Avoiding Point Data Aggregation Issues When Defining Vernacular Regions From Flickr Data

Bryn Macaulay¹, Phil Bartie²a and William Mackaness³b

¹Biological and Environmental Sciences, University of Stirling
²School of GeoSciences, The University of Edinburgh

January 11, 2018

Summary

We argue that vernacular geographies offer a more intuitive way of interacting with geographic information because they accord with our sense of place and how we refer to the geographies around us. Analysis of geolocated flickr tags have formed the basis for automatically defining vernacular regions. Various efforts have sought to filter our spurious regions and to discern meaningful regions at a range of levels of detail. In this comparative paper we present refinements to the traditional tf-idf approach. By adjusting search criteria governing the document size, more meaningful and defensible solutions can be produced at different granularities.

KEYWORDS: Flickr, MAUP, tagging, user-generated content, vernacular geography

1. Overview

Vernacular regions are the common names used to describe places and are used in everyday speech when describing locations or giving directions. Yet despite how central they are to how people talk about space, there has not been a great deal of research into automatically identifying and representing them on maps. This research applies text mining techniques to define vernacular regions automatically from Flickr data, while avoiding the modifiable areal unit problem that has been evident in previous research.

Flickr is a photo and video hosting website that was created in 2004 and has steadily risen in popularity until recently. Users can add tags to photos they upload; tags are user-chosen keywords which often refer to the subject of the photo, its style, events occurring at the time the photo was taken, and where the photo was taken. This creates a vast quantity of data, some may be of little use but within the tags are useful insights such as references to vernacular regions. Developing methods to easily find this information has great potential given the abundance of data in this form.
2. Literature review

2.1. Issues regarding data aggregation

Aggregating data for the purposes of mapping or analysis can lead to a variety of issues. One such issue is the Modifiable Areal Unit Problem (MAUP), which was described at length by Openshaw (1987). The MAUP arises when data are aggregated into arbitrary zones. The essence of the problem is that redrawing the boundaries gives different results and was first recognized in 1938 by C.E. Gehlke & Katherine Biehl, and named by Taylor and Openshaw (1979).

2.2. Defining vernacular regions

While several studies that have attempted to define vernacular regions, (e.g. Jordan 1978; Zelinsky 1980; Good 1981) comparatively few have sourced their data from the World Wide Web. However, studies that have done so show promising results, such as Mackaness and Chaudhry (2013) who defined regions in Edinburgh from Flickr tags and suggested it was a viable method. Additionally, Hollenstein and Purves (2010) found that user generated Flickr tags have roughly city-level accuracy for defining regions. They reported that it was possible with this data to explore the borders of city-centre neighbourhoods.

3. Reasons for investigation

While current literature indicates that gathering data from the web has the potential to produce accurate maps of vernacular regions, the MAUP could lead to inaccuracies when analyzing or displaying the data. Vernacular regions are central to how people perceive and talk about space, therefore the ability to easily define these regions is incredibly useful for a great variety of purposes including development planning, political campaigns, navigation systems and other information services. Furthermore, if meaningful and accurate data can be created from sources such as Flickr tags, this allows the generation of such data with relative ease as in the modern world tagged photos are in abundant supply.

This work develops a new method for defining vernacular regions that does not suffer from the issues presented by the MAUP by avoiding imposing a fixed lattice over the point data, and instead adopting ideas from the Geographical Analysis Machine (GAM) by Openshaw (1987).

4. Approach

4.1. Dataset

The dataset contained data on every geotagged photo taken in or near Edinburgh that was on Flickr at the time the data was collected (27th April 2014). In addition to the location the photo was taken, each point had additional data: a unique number that acted as an identifier for the point; a user id which made it possible to tell whether two photos were taken by the same user but did not reveal the user's name; the time that the photo was taken; and the text tags that the user attached to the photo.
4.2. Term frequency - inverse document frequency

To determine local tag importance it was necessary to process the data in the context of the regional dataset such that general terms that appeared widely across the region (e.g. Edinburgh, Scotland) were not considered locally significant. This was done using the term frequency - inverse document frequency (tf-idf) statistic, which measures how important a word, or term, is in a document compared to the terms in a collection of documents. Its value increases with the frequency of the term’s usage within the document, but decreases with the frequency of the term’s usage throughout the collection to account for the fact that some words will be more common overall. Tf-idf is commonly used for text mining, but here may be used as a way of gauging the importance of a tag at the location it is used, with the assumption that an important tag will be one naming its location. Equation 1 shows the equation to determine the tf-idf value of a tag $t$.

$$\text{tf-idf} = \frac{\text{Number of times } t \text{ appears in the document}}{\text{Total number of terms in the document}} \times \ln \left( \frac{\text{Total number of documents}}{\text{Number of documents containing } t} \right)$$  (1)

4.3. Defining documents

To apply tf-idf to spatial data requires the creation of a document equivalent. Typically a fixed grid of pre-defined cell size is imposed on the region, such that all points falling within a cell are considered as a document. However this imposes a set of regular regions on the point dataset and MAUP issues arise depending on the grid size and starting location. Instead the work presented here was inspired by Stan Openshaw's geographical analysis machine (Openshaw et al. 1987), whereby the underlying dataset is not forced to conform to a pre-defined grid but instead a document was defined as a collection of photos centred around each photo location. The terms that document contained were the tags on the photos that made up the document. Two approaches were used to define the size of each document, based on either radial distance from the central point, or the number of nearest neighbouring points.

