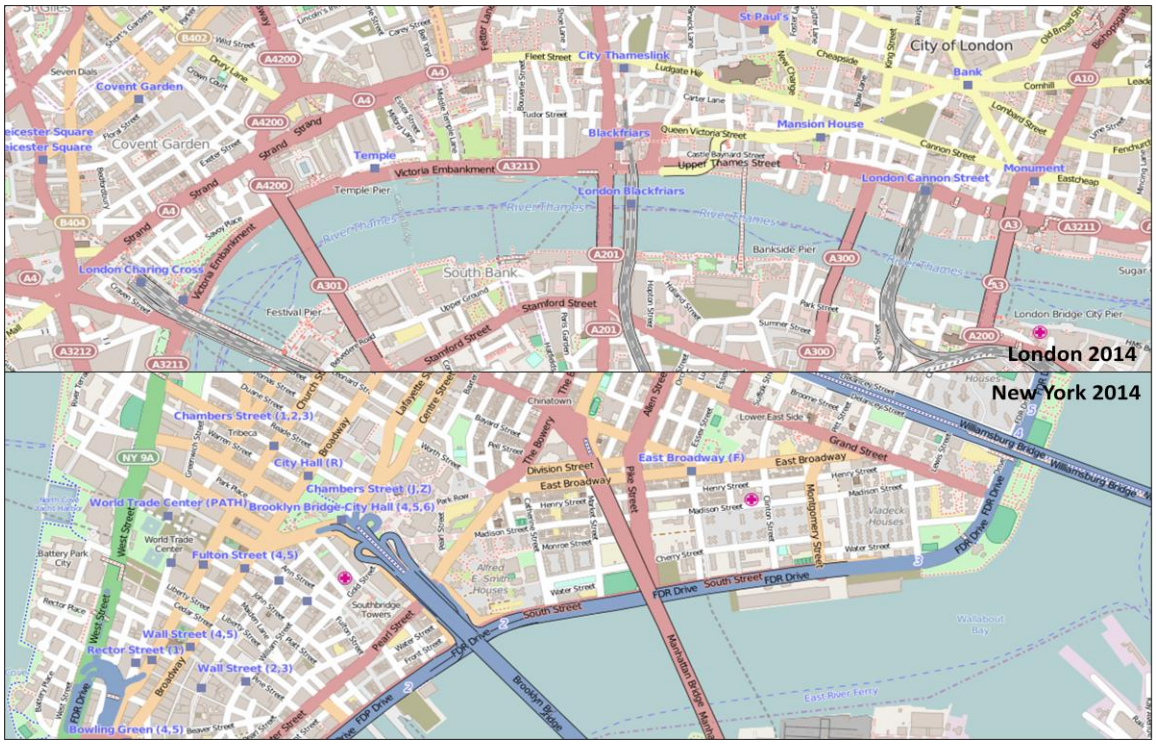
OpenStreetMap User Contributions:

Comparing Communities in the United States and Europe



GEOG 596B: Individual Studies – Capstone Project

Pennsylvania State University

05/06/15

Author:
Jeremy Stratman

Faculty Advisor:
Dr. Jan Oliver Wallgrün

PENNSTATE
1855

Table of Contents

Abstract 3
1. Introduction 4
1.1. OpenStreetMap Overview 4
1.2. Motivation 4
1.3. Objectives 4
2. Literature Review 5
3. Methodology 5
3.1. City Selection 5
3.2. Data Retrieval and Import 8
3.3. Query Design and Execution 9
3.4. Result Analysis 9
4. Community Comparison Results 10
4.1. User Activity 10
4.2. Sources of Data 13
4.3. Community Events 16
4.4. Maturity of Data 18
4.5. Completeness of Data 23
4.6. Focus of Edits 25
5. Characteristics of Successful Communities 26
5.1. City Rankings in Each Category and Overall 26
5.2. Correlations Between Individual Rankings and Other Metrics 27
6. Conclusions and Recommendations 28
6.1. Conclusions 28
6.2. Recommendations to US OpenStreetMap Community 29
7. References 30

Abstract

In just over ten years, OpenStreetMap grew from a handful of contributors in the United Kingdom to a project that mapped the entire globe through almost two million registered users. The goals of this study were to compare the nature of user contributions within cities in the United States to those in Europe, and then extract commonalities from the most successful cities. Europe has the highest activity level in the project and thus served as a potential benchmark for comparison. While previous studies have directly compared numerous cities across mostly quantitative metrics, few have examined temporal data and none have specifically compared samples of US and European cities or profiled the character of communities over time. Metrics to measure contributions included monthly samples of metadata maturity, data completeness, activity of users, event frequency, and other characteristics of user edits. This study looked at 38 total cities across the United States and Europe, equally ranging from approximately 200,000 to 3,500,000 residents living within the study areas. The most successful communities were those that ranked highly in both data completeness and maturity, while also having high user activity and frequent community events. Results show that while US cities did generally get a later start than their European counterparts and continue to lag behind in data maturity and completeness, user activity and community events have increased greatly in the last two years. Cities in the United States have the opportunity to seize upon the recent increased activity and community interest, effectively applying lessons learned from other communities in all known metrics of success.

1. Introduction

This study sought to determine how the activities and contributions of OpenStreetMap communities differed between the United States and Europe, ultimately deriving lessons from the most successful communities that can be applied everywhere. By comparing a sizeable sample of cities from each community across metrics that measured user activity and data quality, measurable differences emerged that helped show where US cities still had room for improvement and where they have made progress over time. Describing the general profile of a successful community was also then possible, resulting in recommendations for communities to improve their participation and output quality.

1.1. OpenStreetMap Overview

OpenStreetMap (OSM) merges the established practice of collaborative mapping with the modern technical ability to share data over digital networks. Wikis via the internet began as early as 1994, culminating with Wikipedia in 2001, as a practical method of sharing data ("Wiki History", n.d.). Crowd-sourced mapping, also commonly known as Volunteer Geographic Information (VGI), has occured for well over a hundred years. One prime example is the Christmas Bird Count, but there have been many others before the age of the internet (Goodchild, 2007). Collaborative mapping is the term of choice for this study, as the term volunteered seems to make assumptions about acceptable contribution motivations.

Steve Coast founded OpenStreetMap in 2004 on these basic principles, adapting a highly customized, wiki-inspired framework for geographic collaboration. His primary motivation was to break down existing proprietary barriers to quality geographic data in the United Kingdom and bring it to the general public. This resonated with members of other European countries, which also generally had few free data sets compared to high cost and proprietary ones ("About OpenStreetMap", n.d.).

All registered users can edit the map or utilize the data, although its use is governed by the Open Data Commons Open Database License (ODbL). This license requires derivatives to also remain under the same license and specifies that the data is owned by all users. Unlike Wikipedia and some other similar models, OpenStreetMap has a generally democratic structure without define hierarchies or approval processes. It is also based on a free tagging system without enforceable standards; best practices are decided upon through online discussions and posted on the OpenStreetMap Wiki ("About OpenStreetMap", n.d.). This lack of defined structure gives much freedom to users, but also makes analysis much more challenging.

1.2. Motivation

The primary motivation for this project was to extract commonalities from successful OpenStreetMap communities and propose actions that mappers in the United States can take to increase participation while ensuring quality data. The primary assumptions being made in this motivation were that US OpenStreetMap communities generally have less participation then their European counterparts and that the source of at least some contributions have been different, discussed further under in Section 2.

1.3. Objectives

The first objective was to determine what significant differences, if any, existed in the nature of user contributions over time between US and European communities. A few important questions to answer were whether the United States simply had a later start than Europe, whether only participation or rate of

growth was different, and whether there were discernable differences in the composition or quality of edits.

The second objective, and arguably the most meaningful for direct feedback, was to define the key traits that the most successful communities exhibit. Success can be described in many ways, but its meaning for this study focuses on both high user participation as well as quality, both in terms of feature and metadata completeness.

2. Literature Review

Numerous studies have evaluated various aspects of OpenStreetMap data and communities, but relatively few have sought to compare the nature of contributions between different geographic areas. Fewer still have looked at data temporally. None have extensively compared the profile of communities in United States to those in Europe.

Despite these gaps in literature, there are several useful insights to be gained from existing studies. With regard to the spatial distribution of edits, urban areas had far greater activity than rural areas (Zielstra, 2010). Many other quality studies have taken place in Europe and indicated generally high accuracy and completeness in areas with active communities, but studies had a heavy focus on just road features (Mondzech, 2011), with some limited analysis of points of interest (Mashhadi, 2011). Temporal studies have occurred as well, but have either been limited in scope or focused more on the daily patterns of contributors (Yasseri, 2013). Hristova's study of the impact of mapping parties on participation concluded that community events drive participation, but this only involved events and activity in London (Hristova, 2013).

Communities in the United States have spent considerable time cleaning up a series of imports early in the history of OpenStreetMap. One study suggested that communities appeared less active after conducting these cleanup efforts, while European communities did not have similar problems (Zielstra, D., Hochmair, H. H., and Neis, P., 2013). The largest and most significant import was the import of the US Census Bureau's Topologically Integrated Geographic Encoding and Referencing (TIGER) data in 2007, whose data was of generally mid to low quality (Willis, 2007). This paper is the first to examine edits, activities, and resulting data temporally across several different metrics in both the United States and Europe.

3. Methodology

The general approach for this project was to select comparable cities in both the United States and Europe, import historic OpenStreetMap and event data for these cities, query the data temporally across numerous metrics, and analyze the results. This section outlines only the general methodology from a procedural standpoint, while Sections 4.1-4.6 describe the backgrounds and justifications of each metric.

3.1. City Selection

The single most important step in this project was the selection of appropriate samples of cities in the United States and Europe. Due to the need to normalize for population across many metrics, comparable

population size within the geographic area analyzed was the primary consideration when selecting cities. Physical size, and thus population density, certainly influences the nature of a community as well and was the secondary consideration. The population size of the greater metro area was the final factor, to best account for local transient populations. Spatial distribution was not explicitly factored into the selection process, but it was reviewed afterward as a control to ensure the sample did include various geographic areas. Due to the inconsistent nature of edits in rural areas within OpenStreetMap, only cities over 150,000 persons were considered (Zielstra, 2010).

Using US Census and United Nations data, all cities with populations greater than 150,000 persons for the United States and Europe were entered into a spreadsheet along with approximations of their metro population size and physical area. Medium and small cities were then filtered by city population, physical size, and metro population. The most statistically similar pairs were highlighted as potential sample cities. The largest cities required exceptional workflows, however, because the quantity of cities with several million residents are limited and not directly comparable using their official administrative boundaries. Combinations of geographic subsets of cities were therefore analyzed to find the best matches. As a result, New York consists of just Manhattan and the Bronx, London includes 13 inner boroughs, Los Angeles includes six inner neighborhoods, and Houston consists of 76 small contiguous administrative districts.

