Transactions in GIS

Research Article

Transactions in GIS, 2013, 17(3): 315–334

Assessing the Effect of Data Imports on the Completeness of OpenStreetMap – A United States Case Study

Dennis Zielstra,* Hartwig H. Hochmair* and Pascal Neis†

**Fort Lauderdale Research and Education Center, University of Florida*
†*Geoinformatics Research Group, University of Heidelberg*

Abstract
The assessment of OpenStreetMap (OSM) data quality has become an interdisciplinary research area over the recent years. The question of whether the OSM road network should be updated through periodic data imports from public domain data, or whether the currency of OSM data should rather rely on more traditional data collection efforts by active contributors, has led to perpetual debates within the OSM community. A US Census TIGER/Line 2005 import into OSM was accomplished in early 2008, which generated a road network foundation for the active community members in the US. In this study we perform a longitudinal analysis of road data for the US by comparing the development of OSM and TIGER/Line data since the initial TIGER/Line import. The analysis is performed for the 50 US states and the District of Columbia, and 70 Urbanized Areas. In almost all tested states and Urbanized Areas, OSM misses roads for motorized traffic when compared with TIGER/Line street data, while significant contributions could be observed in pedestrian related network data in OSM compared with corresponding TIGER/Line data. We conclude that the quality of OSM road data could be improved through new OSM editor tools allowing contributors to trace current TIGER/Line data.

1 Introduction

The ubiquitous availability of GPS enabled devices boosted the development of a plethora of Volunteered Geographic Information (VGI) (Goodchild 2007, 2009) platforms on the Internet in the past few years. In October 2012, it was estimated that there were more than one billion worldwide smartphone users (Bicheno 2012). Some VGI platforms utilize more passive modes of data contribution, such as enabling the geolocation functionality in Twitter feeds, while other platforms require more active participation of data contributors, such as uploading and geolocating images in Panoramio, or contributing and editing map features in OpenStreetMap (OSM) (Heipke 2010). Coleman et al. (2009) distinguish between five groups of data contributors in VGI projects, ranging between "Neophyte" (someone with no formal background in a subject) and "Expert Authority" (someone who has widely studied and long practiced a subject). This wide range of VGI user types can also be found in the OSM project (Budhathoki et al. 2010, Neis and Zipf 2012), which has clear implications on the OSM data quality, especially given the fact that there is no regulatory instance in OSM that sets and enforces minimum quality standards.

Since the initiation of the OSM project in 2004, the number of registered users has increased exponentially and exceeded the one million mark as of January 2013 (OpenStreetMap 2013e). The OSM database hosts one of the largest VGI sources on the Internet with more than 1.7 billion nodes and 170 million ways as of February 2013. The impressive growth of the project,

Address for correspondence: Dennis Zielstra, Geomatics Program, University of Florida, Fort Lauderdale Research and Education Center, 3205 College Avenue, Ft. Lauderdale, FL-33314, USA. E-mail: dzielstra@ufl.edu

doi: 10.1111/tgis.12037

combined with the changes made to the licensing model by Google Maps in early 2012, which limits the usage of free Map APIs (Google 2012), motivated a number of websites and corporations such as Apple iPhoto (Bennett 2012), FourSquare (FourSquareBlog 2012), Craigslist (Cooper 2012) and Flickr (switch2osm 2012) to switch their mapping applications from Google Maps to OSM. Additionally, OSM received its first major funding from the Knight Foundation in September 2012, awarded to MapBox, a Washington DC-based company, which will utilize the funds to improve the OSM community site and develop tools that make it easier for contributors to add data to the OSM database (Gannes 2012).

The traditional way data are contributed to the OSM project is through active members who use their GPS enabled equipment, combined with their local knowledge, to add the most detailed information for their particular geographic region. Contributors commonly also digitize freely available aerial images and annotate the digitized geometries with attributes based on their local knowledge. These contributors are typically motivated by the fact that comprehensive sets of vector-based geodata are not freely available in their country (Budhathoki et al. 2010), so that the mapping of a region starts from scratch. Examples are Germany or the UK. While there is substantial growth in registered OSM members, mapping efforts differ between geographic regions, both worldwide and even within countries. This led to a second approach for OSM data contribution called bulk uploads. Bulk uploads are data imports which establish a base line of geodata so that the project can evolve, especially in countries or regions with a less active OSM contributor community. Despite the fact that OSM discourages data imports since they could heavily affect other contributors' manual data collection and editing efforts (OpenStreetMap 2013b), several datasets have been imported into OSM within the past few years. For example, a TIGER/Line US Census geodata import was conducted in 2007/2008 for the US, when the OSM data coverage in the US was still sparse. While this data import built the foundation for the active OSM community, the poor quality of the imported road data and additional problems with their conversion to the OSM tagging scheme limited OSM data usability and reliability. Because no TIGER/Line import has been conducted since then, a comparison of OSM and TIGER/Line data for the past six years (2007–2012) allows us to assess whether a newer TIGER/Line import to OSM would improve the data quality of OSM road data compared to the currently available OSM data sets. Although there are several commonly used measures of geodata quality (ISO/TC 211 2010), this article focuses on data completeness. More specifically it analyzes data growth for each data set individually and also compares growth rates between the two selected datasets over the years. The analysis will be conducted separately for network segments that are accessible to motorized traffic (highways, residential roads, etc.), and those that are accessible to pedestrians but inaccessible to motorized traffic (hiking trails, footpaths, etc.). We will further analyze how active the OSM community is in editing and updating once imported TIGER/Line features to OSM.

The remainder of the article is structured as follows: The next section reviews previous findings of geodata quality tests for OSM in the US and Europe. This is followed by a section on data retrieval procedures and data processing steps prior to conducting the analysis, and a section that analyses the two data sources for different years. The last section provides a summary and some ideas for future work.

2 Review of Research on OpenStreetMap Data Quality

2.1 Related Studies

Assessing the credibility of VGI sources has become a significant research topic in a variety of disciplines over the past few years (Flanagin and Metzger 2008). The complexity of the OSM road dataset provides opportunities to analyze this dataset with regards to a variety of geodata quality aspects, such as completeness, i.e. error of omission and commission, positional accuracy, attribute accuracy, logical consistency, semantic accuracy, and temporal accuracy (up-to-dateness) (ISO/TC 211 2010).

Past OSM data quality analyses reported in the literature were primarily conducted for European countries, such as the UK, France, Germany and Austria (Girres and Touya 2010, Haklay et al. 2010, Neis et al. 2012, Rehrl et al. 2013, Zielstra and Zipf 2010). In terms of data completeness, most of the conducted analyses revealed a higher OSM data quality in urban areas than in rural areas. This is because of a higher number of active contributors in urban areas, which results for some areas in a higher completeness of certain feature types (e.g. footpaths) compared with commercial or governmental datasets. Haklay et al. (2010) observed that positional accuracy increased with the number of different data contributors in an area, but only up to a certain number. With regards to metadata and attribute accuracy, Mooney and Corcoran (2012a, b) and Neis and Zipf (2012) found that for heavily edited objects there was no significant relationship beween the number of contributors to a feature and the number of tags assigned to it. Mooney and Corcoran (2012a) also analyzed the problem of misspellings in key values that are associated with primary features (e.g. highway, amenity, landuse). These errors can occur when users type in a value as free text instead of using OSM controlled core values provided in a drop-down-list in OSM editors.

