Archives of Photogrammetry, Cartography and Remote Sensing, Vol. 22, 2011, pp. 105-116
ISSN 2083-2214

ASSESSING THE QUALITY OF OPEN SPATIAL DATA FOR MOBILE LOCATION-BASED SERVICES RESEARCH AND APPLICATIONS

Błażej Ciepłuch [1], Peter Mooney [2], Ricky Jacob [3], Jianghua Zheng [4], Adam C. Winstanely [5]

[1, 2, 3, 5] Geotechnology Research Group, Department of Computer Science National University of Ireland Maynooth (NUIM) Maynooth, Co. Kildare. Ireland (bciepluch, peter.mooney, rjacob, adam.winstanley)@nuim.ie
[4] School of Resources and Environment Science, Xinjiang University, Urumqi, 830046, Xinjiang, China - itslbs@gmail.com

KEY WORDS: Databases , GIS , Mapping , Data Mining , Spatial Infrastructures , Internet/Web , Accuracy , Metadata

ABSTRACT: New trends in GIS such as Volunteered Geographical Information (VGI), Citizen Science, and Urban Sensing, have changed the shape of the geoinformatics landscape. The OpenStreetMap (OSM) project provided us with an exciting, evolving, free and open solution as a base dataset for our geoserver and spatial data provider for our research. OSM is probably the best known and best supported example of VGI and user generated spatial content on the Internet. In this paper we will describe current results from the development of quality indicators for measures for OSM data. Initially we have analysed the Ireland OSM data in grid cells (5km) to gather statistical data about the completeness, accuracy, and fitness for purpose of the underlying spatial data. This analysis included: density of user contributions, spatial density of points and polygons, types of tags and metadata used, dominant contributors in a particular area or for a particular geographic feature type, etc. There greatest OSM activity and spatial data density is highly correlated with centres of large population. The ability to quantify and assess if VGI, such as OSM, is of sufficient quality for mobile mapping applications and Location-based services is critical to the future success of VGI as a spatial data source for these technologies.

1. INTRODUCTION

Volunteers involved in the OpenStreetMap project collect spatial information in ways similar to surveyors in the beginning of the 21st century. They are highly motivated to produce maps of their locality and surroundings. Maps produced by these volunteers can be better and more precise as authoritative mapping agency data or commercial datasets. As Goodchild (2007) states in his article which coined the term VGI the dawn of the ubiquitous smartphone and similar devices means that there are "six billions sensors on the earth". There are many variables to consider when one attempts to assess the quality or usability of VGI. Why do people collect this data and share it? What is their motivation, their skills in survey or map production, etc? Overall we believe that VGI datasets should be accessed in, potentially, different ways to traditional GIS data because the VGI datasets themselves are inherently different to professional GIS data. Our paper will investigate the quality of OSM data, as a case study, and the potential use of this data in Location-based

services. We also investigate what the best methods might be in order to properly assess the quality of this VGI data.

2. PRELIMINARY INVESTIGATION

Our initial attempt to quantify the quality of OSM data involved carrying out a survey of web-based mapping in Ireland. Without access to the vector datasets of commercial spatial data vendors we preformed a rigourous visual comparison of OSM with Google maps, Bing and Yahoo maps (Ciepłuch et al, 2010). The primary concept in this investigation was to assess road completeness in five towns in Ireland. All the work was carried out by manually overlaying OSM data on top of tiles from the commercial datasets. We restricted this study to a small areas as this was very labour intensive. However it allowed us to apply our local knowledge to assess if the OSM datasets and the commercial datasets were temporally correct (all roads currently existing were present) and had high attribute accuracy (names of roads were correct, road hierarchy designation correct, etc). But this type of visual, manual-based, investigation was too labour intensive and time consuming. We obtained an OSI (Ordnance Survey Ireland) dataset and decided to develop some methodologies for the automated comparison of an authoritative spatial dataset with OSM data.

3. OBTAINING OPENSTREETMAP DATA

In order to store, manipulate, and use OSM on your own computer the OSM data must be downloaded in OSM-XML format. Several companies (Cloudmade and GeoFabrik for example) make OSM-XML and Shapefiles available on a per-country basis which makes download size smaller and loading time into the PostGIS database quicker. The OpenStreetMap API allows one to download the OSM-XML corresponding to specific regions as specified by a bounding rectangle. Command line tools such as CURL or WGET can be used for this task. Geofabrik provide near real-time access to the OSM-XML and Shapefiles with usually no more than a 2 hour time difference between the global OSM database and the OSM-XML available for download.

