A study of data representation of natural features in OpenStreetMap

Peter Mooney, Padraig Corcoran, Adam Winstanley

Department of Computer Science, National University of Ireland Maynooth,
Maynooth,Co. Kildare. Ireland.
email {peter.mooney, padraig.corcoran, adam.winstanely}@nuim.ie

1. Introduction

OpenStreetMap (OSM) is a collaborative project to create a fully free and openly accessible map of the world. Volunteers in the OSM community collect geographic information and submit to the global OSM database (Ciepluch et al.; 2009). This paper investigates how natural features such as water bodies, forests, etc are represented in OSM. By representation we mean the following - given the type of natural feature, the geographical area of the feature and the spatial polygon representation of that feature, is there sufficient detail (enough sampling points or polygon nodes) present to provide a high quality spatial representation of the natural feature? This representation is important for a number of reasons - most notably in that good representation allows more accurate generalisation and simplification of the data at different scales (Fritz et al.; 2009) and provides better overall structure and quality of the shape model (Baldwin et al.; 1998). Real world geographic features are represented in OSM databases as points, lines, and polygons. Spatial attributes for these features are stored as *tags*. The **natural** tag describes geographic features which occur naturally. The natural tag has a large set of values { bay, beach cave entrance, cliff, coastline, fell, glacier, heath, land, marsh, mud, peak, scree, scrub, spring, tree, volcano, water (lakes, etc. and used to tag an area of permanent water), wetland, and wood }. In a similar fashion the **landuse** tag describes forest, managed forest, or woodland plantation, and preserved woodland which are not actively or regularly forested. Data can be submitted to OSM in three ways. Firstly volunteers can collect GPS traces and then upload them using one of the OSM editors available for this purpose. Secondly there is the option for bulk upload of spatial data (in GIS formats such as ESRI Shapefile) from authoritive sources such as TIGER data in the USA and Corine Land Cover in France. The third method, and potentially the most popular amongst the OSM volunteer community, involves tracing out lines and polygons from aerial imagery. Yahoo! have agreed to let OSM use their aerial imagery for the purposes of tracing. OSM volunteers can use any of the three main OSM editors (Potlatch, Merkaartor, or JOSM), to edit (trace) OSM map data over the Yahoo! imagery. Figure 1 shows a portion of the south-east corner of Dummer - a large lake in southern Lower Saxony (Germany) created in this fashion. The lake has a surface area of $13.5km^2$. The aerial imagery is taken from Google Maps. The red line shows the overlay of the OSM representation (OSM-ID=9086711) of the Dummer See. The OSM representation is a polygon with 109 vertices. Visually it is apparent that

there is a significant error between the actual aerial image (using the Google aerial image as a pseudo ground-truth) and the OSM representation due to severe under-representation of the ground-truth of the natural feature and very inaccurate tracing over aerial imagery using one of the OSM editors.

Figure 1: Using aerial imagery from Google Maps the OpenStreetMap representation of the Dummer See in Lower Saxony, Germany is overlayed

The remainder of this extended abstract is organised as follows. Section 2. describes our experimental setup for the analysis of polygons in OpenStreetMap data. Some results are provided in Section 3.. Section 4. closes the paper by providing some conclusions at this stage of the research and outlining our plans for future work on this problem.

2. Experimental Setup

All OpenStreetMap data was downloaded from the Geofabrik website `http://download.geofabrik.de`. Geofabrik provide up-to-date packages of OpenStreetMap data conveniently separated into countries and regions. All data was downloaded in OpenStreetMap XML (OSM-XML) format on June 26^{th} 2010. The OSM-XML for each country and region chosen was then directly imported to a PostGIS database using the **osm2pgsql** tool. **osm2pgsql** is a free and open source tool for converting OSM-XML to a set of PostgreSQL statements providing a means to build database tables to hold the OpenStreetMap data. **osm2pgsql** allows one to chose which tags are imported from the OSM-XML as the OSM-XML contains every tag for every feature in that region or country. To reduce the amount of spatial attribute data stored in the PostGIS database we chosen only to import a small subset of tags which included information on natural features, source attributions, and naming. A set of 10 packages were downloaded which included 8 countries (Austria, Denmark, Estonia, Iceland, Ireland, Latvia, Scotland, Switzerland) and 2 regions (Lower Saxony (DE), Bretagne (FR)). These countries and regions were chosen in order to provide variance in the OSM databases. Austria, Estonia, Switzerland, Bretagne, and Lower Saxony have bulk uploaded publicly available government spatial data providing national scale coverage. Iceland,

Ireland, and Scotland are a mixture of OSM volunteer data collection and aerial imagery tracing. Denmark is a mixture of all three approaches.