2.2 Data Imports

Little research has so far been conducted for countries in which OSM relied on data imports. Although numerous spatial datasets are available under OSM compatible licenses that are being considered for import into OSM (OpenStreetMap 2013c), the OSM community is undecided on the benefits of data imports for the OSM project, especially for areas such as the US where large governmental datasets are freely available (OpenStreetMap 2013i, 2013j). Other countries such as India and the Netherlands conducted data imports of donated, professional Automotive Navigation Data (AND) into OSM (OpenStreetMap 2013a). However, for none of these data imports has their impact on OSM data quality been analyzed so far.

Several websites exist that help OSM members to evaluate their planned automated and scripted imports or large-scale edits prior to running them against the database. A specific code of conduct was introduced that outlines some of the major steps that need consideration by the member that is planning to import data, such as the potential impacts on to other work that could be affected by the new import, and discussions of the planned work on the mailing lists to gather feedback from more experienced members prior to import (OpenStreetMap 2013b). More detailed suggested guidelines regarding data imports can be found on a separate wiki page (OpenStreetMap 2013k), including factors, such as familiarizing oneself with the history of imports, obtaining proper permission and licenses from the original data owner, and documenting the import process. It is also stated that problems with prior data imports, such as with data that is in GIS data formats that does not translate well to OSM data formats, need to be taken into consideration. Such data import problems are listed on a designated wiki page (OpenStreetMap 2013l).

One deciding factor in considering the appropriateness of a data import is the quality of the dataset in question. The US TIGER/Line nationwide dataset was updated for the 2010 Census through the MAF/TIGER Accuracy Improvement Project (MTAIP) initiated in 2002. The main goal of the project was to align the MAF/TIGER databases with GIS files maintained by state, local and tribal partners and other agencies to improve the spatial accuracy and completeness of the dataset (Best 2005). Zandbergen et al. (2011) measured the positional accuracy of 2000 and 2009 TIGER/Line road networks through comparison with ortho imagery. Results showed that TIGER 2009 data are much improved in terms of positional accuracy compared with TIGER 2000 data, and that TIGER 2009 is consistently more accurate in urban areas than rural areas.

Although in 2006 these improvements were not yet fully integrated into the dataset two OSM contributors decided to import the TIGER/Line 2005 dataset to OSM (Linux.com 2007). After the first attempt failed due to data integrity problems and was cancelled in November 2006 the second attempt started in 2007 and successfully uploaded the 2005 TIGER/Line dataset, which was completed by February 2008 (OpenStreetMap 2013f). The US has now the second largest OSM community in the world among all countries, representing 9% of the total OSM community (Neis and Zipf 2012). However, the small amount of OSM data contributions made up to 2006 and the availability of a free governmental road dataset at the same time seemed to make a TIGER/Line data import especially beneficial for the US at that time.

Only recently a few studies analyzed OSM data completeness in the US. One study compared data completeness between OSM and commercial datasets (TeleAtlas and NAVTEQ) for Florida (Zielstra and Hochmair 2011b). As opposed to Germany, OSM data coverage was found to be generally higher in rural areas when compared with commercial datasets, while OSM coverage was lower in urban areas. The good coverage for OSM in rural areas does not stem from individual user contributions but from TIGER/Line imports which contain more data than commercial datasets in agricultural areas (tracks, gravel roads). Another study showed that in four of five analyzed US cities OSM contained by far more pedestrian-only segments than TIGER/Line data (Zielstra and Hochmair 2011a) due to an active OSM community in cities such as Chicago or San Francisco. A higher density of OSM pedestrian segments compared to TIGER/Line led also to shorter shortest paths in pedestrian routing (Zielstra and Hochmair 2012). A recent analysis on OSM bicycle facilities shows large differences in the completeness of mapped off-street bicycle trails and designated lanes between 70 Urbanized Areas in the US, indicating different levels of OSM mapping activities in the different regions (Hochmair et al. 2013).

3 Study Setup: Data Sources and Data Processing

During the time of the 2005 TIGER/Line data import into OSM, TIGER/Line was stored in a now outdated ASCII data format which was replaced by the more common Esri shapefile format in 2007. The 2005 TIGER/Line dataset is not available for download anymore and could therefore not be integrated within the analysis. The 2006 TIGER/Line dataset, still in ASCII format, was incomplete for some areas, which rendered the dataset unusable for analysis, too. Since the data import into OSM was finished by February 2008 (OpenStreetMap 2013f), a comparative analysis between OSM and TIGER/Line in this article is conducted for the years 2008 through 2012.

The retrieval and processing of the OSM datasets is based on two files that are available at the OSM projects website (OpenStreetMap 2013h). For the majority of the tested years (2006–2011) a full history dump file, dated June 2012, was downloaded and processed through a variety of self-developed tools written in JAVA, as illustrated in Figure 1. We use full history dump files for analyzing OSM development since they contain all historical informa-

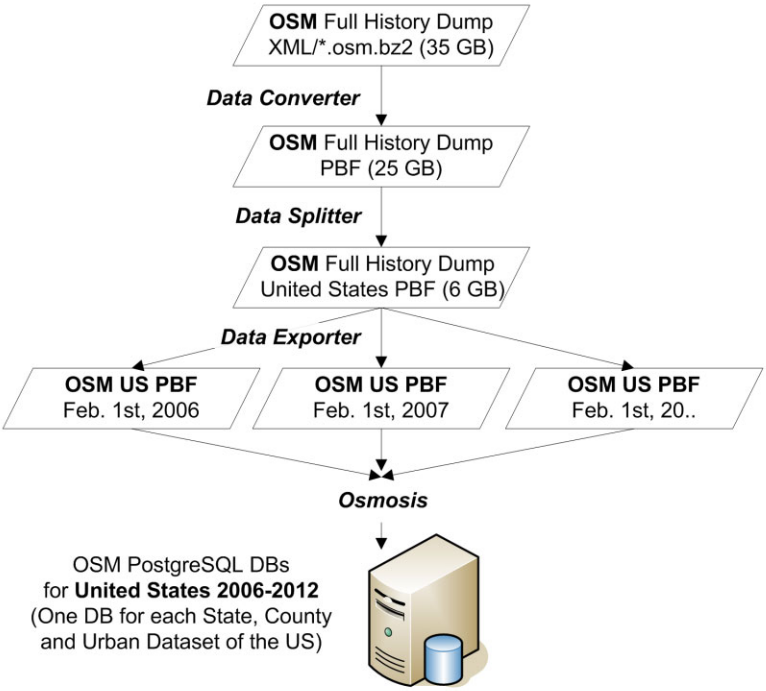


Figure 1 OSM data processing workflow

tion of the dataset, i.e. all versions for nodes, ways, and relations that ever existed. In a first step the downloaded dataset was converted from the standard XML format into the Protocolbuffer Binary Format (PBF) (OpenStreetMap 2013m) which facilitated faster processing. Next, in a data splitting process the dataset was extracted for the 50 US states and the District of Columbia. This was followed by extracting a state-wide dataset for each individual year which was then imported into a PostgreSQL database using the OSMOSIS open-source command line JAVA tool (OpenStreetMap 2013g). In this procedure the 2006 OSM dataset was also included to obtain more detailed information about the data development in the US before the actual data import. The years 2006 and 2007 were not included in the comparison with TIGER/Line data but used only for longitudinal analysis within OSM data.

The second OSM dataset used was a planet dump file dated September 12, 2012. Although the planet dump file in comparison to the history dump file does not include the history of all features, it provides the latest information of the entire OSM database at a specific point in time. The September 12, 2012 OSM dataset was chosen since it already reflects the changes and potential data losses in the OSM database that may have been caused by the redaction bot that was introduced to change the database from a creative commons Attribution-ShareAlike (CC-BY-SA) license to an Open Database License (ODbL) in July 2012. Due to the license change, all data that was created by contributors who did not agree to the new license had to be deleted, which was accomplished with the redaction bot. The impact of these changes varied from country to country. In total, about 99% of the OSM data was retained, whereas larger negative impacts could be determined in Australia, Poland and Egypt (OpenStreetMap Foundation 2013). The planet dump file from September 2012 was also split into different geographic areas and imported into a PostgreSQL database. All data filtering and length measurement procedures were accomplished within PostgreSQL databases with PostGIS extensions using SQL commands.

