Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata

Marco Helbich, Christof Amelunxen, Pascal Neis

GIScience, Department of Geography, University of Heidelberg, Berliner Strasse 48, 69120 Heidelberg, Germany

Abstract

The emergence and ubiquitary availability of geotechnologies yield an explosion of user generated geographical data, utilized for mapping, modeling etc. Using a well mapped German city in OpenStreetMap as an example, this research models the positional accuracy of locations of road junctions, whereas a statistical comparative approach with high precise survey data and commercial Tele Atlas data is conducted. The OpenStreetMap and Tele Atlas data showed similar spatial deviations and both do not coincide with the survey data. Especially, OpenStreetMap suggested spatial heterogeneity in the error distribution, leading to significant clusters of high and low positional accuracy.

1 Introduction

Profound changes have taken place in Geographic Information Science lately (e.g., Goodchild, 2007; Elwood, 2008; Sui, 2008). Until recently, the generation, maintenance and distribution of geographic data had been almost solely the domain of either official land surveying offices or commercial companies. This was due to the immense costs related to the actual surveying and maintenance as well as the efforts involved to share and distribute spatial data. What we see nowadays is a massive increase of geographic data collected and shared by volunteers, working in a collaborative

fashion. The dramatically reduced costs of modern satellite navigation handheld devices have enabled people to privately collect geographic data with ease of use and in precision levels which had formerly been simply beyond reach for the masses. Furthermore, the progress of the internet to the “web 2.0” participatory approach has made collaborative efforts to generate and share content of various kinds very common. This phenomena is widely known as Volunteered Geographic Information (VGI; Goodchild, 2007; Elwood, 2008). In combination with nowadays ubiquitous available Open-Source software (e.g., Google Earth) and miscellaneous web services, Sui (2008, p. 1) calls this revolutionary development the "wikification of GIS", affecting our daily life.

Among a broad list of initiatives working with VGI, OpenStreetMap (OSM) is one of the most promising crowd sourced products. When the project started its primary goal was simply to generate a free map of the world through volunteered participation. Nevertheless, although the generation of maps still is the focus of the project, the collected spatial data is made publicly available and may thus be used for other purposes as well. Additionally, OSM serves as a platform for location based services, including routing, geocoding, accessibility analysis and spatial searches. OpenRouteService.org (Neis and Zipf, 2008) is an example which has successfully implemented a routing service and recently applied in disaster management (Neis et al., 2010).

However, using these data means accepting their limitations - especially concerning the data quality. For a comprehensive overview of quality aspects we refer to Van Oort (2006) whereas in this paper we will focus on a single aspect, namely positional accuracy. It denotes the coordinate deviation of a spatial object compared to its real location (Haklay, 2010). The positional accuracy of the collected data is affected by different influences, e.g. the technological bias like the accuracy of the GPS-receiver used, different data acquisition techniques (e.g., digitizing) or subjective knowledge about the data gathering process. In order to assess the usability of VGI in varying cases of application the positional accuracy of the data to be used has thus to be thoroughly evaluated, because missing and imprecise data effect model calibrations and in the worst case leads to false conclusions.

The main purpose of this research is the statistical analysis of the positional accuracy of three different data sources, namely OSM, Tele Atlas (TA) and survey data (SD), for a (well mapped) medium size German city. For this purpose, we have to make the crucial assumption that our official SD possess the highest accuracy, based on precise surveying techniques (e.g., triangulation). Therefore, SD serve as the reference data, to which the other data sets are relatively evaluated. In this context it must be mentioned that, because routing being its main application, the primarily inten-

tion of the TA dataset is topological correctness and positional accuracy is just a second but of course essential issue. Thus, comparisons can be, but must not be, biased.

2 Related Work

Research concerning different kinds of accuracies (e.g., positional or topological) of VGI has not gained much interest yet. A first descriptive attempt was conducted by Haklay (2010) who analyzed the positional accuracy of OSM compared to commercial data (OS Meridian 2) for the United Kingdom. He analyzed the percentage of overlaps between both data vendors within a buffer distance, as proposed by Goodchild and Hunter (1997). The methodology has been adapted by Zielstra and Zipf (2010) for Germany, comparing the completeness of OSM to TA. Both studies concluded that OSM is a viable alternative data source, but emphasize that there are some limitations in usage concerning its completeness in rural areas.