4. GROUND-TRUTH SHAPEFILE MANAGEMENT

Obtaining ground-truth data in ESRI Shapefile format is probably one of the most likely formats in which such spatial data will be distributed. For this implementation we obtained a 1:5000 scale dataset of the roads of the Republic of Ireland in Shapefile format. Shapefiles can be easily imported into PostGIS (PostGIS, 2011) databases using the PostGIS tool shp2pgsql (shp2pgsql, 2011). This tool creates tables within the database corresponding to the attribute database of the shapefile. The shape geometry is included in the same table as the feature attributes. However we chose a different method in order to improve the flexibility of the PHP software. The shp2osm application (freely available Java library) converts input ESRI Shapefiles to OSM XML. This OSM XML is then imported into the PostGIS database using the osm2pgsql tool. The advantage to this approach is that the tables holding the spatial data from the ESRI Shapefiles is part of the same data model as the OSM data. This allows the development of cross data source compatible PHP

database query code. This method also allows for the update of the OSM data and the ground-truth datasets to happen seamlessly. In Table 1 an example of the attributes rules file for the shp2osm tool is shown. This is a simple structured text-formatted text file which contain the rules about how the ground-truth data should be converted to the OSM database model. The example of "unclassified roads" in Table 1 shows that the attribute "unclassified roads" in the OSI dataset can potentially map to four different types of roads in the OSM dataset.

Tab. 1. Road classification in both OSM and the ground-truth dataset

OSI Road Classification	OSM Road Classification
Motorway (represented by 1 line)	Motorway (represented by 2 lines)
Trunk_Road	Trunk_Road
Primary	Primary
Secondary	Secondary
Tertiary	Tertiary
Unclassified Roads	Unclassified Roads, Residential Roads, Living Streets, Service Roads
Track or dirt roads	Track or dirt roads

4.1 Grid generation

One of the most widely used techniques in spatial analysis is the use of a grid-cell vector layout. Each grid cell is of uniform size (for example 1km square, 5km square, etc) and spatial data from potentially several other datasets from the same geographic region can be queried to establish their relationship to a specific grid-cell. Using a grid-based approach to spatial analysis of OpenStreetMap has been performed by several authors: (Haklay, 2010) used it in his comparative analysis of OpenStreetMap and the Ordance Survey data in the United Kingdom. (Zielstra and Zipf, 2009) used a similar approach for their comparative analysis of OpenStreetMap and Teleatlas data in Germany. Some GIS software packages provide in-built functionality to automate the process of grid generation for arbitrary geographical areas. As stated above our work uses only free and open source GIS software. The QGIS desktop GIS provides functionality to generate a grid automatically as an ESRI Shapefile. We decided to develop a PHP script to automate the process of grid generation for an arbitrary geographical area. The development of the script allows us to run the script as an optional component in a work-flow of PHP programs used for this research. The algorithm is outlined below in Figure 1.

```
input : Grid cell size, coordinates of bounding rectangle (in UTM) -
        (NWEasting,NWNorthing) and (SEEasting,SENorthing)
output: A spatial database table with polygons representing each
        grid cell in UTM coordinates
currX ⟵ NWEasting;
currY ⟵ NWNorthing;
GRID ⟵ 5000;
while currY ≥ SENorthing do
    x1 = currX, y1 = currY;
    while currX ≤ SEEasting do
        Assign other vertices of cell;
        x2 = x1 + GRID, y2 = y1;
        x3 = x2, y3 = y2 − GRID;
        x4 = x1, y4 = y1 − GRID;
        Store Polygon in Database;
        POLYGON(x1y1, x2y2, x3y3, x4y4, x1y1);
        Move to next grid cell along easting;
        x1 = x1 + GRID,currX = x1;
    end
    Move south currY = currY − GRID;
    Reset easting currX = NWEasting;
end
```

Fig. 1. Algorithm sketch which create boxes and perform analysis about data inside them