The geographic information stored in the OSM database is categorized into three data types (called elements), which are nodes, ways and relations. A node represents a point with latitude and longitude coordinate information. Lines such as roads and polygons are stored as ways. The logical or geographic relationships between objects, such as turn restrictions, are stored in relations. Elements are expressed through tags, where each tag consists of a key and a value. A key specifies the feature type of an element (e.g. a highway) or the attribute associated with an element (e.g. speed limit), and the value more specifically describes its accompanying key. OSM uses a total of 26 primary feature keys including building, highway, or landuse (OpenStreetMap 2013d). Filtering the data by particular tags allows the differentiation between car accessible and pedestrian-only way segments in data analysis. In the TIGER/Line dataset the filtering of way segments was accomplished through MTFCC (MAF/TIGER Feature Class Code) attribute values (Census 2012b) which were introduced with the 2007 TIGER/Line dataset. The imported 2005 TIGER/Line data relied on the former CFCC (Census Feature Class Codes) classification schema (Census 2012a) which differs from the new schema by providing additional sub-categories of road type classes, e.g. by distinguishing between separated and unseparated roads within the same road type class. For example the primary road class with CFCC values ranging from A11 to A29 is now classified as MTFCC = S1100.

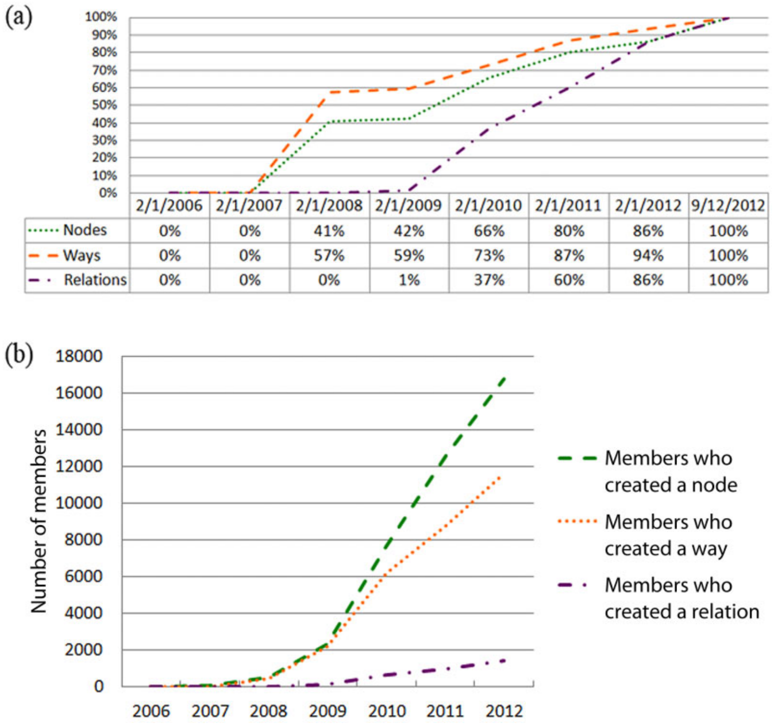
Our approach of length comparison per administrative unit does not incorporate the matching of line segments between the two datasets (Koukoletsos et al. 2012), since this method becomes unfeasible for larger areas such as the entire US. Also, automated scripts would inevitably introduce new errors. However, we believe that the results provide information about the development of the OSM and TIGER/Line datasets after the 2008 import. Both the analysis for states and Urbanized Areas as a whole reduces the effect of potential outlier regions that would appear if applying the method to smaller areas (e.g. 1 km^2 grid fields).

4 Analysis: The Effect of TIGER/Line Imports on US OSM Data Development

4.1 OSM Data Development

Figure 2a shows the increase of nodes, ways and relations in the US before and after the TIGER/Line import. The contributions to the project in 2006 were too minimal to show any increase in the diagram. The bulk upload in 2007/08, indicated by the steep increase of nodes and ways between 2007 and 2008, helped to create the foundation for OSM in the US. In the following years, active contributors added nodes, ways, and relations. The diagram also shows that more new contributions were added between February 2012 and September 2012 than were deleted through the redaction bot following the license change. Figure 2b visualizes the number of contributors that created at least one node, way or relation in the US in the given year, indicating a steady increase in active members.

Next, a more detailed analysis of the differences in total length between consecutive years was conducted for selected OSM feature classes. One of the major challenges associated with data imports into OSM is to correctly match feature classes of the source TIGER/Line data sets with the OSM data schema. Table 1 shows the absolute differences (given in thousands of km) and relative differences between consecutive years for the 24 most common highway classes in the US OSM dataset. While the length of most highway classes increased between consecutive years, it can be noted that the length of residential classes decreased each year (shown in gray). Further inspection of affected road segments revealed that the decrease in



	2/1/2006	2/1/2007	2/1/2008	2/1/2009	2/1/2010	2/1/2011	2/1/2012	9/12/2012
Nodes	0%	0%	41%	42%	66%	80%	86%	100%
Ways	0%	0%	57%	59%	73%	87%	94%	100%
Relations	0%	0%	0%	1%	37%	60%	86%	100%

Figure 2 (a) Development of OSM nodes, ways and relations in the US, and (b) OSM member activity in the US, 2006–2012

residential classes was mostly caused by continuous corrections of OSM tagging errors that occurred during the original TIGER/Line data import. More specifically, while TIGER/Line groups residential, local neighborhood roads, rural roads and city streets into one road class (i.e. CFCC = A41–A49 or MTFCC = S1400, respectively), OSM uses a more refined data schema that provides different highway tag values for these different road classes. The TIGER/Line data import apparently incorrectly assigned all TIGER/Line roads with the aforementioned codes to the residential road class in OSM, although not all of these streets are residential roads. Figure 3a shows the decrease in OSM roads tagged as residential over the past few years while at the same time an increase in the classes track, tertiary and service can be observed. This can be attributed to the retagging of residential tagged roads to other classes. The decreasing difference between 2011 and 2012 for the residential class also indicates that the retagging process has slowed down within the past year.

Although the TIGER/Line import provides a foundation for contributors in the US, the TIGER/Line dataset mainly focuses on the general road network for the US. However, as has been shown earlier for European countries, OSM is a particularly valuable data source for pedestrian- and cyclist-related features. Figure 3b shows the annual length differences for three OSM highway classes representing cycleways, footways and paths. All three classes grew over the past few years, indicating that new contributions for non-motorized traffic were made after the initial TIGER/Line import for motorized traffic data.