Ludwig et al. (2010) alluded to a related issue, criticizing the lack of specific attributes, like maximum speed limits and street names. Neis et al. (2010) compared the length of the mapped street network of commercial data and OSM for the year 2010. They stated that there are nearly no differences in the overall street length between both data sets, but found a difference of 40 percent of street segments capable for routing applications. Similarly, Schmitz et al. (2008) analyzed the routing capabilities of OpenRouteService, based on OSM. Because of topological errors (e.g., unconnected street segments) within the street graph, 3-5 percent of all routing requests were not executable. Chen (2010) deals with topology correctness and completeness of digital maps through the integration of different user-generated (OSM) and commercial data sources (NavTeq, TA), comparing the correspondence of road crossings. His findings, among others, clarify that NavTeq and TA have a higher topological similarity than TA and OSM. Furthermore, urban areas show a higher similarity than rural areas.

Over et al. (in press) extend the range of application of OSM data to the third dimension. In combination with free elevation data (SRTM), the usefulness for 3D visualizations (e.g., buildings) is shown. Further research has addressed VGI for the purpose of geocoding (Amelunxen, 2010). He concludes that the positional accuracy of geocoding results based on OSM data can be equal to or even better than the accuracy provided by the commercial geocoding service offered by Google Maps. Nevertheless, these accuracy levels could only be achieved when OSM data were availa-

ble on house number level which, at the time of the research, had been the case for only about 5 percent of the sample requests within the study area, but is increasing fast.

This brief literature review highlights, mostly in a descriptive fashion, some limitations as well as the potential of OSM data. Further, it clarifies the need of statistical analysis of the positional accuracy of OSM compared to proprietary geographical data, like SD and TA. The present research tackles this issue.

3 Methodology

3.1 Data Processing

Datasets containing linestrings of road segments from all three sources are semantically aligned and loaded into a PostgreSQL/PostGIS spatially enabled relational database. As a first preprocessing step for each dataset, separate road segments sharing the same street name are merged in order to provide a single linestring for either street. The junctions within the datasets are then extracted by determining all point coordinates where exactly two distinct linestrings cross each other. This approach admittedly rules out junctions where three or more streets cross but has been preferred for the sake of clarity.

The concatenated names of the streets crossing each other serve as an identifier for given junction. These identifiers are then used to select and spatially compare corresponding junctions among the datasets. As the identifier has to be unique, this approach additionally requires to rule out those cases where two streets cross each other more than once. Figure 1 illustrates this approach.

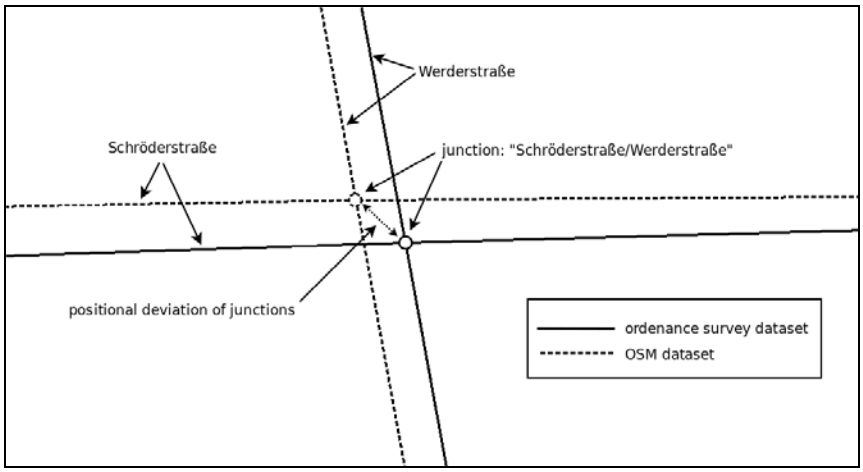

Fig. 1. Extraction and comparison of road junctions.

The deviation of the junction point coordinates from the corresponding points in the defined reference data set is then used as a measure of positional accuracy. Based on this, a scatter diagram of positional errors is investigated to inspect their spatial distribution and in order to detect potential systematic errors.

3.2 Geometrically Evaluation of the Distortion

To get deeper insights of this geometrically distortion of the point patterns a bidimensional regression (Tobler, 1994; Friedman and Kohler, 2003) is calculated. This method allows assessing the transformation parameters between two (plain) maps and point patterns, respectively. Contrary to Tobler's (1965) remark, that bidimensional regression could be particularly useful for geographical analysis, it is rarely applied till these days. Primary, spatial positional accuracies in cognitive maps are analyzed (Lloyd, 1989). Other scope of applications are concerned with the lineage of historical maps (Symington et al., 2002), rubbersheeting as well as corrections of remote sensing images (Tobler, 1994).