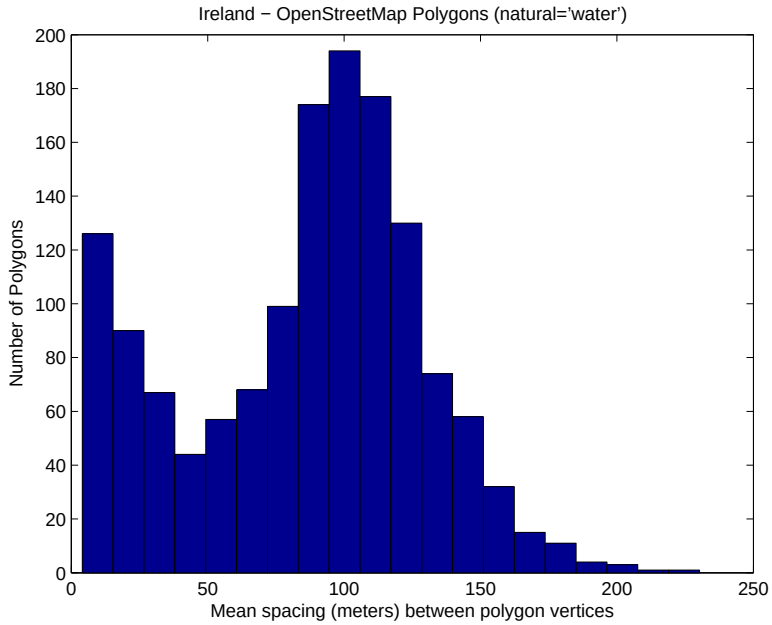

Figure 2: All "water" polygons in Ireland - distribution of mean distance between vertices of polygons

3. Results

In this section we provide a summary of some of the current results from this study. Considering each polygon as a shape embedded in $2-D$ metric space (Ying et al.; 2010) allowed us to calculate many shape-descriptors including: turning angle ratio, convexity, circularity, rectangularity, etc. However in this paper we focus on spatial sample point characteristics of the polygons. Figure 2 shows a statistical distribution of the mean distance between connected vertices of polygons representing "water" features in the OSM Ireland database. Clearly over 50% of water features are represented by polygons where the mean distance between adjacent polygon vertices (sample points) is greater than 50 meters. In Figure 3 the "forest" polygons for OSM Switzerland are shown. Two characteristics of the polygons are represented: (1) the mean distance (meters) between adjacent nodes in a polygon to provide insight into how close the data points actually are and (2) the perimeter of the polygon (in meters) divided by the number of nodes N in the polygon which would indicates a theoretical equidistant spacing between all adjacent nodes in the same polygon. The two distributions in Figure 3 are statistically very similiar. The distribution of mean distance between sample points is greater for larger spacing groups - from $150-200$ meters onwards. The high distribution of P/N for $0-50$ and $50-100$ are a result of polygons with both a small perimeter and a small number of nodes. Table 1 provides an analysis of all OSM polygons tagged as "forest" for the OSM databases for the 10 countries and regions selected while Table 2