Using the OSM full history dump file we can determine how many TIGER/Line features were edited after their import to OSM. Filtering the data by the name of the OSM contributor

Table 1 Absolute and relative annual differences between OSM highway classes in the US

	Feb 2008 to Feb 2009		Feb 2009 to Feb 2010		Feb 2010 to Feb 2011		Feb 2011 to Feb 2012		Feb 2012 to Sep 2012	
OSM – tag	km/1000	%	km/1000	%	km/1000	%	km/1000	%	km/1000	%
bridleway	0.2	92.3	0.2	53.6	0.1	19.9	0.3	37.6	0.7	45.6
construction	0.2	99.7	0.5	68.5	1.2	63.8	1.8	47.7	0.7	16.7
cycleway	2.8	71.0	7.3	65.2	5.7	33.8	6.4	27.3	6.3	21.3
footway	4.6	14.8	6.8	17.9	14.0	26.9	14.1	21.2	9.1	12.0
living_street	0.1	100.0	0.1	57.2	0.1	35.3	0.3	46.6	0.3	31.5
motorway	4.4	2.6	18.6	9.9	5.8	3.0	4.8	2.4	2.8	1.4
motorway_link	1.2	2.7	1.6	3.5	5.3	10.2	3.8	6.8	2.4	4.2
path	2.1	100.0	15.7	88.4	12.3	40.9	14.2	32.0	11.2	20.2
pedestrian	0.1	77.0	0.2	63.2	0.3	41.6	0.3	35.6	0.1	12.2
platform	0.0	0.0	0.0	100.0	0.0	65.2	0.0	58.2	0.0	7.9
primary	0.6	0.2	4.1	1.4	–9.6	–3.5	35.7	11.6	7.4	2.4
primary_link	1.5	90.7	2.3	57.0	–1.0	–32.8	–0.6	–24.8	0.0	2.0
residential	–71.3	–0.8	–240.6	–2.8	–267.4	–3.2	–330.6	–4.2	–172.3	–2.2
road	1.5	100.0	2.2	59.3	0.7	16.0	4.5	50.0	1.6	14.8
secondary	3.0	0.5	6.7	1.2	–6.8	–1.2	–44.9	–8.8	11.3	2.2
secondary_link	0.7	97.3	0.6	47.6	–0.2	–21.4	0.5	30.3	0.2	10.2
service	8.2	1.4	17.2	2.8	27.7	4.2	36.6	5.3	29.3	4.1
steps	0.0	39.5	0.0	52.9	0.0	32.0	0.1	29.9	0.0	12.6
tertiary	26.5	71.9	97.6	72.6	156.6	53.8	125.3	30.1	65.1	13.5
tertiary_link	0.0	98.5	0.0	–24.0	0.1	68.6	0.3	70.1	0.3	40.8
track	36.0	8.2	27.8	6.0	46.8	9.1	146.4	22.2	90.3	12.0
trunk	26.7	85.5	23.2	42.6	40.2	42.5	40.9	30.2	3.0	2.2
trunk_link	1.1	83.3	2.0	61.5	1.0	22.5	1.0	19.1	0.4	7.5
unclassified	4.1	21.1	17.1	47.1	14.4	28.5	17.4	25.6	21.4	23.9

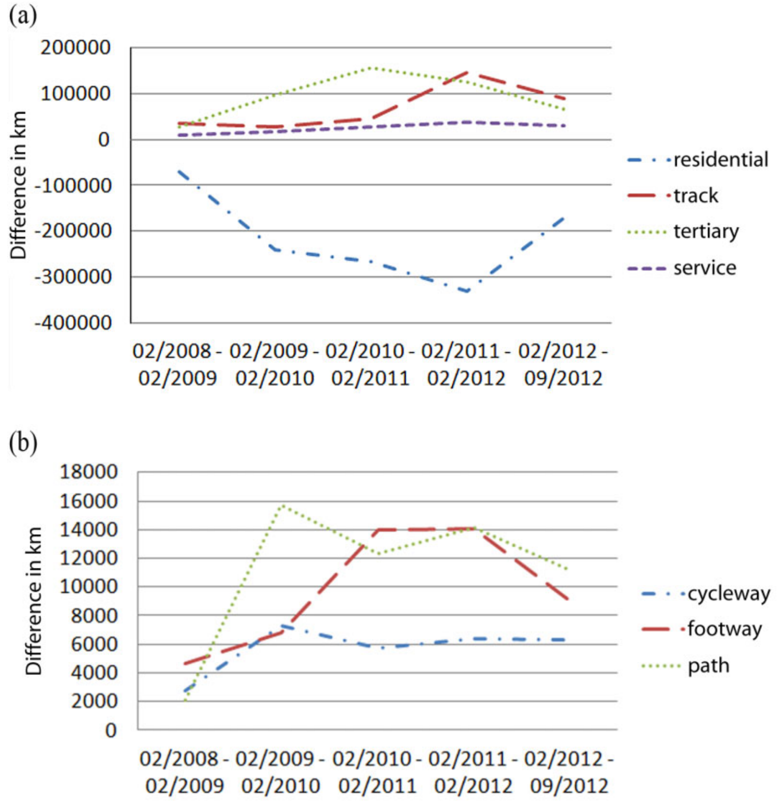


Figure 3 Annual differences between (a) selected OSM highway classes and (b) selected pedestrian- and cyclist-related OSM highway classes in the US 2006–2012

who imported the TIGER/Line dataset and by a version number of one, it was found that about 11.20 million TIGER/Line features were imported for motorized (about 11.19 million features) and pedestrian travel (about 10,000 features). Next we analyzed how many of the imported road features were edited and by whom. Edits can either be made by humans or an automated script (bot). Bots can be identified by counting the number of edits made by a contributor per change set, besides the total number of edits. A very large number of edits per change set (impossible to achieve by a human editor) in combination with a large total number of edits indicates the work of a bot. Utilizing user IDs of human contributors and bots, and the history of versions of each feature with their annotated user IDs, we analyzed how many of the original TIGER/Line imported road features for motorized traffic were edited by bots only, by humans only, or by both (Figure 4). While in general the majority of modified features comes from human edits the bot activity was particularly high in 2010–2011, editing almost 18% of the imported TIGER/Line data. Because the development of these automated scripts is complex, only few OSM contributors make changes to the database with this method.

As shown in Figure 4 the number of edited features ranged between 5 and 6% over the past few years, suggesting some contributor activity towards data quality improvement. Using the same principle as before (i.e. number of edits per changeset), we can also identify additional data imports that ran after or parallel to the original TIGER/Line import. One of them was an import of about 200,000 road features from the Commonwealth's Office of Geo-

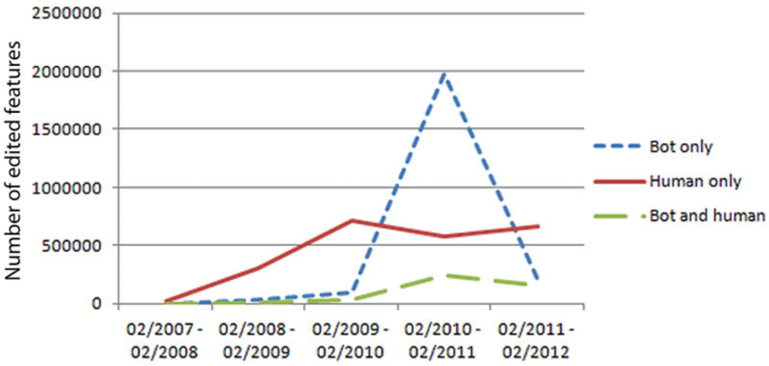


Figure 4 Number of edited Tiger/Line road features after import to OSM

graphic and Environmental Information (MassGIS) for the Massachusetts area between 2007 and 2008. No features for non-motorized road segments were imported into OSM after the TIGER/Line import, proving that the increase of these features in the OSM database for the US is based on contributions by active members.