The present research uses bidimensional regression models as descriptive statistics to determine the correspondence between OSM, TA, and SD. In Euclidean bidimensional regression the vectors of the regression equation are extended to be two-dimensional Cartesian coordinates pairs (x_i, y_i; u_i, v_i), where x_i, y_i are the estimated coordinates from the of OSM and TA data, respectively, and u_i, v_i are the associated dependent reference coordinates of the surveying data. A scaling, translation and rotation parameter

reflect how the estimated point pattern must be transformed to fit back into the reference point pattern. Thus, it is possible to quantify the geometrical relationship between two point patterns. The resulting bidimensional regression equation has following structural form:

$$\begin{pmatrix} u_j \\ v_j \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} x_j \\ y_j \end{pmatrix} + \begin{pmatrix} e_j \\ f_j \end{pmatrix} \quad (1)$$

where the parameters a_1 and a_2 correspond to the ordinary least squares (OLS) intercept term and they carry out the translation. The b_{ij} values conduct the scaling and rotation and can be understood as the slope coefficient in OLS regression. e_j and f_j are the errors.

First, the magnitude of the horizontal (a_1) and vertical (a_2) translation between the reference pattern and the independent pattern is estimated, determining a least squares solution. A positive value of a_1 indicates a west-to-east shift and a negative value indicates an east-to-west shift. Likewise positive values of a_2 are in accordance with a south-to-north shift and vice versa. Second, b_1 and b_2 are used to derive a scale parameter ϕ and angle parameter θ. Former causes the magnitude of contraction or expansion, whereas a ϕ value < 1 indicates a contraction and a ϕ value > 1 means an expansion relative to the reference pattern. The direction of the rotation to get the best fit is determined by the angle parameter θ. A positive θ value indicates a counterclockwise rotation and a negative θ a clockwise one (Lloyd, 1989; Friedman & Kohler, 2003). An overall "quality criterion" is the Distortion Index (DI) introduced by Waterman and Gordon (1984) and discussed in Friedman and Kohler (2003). This index "can be thought of as a standardized measure of relative error" (Lloyd, 1989, p. 110) and has a range between 0 and 100, where a lower value means less distortion.

3.3 Local Spatial Association of positional errors

Because bidimensional regression is a global statistic, it seems necessary to explore spatial heterogeneity in the positional errors as well. Therefore, an appealing method, among others, to detect local patterns of spatial association is the G^*-statistic (Getis and Ord, 1992). The G^*-statistic yields the proportion of the weighted sum of the variable within a distance d from location i as a proportion of the variable aggregated over the entire study region:

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j} \quad (2)$$

where x_j correspond to the value of the observation at j, $w_{ij}(d)$ is the ij element of the spatial weight matrix and n is the number of observations. As a result spatial clusters of high and low values can be evaluated. In our case, a cluster of high values (z-scores) means a clustering of high positional errors and low values (z-scores) are related to an accumulation of low errors, always compared to SD. Significance is tested via a randomization approach.

4 Results

The preprocessing algorithm was able to extract 121 identical road junctions within our three datasets. The resulting point pattern is visualized in Figure 2. It can be seen that the junctions are spatially bounded to urban areas. Taking the above stated hypothesis into account, that the SD serve as a spatially precise reference dataset, the spatial deviation between SD and OSM and TA, respectively, was evaluated.

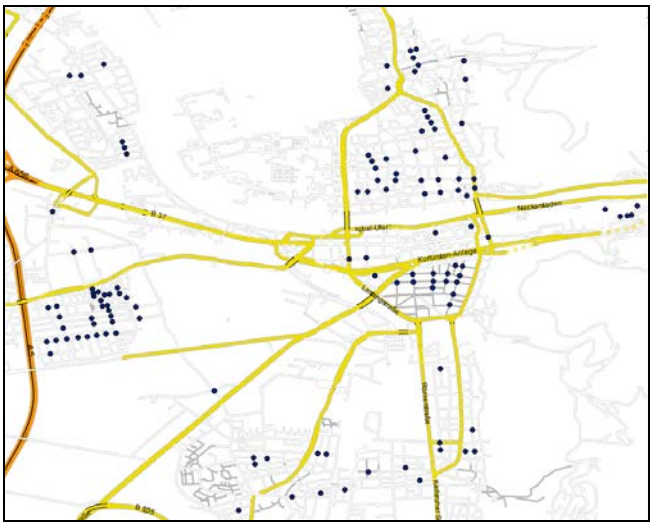

Fig. 2. Study site and identical road junctions (point signatures).