In the initial approach, all photos within a 100m radius of the target point were considered part of the document (Figure 1a). This approach was repeated with a 300m radius to investigate how radius effected the outcome. In the second approach, to produce more consistent size of document, the photos included in a document were the nearest 250 photos to the target point (Figure 1b). As a method of removing outliers and more clearly defining region boundaries, a "second pass" was used here, removing points where the term with highest tf-idf on that point did match that of the surrounding points.

![Figure 1](image)

**Figure 1** Definition of a document in the 100m radius approach (a) and the "nearest neighbour" approach (b)
5. Results

5.1. Differences Between New Methods and Grid

As shown in Figure 2, Figure 3, and Figure 4, a grid-based approach creates areas defined by large square borders, while the new approaches create clusters of points each associated with a tag. These clusters mean that the non-grid approaches, while slower than using an imposed grid, are capable of representing regions that would be too small to be represented by the grid, and come closer to representing the "true" boundaries of the regions.

With the grid method, each grid cell can only be assigned to one region, however, vernacular regions often have overlapping boundaries. The clusters produced by the new methods can more easily represent this, by creating convex hulls of the clusters, as seen in Figure 5.

![Figure 2](image_url)  
*Data Sources: Flickr, OpenStreetMap*

**Figure 2** Comparison between Mackaness and Chaudhry's grid-based output and the output of the method defining documents by a 100m radius for a small selection of tags in the Edinburgh City Centre.
**Figure 3** Comparison between Mackaness and Chaudhry’s output and the output of the method defining documents by a 300m radius for a small selection of tags in the Edinburgh City Centre.

**Figure 4** Comparison between Mackaness and Chaudhry’s output and the output of the “nearest neighbour” method for a small selection of tags in the Edinburgh City Centre.
Figure 5 Convex hulls of tag clusters created by the "nearest neighbour" approach. Clusters shown are those with an area of at least 100 m², and in which the tag they represent was used at least 50 times, by at least 10 unique users, with at least 100 days between first and last usage of the tag.

6. Comparison of New Methods
In the approaches that used a document with fixed radius, a trend emerged - in areas where there were fewer photographs, all tags had a higher tf-idf. This is likely an artefact of the tf-idf equation, since the term frequency involves dividing by the number of terms in the document, having many terms in a document will decrease the tf-idf of terms in that document.

The 300m radius approach produces very large tag clusters that often spill far outside the boundaries of the region they define, as seen in Figure 3. As well as over-representing some regions this can also lead to other regions being covered and hidden, making it unsuitable for small scales.

While these methods meant definitions of regions were much more precise than they would be by just drawing all mentions of a region from the original dataset (Figure 6 in all methods there were cases where the boundaries drawn exceeded the extent of the area they were describing, such as the area of the "castle" tag including an area to the north of the castle itself. This was least pronounced in the "nearest neighbour" approach, though not entirely absent. However, in areas with very high photo density, this approach also produced many more clusters that did not define a vernacular region. This may be partially due to the fact that in densely populated areas, the geographical size of each document was much smaller than with the other two approaches.

![Figure 6](image.png)

*Figure 6* Comparison of area for "grassmarket" as defined by all points in the raw dataset (blue area, green dots) and as defined by the nearest neighbour method (green area).

7. Conclusion

The methods used here are not flawless, however, they were able to successfully map various vernacular regions in the city of Edinburgh, and could do the same given similar data in any location. The new methods show qualities, such as ability to represent smaller or overlapping regions, that are lacking in a grid-based approach. The different methods used each have different strengths and further research should look to further improve them and fix weaknesses where possible.
The method used still includes some arbitrary figures, such as the radius/number of photos that define a document, however it lessens the dependence on arbitrarily defined zones and therefore limits the effect that the MAUP can have on the outcome. Further research could explore ways to further reduce arbitrary elements and make the output more strongly controlled by the data rather than other factors.

8. Acknowledgements

Thanks to Thomme et al (2016) for making the YFCC100M dataset available for research.

9. Biographies

Bryn Macaulay is a recent graduate in Environmental Science from the University of Stirling. His interests are focused on conservation and ecology, but he also has an interest in GIS, which was the focus of his thesis, which this paper expands upon.

Phil Bartie is Lecturer in Geospatial Technology at the University of Stirling. His research is centred in the overlap between models of space, mobile computing, and human-computer interaction. He has a particular interest in improving the effectiveness of interactions with mobile computers while exploring urban space.

William Mackaness is an expert in GIS in the School of GeoSciences at The University of Edinburgh with a particular interest in automated mapping. More recently his research has explored notions of place through vernacular geographies.

References