The input to the algorithm in Figure 1 are the coordinates (in UTM coordinates) of the top-left and bottom-right of the bounding rectangle for the grid-map. The bounding coordinates can be easily calculated by viewing the OpenStreetMap data within a desktop GIS such as QGIS, or by extracting these coordinates (stored in WGS 84) from the OSM XML file and converting them to the desired UTM coordinates. The algorithm creates a closed polygon (grid cell) for each location on the grid and stores these polygons as a geometry in a spatially-enabled table in a database. Each grid-cell is assigned a unique identifier. Our PHP script implementation of the algorithm in Figure 1 stores the grid-cell geometries in a PostGIS table. The QGIS (Quantum GIS, 2011) package can then directly connect to PostGIS and display the grid-cell layout overlayed over other vector layers. Grid generation is fast. For most European countries a 5km grid-cell map can be generated in less than 20 seconds. The slowdown in running time is caused by the bottleneck of thousands of INSERT SQL statements to the

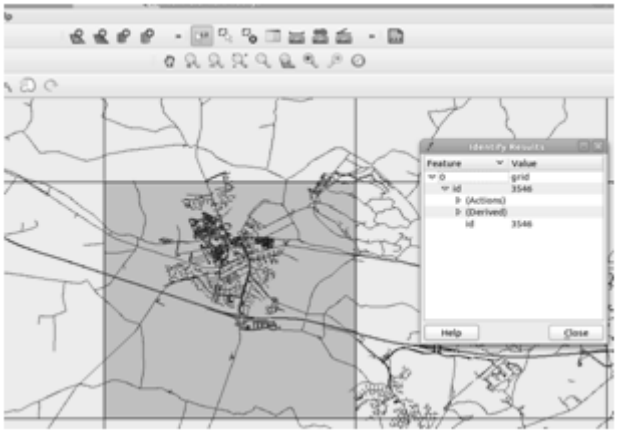

Fig.2. Grid map layout displayed in Quantum GIS

PostGIS database. For larger countries in Europe such as Germany, France, the United Kingdom, and Spain grid generation can take up to 2 minutes. Again the slow down is caused by the number of insert statements required. However grid-map generation should only be required once for a given country or region and is not re-run each time a new analysis is performed unless a different grid alignment or grid-cell size is needed. Figure 2 shows an example of a 5km grid cell map overlayed on an OpenStreetMap roads dataset.

In this section we present some of the types of results which our software can generate from a comparison of an OpenStreetMap database and a ground-truth dataset. There are a number of input parameters for the software including: paths to dataset file locations, path to output folders, buffer sizes, grid-cell size, database schema names, etc. These are stored in a plain text-formatted file. Output from the various components of the software is written to a series of CSV files in tabular format. For quick viewing of the results the output is also written to HTML files as tables.

4.2 Feature Length-based calculations

Tab. 2. Table presenting lengths of all roads in Ireland in both datasets OSM – OpenStreetMap and OSI – Ordnance Survey Ireland. All measurements are in kilometres

Road Feature Type	OSI Dataset	OSM Dataset
Motorways	841	1419
National Roads	2638	2478
Primary Roads	2673	2843
Secondary Roads	12134	11577
Tertiary Roads	24900	12212
Unclassified	91744	20278
Track	42465	1247

In table 2 we present the results of a simple query to calculate the number of kilometres of roads in Ireland. There are a number of interesting observations from these results which could be used as part of a larger overall quality evaluation. In terms of temporal accuracy the OSI dataset was created for 2007. The OSM database for Ireland was downloaded during October 2010. This accounts for the large disparity in OSM Motorways and OSI Motorways due to the completion of a number of major motorway construction projects in Ireland between 2008 and 2010. An issue encountered in this query was the representation of motorway roads in OSM. OSM uses two polylines for dual carriage way motorways. In the 1:5000 dataset from OSI motorways the two dedicated mono-directional carriageways were most probably simplified to one polyline. It was necessary to apply a multiplication factor of 2 to the OSI motorway lines to account for this difference. Primary roads appear better represented on OSM side but this may suggest that in some cases these roads are misclassified by contributors as national or primary roads. Another reason for this could potential be the issue that in 2007 or 2008 these primary or national roads were