4.2 Road Segments for Motorized Traffic

This section identifies in which regions of the US the OSM community actively contributes road segments for motorized traffic. This is based on a comparison of data growth of TIGER/Line and OSM data between the years 2008 and 2012. The comparison is conducted for the 50 US states and the District of Columbia, as well as for 70 Urbanized Areas in the US with a population larger than 500,000. The following attribute filters were applied to the OSM and TIGER/Line data sets, respectively, to obtain segments that are accessible to motorized traffic:

- OSM: [key: highway] [value: motorway, motorway_link, trunk, trunk_link, primary, primary_link, secondary, secondary_link, tertiary, tertiary_link, unclassified, residential, track, living_street]
- TIGER/Line: [key: MTFCC] [value: S1100, S1200, S1400, S1630, S1640]

Table 2 shows the differences in thousands of km between OSM and the corresponding TIGER/Line dataset for motorized traffic for selected states, where TIGER/Line total lengths are subtracted from OSM total lengths (left column for each year). The 2011 TIGER/Line dataset was the most recent dataset available from the Census Bureau at the time of this analysis, thus the comparison in the right-most two columns was based on OSM 2012 and TIGER/Line 2011 datasets.

A negative difference value in Table 2 implies that OSM is missing data compared to TIGER/Line in the analyzed area while for positive values the converse is the case. Additionally, a second value expressing a normalized difference is shown for each year. This value controls for the differences in total street network length of analyzed states. It takes the difference between the two data sources (left values in Table 2 columns) and divides them by the total TIGER/Line 2011 road network length of the state and multiplies the result by 1,000. Based on these normalized difference values a ranking was established that sorts the states from the best (low rank number) to worst (high rank number) in terms of OSM data completeness relative to TIGER/Line.

The results reveal that out of the five highest ranked states only two states, i.e. Nevada and Alaska, show a positive difference in 2012, indicating that these two are the only states

Table 2 Absolute and relative differences between OSM and TIGER/Line data for motorized traffic per state

		OSM Feb. 2008 minus TIGER / Line 2008		OSM Feb. 2009 minus TIGER / Line 2009		OSM Feb. 2010 minus TIGER / Line 2010		OSM Feb. 2011 minus TIGER / Line 2011		OSM Sept. 2012 minus TIGER / Line 2011	
Rank	State	km / 1000	(dif/Σ) x 1000	km / 1000	(dif/Σ) x 1000	km / 1000	(dif/Σ) x 1000	km / 1000	(dif/Σ) x 1000	km / 1000	(dif/Σ) x 1000
1	Nevada	21.9	123.3	19.9	111.7	7.3	41	8.7	48.9	10.8	60.6
2	Alaska	11.0	200.8	11.2	204.0	2.3	42.8	2.5	45.5	2.9	52.2
3	Wyoming	18.6	81.5	22.3	97.8	–3.6	–15.9	–3.1	–13.6	–2.5	–11.1
4	New Mexico	6.1	18.3	12.7	38.2	–22.1	–66.7	–20.3	–61.2	–21.2	–64
5	North Dakota	22.2	87.4	22.1	87.1	–10.9	–43	–14.9	–58.6	–17	–66.8
. . .						. . .					. . .
47	Illinois	–32.3	–78.9	–7.4	–18.0	–126.8	–309.6	–127.1	–310.4	–129.9	–317.1
48	North Carolina	–9.2	–23.8	–8.9	–23.1	–130.9	–339.6	–129.1	–335.1	–126.4	–328.0
49	Ohio	3.3	9.0	0.3	0.8	–131.4	–353.5	–125.3	–336.9	–126.7	–340.9
50	Delaware	–1.2	–51.3	–1.5	–66.6	–9.2	–409.1	–8.9	–394.9	–8.2	–364.5
51	West Virginia	–16.9	–96.9	–16.7	–95.2	–71.0	–405.8	–71.9	–410.7	–71.9	–411.2

where OSM provides more data compared to TIGER/Line. However, a more detailed review of the datasets shows that the positive difference values in both states stem from uncorrected tagging errors that occurred in the re-classification process during the TIGER/Line data import in early 2008. More specifically, during the import, CFCC A41–A49 TIGER/Line classes (residential, local neighborhood roads, rural roads and city streets), were incorrectly tagged as residential in OSM during the import, contributing to a high OSM total length of roads for motorized traffic. While this error was left essentially uncorrected in Nevada and Alaska, in other states the corrective retagging process was progressing fast due to an active OSM community, reducing the OSM total length towards a more correct value.

The data import in 2007/08 was based on an inaccurate and outdated 2005 TIGER/Line dataset. This dataset has, however, been significantly updated by the Census Bureau since then. More specifically, in preparation for the 2010 Census, Census Bureau employees used handheld computers that captured GPS information to improve both address lists and the census road network as part of the aforementioned MTAIP project (Census 2012c). This increase in data collection by the Census Bureau is also reflected in Table 2 where all selected states show a sudden change in length differences between OSM and TIGER/Line in 2010. This gap indicates that OSM data contributions in many US states are not able to keep up with the additions made to the TIGER Line 2010 dataset by the US Census Bureau.

The least complete OSM network data for motorized traffic was found in the midwestern and eastern US states of Illinois, North Carolina, Ohio, Delaware and West Virginia. West Virginia showed the largest difference between the two providers with almost 72,000 km missing data in OSM in 2012. Previous research for Europe has shown that OSM has more active mapping communities in urban areas than in rural areas (Neis et al. 2012, Zielstra and Zipf 2010). Thus we speculate that the lack of OSM data for West Virginia is because of the dominance of rural areas in this state and a lack of an active OSM community as a consequence thereof. Figure 5a visualizes for a sample area in West Virginia the differences between the two data providers. The OSM 2012 dataset in light green (light gray) misses several roads both in more densely populated and in rural areas compared with the TIGER 2011 dataset in orange (dark gray). The lack of active community members in the area did not allow for noticeable improvement of the OSM dataset over the past few years. While the OSM community is generally more active in urban areas, OSM is still behind with data updates even in urban areas. Figure 5b shows a residential area of Jacksonville, FL, where newly developed areas and roads are contained in the TIGER/Line dataset (red/dark gray) but not mapped in the OSM database (light green/light gray).

To control for the effect of rural vs. urban area on OSM data completeness we also computed length differences in the same manner for 70 Urbanized Areas with a population larger than 500,000. Table 3 shows the differences between OSM and TIGER/Line data for the top and bottom five ranked Urbanized Areas in hundreds of km. As for states, a normalized value, now based on total length of TIGER/Line data from 2011 for the Urbanized Area under consideration, was also computed.

All analyzed areas show negative difference values for 2012, indicating less complete datasets for OSM even in Urbanized Areas. The five top ranked Urbanized Areas, three of which are located in California, show absolute differences between 150 km (Fresno, CA) and almost 1,300 km (Houston, TX) in 2012. As before with the state comparison, Table 3 shows a sudden increase in length differences between OSM and TIGER/Line for 2010 due to intensified data collection efforts through the US Census Bureau.

In summary, the results indicate that the OSM community is not very active in updating and correcting road data for motorized traffic once imported.