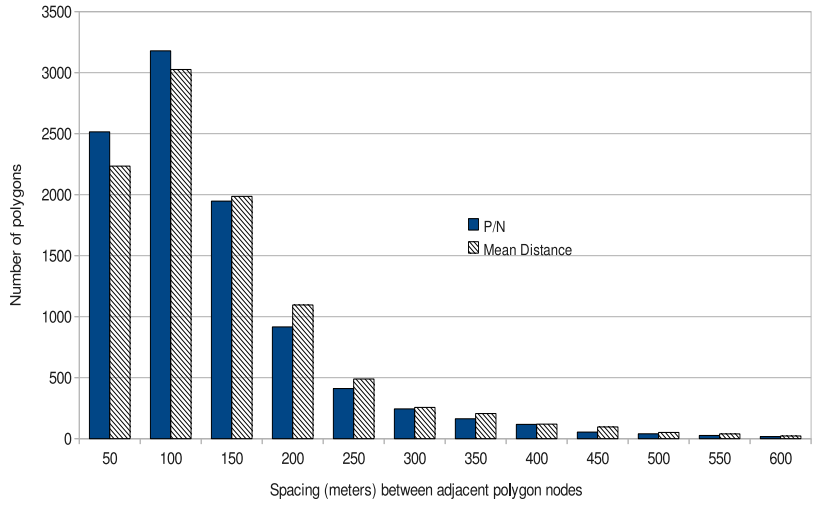

Figure 3: Comparison of "forest" polygons in Switzerland - The mean spacing between adjacent polygon nodes (in meters) and perimeter (meters) divided by the number of nodes are shown

provides an analysis of all OSM polygons tagged as "water" for the OSM databases for the 10 countries and regions selected. Tables are sorted in ascending order of N the number of "forest" polygons in the OSM database for the corresponding country or region. The column min indicates the minimum distance (in meters) between any two sample points in any of the "forest" polygons for that country/region. The column $mean$ indicates the mean distance between adjacent points, while $median$ and $StdDev$ are the median and standard deviation of distance in meters between adjacent sample points. The 95% column provides the 95% percentile of all adjacent distance spacing. There are a number of interesting observations from both Table 1 and Table 2. Firstly the mean spacing meters between nodes on "forest" polygons is much greater than the mean spacing in meters between nodes on "water" polygons. For example Switzerland the mean spacing for "forest" polygons is 105.96 while the mean spacing for "water" polygons is 40.57. There are a number of possible reasons for this: waterbodies are in many cases easier for volunteers to survey with GPS devices - walking around the edge of a lake or pond. The deliniation between a what consitutes a "forest" boundary and the adjacent landcover may confuse many contributors (Comber et al.; 2005). Aerial tracing of forests from poor resolution aerial imagery may mean that less data points are collected. Several authors (Pinto-Coelho et al.; 2010; de Solla et al.; 2005) have shown that in some applications domains such as Biodiversity studies poor spatial sampling rates can greatly impinge on the quality of the scientific assessment.

4. Conclusions and Further Work

Given the dynamic and organic nature of the spatial data contained in the OSM database the statistical results of this study could change dramatically over a relatively short space of time. Edits to current spatial data in the OSM database and upload and submission of newly

Table 1: Forests: Spacing (in meters) between sample points for forest polygons

Country	**N**	**Min**	**Mean**	**Median**	**StdDev**	95.00%
Iceland	21	25.91	105.95	89.21	65.93	230.36
Ireland	388	10.17	157.45	153.02	92.59	291.09
Scotland	1030	9.26	147.25	114.64	115.87	369.55
Latvia	1668	2.76	141.55	118	82.56	319.89
Bretagne	2953	7.92	91.02	89.31	24.96	129.87
Denmark	2959	1.11	94.36	77.52	70.07	224.5
Switzerland	9664	3.96	105.96	86.48	81.29	263.84
Lower Saxony	11713	2.59	100.19	82.54	77.17	245.28
Austria	13176	6.11	90.55	75.03	64.30	204.12
Estonia	13263	5.38	124.67	122.58	34.82	178.81

Table 2: Water Features: Spacing (in meters) between sample points for water polygons

