

Approximation of Neighborhood Boundaries Using Collaborative Tagging Systems

Fabian Wilske

Institute for Geoinformatics, University of Münster. Germany
fabian.wilske@gmail.com

Abstract. Urban neighborhoods are regions which often do not have officially defined boundaries. In addition, there are neighborhoods which are designated by informal names (as in "Downtown" or "Docklands"). Their location and extent (the region's "spatial footprint") is a matter of individual perception. Thus such regions lack an official entry in geographical dictionaries, such as gazetteers. This paper introduces a method to approximate the spatial footprint of neighborhoods by collecting individual opinions about their location and then aggregating these opinions into a common idea of the region's extent. This is done by using spatial metadata that is annotated to user-contributed web content.

1 PROBLEM STATEMENT

Urban neighborhoods often do not have officially defined boundaries. London's inner city consists, for example, of various areas which have informal names or which still bear the names of former (official) boroughs (e.g. Bloomsbury, Fitzrovia, Soho). Indeterminate boundaries are not only the result of the temporal change of the formerly delimited regions. Other possible reasons can be simply incomplete information or boundaries which constitute continuous transitions (Kulik, 2001). See (Hadzilacos, 1996) for an overview concerning the different types of indeterminate or vague boundaries.

To find neighborhoods gazetteers are used. A gazetteer is a collection of geographic place names together with their geographic location and other descriptive information such as the feature type of the place (Hill et al., 1999). If the feature type is some kind of (vaguely bounded) region, e.g. a neighborhood, then the gazetteer only delivers the often arbitrary chosen centroid of the region or the bounding box which is its minimum bounding rectangle. The content of gazetteers is often derived from standard map series produced by national mapping agencies. As such they reflect an official or administrative view on geographic space (Jones et al., 2008). Thus gazetteers cannot directly be used to approximate the spatial extent of regions which have no officially defined boundaries.

In addition to neighborhoods with indeterminate boundaries, there are neighborhoods that are referenced by informal place names. Identifiers such as “West End”, “Downtown” or “Docklands” are components of a so-called vernacular geography (Waters and Evans, 2003; Pasley et al., 2007). Due to the informal name the referenced regions not only lack precise boundaries but also a precise referent, for example a landmark representing the region. Their spatial location and extent (also called the region's „spatial footprint“) is subject of individual perception. Thus such regions lack an official entry in geographical dictionaries, such as gazetteers.

To illustrate the mismatch of place perception the imprecise defined place name of London's “West End” is discussed. According to Wikipedia's description¹ “West End” can either be the entertainment district around Leicester Square and Covent Garden, or the shopping district located on Oxford Street, Regent Street and Bond Street, or even all areas of Central London that lie west of the City of London. Obviously its location is not only subjective but also depends on the user's specific context.

In this paper we introduce a method to approximate the spatial footprint of neighborhoods by making use of their subjective perceptions. One potential use of this approach could be the enhancement of local search services like Google Maps², Yahoo! Local³ or Microsoft's Live Search Maps⁴. The purpose of these services is to find lists of business of a given kind that satisfy some geographical constraint, e.g. “Hotels in West End”. Local search services use gazetteers as their primary source of geographical background knowledge (Schockaert and Cock, 2007). But as stated above, most gazetteers lack information about neighborhoods and other types of vaguely defined regions.

2 COLLABORATIVE TAGGING

A promising approach to approximate the spatial footprint of neighborhoods is to collect individual opinions about their location and aggregate these opinions into a common idea of the region's extent. Therefore, we use geotagged photos publicly available on photo sharing websites like Flickr⁵, Panoramio⁶ or Locr⁷. To retrieve photos and make them explorable by others, users of these websites are encouraged to label them with freely-chosen keywords, called tags. Many users tend to label their photos with one or more place names that reflect the photo's location

¹ http://en.wikipedia.org/w/index.php?title=West_End_of_London&oldid=193677087

² <http://local.google.com/>

³ <http://local.yahoo.com/>

⁴ <http://maps.live.com/>

⁵ <http://www.flickr.com/>

⁶ <http://www.panoramio.com/>

⁷ <http://www.locr.com>

of origin. There is a large amount of photos on these photo sharing websites which are indirectly geo-referenced through place names.

A special kind of tags are geotags. A geotag is a machine readable tag consisting of latitude and longitude coordinates and can be used to directly geo-reference or geocode web content. Photos that are annotated with a place name tag and a geotag can be used to create the spatial footprint of a place name.

