MVP OSM: A TOOL TO IDENTIFY AREAS OF HIGH QUALITY CONTRIBUTOR ACTIVITY IN OPENSTREETMAP

Maurizio Napolitano and Peter Mooney

Abstract

OpenStreetMap's success continues to grow and contributions are not limited to the collection of spatial data using GPS (Global Positioning System) equipment. A very wide range of software tools developed by, and available to, the OSM community means that at present, anyone can also make a contribution through, for example, tracing aerial imagery, directly importing data, or by adding spatial information retrieved from smartphones. Consequently 'the map' has become increasingly rich, but the quality of the data is very often questioned and comes under scrutiny from the GIS and LBS (Location-Based Services) communities. By examining the world map generated from OpenStreetMap, it is relatively easy to identify areas which are more or less well supported in community mapping activities; a very high level of spatial detail in certain areas can indicate the quality of OSM data. MVP OSM is a software tool designed to highlight areas in OpensStreetMap where users (contributors) are dedicated to providing high levels of spatial detail. This usually correlates with the use of a GPS and on-the-ground mapping, or, at the very least, a deep local knowledge of the area and an inherent desire to see it represented in the highest level of detail on OpenStreetMap. The input to MVP OSM is an OSM XML file, which is converted by Python into a file for spatialite (the GIS extension for sqlite). Within spatialite the data is processed to create clusters and using these spatial clusters, the tool can then derive the activity of single or multiple users in that area. Vector layers and heatmaps are generated as output that can be overlaid onto OSM maps. A high level of detail can be considered a good indicator of the quality of OSM data within a given area. The MVP OSM tool hides the details of OSM XML processing, which many researchers find difficult, and processes the data to produce very useful visualizations of contributor activity in any given OSM area.

Keywords

OpenStreetMap, Volunteered Geographic Information, Crowdsourcing, Data Quality, GIS

1 Introduction

The OpenStreetMap (OSM) project is one of the best examples of volunteered geography.[1] The concept behind the project is similar to Wikipedia: everyone contributes to the creation of a freely available map of the world.[2] Contributions are made in many different ways, but unlike Wikipedia, OSM contributors use different tools to update the spatial information in the OSM database. These tools range from the Flash application (Potlatch2) integrated into the main OSM website to external applications (JOSM, Merkaator, QGIS, and so on), which involves using the OSM API.[3] There are also many different means by which spatial data are collected by contributors. The principal methods for contribution are tracing from aerial imagery (of which the imagery providers have given permission for use in OSM) or using GPS tracks that have physically been collected by contributors. Occasionally, certain spatial datasets with suitable licenses allowing use in OSM are imported in bulk. After the initial contribution of spatial data, contributors can then use any of the tools mentioned above to edit the map data. This spatial-data lifecycle results in constant activity, with changes and updates to the world map shown in near real-time on the OSM website. OSM has also proved highly successful in the rapid production of maps useful for the management of natural disasters or human rights. The best examples are the response of OSM after the earthquake in Haiti[4] and the mapping of the slums of Kibera. The role played by OpenStreetMap in helping the post-earthquake recovery in Haiti was recognized within the Disaster Relief 2.0 report by the UN Foundation. The Humanitarian OpenStreetMap Team (or HOT) has been responsible for seven mapping missions to Haiti since the earthquake. When the earthquake struck Haiti, citizens of the Port-au-Prince and emergency services alike found themselves without quality mapping products for the city. Aerial imagery was donated to OSM for Haiti, volunteer mappers traced aerial imagery, and in a matter of hours the OpenStreetMap community had produced maps that were being loaded into GPS devices being used by rescue teams. Similarly, the Kibera slum in Nairobi, Kenya, is not recognized by the local government and so the project ‘Map Kibera’ has created a map of this area with roads, place names, first aid locations, places of worship, water, and so on. These, and other successful case studies, have helped to increase confidence in the OSM project.

Despite these successes and the work by Haklay (2010), Ciepłuch *et al*. (2010), and Mooney *et al*. (2010), there is still some resistance from various communities. The OSM map is incomplete in some places and errors in mapping techniques or tagging of spatial objects are often used as evidence against OSM. The combination of these factors can damage confidence when using the data in geomatics applications and LBS. In many cases, OpenStreetMap data is more current than those of private companies, expectations are that by the very nature of its spatially distributed crowd of contributors, it sometimes compares poorly with professional mapping products in some areas.[5]

To identify regions in OSM where there are very or low numbers of contributors, the MVP OSM tool provides a simple means to discover places where OSM contributors have actively mapped and done so with a particular attention for rich spatial detail. MVP means Most Valuable Player (an acronym commonly used in sport to indicate the best player after a match). In our case, the MVP OSM software tool is used to identify the best ‘players’ (contributors) in the ‘OSM game’. The underlying hypothesis is similar to the analysis of the quality of Wikipedia contributions and knowledge development. The more MVPs there are in a place, the higher the likelihood of finding good data quality.