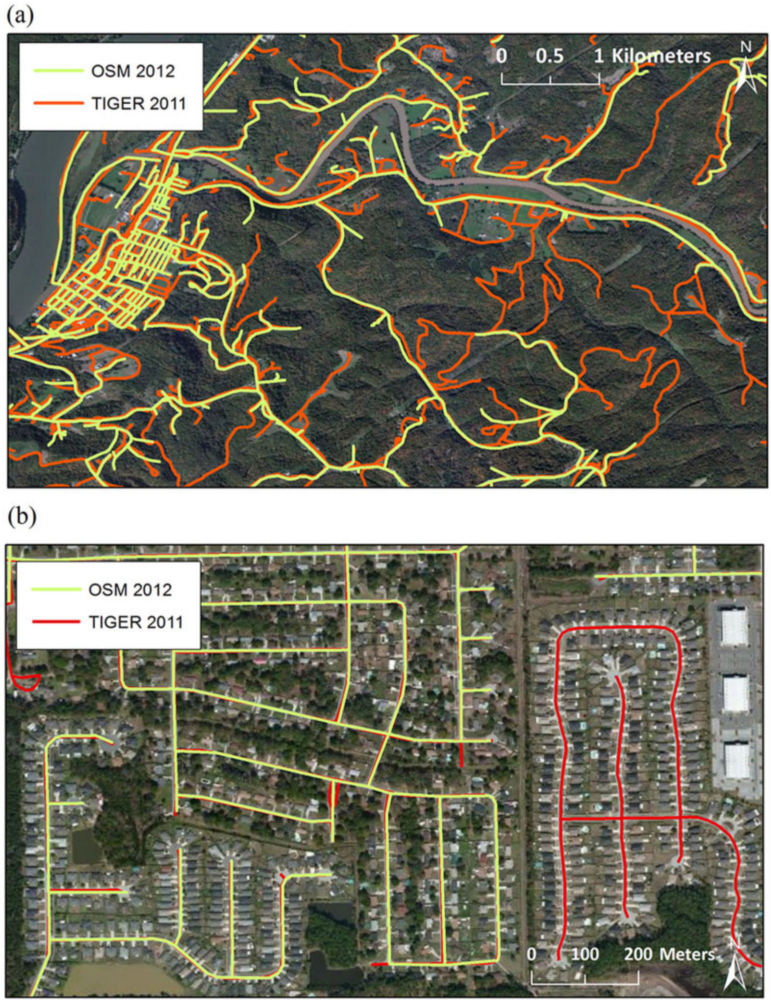


Figure 5 Visualization of OSM and TIGER/Line network data for motorized traffic in (a) West Virginia and (b) Jacksonville, FL

4.3 Path Segments for Pedestrian Navigation

This section analyzes the completeness for network segments that allow pedestrian movement but at the same time are closed to motorized traffic. An SQL query was applied to extract corresponding types of network data from the OSM and the TIGER/Line datasets using the following attribute values:

- OSM: [key: highway] [value: pedestrian AND area='no', path, footway, bridleway, steps]
- TIGER/Line: [key: MTFCC] [value: S1720, S1710, S1820]

The results gathered for the state comparison shown in Table 4 illustrate that the focus of the US OSM community has been on the collection of non-motorized network information after the import of the main road network in 2008. In 2012 almost all states showed a positive difference value between OSM and TIGER/Line indicating a better coverage of non-motorized

Table 3 Absolute and relative differences between OSM and TIGER/Line data for motorized traffic in selected US Urbanized Areas

		OSM Feb. 2008 minus TIGER / Line 2008		OSM Feb. 2009 minus TIGER / Line 2009		OSM Feb. 2010 minus TIGER / Line 2010		OSM Feb. 2011 minus TIGER / Line 2011		OSM Sept. 2012 minus TIGER / Line 2011	
Rank	State	km / 100	(dif/Σ) x 1000	km / 100	(dif/Σ) x 1000	km / 100	(dif/Σ) x 1000	km / 100	(dif/Σ) x 1000	km / 100	(dif/Σ) x 1000
1	Houston, TX	56.0	149.4	54.1	144.2	–4.8	–12.9	–5.4	–14.4	–12.6	–33.7
2	Fresno, CA	0.2	4.1	–1.6	–42.1	–1.1	–30.4	–1.1	–30.1	–1.5	–40.6
3	Sacramento, CA	–2.5	–25.2	–3.2	–32.8	–7.0	–71.2	–6.1	–62.4	–5.3	–54.7
4	Dayton, OH	13.5	195.3	12.7	184.3	–4.7	–67.5	–4.6	–66.9	–4.4	–63.5
5	San Diego, CA	–5.8	–33.1	–3.8	–22.0	–15.1	–86.7	–13.5	–77.5	–11.6	–66.4
. . .						. . .					. . .
66	Las Vegas, NV	–4.0	–38.4	–3.6	–34.4	–29.5	–280.1	–29.1	–276.6	–27.5	–260.8
67	Tulsa, OK	0.1	1.0	–0.5	–7.0	–23.0	–293.9	–22.8	–291.0	–21.2	–270.6
68	McAllen, TX	–5.2	–75.3	–4.5	–64.8	–21.1	–305.0	–20.7	–299.5	–20.2	–291.9
69	Salt Lake City, UT	–5.9	–64.5	–5.8	–63.7	–38.8	–423.7	–38.5	–421.2	–35.4	–386.6
70	Jacksonville, FL	–3.1	–20.9	–3.1	–21.4	–66.2	–453.1	–65.6	–449.1	–64.5	–441.6

Table 4 Absolute and relative differences between OSM and TIGER/Line for pedestrian-only accessible segments per state

		OSM Feb. 2008 minus TIGER / Line 2008		OSM Feb. 2009 minus TIGER / Line 2009		OSM Feb. 2010 minus TIGER / Line 2010		OSM Feb. 2011 minus TIGER / Line 2011		OSM Sept. 2012 minus TIGER / Line 2011	
Rank	State	km	(dif/Σ) x 10000	km	(dif/Σ) x 10000	km	(dif/Σ) x 10000	km	(dif/Σ) x 10000	km	(dif/Σ) x 10000
1	District of Columbia	10.9	49.1	104.9	473.1	124.6	561.8	151.1	681.2	211.1	951.9
2	Massachusetts	–122.9	–13.3	–12.0	–1.3	561.4	60.7	1284.9	138.8	3241.2	350.2
3	Colorado	–25.9	–0.9	182.3	6.5	2154.3	76.7	4626.3	164.7	8346.5	297.1
4	New Hampshire	34.8	7.2	76.4	15.8	213.6	44.1	881.1	181.8	1391.7	287.2
5	California	189.0	2.8	2407.4	35.3	6632.0	97.4	11178.6	164.1	16617.2	243.9
. . .						. . .					. . .
47	Louisiana	–3.2	–0.2	63.1	3.1	45.6	2.2	102.8	5.0	165.6	8.0
48	Arkansas	0.9	0.0	1.4	0.1	–2.0	–0.1	11.9	0.4	160.6	5.9
49	Mississippi	–1.1	–0.1	0.4	0.0	70.1	3.3	71.0	3.3	108.5	5.1
50	West Virginia	30.3	1.7	31.8	1.8	128.2	7.3	166.6	9.5	51.6	2.9
51	Alaska	–249.1	–45.6	–189.0	–34.6	–191.8	–35.1	–248.8	–45.5	–116.0	–21.2

segments in OSM. This difference amounts to over 16,600 km for California, which has the highest overall difference in all years. Table 4 shows also a normalized difference value as before. Some of the top-ranked states, i.e. California, Colorado, and Washington DC, contain Urbanized Areas that were found in earlier studies also to be top-ranked in terms of OSM mapping activities for bicycle trails, including Denver-Aurora, Portland, San Francisco, and Washington DC (Hochmair et al. 2013). Thus the OSM community appears to be contributing both bicycle and pedestrian related infrastructure in similar ways.

The five lowest ranking states showed only small positive differences of less than 170 km. Alaska, as the only state, even shows a small negative difference value of –116 km for 2011/2012 which implies missing pedestrian data in OSM compared with TIGER/Line. The data increase in the 2010 TIGER/Line dataset did not affect the differences in pedestrian-related information as it did in the previous car-related information analysis, indicating that the main focus of the government agency is still on the collection of vehicular road network geometries and not on pedestrian data.

To illustrate the OSM data collection efforts for pedestrian network features, Figure 6a highlights trails in Yosemite National Park, California. In addition to the imported TIGER/Line data (orange, or dark gray) which overlaps with the OSM dataset in the eastern and western part, OSM provides additional hiking trail features in the center section (light green, or light gray), which were either actively collected or traced from satellite images by the OSM community.

