Towards Quality Metrics for OpenStreetMap

Peter Mooney*
Department of Computer Science
NUI Maynooth
Co. Kildare. Ireland
p.mooney@epa.ie

Padraig Corcoran
Department of Computer Science
NUI Maynooth
Co. Kildare. Ireland
padraigc@cs.nuim.ie

Adam C. Winstanley
Department of Computer Science
NUI Maynooth
Co. Kildare. Ireland
adamw@cs.nuim.ie

ABSTRACT

Volunteered Geographic Information (VGI) is currently a "hot topic" in the GIS community. The OpenStreetMap (OSM) project is one of the most popular and well supported examples of VGI. Traditional measures of spatial data quality are often not applicable to OSM as in many cases it is not possible to access ground-truth spatial data for all regions mapped by OSM. We investigate to develop measures of quality for OSM which operate in an unsupervised manner without reference to a "trusted" source of ground-truth data. We provide results of analysis of OSM data from several European countries. The results highlight specific quality issues in OSM. Results of comparing OSM with ground-truth data for Ireland are also presented.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial databases and GIS

Keywords

Quality of Spatial Data, Shape Representation, OpenStreetMap

General Terms

Human Factors

1. INTRODUCTION

With the expanding availability and accessibility of GIS data and their various applications, often different from the purpose of the original data set, the characterization and quality evaluation of GIS data sets has become increasingly important. The ever increasing volume of georeferenced data being generated, transferred, and utilized and the amount of uncertainty embedded in spatial databases has become a major issue of crucial theoretical importance and practical consideration. At the same time Volunteered Geographical Information (VGI) is becoming a very important source of geographical information [10]. OpenStreetMap (OSM) is one of best known sources of VGI. The majority of OSM data is collected by "non specialists" and "amateur geographers" [6] giving rise to serious concerns in the professional GIS community surrounding the quality of OSM. Few examples appear in the literature where OSM data has been used for GIS modeling, spatial analysis, or spatial statistics. Over *et. al* [13] describe the development of $3D$ models for cities using OSM data combined with Digital Terrain Model (DTM) data but the authors comment that while "in Germany the OSM street network database is nearly complete OSM data is not widely used in the geoinformatics". Boin and Hunter [2] state that for data consumers to use VGI they "need some measure of the quality of the data to make informed choices towards reducing or absorbing possible uncertainty in the spatial data". In this paper we investigate the development of quality metrics for OSM: both for use in isolation (analysis of OSM data without ground-truth comparisons) and alternatively with access to ground-truth data. Experimental results show some serious problems with OSM data. OSM was founded in 2004 and has grown from modest beginnings to over over $200,000$ contributors at the end of 2009. Haklay [11] shows that in England by March 2010 OSM coverage of England had grown to 69.8% from 51.2% a year previous. Zielstra and Zipf [16] comment that in Germany in 2009 the amount of OSM data increased by 20% in under three months. Ciepluch *et. al* [5] give a detailed overview of the steps involved in establishing an OSM database and server system. The standard means of collecting and uploading data to OSM is by: (1) collecting data using GPS devices or (2) tracing outlines of polygons, polylines, etc from publicly available aerial imagery. Yahoo! have agreed to let OSM use their aerial imagery for the purposes of OSM tracing. Landsat satellite imagery, produced by NASA, can also be used as a source for OSM. These two processes of data upload to OSM are often rapidly accelerated when import activity extends to importing government or mapping agency spatial data provided it is usable under the OSM license. The paper is organized as follows. A discussion of spatial data quality with specific emphasis on the VGI domain is provided in section 2. The experimental analysis of OSM data is outlined in section 3. The paper closes with section 4 where we provide some discussion of the results from section 3 and the possible implications of these results. Issues for ongoing and future work are also outlined.

*Also at: Environmental Research Centre, EPA, Ireland.

ACM GIS '10 , November 2-5, 2010. San Jose, CA, USA
(c) 2010 ACM ISBN 978-1-4503-0428-3/10/11...$10.00.