The number of selected cities needed to be substantial enough to account for large variations between individual cities, but not an overwhelming number due to some of the manual processes involved in this project. 38 cities were acceptably comparable from the initial pool of cities with populations over 150,000, so this entire population was selected as being adequate. Table 1 lists all selected cities with respective populations, areas, and population densities. US cities include their respective state, while European cities include their country. Figures 1 and 2 show the spatial distribution of cities; while this was not a selection criterion, they are generally well distributed in their respective geographic areas.

US City	Population	Area (km^2)	Density (persons/km^2)	European City	Population	Area (km^2)	Density (persons/km^2)
New York, NY*	3,040,000	168	18,095	London, UK*	2,924,000	301	9,714
Los Angeles, CA*	3,760,000	992	3,790	Berlin, DE	3,502,000	891	3,930
Chicago, IL	2,719,000	606	4,487	Madrid, ES	3,234,000	605	5,345
Houston, TX*	1,593,000	805	1,979	Hamburg, DE	1,799,000	755	2,383
Philadelphia, PA	1,553,000	369	4,209	Munich, DE	1,378,000	310	4,445
San Jose, CA	999,000	457	2,185	Seville, ES	702,000	140	5,017
Washington DC	646,000	176	3,673	Frankfurt, DE	692,000	248	2,788
Denver, CO	650,000	400	1,624	Oslo, NO	593,000	454	1,306
Boston, MA	646,000	125	5,168	Dusseldorf, DE	592,000	217	2,730
Baltimore, MD	622,000	209	2,977	Rotterdam, NL	610,000	325	1,878
Seattle, WA	652,000	217	3,006	Glasgow, UK	599,000	175	3,422
Portland, OR	609,000	345	1,767	Bremen, DE	548,000	326	1,682
Albuquerque, NM	556,000	490	1,136	Nantes, FR	580,000	524	1,107
Atlanta, GA	448,000	342	1,309	Gdansk, PL	457,000	260	1,757
Miami, FL	418,000	92	4,540	Zurich, CH	367,000	88	4,168
Honolulu, HI	375,000	177	2,117	Nice, FR	345,000	72	4,790
St Paul, MN	295,000	135	2,184	Mannheim, DE	315,000	144	2,187
Richmond, VA	214,000	156	1,373	Saarbrucken, DE	176,000	167	1,055
Providence, RI	178,000	60	2,967	Liege, BE	192,000	70	2,737

Table 1. US and European Cities Selected for Comparison

* Subsection of city used in order to best match all selection criteria.

Figure 1. Spatial Distribution of European Cities (Map by Mapbox)

Figure 2. Spatial Distribution of US Cities (Map by Mapbox)

3.2. Data Retrieval and Import

The bulk of the data for analysis came from the OpenStreetMap full history dump file, a single experimental XML-based file that contains every version of almost all nodes created after October 2007, and most data prior to then. Nodes redacted after a 2012 license change are also not included, but these changes are too limited to impact the overall results of this study. The file's compressed size exceeded 52 gigabytes at the time of its download ("Planet.osm/full", 2014). Also required was a smaller Changeset file that contains bounding boxes of each user editing session for the same historic time period as the full history file, and is just under one gigabyte in size.

In order to be analyzed, the changeset and feature data had to first be imported into PostGIS, done using MaZderMind's osm-history-renderer python scripts ("MaZderMind/osm-history-splitter."). The large file sizes required the data to be parsed out by city in order to be computationally manageable, even using a dedicated Linux partition on a high end personal computer. City outlines were derived from administrative boundaries from either the same sources as population data or OpenStreetMap itself, forming the mask for splitting out data into each database. Ideally, a single database could have been used for querying all cities, but dedicated databases for each city allowed queries to run more quickly. Figure 3 shows a snapshot of individual databases and one city's resulting tables in pgAdmin III.

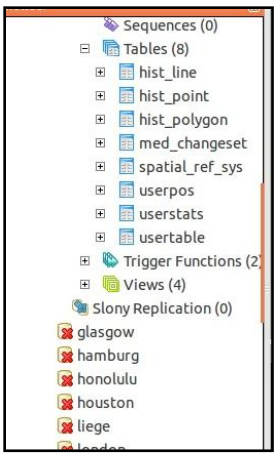


Figure 3. Individual City Databases and Tables in pgAdmin III

Non-spatial information about events had multiple sources, including the OpenStreetMap wiki, Maptime.io, and individual city websites. The latter required searching for each city and the term "OpenStreetMap" via online search engines. Several results in other languages were found, so while there is some risk that the search engine did not find all potential results due to language barriers, the risk seems relatively low. These results only reflected documented events, which were then added to a single spreadsheet for further analysis.

3.3. Query Design and Execution

With the raw XML data imported into PostGIS, queries could be designed and executed. Due to the existence of individual city databases and the desire to extract monthly snapshots of each query for each city, python helper scripts were required to loop through city names and desired date intervals as opposed to making manual queries. CSV files of the required dates and city names provided the conditions for these python loops, and thematically grouped python scripts were written to actually execute the queries and output results into CSV formats for follow-on analysis in an Excel spreadsheet dashboard.

After successfully testing the mechanism for querying, each individual query had to be crafted for all metrics to be analyzed. Specific explanations for the choice of each metric and results from the applicable queries are discussed in Section 4.

3.4. Result Analysis

In order to meet both project objectives, analysis was split into two main steps. First was the comparison of European to US cities across numerous metrics within a few broad categories to ascertain significant differences. The specific queries were grouped into the following categories of analysis:

- User Activity
- Sources of Data
- Community Events
- Maturity of Data
- Completeness of Data
- Focus of Edits

Following analysis of detailed query results within each category, several selected metrics within four of the above categories were chosen to represent the general measures of success across all community, regardless of geography. Categories selected included User Activity, Community Events, Maturity of Data, and Completeness of Data. Each city was ranked by these metrics, aimed at broadly identifying the cities at either extreme, and then these four rankings were averaged to produce a single ranking of cities by measures of success. Specific rank numbers were not as important as the overall conclusions that one can derive from these commonalities and general trends.

4. Community Comparison Results

4.1. User Activity

The first means of comparing OpenStreetMap communities was through the analysis of user activity. For the purpose of this study, each unique user ID contributing at least one edit to any feature within the footprint of a given city was considered a user of that city.

4.1.1 Unique Contributing Users per Month

The number of unique contributing users per month is an effective measure of user activity in a given city over time. Querying the total number of unique userID occurrences each month is relatively straightforward and yields the number of different accounts that have contributed to that area each month. It does not capture the recurrence of specific userIDs month over month or filter out accounts that may have made automated edits or imports across a wide geographic area. Analyzing individual cities over time, and in aggregate, help to mitigate most influence by non-human account activity. Higher values here are assumed to correspond with higher human activity in the given city.

Results of the cities averaged across the United States and Europe can be seen in Figure 4. Most notably, both the raw number of unique contributing users and values normalized per 100,000 residents are included. The overall curves are very similar because the selection of cities included very comparable cities, in terms of population. As can be seen, the cities in the European sample have considerably more contributing users, almost three times as many for several years. These numbers confirm results from the Literature Review; European cities do generally have more users.

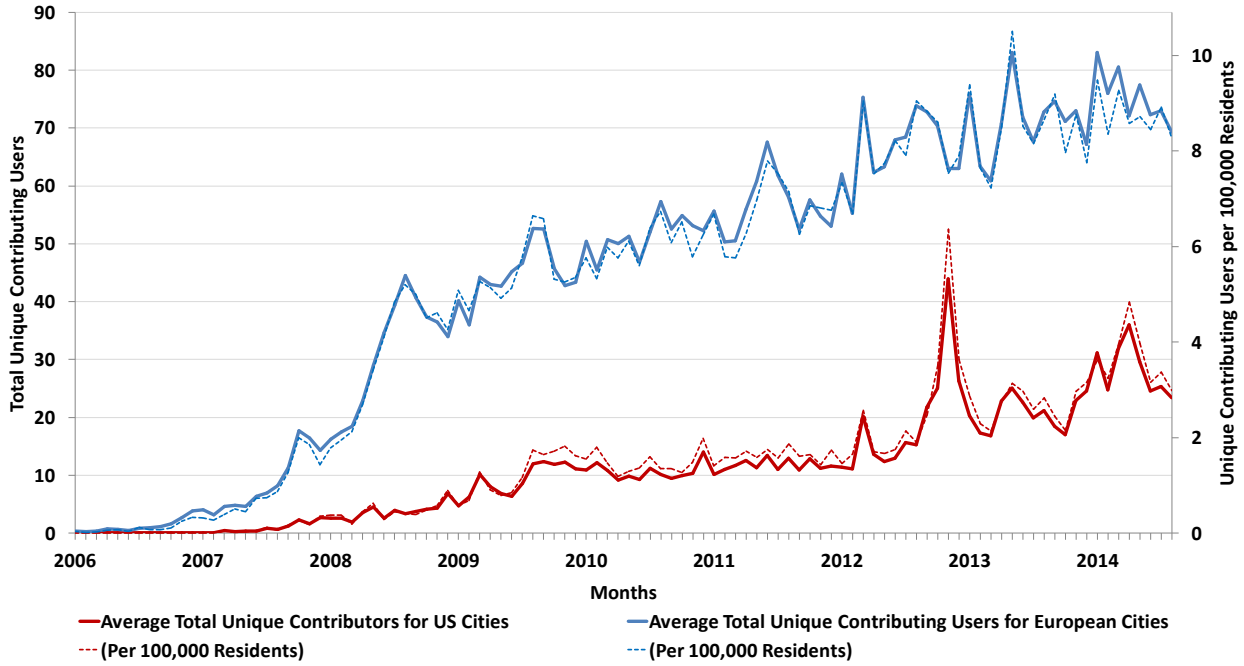


Figure 4. Average Unique Contributing Users per Month, Total and Normalized for Population

With the above results being averages, it was necessary to ensure that a few select cities were not heavily influencing the values. The median of each group was also analyzed as a first step to check for this possible influence, but it yielded similar results. Individual cities were each reviewed, and Figure 5 shows the results of three individual comparisons between US and European counterpart cities of similar sizes.