Another example is shown for the area around the National Mall in Washington, DC (Figure 6b), with pedestrian paths in OSM highlighted in light green (light gray).

Two of the five Urbanized Areas with the least pedestrian related information in OSM (Table 5) are located in Texas, one of which (McAllen) did not show any OSM data contribution of pedestrian network segments over the past five years. In summary it can be stated that the coverage of pedestrian network data in OSM is higher than for TIGER/Line data, which is the opposite of what has been found for roads that are accessible to motorized traffic.

5 Discussion and Conclusions

Whether or not recurring data imports benefit the OSM project has been an ongoing discussion in the OSM community. The presented research provides a first insight into the effect of data imports on OSM data quality in the US through a longitudinal analysis of OSM and TIGER/Line development over the past few years. By analyzing the TIGER/Line import of 2007/08 it was found that the imported 2005 TIGER/Line dataset was outdated and erroneous, whereas recent efforts by the US Census Bureau led to a significant improvement of TIGER/Line road data in 2010. While this improvement of data quality would primarily benefit regions with a less active OSM user community when used in recurring data imports of TIGER/Line data, the tagging error as a result of different road classification schemas in TIGER/Line and OSM data would remain. Thus, while for many areas the errors were mitigated through a crowd-based retagging process, the classification errors would re-occur with a new TIGER/Line import. This would mean the loss of user corrections that were applied to road classification tags in OSM over recent years. This fact alone makes a new TIGER/Line data import to OSM highly impracticable. Although this problem would not mean a significant loss of data for areas with a less active OSM community, such as Alaska and West Virginia, it would affect most regions. The particularly active OSM data contribution of pedestrian related segments over the past years would also be affected in case of a recurring

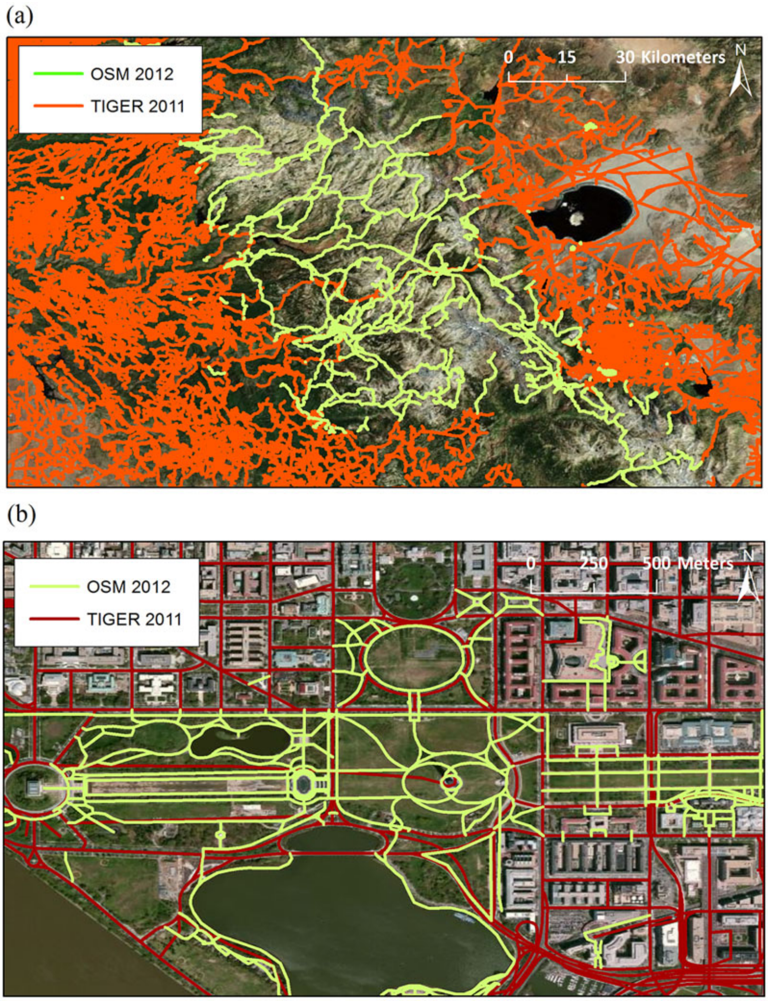


Figure 6 Visualization of OSM and TIGER/Line pedestrian network data in (a) Yosemite Park, California and (b) the National Mall area, Washington DC

TIGER/Line data import in different ways. For example, pedestrian related segments that have been snapped to a road geometry in an earlier TIGER/Line version could become disconnected from that road segment through later updates. Like OSM, other collaborative mapping projects such as Google Map Maker, Waze, or Cyclopath could rely on high-quality, freely available datasets as a reference when editing map data. However, the problem of data imports conflicting with community based-edits would remain here as well.

The data analysis also revealed that the US OSM user community does not focus on improving the completeness of the originally imported TIGER/Line dataset, although, as shown in Figure 4, the number of edited TIGER/Line imported OSM road features has increased over the past few years. Network data for motorized traffic, in particular, is inaccurate or missing in OSM when compared with current TIGER/Line data, independent of state or Urbanized Area. Given the arguments against recurring bulk uploads discussed before, we

Table 5 Absolute and relative differences between OSM and Tiger/Line for pedestrian-only accessible segments in selected US Urbanized Areas

		OSM Feb. 2008 minus TIGER / Line 2008		OSM Feb. 2009 minus TIGER / Line 2009		OSM Feb. 2010 minus TIGER / Line 2010		OSM Feb. 2011 minus TIGER / Line 2011		OSM Sept. 2012 minus TIGER / Line 2011	
Rank	State	km	(dif/Σ) x 10000	km	(dif/Σ) x 10000	km	(dif/Σ) x 10000	km	(dif/Σ) x 10000	km	(dif/Σ) x 10000
1	Portland, OR	–0.9	–0.7	4.5	3.3	99.4	72.3	216.9	157.7	1491.9	1084.9
2	Denver, CO	–11.7	–7.7	25.7	17.0	681.9	449.7	1073.4	707.9	1359.4	896.5
3	Minneapolis, MN	–2.4	–1.0	33.7	14.5	204.8	88.0	620.1	266.4	1665.8	715.5
4	Washington D.C.	45.3	15.4	287.2	97.4	486.6	165.1	1052.8	357.2	1814.5	615.6
5	Omaha, NE	–1.3	–2.1	–8.2	–13.0	–0.7	–1.1	84.9	134.9	386.8	614.8
. . .						. . .					. . .
66	El Paso, TX	0.0	0.0	0.0	0.0	1.9	3.2	6.4	10.6	14.3	23.8
67	Birmingham, AL	0.0	0.0	–0.9	–0.9	3.8	3.8	10.7	10.7	22.4	22.4
68	Fresno, CA	18.1	48.9	18.1	48.9	0.4	1.1	1.7	4.5	5.2	14.0
69	Akron, OH	–0.7	–1.2	5.5	9.2	0.7	1.2	14.7	24.3	6.1	10.1
70	McAllen, TX	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

conclude that one option to improve the quality of OSM road data could be the development of OSM editing tools that allow the tracing of current TIGER/Line data to correct OSM geometry errors caused by the initial data import. This would keep OSM user contributions and edits unaffected, but allow the OSM community to benefit from recurring TIGER/Line updates provided by the US Census Bureau. First attempts have already been made, with some success, by adding overlays of TIGER/Line 2011 data to popular OSM editing tools, such as Potlatch 2 and JOSM. It remains to be seen if the newly awarded Knight Foundation grant will help with the development of improved and more intuitive tools to increase OSM data quality in the US and to make data contributions a simpler task even for new or casual members.