2. DATA QUALITY AND VGI

In this section we give an overview of the current literature on spatial data quality in VGI. Flanagin and Metzger [7] state that as the amount of VGI continues to grow "the issues of credibility and quality should assume a prominent place on the research agenda". Bulterman [4] suggests that the "complete disregard for documentation of data resources" has made it almost impossible for one to perform a fitness for use/purpose evaluation on data resources. For GIS data the lack of documentation of quality controls, measurement methods, etc may actually be an artifact of previous practices within the professional GIS community. Goodchild [9] remarks that in GIS it is often "common to remove any information that might link GIS layers to original measurements and thus to present data in a way that makes any conventional error analysis impossible".

Without some quantitative measures of accessing the quality of the OSM data the GIS community has been slow to consider OSM as a serious source of data. Flanagin and Metzger [6] remark that for VGI in general the "professional and scientific gate-keeping that usually filters and reviews data may not be present in sufficient forms and subsequently can lead to information which is prone to being "poorly organized, out-of-date, incomplete, or inaccurate. Some results of OSM data quality analysis are beginning to appear in the literature. Haklay [11] describes a comparison of the road network in OSM for England with the road network in the Ordnance Survey UK Meridian dataset and concludes that OSM is "as good if not better than the Meridian dataset in terms of positional accuracy". However he emphasises "serious issues about completeness". The recent study by Zielstra and Zipf [16] of OSM and TeleAtlas for Germany shows that "while professional data is not without it's faults the coverage of OSM in rural areas is too small to be seriously considered a sophisticated alternative for *any* applications". However the study does conclude that for larger cities (Berlin, Frankfurt, Munich) the data diversity is so rich that "OSM is replacing proprietary data for many projects". In Over *et al* [13] the authors comment that the quality control of OSM differs fundamentally from professionally edited maps. The community-based approach allows anyone to upload and alter map data. Due to the huge number of editors , errors and conflicts are usually quickly resolved. In urban areas changes in the road network appear in the OSM data set long before appearing in other map data providers.

3. RESULTS AND ANALYSIS

In this section we provide the results of some analysis of OSM data for Ireland and several other European countries.

3.1 Experimental setup

All OSM data was downloaded, in OSM-XML format, from the Geofabrik service [8] and are correct as of September 1^{st} 2010. Eleven countries and regions are studied in this paper: Wales UK(W), Bretagne in France FR(B), Ireland IE, Latvia LV, Switzerland CH, Denmark DK, Estonia (EE), Iceland IS, Austria AT, Scotland UK(S), Spain (ES), and Lower Saxony in Germany DE(LS). Ireland is used as a case-study with ground-truth data, Austria and Lower Saxony contain publicly available government-generated spatial data, while Bretagne contains Corine Land-Cover mapping data. Scotland, Wales, and Latvia as they are of comparable size to Ireland. OSM in Estonia contains full national coverage of natural features from publicly available government-generated spatial data. Spain is chosen as a large country with comparably poor OSM coverage. Ordnance Survey Ireland (OSI) data at $1:5000$ scale of the lakes in Ireland was used.

Table 1: Spacing (m) between nodes in water polygons

Loc	N	$\bar{s}$	$\tilde{s}$	μ	95.00%
FR(B)	1109	36.56	25.54	31.19	91.47
DE(LS)	6992	40.45	25.89	43.25	125.56
CH	1620	40.57	26.26	44.73	122.24
AT	3906	40.95	29.21	40.02	114.18
DK	2316	43.13	27.21	46.51	131.15
EE	923	63.07	38.11	74.57	156.5
ESP	1580	65.19	37.45	74.76	208.54
UK(S)	4382	66.64	57.34	49.04	159.39
UK(W)	436	67.04	58.6	53.32	151.49
PL	12063	76.37	57.49	59.75	188.79
IS	3571	76.99	79.1	43.54	168.74
IE	1342	85.52	91.68	52.97	149.68
LV	1343	101.4	91.01	74.56	230.2

3.2 Polygon Respresentation in OSM

In this section we analyze how OSM polygons are sampled by contributors and how this impacts on the data representation of natural features such as lakes, ponds, and forests. In table 1 a summary of an analysis of the spacing, in meters, between samples points in OSM polygons representing water features in the 11 different countries and regions in Europe is shown. N represents the number of water features, $\bar{s}$ the mean spacing s between polygon nodes for all polygons, $\tilde{s}$ the median spacing, μ the standard deviation, and the 95^{th} percentile. The same analysis is provided in table 2 for polygons representing forests and woodland features. Both tables are sorted by $\bar{s}$ in ascending order. The top 5 ranked databases: (FR(B), DE(LS),CH,AT, and DK) are countries and regions where bulk data imports of government and mapping agency data to OSM were performed. The ranking changes slightly for forest features in table 2. In almost every case $\bar{s}$ for water polygons is less than $\bar{s}$ for forest polygons in the same country. This could indicate that it is easier for volunteers to physically sample water features or trace their outline from aerial imagery. The precise bounds of a forest/woodland can be difficult to measure.