2 Recent Investigations into OpenStreetMap Data Quality

OpenStreetMap data are already used by some companies for business applications (e.g. navigation software[6]). There have also been academic studies which have compared the data quality between OSM and national mapping agencies[7] and also professional mapping companies.[8] Both yielded very positive results regarding the quality of OSM data.

In these comparisons, the primary focus has been on the geometrical accuracy of geographical features. Provided there is ‘ground-truthing’ or authoritative datasets to with which to test against, such comparisons are possible. However, these datasets are not always available, causing other researchers to begin to investigate if there are ways to assess the quality of OSM data, based solely on these data. Mooney *et al.* (2010) were amongst the first to propose this,[9] and their ideas were extended by other researchers. For example, the project called OSMatrix[10] shows different cartographic representations to map user activity in OpenStreetMap. Another relevant study, by van Exel *et al*. (2010),[11] explored the concept of ‘crowdquality’ by identifying two important characteristics of OSM which are interdependent: the quality of the user and the quality of geographic information generated.

They then focused on what caused contributors to become active ‘mappers’, in particular, establishing that the reasons are normally centered around what concerns the mappers as citizens: local issues, local knowledge, experience, and recognition in the OSM community.

OSM data can be downloaded from a variety of sources. OpenStreetMap XML Data files are regular text files, easily editable in any text editor. OpenStreetMap Protobuf Data files are binary files, which take up less space (and so are quicker to download and process) but are not editable. The spatial data is divided into nodes, relationships, and ways. Tags (key value pairs) can be associated with nodes, ways, and relations. The list of these tags is available in the project wiki page and managed by the community. The entire world dataset for OSM is managed in a single file called ‘planet.osm’.

3 How MVP OSM Identifies Users who Contribute High Quality Data

3.1 General Approach

The work of van Exel *et al*. (2010) provided the motivation to begin thinking about the development of MVP OSM and how information about contributors could be extracted: local knowledge, mapping experience, and community recognition from data stored in OSM. As Mooney *et al.* (2010) suggested, because access to ground-truth data is not always possible, for example in making geometrical comparisons against OSM data, it is necessary to search for quality indicators that are actually inherent to OSM. Therefore, in MVP OSM an OpenStreetMap contributor is an MVP if they score high on: local knowledge, mapping experience, and community recognition. The definition of a score within MVP is calculated automatically. Firstly, the concept of experience can be derived by computing the time spent on the OSM project (number of edits and months). Then, to gauge the property of local knowledge, we selected 44 keys (below) from the official list of ‘Map Features’ on the OSM Wiki. The keys were selected very carefully as the key choice is necessary to distinguish spatial information that could not easily have been extracted by tracing aerial imagery.

abandoned
foot
recycling:plastic_packaging

access
foot:backward
recycling:scrap_metal

access:bicycle
footway
recycling:white_goods

access:bus
hiking
recycling:wood

access:foot
horse
step.condition

access:hgv:max_length
incline
step_count

access:motorcar
oneway
step.height

amenity
recycling:batteries
step.length

bridge
recycling:cans
surface

bicycle
recycling:clothes
surface.material

bicycle:backward
recycling:engine_oil
traffic_calming

dispensing
recycling:glass
traffic_sign

disused
recycling:paper
trail_visibility

drinkable
recycling:plastic
visibility

embankment
recycling:plastic_bottles

Simple examples of the richness of spatial data are the labelling and classification information on recycle and garbage cans on the street, or, in residential areas, the number of steps in a flight of stairs, the information text on street signs, and so on. Such information would not easily (if at all) be gleaned from the aerial imagery made available for tracing in OSM. Conversely, the names of streets were excluded, as anyone can retrieve this information from a street guide.

While these 44 keys are only a subset of the entire listing of keys on the OSM Wiki Map Features page, they greatly increases the possibility that the contributor who created or edited the tags (key value pairs) physically went to that area, is a local resident, or has photographic evidence of these features. The concept of recognition is a little different to that of experience. Recognition is interpreted in MVP OSM as the frequency in which a contributor updates data within a given area. The Internet has enabled projects that openly accept contributions from a global audience. Many such projects, including open-source software, OSM, and Wikipedia, do so without any obvious reward for their contributors. Many contributors do so for personal satisfaction and recognition amongst their peers. It has been shown in several studies that in OSM a large percentage of edits are made by a 'relatively' small group of dedicated volunteers. These contributors are normally well known within their own OSM communities and often become well known within the Global OSM community. Such contributors often gain the respect of their fellow contributors due to their dedication to the project, quality of contributions, and length of time in the OSM community.