The mean deviation error is approximately one meter smaller in the OSM dataset, compared to TA (Table 1). A two sample Welch's t-test con-

firms significant differences between both mean values (t = -3.037, p = 0.003). The Fligner-Killeen-test is used to proof homogeneity of both variances. The result clearly rejects the null hypotheses (FK = 57.644, p < 0.001) and there are significant differences between the OSM and TA error variances. Moreover, OSM scatters more around the mean than TA, but comprising the directional scattering around the true position of the road junctions, as shown in Figure 3, it is noticeable that TA error clearly scatters more westward around the "true" position, than OSM does. The two varying mean centers of each point pattern support this finding and refer to a possible systematically variation.

Table 1. Descriptive statistics of the error deviation (in meters) between reference data and OSM and TA

	OSM	TA
Min.	0,220	2,759
Max.	18,694	13,607
Mean	5,229	6,145
Std. dev.	3,037	1,300

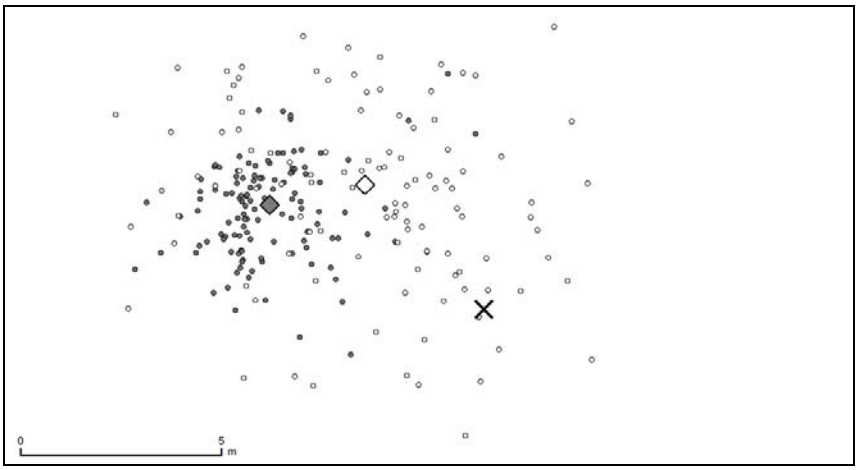

Fig. 3. Directional scattering around the "true" position of the road junctions. Darker points represents TA junctions, brighter ones OSM junctions, the cross marks the true position, and the rectangles show the spatial means of the error distributions

Hence, the geometrically distortion of the OSM and TA point pattern are compared to the one of SD on a global level, using the bidimensional regression framework. Thus, the OSM and TA pattern are regressed on the SD reference pattern, leading to the estimated parameters shown in Table

2. These parameters indicate how the OSM and TA pattern, respectively, must be transformed to get the SD pattern. Overall the OSM and TA pattern have the same geometrical distortion, hence having the same parameter signs. Both patterns are shifted east-to-west as well as south-to-north. Furthermore, the contraction or expansion parameter is negligible, because differences occur only after the five decimal place. θ refers to a clockwise rotation of the OSM and TA pattern, whereas OSM is marginally more rotated. The DI suggests that the relative error is slightly lower and thus the TA pattern corresponds more to the reference pattern. Nevertheless, the gaining of knowledge positional accuracy is not overwhelming and henceforth local statistics are used.

Table 2. Estimated Parameters (rounded) of the bidimensional regression (SD dependent variable, OSM or TA independent variable)

	a_1	a_2	b_1	b_2	ϕ	θ	DI
OSM	-35,844	0,690	1,000	-0,000	1,000	-0,011	0,178
TA	-2,561	19,351	1,000	-0,000	1,000	-0,009	0,095

Mapping the positional errors (Fig. 4) gives a first indication of spatial heterogeneity, but this impression needs some statistical validation. Therefore, to explore areas with high and low accuracy, the G^*-statistic is calculated. We applied the zone of indifference option for conceptualization of the spatial relationships between the entities, which is a combination of the inverse distance and fixed distance band model, leading to a neighborhood search threshold of 967 meters. Points with high z-scores and p-values below 0.05 indicate spatial clustering of high positional errors (approx. beyond +/- 2 standard deviations) and vice versa. Values between +/- 2 standard deviations suggest no significant clustering.