3.3 Tagging and Documentation

When data is uploaded or edited in OSM users can *tag* or annotate this data. The OSM community has a democratically accepted ontology of tags described on the OSM wiki[12]. Provided a tag has a set of verifiable values it can be part of the ontology. A number of special tags in the ontology are provided to allow annotation of OSM data. These include: `source`, `description`, `attribution`, and `source-url`. We analysed the usage of these annotation tags on lines and polygons (ways) for all eleven countries and the results are tabulated in table 3. N is the total number of ways, T is the number of ways which have tags, and `tags(T)` is the number of these ways which have source description and

Table 2: Spacing (m) between nodes in forest polygons

Loc	N	$\bar{s}$	$\tilde{s}$	μ	95.00%
UK(W)	436	67.04	58.6	53.32	151.49
AT	13176	90.55	75.03	64.30	204.12
FR(B)	2953	91.02	89.31	24.96	129.87
DK	2959	94.36	77.52	70.07	224.5
ESP	748	99.55	75.13	75.76	244.40
DE(LS)	11713	100.19	82.54	77.17	245.28
IS	21	105.95	89.21	65.93	230.36
CH	9664	105.96	86.48	81.29	263.84
PL	33033	107.01	92.12	61.29	226.29
EE	13263	124.67	122.58	34.82	178.81
LV	1668	141.55	118	82.56	319.89
UK(S)	1030	147.25	114.64	115.87	369.55
IE	388	157.45	153.02	92.59	291.09

attribution tags. The overall usage of metadata enhancing tags is disappointing. With the exception of AT the usage of source description and attribution tags is almost negligible even for countries with third-party bulk contributed data. Brando and Bucher [3] argue that the quality of VGI is enhanced if proper metadata or tags are created and maintained which detail: types of changes and edits, methods of survey and collection, and finally a fitness for purpose statement. These tags provide would-be users of OSM data a means of evaluating the data's fitness for purpose, lineage, and fitness for usage. However, as argued by Skageby [14] argues that the "rewards of tagging are very hard to calculate for content contributors".

Table 3: The use of data source attribution tags for all ways (lines and polygons) in a given country/region

Loc	N	T	tags(T)	Total
AT	525258	283858	139795	49.2%
DK	181352	36155	1777	4.9%
UK(S)	162908	61115	2743	4.5%
UK(W)	98495	35989	1353	3.8%
DE(LS)	598852	212229	7154	3.4%
PL	556439	246767	5526	2.2%
IS	22193	9738	198	2.1%
IE	142289	46042	650	1.4%
ES	591336	176529	1453	0.8%
CH	573743	291968	1800	0.6%
EE	183124	157494	397	0.3%
FR(B)	423302	338305	414	0.1%

3.4 Shape Similarity Tests

The data available for download from OSM is at full resolution and has not undergone any simplification or generalisation. We now present results from determining the shape similarity of OSM polygons representing lakes and the corresponding lakes in the OSi lakes dataset. To determine the shape similarity between two polygons we implemented the turning-function shape similarity metric of Arkin et al. [1]. The boundary of a polygon A can be represented by a turning-function $\Theta_A(s)$ and the polygon is rescaled such that the total perimeter is 1. The similarity