3.2 Methodology

The analysis began by carefully selecting a number of contributors to Italian OpenStreetMap data. As some of the contributors to MVP OSM are Italian, this allows direct comparisons between a local knowledge of OSM Italy and any results generated by MVP OSM to be made. Whilst the chosen contributors were mainly located within the North of Italy, in many cases their contributions covered many parts of Italy and so were not confined to this area.

The next phase of the methodology followed classical interview techniques where contributors were asked a number of questions about their relationship with OSM and the nature of their contributions to OSM in Italy. During the interview, the contributors confirmed several points: that they had indeed contributed to OSM in Italy; had imported some Italian geodata; that the areas shown to them were areas that they knew well (through family, work, or vacation); that they had attended or organized mapping parties; and that they broadly agreed with the selection of our 44 keys from the OSM Map Features collection. The interview responses confirmed that the 44 chosen keys would serve as good set of indicators in attempting to understand, in an automated manner, if a given OSM contributor had indeed physically visited (or at least had a good local knowledge) a particular area or region.

In order to ensure that the generated results could be verified, it was decided to increase the number of OSM contributors in the study but reduce the geographical area under investigation. To this end, it was decided to focus on OSM contributors located in the Trentino-Alto Adige region in Northern Italy (13,000 km^2). This region is extremely mountainous and covers a large part of the Dolomites and the southern Alps, having land borders with Austria and Switzerland and also internal administrative boundaries with the Italian regions of Lombardy and Veneto. Being very familiar with the region and having a deep local knowledge of the area, we felt that the selection of our geographical study area was very important, particularly in the interview stage of the experiment. The number of users who have made the last snapshot of OpenStreetMap for Trentino Alto-Adige is 1,253 (2,465,872 points), with 600 users (13,675 points) who paid attention to the 44 keys.

In order to identify the areas represented in polygons instead of point clouds, we created a grid representation of the selected territory, using 1 km² cells. The OSM data within each of these cells is then organized by contributor, giving the number of OSM nodes contained in the cell and the difference (in days) between the first node and the most current node being placed in that cell by that particular contributor.

The first variable analysed was ‘time’, allowing an investigation of the concepts of experience and recognition. As stated previously, many contributors to OSM and Wikipedia (and similar crowd-sourced knowledge-generation projects) have their own intrinsic motivation – a significant factor which influences the individual willingness to share knowledge – and personal enjoyment was a key main motivation for knowledge-sharing. Many OSM contributors feel a certain loyalty to, and connection with, OSM. Therefore, they are very conscientiousness in their contributions; in particular, keeping OSM updated around the areas where they live, work, visit for the weekend, or were brought up, and so on. This concept has been well studied in Wikipedia, where contributors who take care of specific pages have what is often referred to as ‘Pet Pages’. So, in contributions to OSM, this meaning could be extended to those contributors having ‘Pet Locations’ (from the Wikipedia study of Liberman and Lin[12]). Therefore, the distinction between a pet location and a normal location can be determined by the frequency of updates performed by a contributor.

The OSM Wiki holds a list of the dates and locations of the OSM Mapping Parties held in a particular region. Analysis of the areas which have held mapping parties revealed that the period from the first to the last node inserted is about a week. Therefore, by removing cells with fewer than seven days of activity, the pet locations for each user can be identified. One final issue remains to be analysed, and, again, the variable ‘time’ is helpful. Users can, over time, lose interest in OpenStreetMap. We wanted to identify areas where users are most active and where the data are then frequently updated. For this reason, only contributors who had been active in the past three months were chosen, and, to identify contiguous areas, only cells adjacent to those demonstrating activity by the same user were selected. In generating these polygons, the minimum and maximum distance between the first and the last entry; the user identifier; and the maximum and minimum number of nodes included in a cell were associated with each geometric test. In the next section, some of the results of the implementation of this methodology using the MVP OSM software will be discussed.

4 Results and Discussion

The input to MVP OSM is an OSM XML file. This is converted by Python[13] into a file for spatialite (the GIS extension for sqlite). The results of the analysis are stored in a table that can subsequently be visualized as a layer in GIS software and QGIS was used for visual analysis of these results.

At each stage in the analysis, the contributors of each area were interviewed. The first stage was to analyse areas relating to the activity of each individual contributor and to display each of these areas using four classes according to the maximum number of days of activity through a colour ramp ranging from red (most days) to yellow (fewest days) (Fig.1). The interviews confirmed that the areas of highest activity are the places where users spend most of their time (e.g. home and/or workplace), and, that these are considered to be their own ‘pet location’ by themselves.

