Exploring Quality and Usability of OpenStreetMap Data

Christopher Braune[1] and Jens Klump[2]
cbraune@gfz-potsdam.de
(1) Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences
(2) CSIRO Earth Science and Resource Engineering

Poster-EGU2014-6794

Introduction

OpenStreetMap (OSM) is an international project which provides free geographic data by creating and distributing continuously new geodata on a strictly voluntary basis [1]. There is a significant data growth in the OSM project [2]. As a result, the interest in OSM data for various applications (such as navigation) increases. However, the data collection is not regulated. OSM users can add new objects, but also semantic information in form of freely selectable tags. The free use of new data is a typical characteristic of crowdsourced data. This is both – boon and bane. OSM provides a wide range of information. This is accompanied by a data heterogeneity, which is difficult to estimate, qualitatively as well as quantitatively.

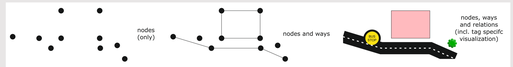

Figure A: *OSM data structure: nodes, ways and relations. Nodes are the geometric primitives of the OSM data. Ways are lists of specific nodes and relations are used for creation of complex geoobejcts like polygons.*

Typical OSM data sources:

- GPS (tracks and points)
- Maps with passed copyright
- Satellite images (NASA Landsat e.g.)
- On-site-knowledges

Aims

The aim of this study is to explore the global OSM dataset and to define the quality of the dataset through its intrinsic properties instead of comparing its spatial location accuracy to a reference dataset. For instance, one hypothesis states that the spatial object density of OSM will correlated with population density. Other studies have already confirmed this correlation for specific study areas [3], but not on a global scale. The main aspect of this study is to explore the semantic properties of tags used to describe OSM elements. In addition to the investigation of the OSM data, information about the usability of various tools and application software can be obtained in this project. The processing of original OSM data depends on the software.

Quality

The term "quality" is always bound to a specific purpose and describes the relative relationship between a data base and the target purpose (according to DIN EN ISO 9000).
The definition of study parameters in terms of quality and usability is oriented to the ISO 19113: Geographic information - Quality principles [4]:
- completeness
- logical consistency
- positional accuracy
- temporal accuracy
- thematic accuracy

Technical Hurdles

Frequently, the large amount of data in the global OSM data set becomes a problem. Programs that work with smaller regional datasets often fail on the complete data set. GRASS does not allow the full import of data in one step. Furthermore, GRASS does not support the "bigint" (big integer) data type. This type is used by OSM as an ID column. Currently, an additional integer ID in the database must be created to integrate the data. In addition, osmosis can only handle the uncompressed global data set.

Generally common tools for OSM data experience problems with the handling of object tags because unknown key-value combinations (elements of tags) are not supported in some software applications. This means that these objects may be missing in the data subset used by a specific application. This behavior led to the consideration of the characteristics of semantic tags and to the analysis of untagged objects.

Implementation

The project includes the implementation of suitable hard- and software environments (Figure B). The core forms a PostGIS database (Figure C). The osmosis command line tool (written in java) is used to convert the hierarchically organized OSM data in the relational model of the PostGIS database. Due to the resulting database size of 600 GB a powerful client-server system is being used.

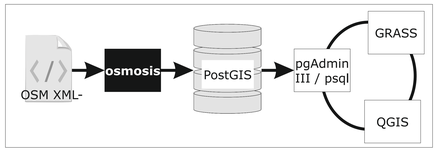

Figure B: *Representation of the used infrastructure: the OSM XML data are transferred into a relational PostGIS database via the osmosis tool. The access to the data is made by SQL queries (pgAdminIII and psql). The presentation and spatial analysis is performed primarily in a desktop GIS (GRASS and QGIS).*

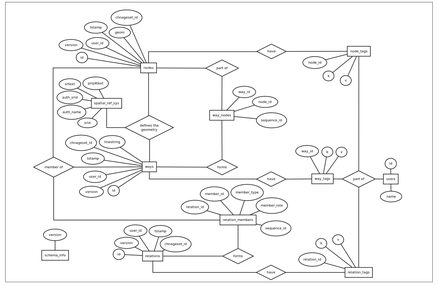

Figure C: *Simplified representation of the used (PostGIS) database schema. The cardinalities between the database relations (tables) are not shown.*

Results

The investigation of a potential correlation between the population density and the OSM object density can be made for limited study areas (for example Berlin-Brandenburg). The global approach has failed so far because of the GRASS data import.

The investigation of the tags illustrates the semantic diversity but also the heterogeneity of the OSM data. Already the investigation of the keys (fig. E) clearly shows that a very large part of the information are rare or unique. If actually untagged objects exist has not yet been clarified sufficiently.

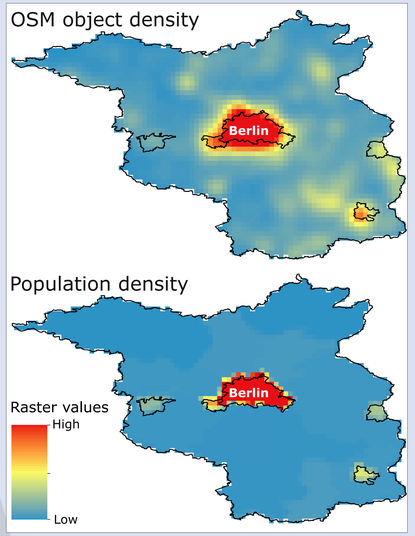

Figure D: *Visual comparison between the OSM density grid and the population density grid. The OSM grid was created by the GRASS function "v.kernel". The population density is a data set of SEDAC (Socioeconomic Data and Applications Center).*

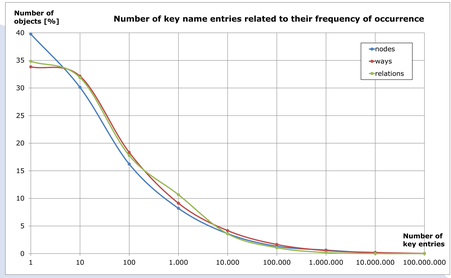

Figure E: *Occurrence of the key entries of the OSM base elements (nodes, ways, relations) in relation to the total amount of all OSM objects. All OSM base elements show a similar trend. 30 to 40 percent of the keys occur once. Therefore specific semantic information can only be obtained manually.*

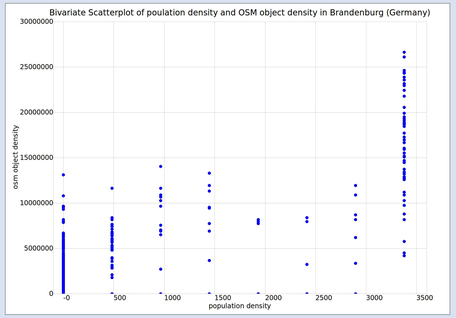

Figure F: *Bivariate scatterplot (generated in GRASS) of grids which are shown in Figure D. The data set of Berlin-Brandenburg (Germany) simply shows a tendency but no clear correlation between the studied variables.*

References

[1] Offical OpenStreetMap Wiki, URL: http://wiki.openstreetmap.org (January 2014)
[2] Offical OpenStreetMap Wiki, URL: http://wiki.openstreetmap.org/wiki/Stats (March 2014)
[3] Neis, P., Zielstra, D., Zipf, A., 2011, The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007-2011
[4] International Standard: Geographic information - Quality principles (ISO 19113), First edition 2002-12-01