tagged as such by their original contributor. Some of these roads may have changed designation since then but the OSM database has not been updated since to reflect these changes. In Secondary roads there are 500km more from OSI side. This could be a result of better spatial representation of the real-world geography of the roads or a manifestation of the "white areas" on the OSM map where roads have not yet been mapped by OSM contributors. The potential weakness of the OSM dataset in regards to road network representation is highlighted in the final three classifications of roads. Unclassified roads really only appear currently in OSM within major centres of population. The comparison of track roads in OSI and OSM is an example of where a OSI can provide national level coverage of a geographical features which is correspondingly difficult to OSM volunteers to match. Track roads include many of Ireland boreens (from the Irish *bóithrín*, meaning a small, narrow, rough, rural road), forest roads, mountain tracks etc. are missing from OSM at present. There are a number for reasons for this including: most mapping activity in OSM is happening in the urban areas and there are simply not enough contributors in rural areas; many track roads are often difficult to access on foot and may even be subject to differences in designation as a public or private right-of-way.

Tab. 3. Number of Kilometres of Railway lines in both OSM and OSI dataset

Dataset	Rail Lines	Disused Rail	Industrial Rail
OSM	1878	59	5
OSI	2092	1449	499

A similar analysis to the discussion above was performed for a comparison of railway lines in Ireland in both the OSM and OSI dataset. Due to the access restrictions to railway line infrastructure it is likely that the OSM railway lines were traced from freely available aerial imagery (Yahoo!). However there is still a small overall difference between the two datasets. The OSI dataset stores a large amount of disused railway line. This is most likely as a result of it's wider application usage in environmental planning and heritage management in Ireland. In the case of industrial railway (those predominantly around ports and railway lines on peat bogs transporting peat to electricity generating stations) OSI appears to represent all railways under this designation. OSM only has 5km of this rail represented in the database. The reasoning behind the very poor OSM results from the analysis for disused and industrial rail lines are probably a result of: a perceived lack of real-world application of such geographical features by OSM volunteers collecting and generating spatial data; difficulty in obtaining access to industrial rail sites; the fact that much disused railway is overgrown and hidden by bush and scrub now which makes both manual sampling with a GPS or tracing over aerial imagery difficult. Overall this may indicate that currently there are no volunteers in OSM in Ireland interested in mapping these railway features.

4.3 Grid-based analysis

In this section we outline the results of applying the grid-based analysis functionality of our software for the comparison of the two datasets. This analysis uses the grid generation algorithm described in section 3. For the purposes of this analysis a grid size of 5km was used. However depending on the requirements of the analysis the software can generate grids of arbitrary size by setting the grid size input parameter. In table 4 we present the results of a grid-based analysis of the locations (within grid cells) of the spatial data for the various classifications of roads in both datasets.

Tab. 4. 5km Grid-cell based analysis of roads in OSM and OSI datasets. The number of cells where OSM has more km of road is shown in the OSM column while the number of cells where OSI has more km of road is shown in the OSI column.

	#OSM	#OSI	Total Cells
Motorways	84	41	125
National Roads	215	230	445
Primary Roads	251	217	468
Secondary Roads	622	690	1310
Tertiary Roads	312	1189	1501
Unclassified Roads	54	1585	1639
Track Roads	10	1622	1632

There are some interesting observations that can be made from Table 4 which corroborate the results from section 4.1. Track roads appear in only 10 cells from the X cells in Ireland. The number of OSM cells containing more length of motorway feature is almost twice that of the number of OSI cells. However this has been explained above. Overall the OSI data shows a uniform and homogeneous representation of the Irish national road network beginning at the highest level (Motorways) and working down to the lowest priority level roads. Our current analysis shows that OSM is biased toward higher priority roads and less focused on tertiary, unclassified, and track roads. This is ironic given that one of the major strengths of the OSM crowd-sourcing model is the ability to engage OSM volunteers to map their own localities. The results in this table again confirm that much of the mapping activity in OSM is taking place within urban locations. However this is set against a poor representation of unclassified roads which include: living streets, access roads, and pedestrian only roads.

The final set of results for the grid-based analysis performed by our software is shown in Table 5. The results in Table 5 quantify the number of cells where some condition based on the length of road per road classification is computed. For example the column 1km to 5km and row “motorway” indicates the number of grid-cells where the differences in overall length of OSM motorway and OSI motorway was between 1km and 5km. Some interesting observations include: the comparative lengths of secondary roads is very inhomogeneous.