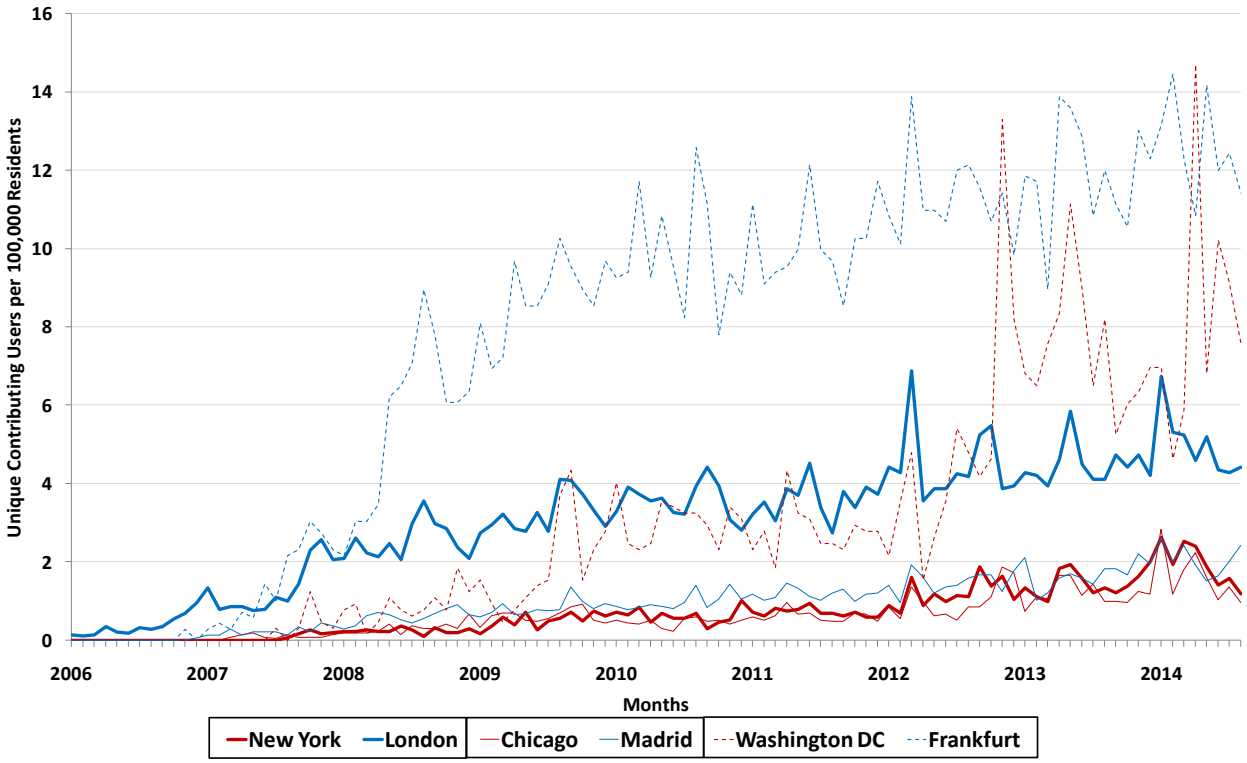


Figure 5. Monthly Users per 100,000 Residents for Select Pairs Cities with Comparable Populations

In this particular figure, London and Frankfurt both have approximately three times the user activity, although Washington DC has begun closing that gap in the last few years. Washington DC has the most contributors per month per 100,000 residents in the US, but its counterpart is only the fourth highest in Europe. Madrid has the least activity when normalized for population in Europe, but it still manages to stay above Chicago's activity. The same trend continued across cities. In fact, Seattle was the only US city that consistently exceeded its European counterpart, averaging ten percent more users than Glasgow in the past five years. This further supports that the averages in Figure 4 hold up under scrutiny.

4.1.2 Rates of Growth in User Activity

The raw number of users, even when viewed over time, does not tell the entire story of a community. One must also consider its growth rate in terms of unique contributing users; a community that had a late start may be behind in the first metric, but similar or greater growth may also indicate equal success. Table 2 shows that while the US was very slow to increase its number of users initially when compared to Europe, the average growth in mappers per year was equal from 2012 to 2014 between both communities. The rate of growth is not enough to close the gap between the two communities yet, but it highlights the increasing interest in OpenStreetMap within the United States.

	2006-2008	2009-2011	2012-2014
US Cities	2.3	4.6	5.6
European Cities	15.1	6.3	5.6

Table 2. Average Annual Growth in Number of Active Monthly Users per City

4.1.3 Median User Changeset Location (Non-Temporal)

Rather than attempt to derive user location using the methodology of Neis (Neis, P., and Zipf, A., 2012), user location data for this study only reflects the median center coordinate of all of their changesets. Typically, the resulting coordinate should be in vicinity of the location a user edits most often. Figure 6 depicts the resulting percentage of user median changeset locations within a given distance from their respective cities. Users editing European cities are far more likely to contribute the majority of their edits closer to the city being analyzed. While this methodology is simplistic in nature, the disparity may highlight one key difference between the United States and Europe: physical size. The distance between cities and even metro areas is far greater in the United States. It is difficult to determine if this influences the sense of community, or actual contributions, but it is an important difference.

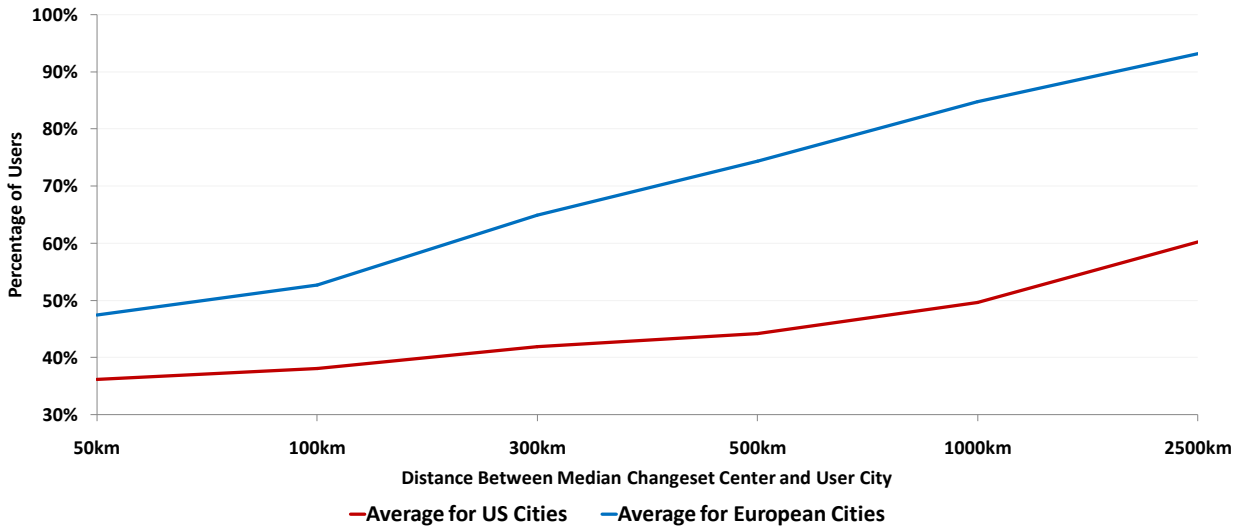


Figure 6. Average Percentage of Median User Changeset Location within a Given Distance from Cities

4.2. Sources of Data

Given the known major difference in the source of road network data in the United States, comprehensively evaluating the sources of data that each community most heavily relied upon seemed prudent. The source of data can heavily influence its quality. In OpenStreetMap data, "Source" is a tag that should always be included upon feature creation, per community standards. As with all tags, however, it is optional and freeform in nature. As a result, the associated values vary greatly between features, and some interpretation was required through automated and manual methods. For this study, sources were categorized as either coming from the import of existing data, physical survey of features in person or by GPS, or the tracing of imagery. It would not have been feasible in this project to analyze every source tag of every feature, so only the most common values each month for each city were extracted. These represented the source of data that dominated edits for that city in that particular month. Imagery as a source was almost equally utilized between the US and Europe, so it is not included in the following sections.

4.2.1 Percentage of Months with Majority of Edits Sourced from Imports

Aside from the US government TIGER data import within the United States, little was known about the prevalence of imports beyond the topic as a hotly debated subject across the entire OpenStreetMap community ("Import", 2014). This study attempted to also identify individual imports by number of edits in a given changeset, looking at unnaturally high numbers of edits, but this proved unreliable. A more sophisticated approach may be warranted in future studies, but analysis of imports here was limited to months with edits dominated by an imported source.

Identifying source tags for imports was a challenge also. General keywords could not identify these, as almost all tags only reference the name of the data source. Examples of this include 'massgis', 'cadastre-dgi-fr', 'TIGER..', and others. These had to be manually identified, and a list of keywords for each general type of import created to then automatically recode all matches as imports. Any values not identified through this process, or similar methodology applied to imagery and surveyed sources, were manually recoded.

Figure 7 shows two series for the US, one with TIGER imports included, and one without. The TIGER import represents the priorities and actions of a minority of users within the US OpenStreetMap community, so looking at imports with those removed seemed to be a fairer comparison. For most of the history of OpenStreetMap, the US has indeed relied upon imports more often the Europe, but only by a small margin. In the past year, use of imports has been similar between both communities.

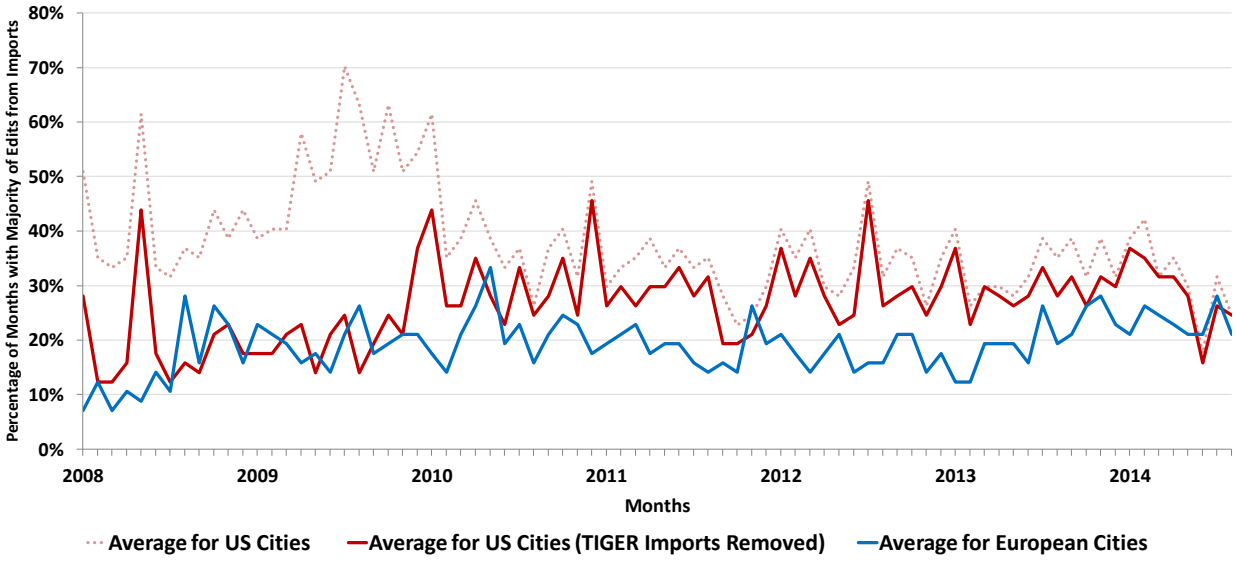


Figure 7. Percentage of Months with Majority of Edits Sourced from Imports

4.2.2 Percentage of Months with Majority of Edits Sourced from Surveys

In the same manner as with imports, surveys were identified through an automated process that searched for keywords first, and highlighted unidentified tags for manual review. Fortunately, surveyed sources had many more key words in common, such as 'gps', 'survey', 'physical mapping', and others.