Country	**N**	**Min**	**Mean**	**Median**	**StdDev**	95.00%
Estonia	923	1.41	63.07	38.11	74.57	156.5
Bretagne	1109	2.14	36.56	25.54	31.19	91.47
Ireland	1342	4.03	85.52	91.68	52.97	149.68
Latvia	1343	2.83	101.4	91.01	74.56	230.2
Switzerland	1620	0.77	40.57	26.26	44.73	122.24
Denmark	2316	0.67	43.13	27.21	46.51	131.15
Iceland	3571	1.11	76.99	79.1	43.54	168.74
Austria	3906	0.67	40.95	29.21	40.02	114.18
Scotland	4382	0.72	66.64	57.34	49.04	159.39
Lower Saxony	6992	1.01	40.45	25.89	43.25	125.56

collected and captured spatial data can happen quickly. Many of the features analysed are under-represented (in terms of the number of points used to represent their polygon in OSM) while it can be argued that other features are slightly over-represented (small urban green spaces and golf courses are often sampled at very high GPS epoch-rates). Under representation is an artefact of several aspects of OSM data collection: differing levels of GIS skills amongst OSM volunteers (Qian et al.; 2009), problems in surveying physically inaccessible features such as lakes, quarries, etc and differences in accuracy of equipment and methods used to survey and capture data. Tracing over aerial imagery is problematic if care is not taken to carefully recreate the spatial outline of the underlying shape with low resolution aerial imagery used for tracing. In most cases over-representation, once it does not impact on data quality, is not problematic. However under-representation of natural features has more serious consequences OSM such as rendering it unsuitable, at present, for use in certain earth science applications such as: ground-truthing of remotely sensed imagery (Baraldi et al.; 2005) where, for example, obtaining good ground-truth data is crucial for quantitative analysis of landcover classification techniques (Corcoran et al.; 2010). An important issue for further work is the establish the quality of the OSM representation of "water" and "natural" polygons (and other features) against an established ground-truth dataset. At the GIScience conference we will provide results of a quantitative comparision between Irish OSM data and grouth-truth spatial data. This comparision will help establish if some of the representation issues mentioned in this paper are present in the Ordnance Survey datasets or are uniquely an artefact of the OSM data collection methodology. We will also investigate if representation issues are strongly correlated with the data collection methodology.

Acknowledgements

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. Peter Mooney is funded by the Irish Environmental Protection Agency STRIVE programme (grant $2008 - FS - DM - 14 - S4$). Padraig Corcoran gratefully acknowledges the support of the Department of Computer Science NUIM.

References

Baldwin, B., Geiger, D. and Hummel, R. (1998). Resolution-appropriate shape representation, pp. 460 –465.

Baraldi, A., Bruzzone, L. and Blonda, P. (2005). Quality assessment of classification and cluster maps without ground truth knowledge, *Geoscience and Remote Sensing, IEEE Transactions on* **43**(4): 857 – 873.

Ciepluch, B., Mooney, P., Jacob, R. and Winstanley, A. C. (2009). Using openstreetmap to deliver location-based environmental information in ireland, *SIGSPATIAL Special* **1**: 17–22.

Comber, A., Fisher, P. and Wadsworth, R. (2005). What is land cover?, *Environment and Planning B: Planning and Design* **32**(2): 199–209.

Corcoran, P., Winstanley, A. and Mooney, P. (2010). Segmentation performance evaluation for object-based remotely sensed image analysis, *International Journal of Remote Sensing* **31**(3): 617–645.

de Solla, S. R., Shirose, L. J., Fernie, K. J., Barrett, G. C., Brousseau, C. S. and Bishop, C. A. (2005). Effect of sampling effort and species detectability on volunteer based anuran monitoring programs, *Biological Conservation* **121**(4): 585 – 594.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F. and Obersteiner, M. (2009). Geo-wiki.org: The use of crowdsourcing to improve global land cover, *Remote Sensing* **1**(3): 345–354.

Pinto-Coelho, R. M., Brighenti, L. S., Bezerra-Neto, J. F., Jr., C. A. M. and Gonzaga, A. V. (2010). Effects of sampling effort on the estimation of spatial gradients in a tropical reservoir impacted by an oil refinery, *Limnologica - Ecology and Management of Inland Waters* **40**(2): 126 – 133.

Qian, X., Di, L., Li, D., Li, P., Shi, L. and Cai, L. (2009). Data cleaning approaches in web2.0 vgi application, pp. 1 –4.

Ying, F., Mooney, P., Corcoran, P. and Winstanley, A. (2010). Polygon processing on openstreetmap xml data, *in* M. Haklay, J. Morely and H. Rahemtulla (eds), *Proceedings of the GIS Research UK 18th Annual Conference*, University College London, London, England, pp. 149–154.