Almost 70% of the grid cells containing secondary roads have differences in length of less than 0.5km for that cell. Yet for the same road classification there are almost 25% of grid cells with a difference of 1km or more.

Tab. 5. Additional grid cell analysis comparing the differences in lengths of road features between the two datasets

	< 0.5km	**0.5km to 1km**	**1km to 5km**	**5km to 10km**	**10km to 20km**	**> 20km**
Motorway	2	4	4	32	42	0
National	5	26	86	67	20	1
Primary	392	14	42	17	3	0
Secondary	801	63	212	194	40	2
Tertiary	94	47	351	326	415	265
Unclassified	27	16	76	69	169	1282
Track	27	14	95	140	455	901

4.4 Buffer Analysis

In this section we describe another component of the query functionality of our software which is designed to perform a buffer analysis on the road features. The size of the buffer required is a modifiable parameter. For the purposes of this example we have set the buffer as 10 meters. The software can then compute the number of km of OSM road features (for each road class) which lie outside a 10 meter buffer of the corresponding OSI road feature. This analysis is the most computationally demanding component. When computation was completed the attributes for each road feature (if they were inside or outside the buffer) were added to the OSM road feature table by our PHP script. Figure 2 shows an example of the visual output from this component as visualised in the QGIS desktop GIS using this information. The map in Figure 2 shows the Maynooth near Dublin urban region in Ireland. The total amount of OSM road feature outside the buffer is quantified in Table 6. Many OSM road features are outside the 10 meter buffer of OSI. There are a number of important observations. For example the secondary roads are outside the buffer in disjoint clusters. This may be a consequence of different OSM contributors sampling the roads using difference GPS devices or relying on tracing of aerial photography through one of the OSM editors. It is also interesting to note, in conjunction with Table 6, that almost all OSM motorways are outside the buffer.

Tab. 6. The total length in km of OSM roads which are outside a 10 meter buffer of the corresponding roads in the ground-truth dataset

Road Classification	Length of OSM road (km) outside 10m buffer of ground-truth dataset
Motorways	118
Trunk Roads	67
Primary Roads	9
Secondary Roads	166
Tertiary Roads	227
Unclassified , Residential etc. Roads	1100
Track roads	103

4.5 Heat Map generation

As stated in the previous section the software can also write directly to the PostGIS table containing the grid cell map for the region under analysis. One of the sample outputs from this analysis is shown in Figure 3. This figure shows a heat-map based on a 5km grid for OpenStreetMap Ireland. It describes in which dataset contains more roads length in km in OSM or OSI for a particular box. This could be used as an indicator of where roads are better mapped. For this particular map we can identify that tertiary roads are generally better mapped in majority of Ireland by OSI with exception of Galway – Wesptort – Athlone area in the West and within Dublin city and region on east coast. Generally the maps produced show clearly that OSM in Ireland is equivalent in coverage (within an acceptable threshold) for Major roads from Motorways down until Secondary roads in the hierarchy of roads as we seen in the tables above and in Figure 4. However, after this OSM completeness dramatically decreases. We can clearly see that these roads of more minor importance are only well represented in bigger agglomerations where most of VGI mapping activity occurs. This is not surprising. VGI, in general, tends to be be found in much higher quantities in cities and urban areas. Where there is OSM representation of lower designation roads in Ireland (Figure 4) this could indicate that this mapping is related to an OSM contributor living or working in this area.

The final figure on this paper shows an example of contributor activity. Figure 5 shows a gridded map of Nottinghamshire and depicts the density of users edits in this county. We can see that the highest density of users in are located close to towns , such as Nottingham. Areas of high population density appear to have more data, in general (not just roads), contributed to the OSM database. It is also interesting to note that contributor activity can be traced radially out from Nottingham where contributors follow the path of some major motorways.