3 METHOD

Obviously there is a need to acquire knowledge about the common perception of the extent of vague regions (Jones et al., 2008). This section describes a method to identify neighborhoods and approximate their vague boundaries by using metadata which are annotated to user-contributed photos.

3.1 Usage of some Terms

Before we describe the method we have to point out the usage of some terms:

A *Perceived Region* or *Place Name Region* $PR(p)$ of a certain place name contains all locations of photos that are tagged with a certain place name p .

A *Well-defined Region* $WR(PR)$ is a well-defined polygonal equivalent of a place name region. A well-defined region neither has to be unique nor does it have to exist for any place name. It is possible that there exist different well-defined regions for one place name region or that a place name region has no well-defined equivalent at all.

RCC-5 is an spatial algebra introduced by (Cohn and Gotts, 1996) that defines topological relations between regions. We use the relation $X\{ProperPart\}Y$ or $X\{PP\}Y$ if we want to say that a region X is entirely contained in region Y .

3.2 Identification of Perceived Regions

The first step on our approach is the identification of a place name region by detecting its geographical center. An appropriate measure of central tendency is the spatial median of a set of points. The spatial median is also known as the point of minimum aggregate travel (MAT) and minimizes the sum of absolute distances to all other points of the set (Longley et al., 2005):

$$MAT = \underset{x \in X}{\operatorname{argmin}} \frac{1}{n-1} \sum_{i=1}^n d(x, x_i)$$

where d is the Euclidean distance.

3.3 Approximation of Vague Boundaries

Next we approximate the (vague) boundary of a place name region. One established approach which formalizes regions with vague boundaries is the “egg-yolk” representation of regions with indeterminate boundaries” by (Cohn and Gotts, 1996). A vague region is represented as a pair of concentric regions with determinate boundaries: A core region (the “yolk”) that contains all locations that definitely belong to this region and a surrounding hull region (the “egg”) that contains all locations whose membership to the region is indeterminate. Every location not in any of both is definitely outside any boundaries of the vague region. The “egg” and “yolk” of an egg-yolk-pair are taken to represent possible “crispings” or precise versions of a vague region (Hazarika and Cohn, 2001). Any acceptable crisping must lie between the inner and outer limits defined by the yolk and the egg (see Figure 1). There is no need to specify exactly where the limits of the crisping lie. They just have to be somewhere in the “white” of the egg.

We use the egg-yolk model to represent spatial vagueness. In the following sections a method is presented to determine the egg and the yolk of a vague region.

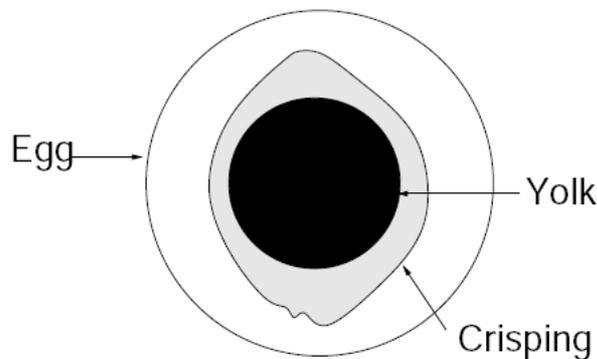


Figure 1: A classical egg-yolk structure, after (Hazarika and Cohn, 2001)

3.4 The Egg

We assume that the MAT is the most central location of our place name region and therewith definitely part of a region with a given place name. To get a reasonable choice about the egg and the yolk of each region we further assume that the average distance from one location to all other locations of photos

$$\bar{d}(x) = \frac{1}{n-1} \sum_{i=1}^n d(x, x_i)$$

gives us a measure of confidence whether a location belongs to a certain place name region or not. The larger the average distance from one location to all other locations the smaller the likelihood that the location is part of a place name region. We order the locations of each interesting place name region $PR(p)$ ascending by $\bar{d}(x)$ to decide whether a location belongs to the place name region, and in that case, whether it depends to the core or to the hull of a place name region. The hull should contain all locations of photos that are tagged by a certain place name. To determine such a region and therewith the egg in the sense of the egg-yolk-model it is necessary to construct a polygonal boundary that encloses $\text{hull}(PR)$. Therefore, we use the convex hull of a set of points X , which is defined as the smallest convex polygon that encloses X .