[FIGURE 1 ABOUT HERE]
Figure 1 ‘Pet location’ for an OSM contributor, using a colour ramp to display maximum activity (red) to minimum activity (yellow)

Using the concept of ‘Linus Law’[14] (the more people that are involved in a project, the higher the quality of the product) a ‘heat map’ of the area where most ‘pet locations’ overlap was created. Visual analysis of the map, except for some small areas, appeared to indicate that the areas of highest quality are the most populated areas. Consequently, these areas have higher OSM detail. The most loyal users to the OSM project like to keep updating and checking the map in and around their ‘pet location’. In the course of the interviews, some interesting, first-hand experiences were obtained regarding the speed at which errors were resolved, and how new roads were added in any given OSM area. One of the most interesting results related to the effect that mapping parties had on participation in OSM; the activity generated by mapping parties does not appear to engage new users of that area over a longer period. In other words, while holding a mapping party undoubtedly leads to the generation of very rich spatial data for that area (and consequently very beautiful cartographic rendering of these data), there appears to be a fall-off in the number of updates to those areas after the mapping party has taken place. It appears that mapping parties encourage OSM contributors from elsewhere to converge upon a particular area, but does not necessarily engage OSM contributors from that area itself.

5 Conclusions and Avenues for Further Study

This paper has proposed a method to identify important contributors to OSM who are very conscientious about their contributions to the OSM project and who often have ‘pet locations’; areas in which they take great care to ensure quality OSM data. It was decided to focus upon OSM contributors located in the Trentino-Alto Adige region (13,000 km^2) in Northern Italy. This region is extremely mountainous and covers a large part of the Dolomites and the southern Alps. The number of OSM contributors within this region is not very large, although, as revealed in the analysis and learned through conducting interviews, these contributors are vitally important to the quality of data in OSM.

The methodology involves the use of Python script, which can be used by other researchers in other OSM areas. To our current understanding there is no similar tool which can identify contributors in this way. One of the possible applications of this method is the identification of OSM contributors who have sufficient knowledge and experience of a given area to potentially be involved in projects to improve the planning of public services, for example, in a city.

In the future we want to develop this tool into a web application and continue with our analysis in selecting a greater area. However, the key element of this work will always be based upon interviewing OSM contributors. After investigating a number of areas, an interesting task for future work would be to compare the quality of OSM data for ‘pet location’ areas with spatial data from other companies or organizations.

Notes

[1] Goodchild, M. (2007) “Citizens as Sensors: The World of Volunteered Geography” *GeoJournal* 69 (4) pp.211–221.

[2] Haklay, M. and Weber, P. (2008) “OpenStreetMap: User-generated Street Maps” *IEEE Pervasive Computing* 7 (4) pp.12–18.

[3] *Ibid.*

[4] Zook, M., Graham, M., Shelton, T., and Gorman, S. (2010) “Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake” *World Medical & Health Policy* 2 (2).

[5] Ciepłuch, B., Jacob, R., Mooney, P. and Winstanley, A. (2010) “Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps” *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* 20–23rd July 2010 (p.337).

[6] O’Hear (2010) “Navmii jumps on the OpenStreetMap bandwagon, releases free Sat-Nav for iPhone” Available at: *http://eu.techcrunch.com/2010/09/27/navmii-jumps-on-the-openstreetmap-bandwagon-release-free-sat-nav-for-iphone/*

[7] Haklay, M. (2010) “How good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets” *Environment and Planning B: Planning and Design* 37 (4) pp.682–703.

[8] Zielstra, D. and Zipf, A. (2010) “A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany” *Presented at the 13th AGILE International Conference on Geographic Information Science,* Guimarães, Portugal, 11th May 2010.

[9] Mooney, P., Corcoran, P. and Winstanley, A.C. (2010) “Towards Quality Metrics for OpenStreetMap” *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* New York, USA (pp.514–517).

[10] Roick, O., Hagenauer J. and Zipf, A. (2011) OSMatrix “Grid-based Analysis and Visualization of OpenStreetMap” *Presented at the State of the Map EU Conference* Vienna, Austria, 15th–17th July 2011.

[11] van Exel, M.; Dias, E. and Fruijtier, S. (2010) “The impact of crowdsourcing on spatial data quality indicators” *Presented at GIScience2010, Zurich, 14th–17th September 2010.*

[12] Liberman, M. and Lin, J. (2009) “You are where you edit: Locating Wikipedia contributors through edit histories” *Presented at the International Conference on Weblogs and Social Media, San Jose, California, 17th–20th May 2009.*

[13] MVP OSM source code available at: *https://github.com/napo/mvp-osm* (MIT Licence).

[14] Raymond, E. (1999) *The Cathedral and the Bazaar* Sevastopol, California: O’Reilly Media