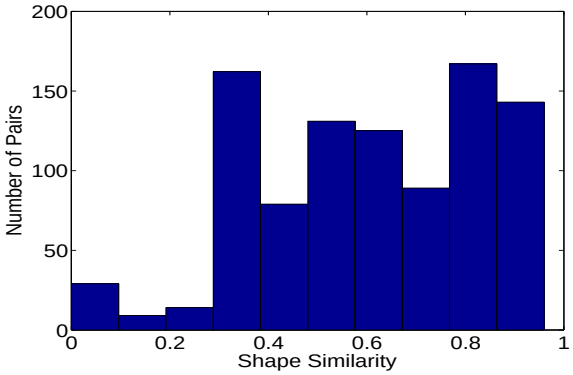

Figure 1: Shape similarity of 900 pairs of OSM and OSI lake polygons

between A and B can be determined by the distance between $\Theta_A(s)$ and $\Theta_B(s)$ according to a given metric function space [1]. Equation (1) defines the L_p distance between A and B where $\|.\|_p$ denotes the L_p norm. It is necessary to solve equation (1) and Arkin et al. showed that $d_2(A,B)$ can be computed by initially finding the optimal θ and then solving for t. This metric returns values in $[0,\infty]$ which are normalized to the range $[0,1]$ by Equation 2. In normalized form a value of 1 corresponds to identical polygons while as this value approaches 0 the polygons become more dissimilar.

$$d_p(A,B) = \left(\min_{\theta \in R, t \in [0,1]} \int_0^1 |\Theta_A(s+t) - \Theta_B(s) + \theta|^p ds \right)^{\frac{1}{p}} \quad (1)$$

$$nd_p(A,B) = \frac{1}{1+d_p(A,B)} \quad (2)$$

Using the nd_2 metric of Equation 2 we calculated the shape similarity between 900 corresponding pairs of polygons from the OSM and OSI datasets. Visual analysis, by three participants, of a randomly selected subset of 100 pairs found that a similarity value of 0.8 or greater corresponded to very similar polygons. On the other hand, a similarity value of 0.5 or less corresponded to very dissimilar polygons. There are a total of 12080 OSI polygons and 1722 OSM polygons. Only 900 OSM polygons were directly comparable with the OSI due to fragmentation of large waterbodies into several smaller polygons and incorrect mapping of lakes. Results of the shape similarity analysis is presented in Figure 1 as a histogram. Just over 30% of OSM polygons have shape similarity of 0.8 or greater with their corresponding OSI polygons. More than 52% of OSM polygons can be considered as completely dissimilar to their corresponding OSI polygons most likely due to the under-representation (poor sampling) of OSM polygons. The larger the number of nodes in a polygon the more complex the shape of the turning function. Computing shape similarity of an OSM polygon P against the corresponding OSI polygon Q where the $nodes(Q) >> nodes(P)$ will yield a low value for $nd_p(P,Q)$. We analyzed the number of nodes in both the OSI and OSM Lakes dataset. Of the 12080 OSI polygons 75% of polygons contain between 0 and 50 nodes. For the 1722 OSM polygons this rises to almost 92% of polygons. The OSM polygons, the majority of which contain between 50 and 100 nodes, represent waterbodies with a range of areas from 0.13 Hectares to 410.9

Hectares. For OSI polygons, with the same number of nodes, this corresponds to waterbodies with areas within the range from 0.09 Hectares to 70.77 Hectares and highlights a more rigourous and accurate physical sampling of these natural features.

4. SUMMARY AND CONCLUSIONS

This paper has provided an overview of spatial data quality in VGI and OSM and is a first step towards the development of quality metrics for OSM data. Analysis in section 3 investigated OSM under several headings: data coverage, feature representation, sampling practices, annotation (metadata) of contributed data, and comparison with a ground-truth dataset. There are a number of important outcomes. Cities and towns lend themselves to easy data gathering while the mapping of rural areas or rugged terrain requires some appreciation of land cover classification and rigourous sampling. This is set against the backdrop of what Haklay [11] calls "excitement of engagement" to map certain areas in which circumstances OSM volunteers rushing to map features may inadvertently under-represent natural features. A US Geological Survey [15] data quality workshop concluded that "all quality considerations are use-case sensitive in VGI where quality depends on what the (VGI) data will be used for". Haklay [11] and Zielstra and Zipf [16] show OSM has many possibilities to obtain good data quality, reasonable and useful coverage, and an effective basis for GIS analysis, without the overheads of paying large fees for proprietary data. GIS experts considering using OSM can do so with the understanding that the data is of variable quality and this should be built into their analysis. While it is early in the lifetime of VGI and OSM many experts agree (outputs from USGC Workshop[15]) that "the quality of VGI might be quantifiable *some day* and definite statements will be possible about VGI data quality". There are a number of issues for future work. First, to obtain a full view of representation of natural features across Europe it may help to assess if there are local, regional, or national trends of differences to how mapping is performed. Second, there is the temporal aspect to consider - how often do features get updated and does this only happen to the larger *more popular* features. Blighted for years by issues of uncertainty and data quality the GIS community will require strong evidence of the quality OSM data. Finally, OSM specific quality metrics are required. What if it is not possible to obtain a ground-truth dataset to *measure* VGI (in our case OSM) against? Another issue is communication of quality. How should the *quality* of VGI be communicated to potential users? A closer survey of the VGI requirements of the GIS community is required.