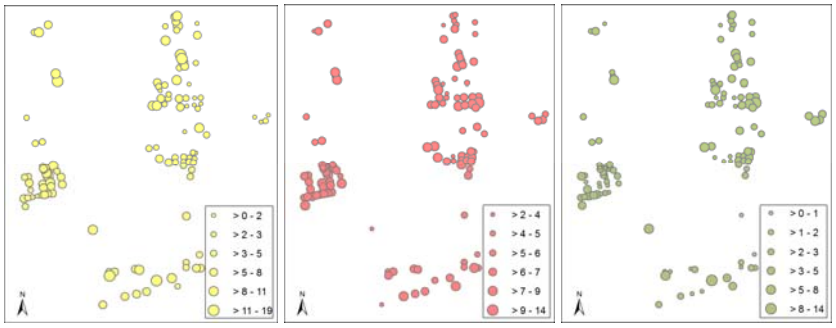

Fig. 4. Absolute deviation in meters between OSM and SA (left) and TA and SA (middle). Absolute value of deviation (in meters) between OSM and TA (right).

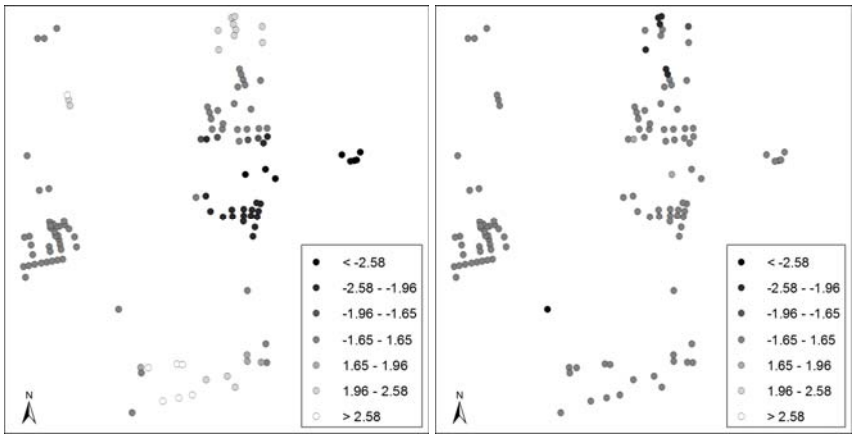

Fig. 5. Results of the *G**-statistics of OSM (left) and TA (right) in standard deviations. Values beyond +/- 2 standard deviations have significant *p*-values ($p<0.05$).

Both maps in Figure 5 show some significant clusters of low values, corresponding to clusters with high positional accuracy. In the case of OSM (Fig. 5 left), this cluster is primarily situated in the center of the map. 27 out of 121 observations have a significant negative *z*-score ($p < 0.05$). The opposite is valid for TA (Fig. 5 right), whereas these areas are located in the northern part of the map (7 significant observations). Positive values beyond 2 standard deviations are interpreted as badly mapped areas. In this regard, OSM shows some limitation, because such areas are present in the northern as well as southern part of the study site, whereas TA is not affected by limited position accuracy, compared to SA. Comparing the amount of such observations confirms this, OSM has 10 times more significantly imprecise mapped observations (OSM: 21, $p < 0.05$; TA: 2, $p < 0.1$). In general, TA map gives a more homogeneous impression of the position accuracy error.

5 Conclusions

The present paper is devoted to the comparison of positional accuracy of volunteered geographic information and proprietary geospatial data, using the case study of an well-mapped German city. On the one hand bidimensional regression analysis is applied to evaluate the global geometries of the patterns and on the other hand clusters of high and low precision are detected.

The results showed that both data sets, OSM and TA, have a highly positional accuracy and may be used for small and medium scale mapping applications. However, the bidimensional regression estimates referred to highest correlation between OSM/TA and their true position, but TA data had less distortion than OSM. The *G**-statistic resulted in some clusters with high a low positional accuracy, interpretable as spatial heterogeneity. Furthermore, the OSM areas of high accuracy are primarily located in the highly populated urban centers, leading to the conclusion that these areas are subject to a higher validation rate and consequently, errors are corrected more quickly than in rural areas. These findings are similar to those reported by Chen (2010), where urban areas have a higher (topological) accuracy. Hence, future comparisons between urban and rural areas seems fruitful, because rural areas are mapped with significantly less completeness (Zielstra and Zipf, 2010) but the continuously tremendous growth of OSM data may shrink this disparity. OSM as well as TA showed similar spatial distortion, which raises the question whether the SA are affected by inaccuracy.