Figure 8 shows a major difference between the US and Europe. Users in European cities appear to be much more likely to use surveying as a data source than the US. This gap has also been closing in recent years, but it is still rather large.

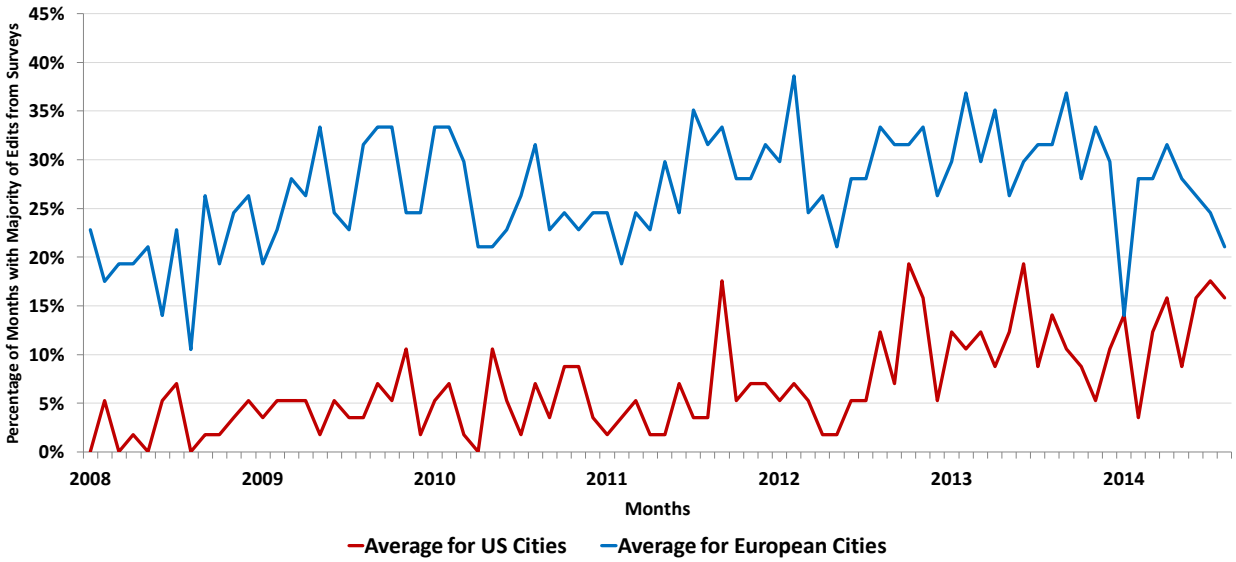


Figure 8. Percentage of Months with Majority of Edits Sourced from Surveys

4.2.3 Percentage of Months with Majority of Edits Unattributed with a Source

As already mentioned, the source tag is optional when adding features to OpenStreetMap, despite being strongly encouraged. Figure 9 shows that both communities were less diligent in the past about including this tag than they are today, with the US less likely to do so then Europe. The lack of a tag makes analysis of the other methods a little less certain, but it primarily highlights an inability to follow best data practices.

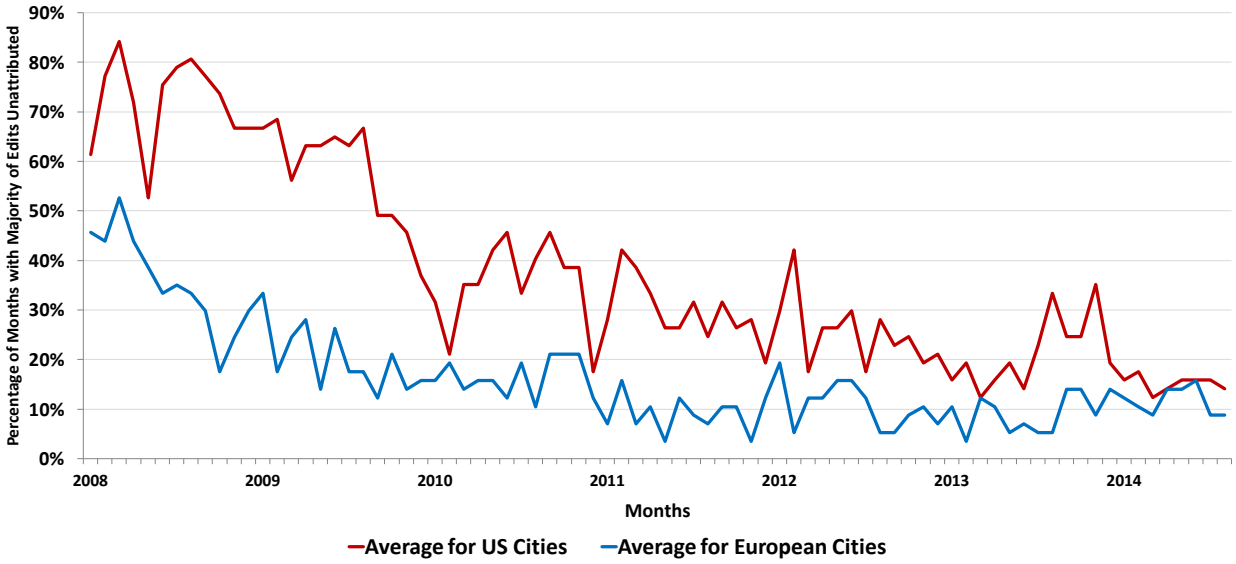


Figure 9. Percentage of Months with Majority of Edits Unattributed with a Source

4.3. Community Events

Characterizing the various communities requires going beyond the map data itself, with the most meaningful measure being the frequency of OpenStreetMap related events. Events are defined here as physical meetings whose purpose was to discuss or contribute to OpenStreetMap, or explicitly socialize as an OpenStreetMap community. These events were categorized as either local events targeting city communities or major events involving a national or international audience.

4.3.1 Major International or National Events

The primary OpenStreetMap event involving the entire international community is the annual State of the Map (SOTM). Until 2011's Denver conference, however, it was always held in Europe. To date, only three out of eight international SOTM events have been held outside Europe. 2014's event is not depicted below and was held in Argentina. In the United States, 2010 saw the first offshoot of the primary SOTM event, largely because of the inaccessibility of the European conference locations. In 2011, the European community did the same thing when the SOTM primary event was held in Denver. These regional events have continued consistently in the US, and sporadically in Europe ("State of the Map.").

Figure 10 overlays these major events on the same monthly contributing users depicted in Figure 4, for the purpose of assessing changes in user activity that may be correlated with the major events. Events are depicted as vertical bars with an abbreviation of the host country above it. While it could be argued that the international SOTM events may be correlated with increases in activity in a few instances, and the same may be true of the 2011 SOTM EU event, the SOTM US events of 2012 and 2014 have the strongest correlations to user activity. In 2012, activity across 14 of 19 US cities increased during or immediately following the SOTM US event. In 2014 this correlation was not as dramatic, seeing 10 of 19 cities increase their activity during the month of the event. In neither instance was the increased activity maintained.

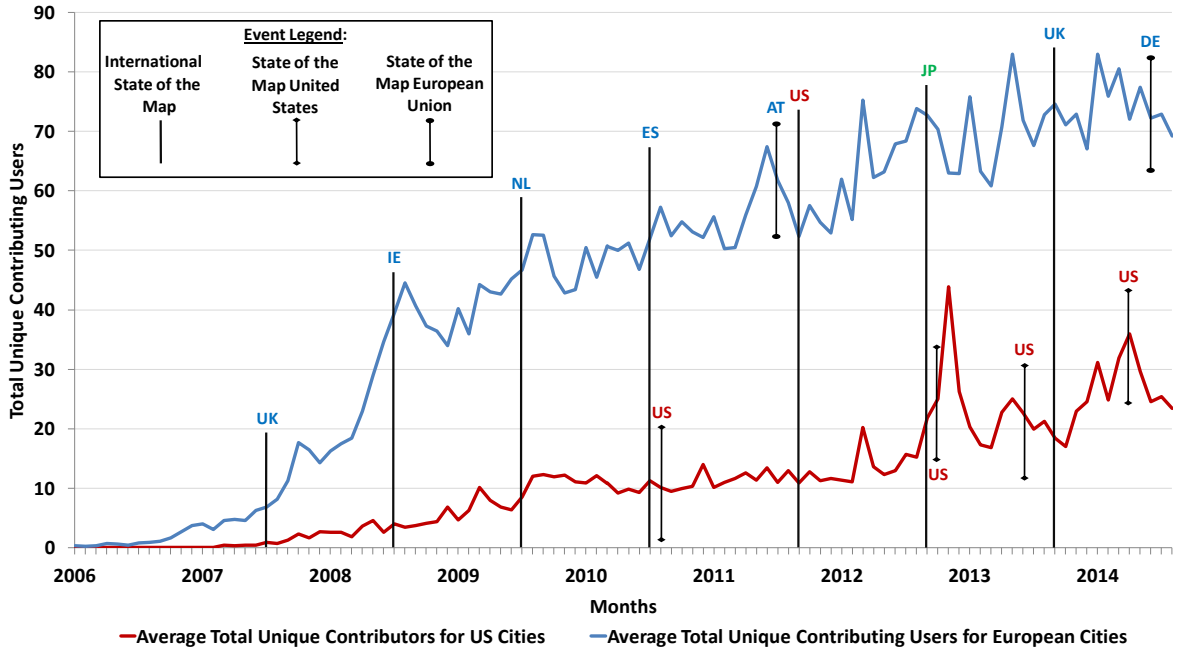


Figure 10. Major 'State of the Map' Events Overlaid on Monthly Contributing Users

This result indicated that major events did not directly correlate with this metric of activity, and further analysis showed that it did not correlate with other metrics either. With that said, the impact of major events has not been insignificant. During the 2013 SOTM US event, a rebranding of the OpenStreetMap communities occurred through the introduction of Maptime chapters. Maptime.io is a single portal with chapters at the city level seeking to organize OpenStreetMap communities ("What is Maptime?", 2015). This has been a major driver of the increase in US local events seen in the next section.

