Elsevier.

Nagelkerke, N. J. D. (1991) A note on a general definition of the coefficient of determination. *Biometrika,* 78**,** 691-692.

O'Hare, N. & V. Murdock (2013) Modeling locations with social media. *Journal of Information Retrieval,* 16**,** 30-62.

Openshaw, S. 1984. *The Modifiable Areal Unit Problem*. Norwich: Geo Books.

Ordnance Survey. 2007. OS MasterMap Topography Layer: User guide and Technical specification.

Overell, S. & S. Rüger (2008) Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science,* 22**,** 265–287.

Pasley, R., P. Clough, R. S. Purves & F. A. Twaroch. 2008. Mapping geographic coverage of the web. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. Irvine, California: ACM.

Purves, R., P. Clough & H. Joho. 2005. Identifying imprecise regions for geographic information retrieval using the web. In *GISRUK*. University of Glasgow, Glasgow.

Purves, R. S. 2011. Methods, Examples and Pitfalls in the Exploitation of the Geospatial Web. In *The Handbook of Emergent Technologies in Social Research,* ed. S. N. Hesse-Biber, 592 -624. Oxford: Oxford University Press.

Rattenbury, T. & M. Naaman (2009) Methods for extracting place semantics from Flickr tags. *ACM Trans. Web,* 3**,** 1-30.

Russel, S. & P. Norvig. 2003. *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Sarin, S., T. Nagahasi, T. Miyosawa & W. Kameyama. 2007. On Automatic Contextual Metadata Generation for Personal Digital Photographs. In *9th International Conference on Advanced Communication Technology*, 66-71. GITS, Waseda Univ., Saitama

Schmitz, P. 2006. Inducing ontology from Flickr tags. In *Proceedings of the Workshop on Collaborative Web Tagging at WWW2006*. Edinburgh, Scotland.

Serdyukov, P., V. Murdock & R. van Zwol. 2009. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09*, 484-491. New York, NY: ACM.

Sheppard, E. & R. B. McMaster. 2004. Scale and Geographic Inquiry. In *Scale and Geographic Inquiry: Nature Society and Method,* eds. R. B. McMaster & E. Sheppard, 1-22. Malden, MA: Blackwell Publishing.

Smart, P. D., C. B. Jones & F. A. Twaroch (2010) Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service. *Geographic Information Science Lecture Notes in Computer Science,* 6292/2010**,** 234-248.

Smith, B. & A. C. Varzi (2000) Fiat and Bona Fide Boundaries. *Philosophy and Phenomenological Research,* 60**,** 401-420.

Toyama, K., R. Logan & A. Roseway. 2003. Geographic location tags on digital images. In *Proceedings of the Eleventh ACM international Conference on Multimedia- MULTIMEDIA '03.*, 156-166. Berkeley, CA, USA, November 02 - 08, 2003: ACM, New York, NY.

van Smaalen, J. 2003. Automated Aggregation of Geographic Objects: A New Approach to the Conceptual Generalisation of Geographic Databases. In *Data & Knowledge Engineering*. the Netherlands: Wageningen University.

Winget, M. 2006. User-defined classification on the online photo sharing site Flickr. In *Proceedings of the 17th ASIS&T SIG/CR Classification Research Workshop*. Austin, TX.

Winter, S. (2001) Ontology: buzzword or paradigm shift in GIS Science? *International Journal of Geographical Information Science (special issue),* 15**,** 679-687.