5. ACKNOWLEDGMENTS

Ordnance Survey Ireland (OSi) data is supplied as part of the STRAT-AG project which is funded by a SRC grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support. Peter Mooney is funded by the Irish EPA STRIVE programme (2008-FS-DM-14-S4). Padraig Corcoran gratefully acknowledges the support of the Dept. of Comp. Sci. NUIM.

6. REFERENCES

[1] E. M. Arkin, L. P. Chew, D. P. Huttenlocher, K. Kedem, and J. S. B. Mitchell. An efficiently computable metric for comparing polygonal shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(3):209–216, 1991.

[2] A. T. Boin and G. J. Hunter. What communicates quality to the spatial data consumer? In A. Stein, W. Bijker, and W. Shi, editors, *Quality Aspects in Spatial Data Mining*, pages 140–147. CRC Press, 2008.

[3] C. Brando and B. Bucher. Quality in user generated spatial content: A matter of specifications. In M. Painho, M. Y. Santos, and H. Pundt, editors, *Proceedings AGILE 2010: The 13th AGILE International Conference on Geographic Information Science*. Springer Verlag, Guimarães, Portugal, 2010.

[4] D. C. A. Bulterman. Is it time for a moratorium on metadata? *IEEE MultiMedia*, 11(4):10–17, 2004.

[5] B. Ciepluch, P. Mooney, R. Jacob, and A. C. Winstanley. Using openstreetmap to deliver location-based environmental information in ireland. *SIGSPATIAL Special*, 1(3):17–22, 2009.

[6] A. J. Flanagin and M. J. Metzger. Site features, user attributes, and information verification behaviours and the credibility of web-based information. *New Media & Society*, 9(2):319–342, 2007.

[7] A. J. Flanagin and M. J. Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72:137–148, 2008.

[8] Geofabrik. What is OpenStreetMap. Online at `http://www.geofabrik.de/geofabrik/openstreetmap.html` - checked June 2010, June 2010.

[9] M. F. Goodchild. A general framework for error analysis in measurement-based GIS. *Journal of Geographical Systems*, 6(4):323–324, 2004.

[10] M. F. Goodchild. Neogeography and the nature of geographic expertise. *Journal of Location Based Services*, 3(2):82–96, 2009.

[11] M. Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and Planning B: Planning & Design*, 4(37):682–703, 2010.

[12] OpenStreetMap. *Map Features* page. Online at `http://bit.ly/HqeO7` : Sept 2010, March 2010.

[13] M. Over, A. Schilling, S. Neubauer, and A. Zipf. Generating web-based 3d city models from openstreetmap: The current situation in germany. *Computers, Environment and Urban Systems*, In Press, Corrected Proof:–, 2010.

[14] J. Skageby. Semi-public end-user content contributions–a case-study of concerns and intentions in online photo-sharing. *International Journal of Human-Computer Studies*, 66(4):287 – 300, 2008.

[15] USGS. The u.s. geological survey volunteered geographic information workshop. Workshop held at Herndon, VA, USA. Online at `http://cegis.usgs.gov/vgi/index.html`, Jan 2010.

[16] D. Zielstra and A. Zipf. A comparative study of proprietary geodata and volunteered geographic information for germany. In M. Painho, M. Y. Santos, and H. Pundt, editors, *AGILE 2010: 13th AGILE International Conference on Geographic Information Science*. Springer Verlag, Guimarães, Portugal, 2010.