4.3.2 Local City Events

For local events, this study only sought to determine whether cities had any such events in a given month and not necessarily the total number, types, or durations of those events. The most active cities may have had multiple events per month, but given the disparity of frequency across the entire sample, just assigning each month a Boolean value related to the presence of events minimized influence by the extreme outliers.

Figure 11 indicates that the European community has steadily increased the number of months with events across this sample of cities since the inception of OpenStreetMap, while the US peaked and then declined from 2008 to 2010. It was not until 2013, the same time that Maptime was initiated, that the US communities began to meet more frequently. 2014 saw a very sharp increase within the US, directly tied to Maptime events, and 2015 is already on track to equal or exceed average number of months with meetings in Europe. The 2015 data is a combination of events held through March 2015 and planned events.

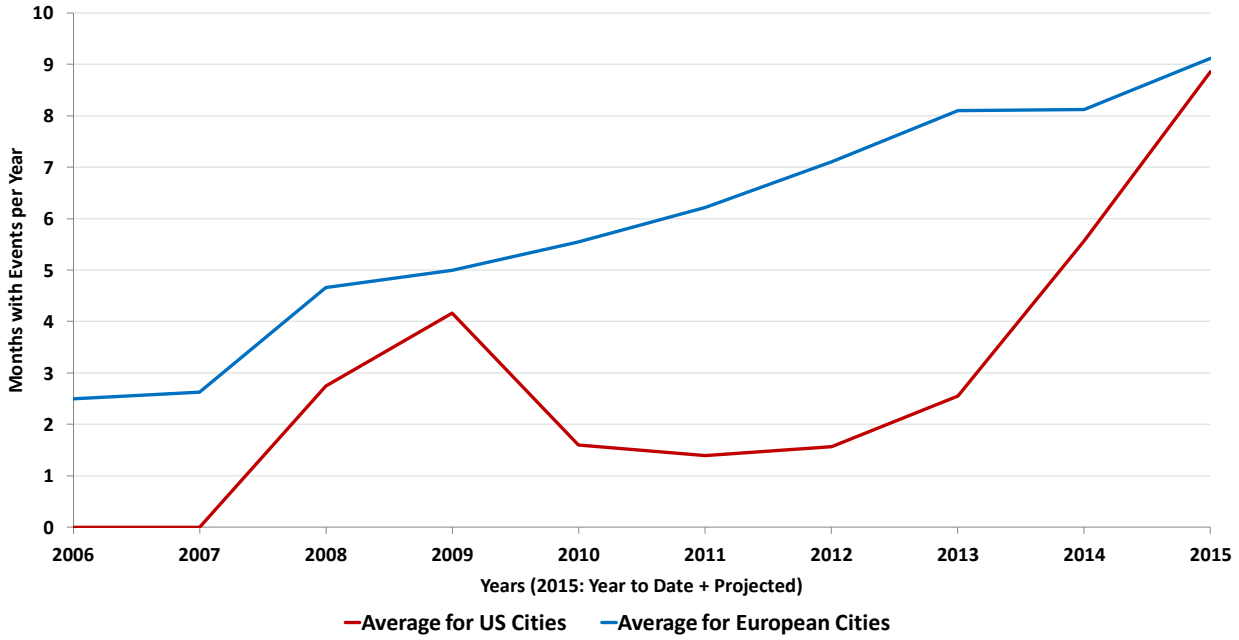


Figure 11. Average Months with Events per Year

4.4. Maturity of Data

In reviewing the map data itself, two key components of its quality are maturity of attributes and overall completeness. While there are many other ways to measure quality, such as assessing spatial or attribute value accuracies by comparing against ground truth data, these would be more an assessment of skills in execution rather than overall community behavior.

Maturity refers simply to the presence of certain tags that represent various stages of metadata completeness, from basic to advanced. To be useful, chosen tags needed be universal to all features being assessed in order to derive meaning from their presence or absence.

Examples best illustrate the concept of maturity. Roads served as one feature type for analysis, with the assumption that nearly all roads have both names and speed limits. A name is typically the most identifying and commonly known attribute about a road, and its inclusion was assessed to represent basic maturity. While most roads have speed limits associated, their inclusion in the data is not necessary for visualization on most maps and represented a more advanced maturity. Food and retail points of interest (POIs) were the second feature type analyzed in this same basic manner.

4.4.1 Maturity of Road Features

As already discussed, the presence of a 'name' tag within each road feature was used as a measure of basic maturity. Road features include those tagged with the 'highway' key that are intended to be paved and for primary use by motorized traffic. Figure 12 confirms that the presence of a name is indeed common to these types of features, with both communities reaching a natural ceiling that likely represents all named streets. The small gap between percentages within each community is likely due to differences in either the interpretation of road classifications or an actual difference in the use of road names. The steep increase in the US data clearly shows the import of TIGER data, which contained road names, compared

to more natural growth of these networks in Europe. All communities undoubtedly value names as an attribute of road features.

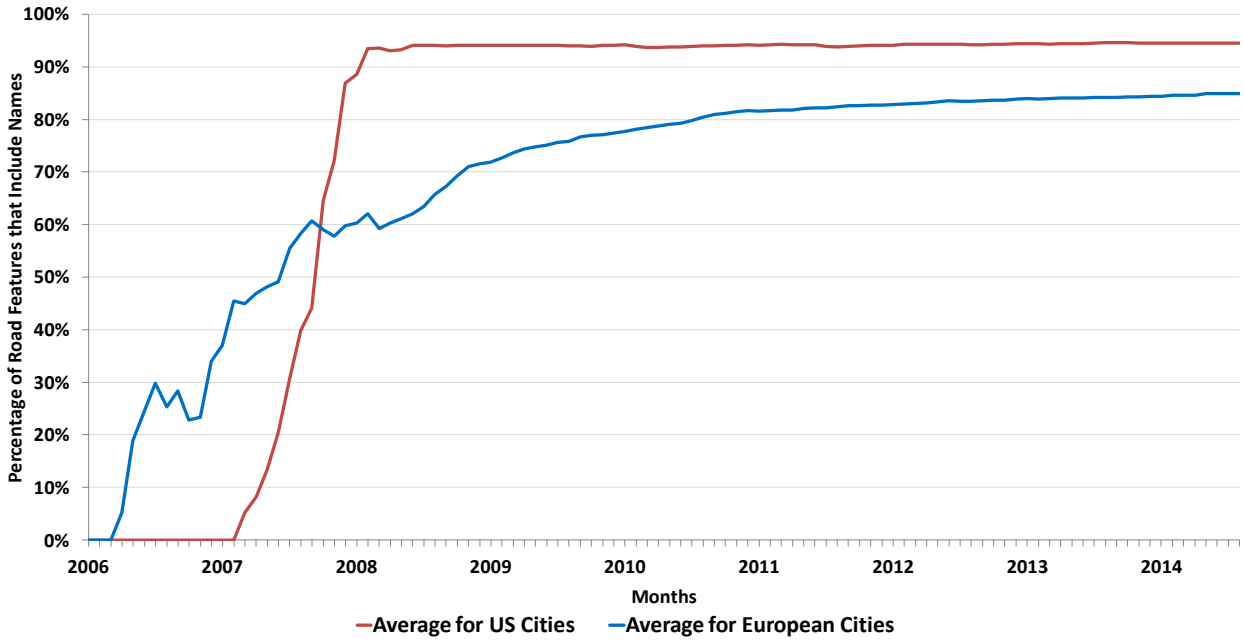


Figure 12. Percentage of Road Features that Include Names

Evaluating the inclusion of speeds limits as an indicator of advanced maturity reveals a much different story. Figure 13 highlights a clear disparity between US and European communities in the use of this tag. While neither community focused heavily on the inclusion of this metadata prior to 2009, its inclusion grew steadily in Europe to almost half of features today. Growth only began in the US in 2012, and still at a much slower rate. The clearest benefit of speed limit metadata is in the accuracy of routing software, not likely in the visualization on a map. It is possible that the use of OpenStreetMap as a routing platform has more use, or is seen as having more potential, in Europe than the United States. Regardless of reasoning, there is unequal value placed on this attribute.

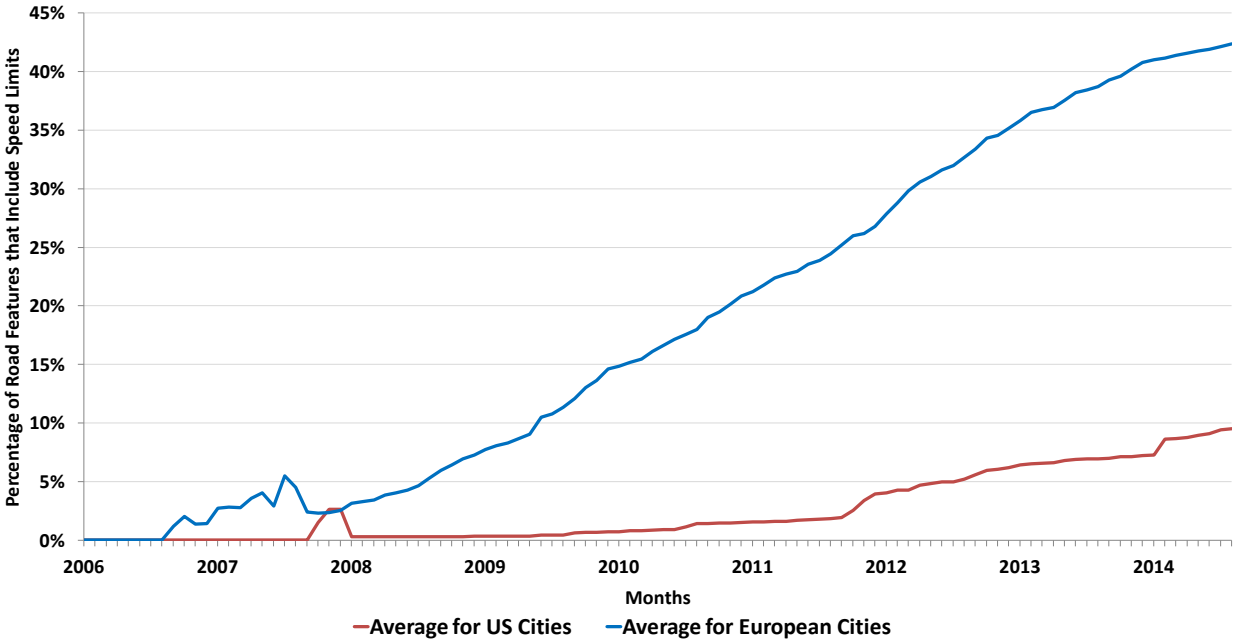


Figure 13. Percentage of Road Features that Include Speed Limits

4.4.2 Maturity of POI Features

The 'amenity' tag is the broadest categorization that includes almost every type of POI and many others, but it is this broad nature that excluded it as a suitable candidate for analyzing attribute maturity. Amenities in OpenStreetMap include parking lots, schools, churches, commercial establishments, and others, none of which have universally common attributes. Most POIs specific to the food and retail industries do have the following attributes: names, addresses, publicly available contact information, and defined hours of business. Like roads, the name attribute is considered basic, while the other three are increasingly advanced and not typically visualized on a map. Table 3 displays two example SQL queries that depict the types of establishments that made up these POIs, separately run for food and retail, but ultimately combined together in results.