Future research could expand this type of analysis to other countries with data imports, such as AND for the Netherlands or India. Also, a more comprehensive analysis of the type of data, i.e. feature types and attributes, which was contributed by the active OSM members after the data import, could reveal in more detail how data imports affect the contribution patterns of the OSM community.

References

Bennett J 2012 Welcome, Apple! WWW document, http://blog.osmfoundation.org/2012/03/08/welcome-apple/

Best C 2005 An overview and update: MAF/TIGER Enhancement Program. In *Proceedings from the ICIT Conference*, Des Moines, Iowa (available at http://www.icit.state.ia.us/Committees/edu/Events/PastEvents/MidYear05/presentations/CensusMAF-TIGEREnhancement.pdf)

Bicheno S 2012 Global Smartphone Installed Base Forecast by Operating System for 88 Countries: 2007 to 2017. WWW document, http://www.strategyanalytics.com/default.aspx?mod=reportabstractviewer&a0=7834

Budhathoki N R, Nedovic-Budic Z, and Bertram D 2010 An interdisciplinary frame for understanding volunteered geographic information. *Geomatics* 64: 313–20

Census 2012a Census Feature Class Codes (CFCC). WWW document, http://www.census.gov/geo/www/tiger/appendxe.asc

Census 2012b MAF TIGER Feature Class Code (MTFCC) Definitions. WWW document, https://www.census.gov/geo/www/tiger/tgrshp2010/TGRSHP10SF1AF.pdf

Census 2012c TIGER Overview. WWW document, https://www.census.gov/geo/www/tiger/tgrshp2012/TGRSHP2012_TechDoc_Ch2.pdf

Coleman D J, Georgiadou Y, and Labonte J 2009 Volunteered geographic information: The nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research* 4: 332–58

Cooper D 2012 Craigslist Quietly Switching to OpenStreetMap Data. WWW document, http://www.engadget.com/2012/08/28/craigslist-open-street-map/

Flanagin A J and Metzger M J 2008 The credibility of volunteered geographic information. *GeoJournal* 72: 137–48

Foursquareblog 2012 Foursquare is Joining the OpenStreetMap Movement! WWW document, http://blog.foursquare.com/2012/02/29/foursquare-is-joining-the-openstreetmap-movement-say-hi-to-pretty-new-maps/

Gannes L 2012 OpenStreetMap gets first Major Funding from Knight News. WWW document, http://www.knightfoundation.org/press-room/press-mention/openstreetmap-gets-first-major-funding-knight-news/

Girres J-F and Touya G 2010 Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS* 14: 435–59

Goodchild M F 2007 Citizens as sensors: The world of volunteered geography. *GeoJournal* 69: 211–21

Goodchild M F 2009 NeoGeography and the nature of geographic expertise. *Journal of Location Based Services* 3: 82–6

Google 2012 Introduction of Usage Limits to the Maps API. WWW document, http://googlegeodevelopers.blogspot.com/2011/10/introduction-of-usage-limits-to-maps.html

Haklay M, Basiouka S, Antoniou V, and Ather A 2010 How many volunteers does it take to map an area well? The validity of Linus' Law to volunteered geographic information. *Cartographic Journal* 47: 315–22

Heipke C 2010 Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing* 65: 550–57

Hochmair H H, Zielstra D, and Neis P 2013 Assessing the completeness of bicycle trails and designated lane features in OpenStreetMap for the United States and Europe. In *Proceedings of the Ninety-second Annual Meeting of the Transportation Research Board*, Washington D.C.

ISO/TC 211 2010 Geographic Information/Geomatics Standards Guide. WWW document, http://www.isotc211.org/

Koukoletsos T, Haklay M, and Ellul C 2012 Assessing data completeness of VGI through an automated matching procedure for linear data. *Transactions in GIS* 16: 477–98

Linux.Com 2007 OpenStreetMap Project Imports US Government Maps. WWW document, http://archive09.linux.com/feature/119493

Mooney P and Corcoran P 2012a The annotation process in OpenStreetMap. *Transactions in GIS* 16: 561–79

Mooney P and Corcoran P 2012b Characteristics of heavily edited objects in OpenStreetMap. *Future Internet* 4: 285–305

Neis P and Zipf A 2012 Analyzing the contributor activity of a volunteered geographic information project: The case of OpenStreetMap. *ISPRS International Journal of Geo-Information* 1: 146–65

Neis P, Zielstra D, and Zipf A 2012 The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet* 4: 1–21

OpenStreetMap 2013a AND Data. WWW document, http://wiki.openstreetmap.org/wiki/AND_Data

OpenStreetMap 2013b Automated Edits Code of Conduct. WWW document, http://wiki.openstreetmap.org/wiki/Automated_Edits/Code_of_Conduct

OpenStreetMap 2013c Import/Catalogue. WWW document, http://wiki.openstreetmap.org/wiki/Import/Catalogue

OpenStreetMap 2013d OpenStreetMap Map Features. WWW document, http://wiki.openstreetmap.org/wiki/Map_Features

OpenStreetMap 2013e OpenStreetMap Statistics. WWW document, http://www.openstreetmap.org/stats/data_stats.html

OpenStreetMap 2013f OpenStreetMap TIGER. WWW document, http://wiki.openstreetmap.org/wiki/TIGER

OpenStreetMap 2013g OSMOSIS Tool. WWW document, http://wiki.openstreetmap.org/wiki/Osmosis

OpenStreetMap 2013h Planet OSM Files. WWW document, http://planet.openstreetmap.org/

OpenStreetMap 2013i Talk-US-Mailing List: Imports and Mass Edits in the US. WWW document, http://lists.openstreetmap.org/pipermail/talk-us/2012-December/009899.html

OpenStreetMap 2013j Talk-US-Mailing List: More on TIGER: Where It's Likely Safe to Import. WWW document, http://lists.openstreetmap.org/pipermail/talk-us/2012-December/009894.html

OpenStreetMap 2013k Import/Guidelines. WWW document, http://wiki.openstreetmap.org/wiki/Import/Guidelines

OpenStreetMap 2013l Import/Past Problems. WWW document, http://wiki.openstreetmap.org/wiki/Import/Past_Problems

OpenStreetMap 2013m OpenStreetMap PBF Format. WWW document, http://wiki.openstreetmap.org/wiki/PBF_Format

OpenStreetMap Foundation 2013 Automated Redactions Complete. WWW document, http://blog.osmfoundation.org/2012/07/26/automated-redactions-complete/

Rehrl K, Gröchenig S, Hochmair H H, Leitinger S, Steinmann R, and Wagner A 2013 A conceptual model for analyzing contribution patterns in the context of VGI. In Krisp J M (ed) *Progress in Location Based Services*. Berlin, Springer Lecture Notes in Geoinformation and Cartography: 373–88

Switch2osm 2012 switch2osm: Case Studies. WWW document, http://switch2osm.org/case-studies/

Zandbergen P A, Ignizio D A, and Lenzer K E 2011 Positional accuracy of TIGER 2000 and 2009 road networks. *Transactions in GIS* 15: 495–519

Zielstra D and Hochmair H H 2011a A comparative study of pedestrian accessibility to transit stations using free and proprietary network data. *Transportation Research Record* 2117: 145–52

Zielstra D and Hochmair H H 2011b Digital street data: Free versus proprietary. *GIM International* 25: 29–33

Zielstra D and Hochmair H H 2012 Using free and proprietary data to compare shortest-path lengths for effective pedestrian routing in street networks. *Transportation Research Record* 2299: 41–7

Zielstra D and Zipf A 2010 OpenStreetMap data quality research in Germany. In *Proceedings of the Sixth International Conference on Geographic Information Science (GIScience 2006)*, Zurich, Switzerland