Fig. 3. Grid map of Republic of Ireland representing Tertiary roads - black pixels representing areas where OSM has better coverage. Gray pixels are where OSI dataset is superior

Fig. 4. Grid map of Republic of Ireland representing Unclassified roads (lower grade roads) connecting farms and small villages. black pixels representing areas where OSM has better coverage. Gray pixels are where OSI dataset is superior

5. CONCLUSIONS

In this paper we have described an analysis and visualisation of some data quality related issues for VGI (case-study of OSM) when compared with an authoritative dataset (OS Ireland). All of the software tools used are open source. Without great difficulty these tools were used to perform some complex spatial analysis and then visualisations. The analysis of the quality of OSM data here in the paper shows that the visualisation of the

results of these analysis could be an important method in understanding the quality and potential fitness for use of the OSM dataset. The heat map representations of various aspects of the OSM dataset presents us with a good overall picture of the characteristics of the VGI data. OSM is very like Ordnance Survey Ireland where both have a very detailed coverage of towns and cities (Girres and Touya, 2011). The OSM picture of rural areas is much poorer and more fragmented. This issue is one which will need to be addressed with urgency by the VGI community. Another issue for further work is the development of algorithms and methodologies to show that VGI is safe to use for a given application or that we are "certain" within a given threshold that this is "good enough" quality (as mentioned by Mooney et al (2010)). From our analysis here we can conclude that within city areas VGI (in our case OSM) provides a very realistic and accurate alternative to commercially available spatial data. This is echoed in several similar studies such as Ludwig et al (2011). For those developing Location-based Services (LBS) within city and urban areas this is particularly good news given that OSM is free to use for any application (Over et al, 2011). However, the developers of LBS must consider the geographical extent of their applications. If they require spatial data (and more precisely metadata and attribute information) about a particular location then OSM, in its current form, may fail these requirements for rural and less densely populated areas.

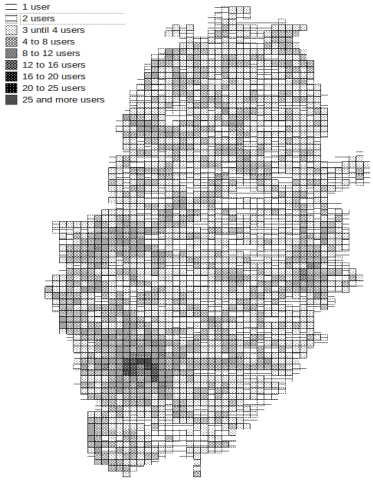

Fig. 5. Grid map of Nottinghamshire representing density of users who are contributing VGI. Most mapping activity occurs in Nottingham city

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the funding for this research provided by the Environmental Protection Agency Ireland's STRIVE programme. The work is part of a larger project Geoinformatics Services for Improved Access to Environmental Data and Information (2008-FS-DM-14-S4) is a 5 year project from 2008 – 2013. Dr. Mooney is PI for this project. The Location-based Services strand of the Science Foundation Ireland funded STRAT-AG programme is co-led by Dr. Adam Winstanley. The support of STRAT-AG is also gratefully acknowledged.

7. CITATIONS

Ciepłuch, B.; Jacob, R.; Mooney, P. & Winstanley, A. (2010), Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps, *in* 'Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences 20-23rd July 2010'.

Girres, J.-F. and Touya, G. (2010), "Quality assessment of the french openstreetmap dataset", Transactions in GIS , Vol. 14, p. 435-459.

Goodchild, M. F. 2007, 'Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0', *International Journal of Spatial Data Infrastructures Research* Vol.2, 24-32.

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

Ludwig, I., Voss, A. and Krause-Traudes, M. (2011), A comparison of the street networks of navteq and osm in germany, in S. Geertman, W. Reinhardt and F. Toppen, eds, 'Advancing Geoinformation Science for a Changing World', Vol. 1 of Lecture Notes in Geoinformation and Cartography, Springer Berlin Heidelberg, pp. 65–84.

Mooney, P., Corcoran, P. and Winstanley, A. C. (2010), Towards quality metrics for openstreetmap, in 'Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems', GIS '10, ACM, New York, NY, USA, pp. 514–517.

Over, M., Schilling, A., Neubauer, S. and Zipf, A. (2010), "Generating web-based 3d city models from openstreetmap: The current situation in germany", *Computers Environment and Urban Systems* , Vol. 34, Elsevier Ltd, pp. 496–507.

Zielstra, D. & Zipf, A. 2010, A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. 13th AGILE International Conference on Geographic Information Science. Guimares, Portugal.

PostGIS 2011, An addition to Postgree SQL database for geographical content http://postgis.refractions.net/

shp2pgsql - shapefile to postgis loader importing tool http://wiki.openstreetmap.org/wiki/Shp2osm

Quantum GIS – free open source GIS application http://www.qgis.org/