SQL Query for Food POIs that Include Addresses (With Python Variables)
"SELECT * from hist_point
WHERE '" + line["year"] + "-" + line["month"] +"-01 00:00:00'::timestamp without time zone >= valid_from AND '" + line["year"] + "-" + line["month"] + "-01 00:00:00'::timestamp without time zone <= COALESCE(valid_to, '9999-12-31 00:00:00'::timestamp without time zone)
AND ((hist_point.tags -> 'amenity'::text) = ANY (ARRAY['cafe'::text, 'bar'::text, 'restaurant'::text, 'fast_food'::text, 'pub'::text, 'bbq'::text, 'biergarten'::text, 'beer_garden'::text]))
AND ((hist_point.tags -> 'addr:housenumber'::text) IS NOT NULL OR (hist_point.tags -> 'addr:housename'::text) IS NOT NULL OR (hist_point.tags -> 'addr:street'::text) IS NOT NULL OR (hist_point.tags -> 'addr:full'::text) IS NOT NULL)"
SQL Query for Retail POIs that Include Contact Information (With Python Variables)
"SELECT * from hist_polygon
WHERE '" + line["year"] + "-" + line["month"] +"-01 00:00:00'::timestamp without time zone >= valid_from AND '" + line["year"] + "-" + line["month"] + "-01 00:00:00'::timestamp without time zone <= COALESCE(valid_to, '9999-12-31 00:00:00'::timestamp without time zone)
AND (((hist_polygon.tags -> 'amenity'::text) = ANY (ARRAY['shop'::text, 'shops'::text, 'shopping'::text, 'retail'::text])) OR (hist_polygon.tags -> 'shop'::text) IS NOT NULL)
AND ((hist_polygon.tags -> 'website'::text) IS NOT NULL OR (hist_polygon.tags -> 'phone'::text) IS NOT NULL OR (hist_polygon.tags -> 'email'::text) IS NOT NULL OR (hist_polygon.tags -> 'fax'::text) IS NOT NULL OR (hist_polygon.tags -> 'contact:phone'::text) IS NOT NULL OR (hist_polygon.tags -> 'contact:fax'::text) IS NOT NULL OR (hist_polygon.tags -> 'contact:email'::text) IS NOT NULL OR (hist_polygon.tags -> 'contact:website'::text) IS NOT NULL)"

Table 3. Example SQL Queries for Food and Retail POIs

Using a percentage of features tagged with given attributes was no concern for roads, but it raised potential issues in the analysis of POIs. Primarily, it was possible that cities with very few POIs mapped may skew the percentages heavily. As seen in Section 5.2, decreased completeness did not actually correlate to increased maturity, and these concerns were not justified.

Figure 14 shows the percentage of POIs with names, resulting in a similar ceiling as all named POIs include this attribute. Unlike road names, POI names grew at a more similar pace for each community, with the US simply starting later. Both communities seem to value names to be associated with POIs.

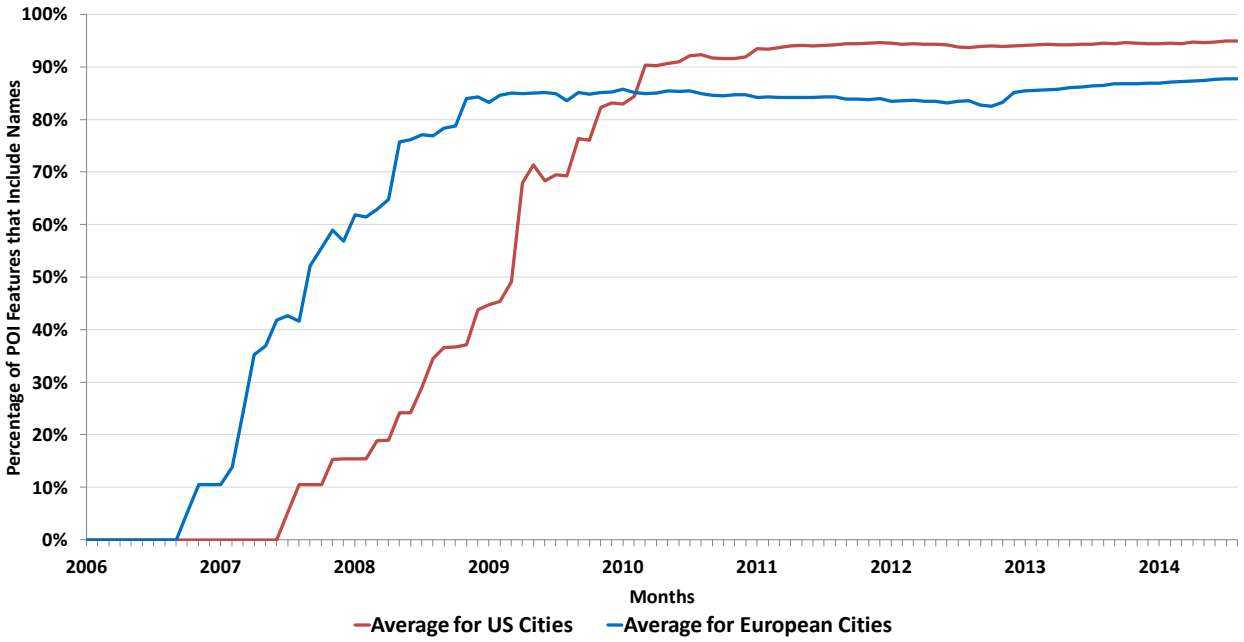


Figure 14. Percentage of POI Features that Include Names

In Figure 15, both communities have shown similar increasing emphasis placed on the inclusion of physical addresses as well as contact information. As with names, the US seems to have simply had a later start, reaching parity with European counterparts at approximately 30% for addresses. The US stills lags behind by about 3% for contact information, but both communities have reached approximately 20% of features tagged with these attributes. Addresses and contact information are certainly less valued than names, but continue to grow at similar rates.

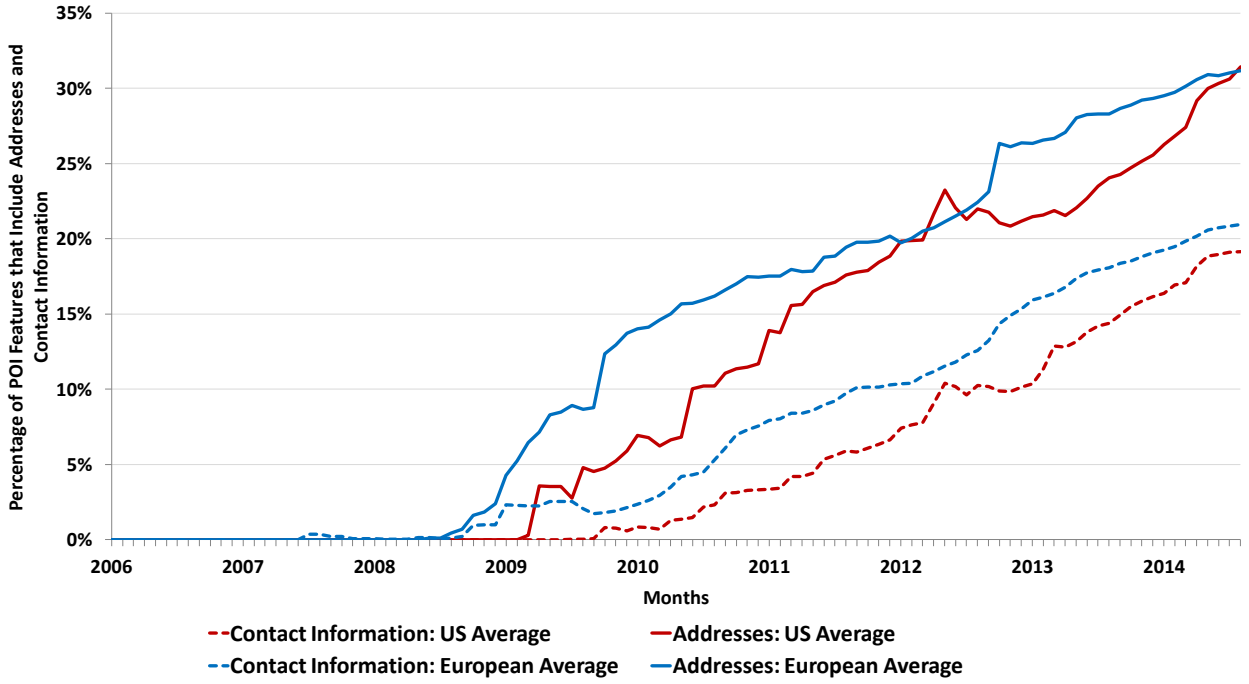


Figure 15. Percentage of POI Features that Include Addresses and Contact Information

The percentage of POI features the include hours of business in Figure 16 shows a pattern more similar to road speed limit attributes, albeit with a much lower overall percentage and less significant gap between communities. European cities do apparently value this attribute more than those in the US, but still fewer than 12% of their POI features include it.