Finally, future research is needed to get confidence, especially other reference datasets and more case studies must be analyzed and other methodological approaches must be tested.

References

Amelunxen, C. (2010) An Approach to Geocoding Based on Volunteered Spatial Data. in Zipf A. et al. (Eds.): Geoinformatik 2010. Die Welt im Netz, 2010, pp. 7-12.

Chen H. (2010) Entwicklung von Verfahren zur Beurteilung und Verbesserung der Qualität von digitalen Karten (PhD Thesis), University of Stuttgart, Germany.

Friedman, A. and Kohler, B. (2003) Bidimensional Regression: A Method for Assessing the Configural Similarity of Cognitive Maps and Other Two-Dimensional Data. Psychological Methods, 8, pp. 468-491.

Elwood, S. (2008) Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS. GeoJournal, 72, pp. 173-183.

Goodchild, M. F. (2007) Citizens as Sensors: the World of Volunteered Geography. GeoJournal, 69, pp. 211-221.

Goodchild, M. F. and Hunter, G. J. (1997) A Simple Positional Accuracy Measure for Linear Features. International Journal of Geographical Information Science, 11, pp. 299-306.

Getis, A. and Ord, J. K. (1992) The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis, 24, pp. 189-206.

Haklay, M. (2010), How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B, 37, pp. 682- 703.

Lloyd, R. (1989). Cognitive Maps: Encoding and Decoding Information. Annals of the Association of American Geographers, 79, pp. 101-124.

Neis, P. and Zipf, A. (2008), OpenRouteService.org is Three Times "Open": Combining OpenSource, OpenLS and OpenStreetMaps. GIS Research UK, Manchester.

Neis, P.; Zielstra, D.; Zipf, A. and Struck, A. (2010) Empirische Untersuchungen zur Datenqualität von OpenStreetMap - Erfahrungen aus zwei Jahren Betrieb mehrerer OSM Online-Dienste. in Strobl, J. et al. (Eds.): Angewandte Geoinformatik 2010: Beiträge 22. AGIT-Symposium Salzburg, 2010, pp. 420-425.

Neis, P.; Singler, P. and Zipf, A. (2010) Collaborative Mapping and Emergency Routing for Disaster Logistics - Case Studies from the Haiti Earthquake and the UN Portal for Afrika. in Car, A. et al. (Eds.): Geospatial Crossroads @ GI_Forum 2010. Proceedings of the Geoinformatics Forum Salzburg, 2010, pp. 239-248.

OpenStreetMap (2010), The Free Wiki World Map. http://www.openstreetmap.org/, Last date accessed October 10th, 2010.

Over, M; Schilling, A; Neubauer, S. and Zipf, A. (in press) Generating web-based 3D City Models from OpenStreetMap: The Current Situation in Germany. Computers, Environment and Urban Systems.

Schmitz, S.; Neis, P. and Zipf A. (2008) New Applications based on Collaborative Geodata - The Case of Routing. XXVIII INCA International Congress on Collaborative Mapping and SpaceTechnology, Gandhinagar, Gujarat, India.

Strunck, A. (2010) Raumzeitliche Qualitätsuntersuchung von OpenStreetMap (Master Thesis), University of Bonn, Germany.

Sui, D. (2008) The Wikification of GIS and its Consequences: Or Angelina Jolie's New Tattoo and the Future of GIS. Computers, Environment and Urban Systems, 32, 1-5.

Symington, A.; Charlton, M. and Brunsdon, C. (2002) Using Bidimensional Regression to Explore Map Lineage. Computers, Environment and Urban Systems, 26, pp. 201-218.

Tobler, W. (1965) Computation of the Correspondence of Geographical Patterns. Papers in Regional Science Association, 15, pp. 131-139.

Tobler, W. (1994) Bidimensional Regression. Geographical Analysis, 26, pp. 187-212.

Van Oort, P. (2006) Spatial Data Quality: From Description to Application (PhD Thesis), Wageningen University, The Netherlands.

Waterman, S. and Gordon, D. (1984). A Quantitative-Comparative Approach to the Analysis of Distortion in Mental Maps. The Professional Geographer, 36, pp. 326-337.

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