3.5 The Yolk

The challenge is now to find a reasonable choice about the core/yolk of a place name region, in other words to determine the locations that definitely belong to a region. Training regions with well-defined boundaries and their corresponding perceived regions are used to approximate the optimal threshold for the border between the yolk and the white. Boundary data for neighborhoods in the U.S., which are freely available from the real estate service company Zillow⁸, are used to define these training regions. Note that these data do not reflect an official or unique view on neighborhood boundaries and that they are defined for the purpose of real estate business. In contrast, one might expect that many of the photos we use are taken by visitors and tourists who tend to tag their photos using place names they have found in guide books. Hence, tourists' perception of the places they are visiting is not necessarily congruent with residents' perception of their neighborhoods.

To approximate the yolk of a vague region a maximum core region $\text{core}(PR)$ for each perceived region is determined. The convex hull of the core region is then entirely contained in the crisp boundaries of a corresponding well-defined region. We let “grow” perceived core regions, initialized by the MAT, until they grow beyond the boundaries of the corresponding well-defined region. Or, in more formal words: For each training region we take the set $PR(p)$, that is ordered by $\bar{d}(x)$, and investigate how many locations of $PR(p)$ are (or are not) contained in the corresponding well-defined region $WR(PR)$. Therewith we try to find a threshold quantile so that the convex hull of all locations below that quantile is entirely contained in $WR(PR)$.

⁸ <http://www.zillow.com/labs/NeighborhoodBoundaries.htm>

The growing process is illustrated in Figure 2 for the well-defined region "Upper West Side", New York City: The quantile of order 0.836 is the largest for which all locations below that quantile are inside the well-defined boundaries (Figure 2, bottom-left). By adding one more location to $core(PR)$, $ConvexHull(core(PR))$ grows beyond the boundaries of the corresponding well-defined-region (Figure 2, bottom-right).