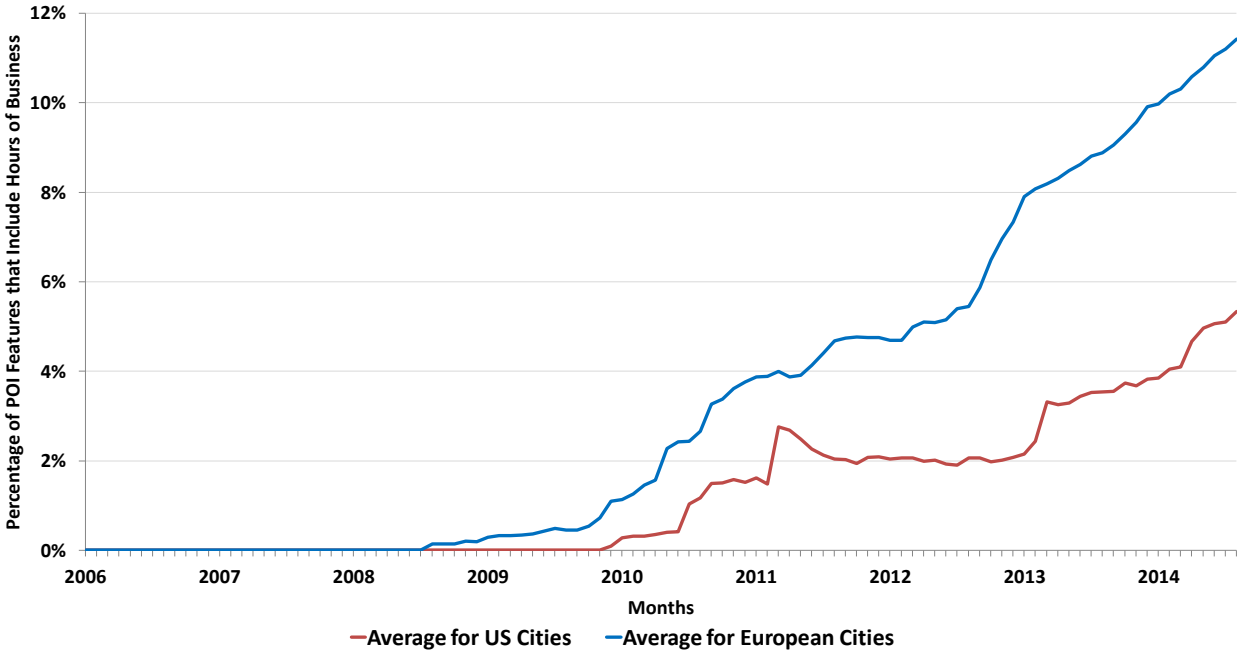


Figure 16. Percentage of POI Features that Include Hours of Business

4.5. Completeness of Data

Overall completeness in the mapping of physical features is one of the most basic measures of quality, but a comprehensive assessment requires ground truth data. Most such data is proprietary for the cities sampled, however, if it even exists. As a result of this limitation, and in keeping with the goal of extracting broad insights across communities, a simpler ratio of number of features to population size was used. The basic facilitating assumption is that the total number of features analyzed should be roughly equivalent between cities, especially when viewed across a sample of this size. The following groupings of features were analyzed for completeness: total number of all features, number of all features tagged as a building, number of all features tagged as an amenity, number of point features tagged as public transportation, and number of features tagged as bike-related. The bike related features proved to be too variable between cities to be useful for making broad comparisons and is not included.

4.5.1 Total Features

Total features normalized for population, depicted in Figure 17, highlight a consistent lag by US cities along a stair step pattern. In fact, the US averages here are indeed driven by a few select cities that have had massive data imports, such as Chicago and Boston. This explains the generally uneven pattern of US feature growth and may reduce the true significance of what may appear to be somewhat parallel growth.

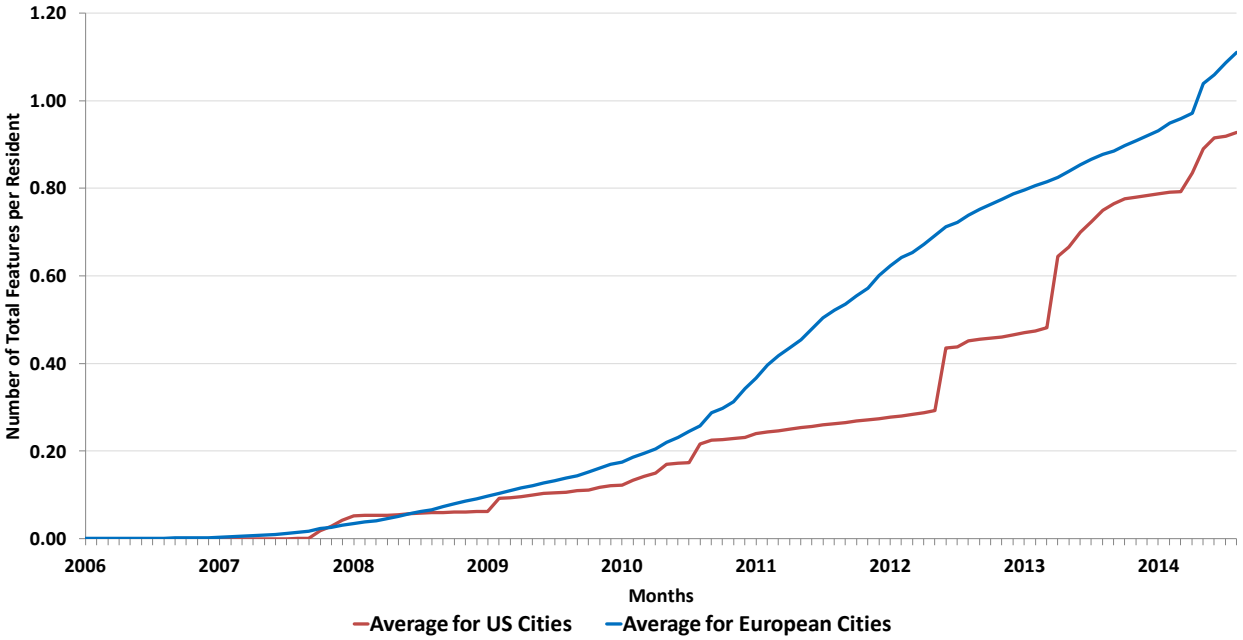


Figure 17. Number of Total Features per Resident

4.5.2 Amenity and Building Features

A look into the completeness of features tagged as amenities and buildings shows a wider gap than total number of features. The pattern of US features in Figure 18 further emphasizes the individual, massive imports of building data by a few US cities and shows both slower feature growth rates as well as lower total numbers. Unless European cities actually have significantly more buildings and amenities, data completeness in OpenStreetMap is much greater for European cities than for US cities.

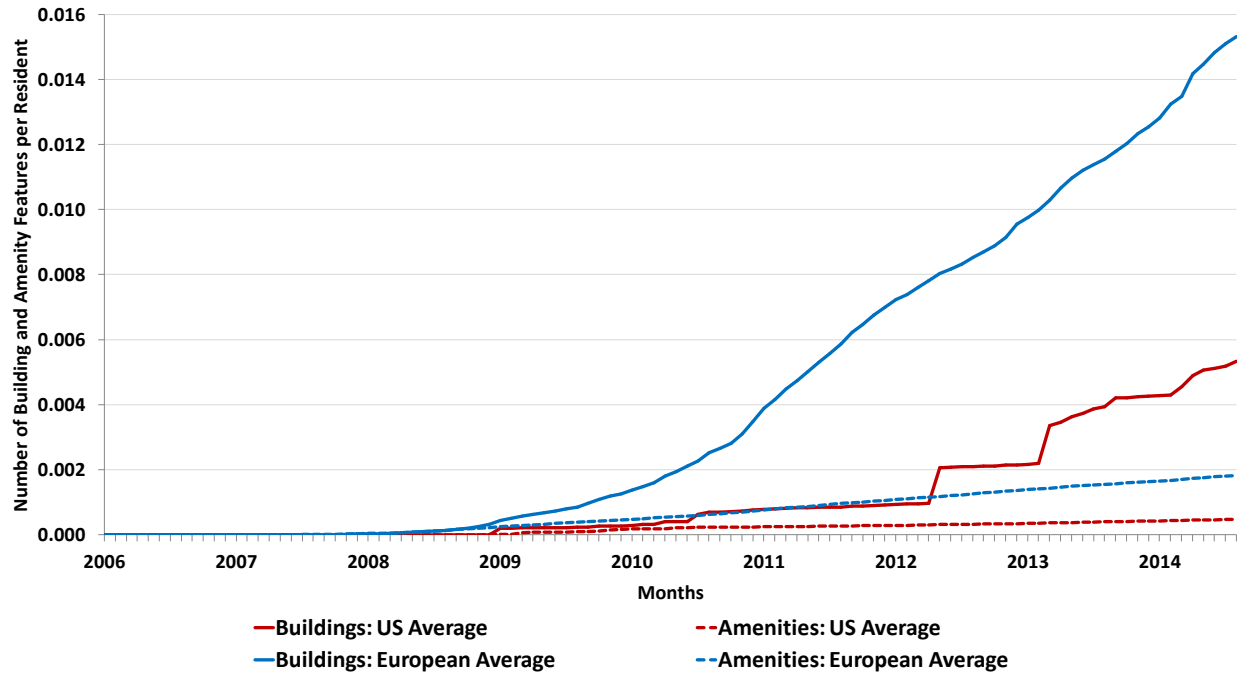


Figure 18. Number of Building and Amenity Features per Resident

4.5.3 Public Transportation Features

Figure 19 shows an even more significant difference in the number of point features tagged as public transportation. Only point features were chosen to compare the simplest level possible, giving locations that may have only tagged bus stops instead of actual linear routes greater consideration. While the prevalence of public transportation may vary more than the other metrics between individual cities, it seemed to still be a relevant comparison when only evaluating point features. The size of the gap between communities is not likely explained solely by European cities having more public transportation infrastructure.