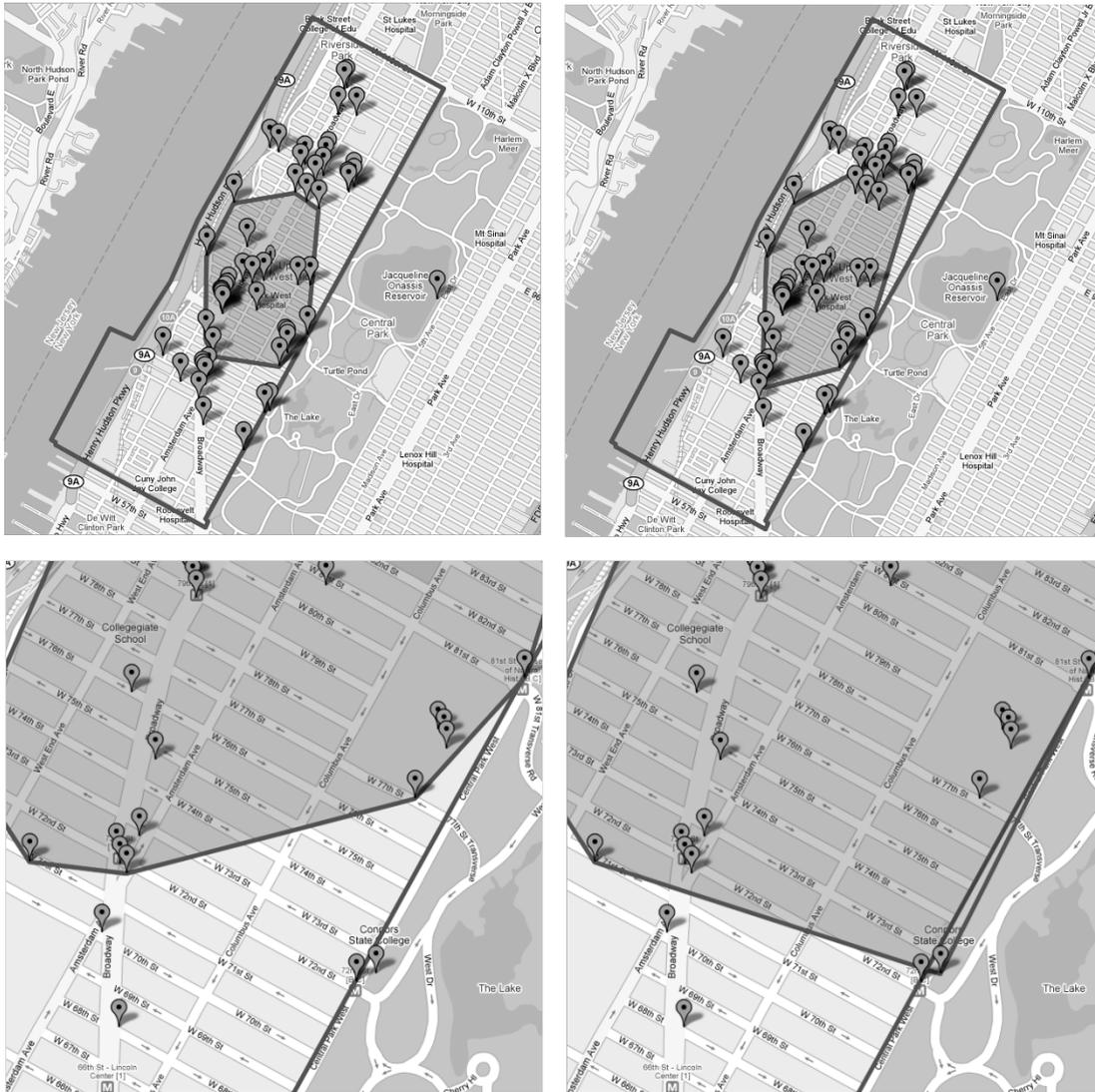


Figure 2: Growing process of the perceived core region "Upper West Side" where the order of the quantile is 0.50 (top-left), 0.75 (top-right), 0.836 (bottom-left) and 0.855 (bottom-right).

Next we determine the median quantile of all PR/WR pairs out of our training data set. This is the quantile below which the condition $ConvexHull(core(PR)) \{ProperPart\} WR(PR)$ holds for 50% of region pairs out of our training data set. We will use this quantile to approximate the yolk of vague regions that do not have corresponding well-defined regions.

4 UNSOLVED PROBLEMS AND CONCLUSION

There are several problems with photo sharing websites as well as with collaborative tagging systems in general. First, a requested place name can also have non-geographic meanings (homonyms). For example, “Finsbury” which is a London neighborhood as well as a famous brand of gin. Second, there are distinct places that have the same name (polysemes), as in London, England vs. London, Ontario. Third, the level of granularity at which a place name is chosen, varies among users. For example, different users tag the same photo either with “New York”, “Manhattan” or “Upper West Side“.

Despite these problems, which are well acknowledged in the field of Geographic Information Retrieval (e.g. by Amitay et al., 2004; Pasley et al., 2007; Jones et al., 2008), we believe the present approach is a useful first attempt to reflect common perception of vaguely defined regions. Current work has already shown that the combination of geocoded web content with freely chosen place name tags is a promising approach to judge about the spatial footprint of neighborhood regions. Further experiments are needed to confirm that the critical choice of the threshold quantile reasonably distinguishes the yolk from the egg of a perceived region. A region search engine is being implemented to test the introduced method.

REFERENCES

- Amitay, E., N. Har'El, R. Sivan and A. Soffer (2004). "Web-a-where: geotagging web content" SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. New York, ACM Press: 273-280.
- Cohn, A. G. and N. M. Gotts (1996). "The `Egg-Yolk' Representation of Regions with Indeterminate Boundaries" Geographic Objects With Indeterminate Boundaries. P. A. Burrough and A. U. Frank. London, Taylor & Francis: 171-187.
- Hadzilacos, T. (1996). "On Layer-Based Systems For Undetermined Boundaries" Geographic Objects With Indeterminate Boundaries. P. A. Burrough and A. U. Frank. London, Taylor & Francis: 237-255.
- Hazarika, S. M. and A. G. Cohn (2001). "A Taxonomy for Spatial Vagueness - An Alternative Egg-Yolk Interpretation" COSIT/FOIS Workshop on Spatial Vagueness, Uncertainty and Granularity.
- Hill, L, J Frew and Q. Zheng (1999). "Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library" D-Lib Magazine 1(5).

- Jones, C. B., R. S. Purves, P. Clough, and H. Joho (2008). "Modelling Vague Regions with Knowledge from the Web" *International Journal Geographic Information Systems (IJGIS)* (In Press).
- Kulik, L. (2001). "A Geometric Theory of Vague Boundaries Based on Supervaluation" *COSIT 2001: Proceedings of the International Conference on Spatial Information Theory*: 44-59.
- Longley, P. A., M. F. Goodchild, D. J. Maguire and D. W. Rhind (2005). "Geographic Information Systems and Science". John Wiley & Sons, 2nd edition.
- Pasley, R. C., P. D. Clough and M. Sanderson (2007). "Geo-tagging for imprecise regions of different sizes" *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*: 77-82.
- Schockaert, S. and M. D. Cock (2007). "Neighborhood Restrictions in Geographic IR" *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, ACM Press: 167-174
- Waters, T. and A. J. Evans (2003). "Tools for web-based GIS mapping of a "fuzzy" vernacular geography" *Proceedings of the 7th International Conference on GeoComputation*.