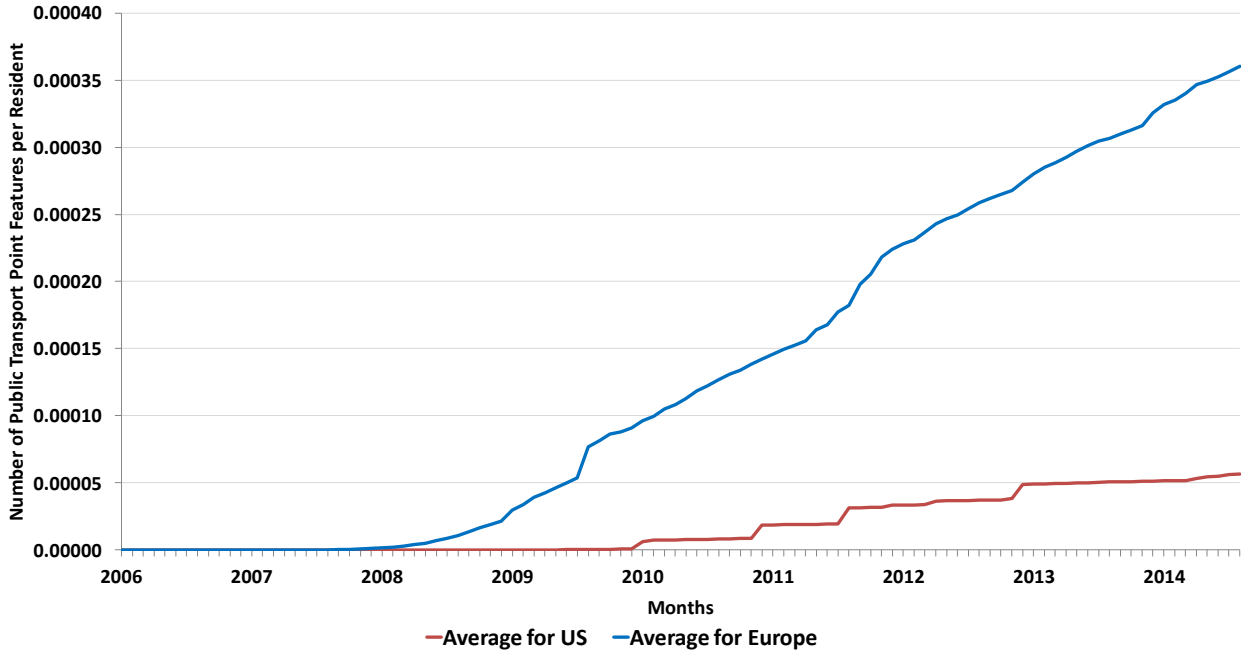


Figure 19. Number of Public Transportation Point Features per Resident

4.6. Focus of Edits

A final point of comparison between communities was the percentage of edits by associated geometry types, which included point, line, and polygon features. All thematic feature types examined under completeness and maturity were also evaluated by percentage of edits, but the few meaningful results are only included in Section 5. Figure 20 shows the percentage of edits by feature geometry, with the most interesting results involving lines and polygons. European cities started shifting a higher percentage of their edits from lines to polygons after 2008, while the US cities have only begun to make slight increases in their focus on polygons. Arguably, polygons are more complex features and a focus on these may be an indicator of maturity as well.

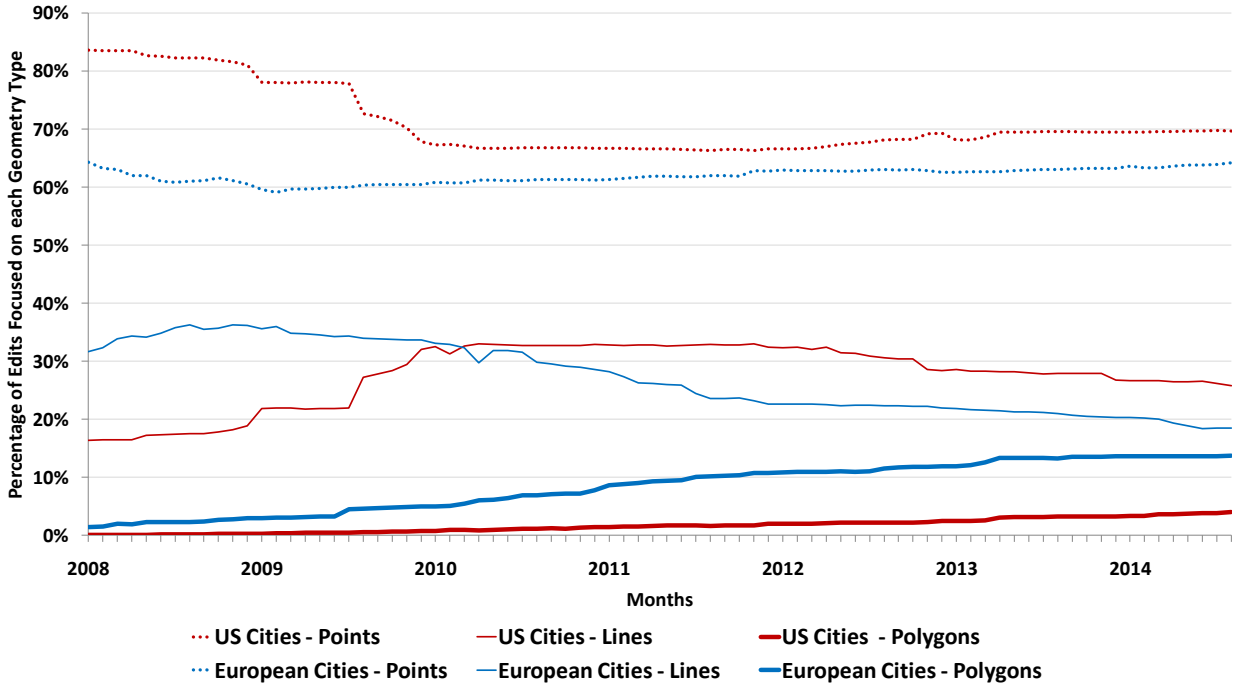


Figure 20. Percentage of Edits Focused on each Geometry Type

5. Characteristics of Successful Communities

Just comparing the US and European communities was not enough to extract meaningful lessons. To do this, specific metrics within four of the above categories of comparison were selected to rank each city in those categories. These metrics were chosen as being able to represent some measure of success across all communities and included User Activity, Local Event Frequency, Maturity of Data, and Completeness of Data. Cities were ranked according to specific metrics that made up each category. In all cases except for Local Event Frequency, averages for the last five years of data were used to give a more complete story of each community and minimize the influence of large surges of limited duration in any category. The resulting four sets of rankings were then correlated to assess their validity in describing what successful communities have in common.

5.1. City Rankings in Each Category and Overall

For User Activity, five year averages of unique contributing users per month and a rolling 12 month growth rate accounted equally for the ranking. For Local Event Frequency, rankings resulted from the total number of months in the last five years that contained at least one local event. There were multiple tied rankings in Local Event Frequncy due to the limited potential options of zero to 60 months that contained event activity.

For Maturity of Data, the five year averages of percentages for the following metrics were equally weighted for a combined ranking: road features with speed limits, POIs with contact information, POIs with addresses, and POIs with business hours. These four represent more advanced maturity than the basic inclusion of names and were therefore better measures of community success.

For Completeness of Data, the units of measure were not directly comparable across total features, public transportation features, building features, and amenity features. As a result, each city was ranked individually in these categories and the rankings were averaged. As with events, there were tied rankings in this data due to similar averages of rankings with finite 1-38 values. Bike features were not used due to the high variability between cities.

The overall city ranking resulted from the averages of all four individual category rankings. While this method is simplistic in nature, it was effective at broadly identifying the cities most successful in all four categories. Table 4 lists the resulting rankings for cities, sorted by overall rankings within their respective communities. Each ranking is associated a color scheme for easy visualization, with first being green and 38th being red. The overall rankings did show a three way tie for sixth, due to the finite possibilities of four rankings of 1-38 averaged together. Had this methodology weighted factors differently, which was outside the scope of this project, this would not have likely occurred.

	Overall	User Activity	Event Freq.	Maturity	Complete-ness		Overall	User Activity	Event Freq.	Maturity	Complete-ness
Washington DC	10	16	8	12	4	Berlin, DE	1	2	1	3	1
Seattle, WA	11	10	12	14	13	Munich, DE	2	1	4	1	2
Portland, OR	14	11	20	22	15	Hamburg, DE	3	4	2	4	3
Boston, MA	15	20	14	16	20	Dusseldorf, DE	4	8	5	7	4
New York, NY	17	19	11	18	27	Bremen, DE	5	13	6	5	7
Denver, CO	18	14	14	21	29	London, UK	6	5	3	19	6
Chicago, IL	21	24	14	27	18	Frankfurt, DE	6	3	10	11	9
Atlanta	22	30	21	17	17	Mannheim, DE	6	9	7	6	11
Los Angeles, CA	25	23	17	33	22	Zurich, CH	9	7	17	8	7
San Jose, CA	26	26	13	36	21	Nantes, FR	12	6	26	10	10
Philadelphia, PA	27	25	17	24	34	Saarbrucken, DE	13	18	24	9	12
Baltimore, MD	29	31	22	23	27	Oslo, NO	16	17	29	13	14
St Paul, MN	31	33	29	34	24	Madrid, ES	19	15	22	26	16
Albuquerque, NM	32	29	32	29	32	Rotterdam, NL	20	32	24	2	22
Houston, TX	33	35	26	31	31	Gdansk, PL	22	12	29	15	29
Providence, RI	35	36	32	25	37	Glasgow, UK	24	27	9	28	24
Richmond, VA	36	34	32	35	35	Liege, BE	27	22	32	20	26
Miami, FL	37	38	32	30	38	Nice, FR	30	21	32	32	19
Honolulu, HI	38	37	32	38	36	Seville, ES	34	28	26	37	33

Table 4. City Overall Rankings and Rankings in Individual Categories

5.2. Correlations Between Individual Rankings and Other Metrics

Rankings in each of the four categories of comparison have meaning individually, but it was important to assess what their meaning may be when combined into an overall ranking. Using a Pearson correlation coefficient matrix, Table 5 shows that all four categories had significant positive correlations with one another. Cities that ranked highly in one category typically ranked highly in all categories; using these together in a single ranking is a meaningful way to extract commonalities between the highest and lowest ranked cities. These categories themselves are, in fact, some of the most meaningful commonalities of successful communities. The correlation between completeness and maturity is notable, as this confirms that cities can and do have advanced metadata across very complete datasets. Frequent events do relate to the other metrics of success, and the level of user activity is very closely tied to the overall completeness of a city's data.

	User Activity	Event Frequency	Maturity
Event Frequency	0.68	-	-
Maturity	0.74	0.54	-
Completeness	0.85	0.72	0.75

Table 5. Pearson Correlation Coefficients Between Individual Category City Rankings

In addition to correlations between individual rankings, those were correlated with many other metrics not included in the rankings in order to identify further similarities of successful cities. Table 6 shows the Pearson coefficients between overall ranking and several metrics related to Sources of Data and Focus of Edits. The most successful communities were also the most likely to follow best practices of tagging the majority of features with a source, most likely to use personal surveys as a source of data, and slightly less likely to use data imports as a source. These same cities also focused a considerably higher percentage of their edits on polygon features and those tagged with amenities, with much less focus on editing roads as an overall percentage. The results of both Tables 5 and 6 provide the basis for profiling the nature of a successful OpenStreetMap community.

Sources of Data:	
0.79	More Features Tagged with 'Source'
0.60	More Months of Edits Dominated by Surveyed Sources
0.29	Fewer Imports
Focus of Edits:	
0.76	High Percentage of Edits Involving Polygon Features
-0.66	Low Percentage of Edits involving Road Features
0.42	High Percentage of Edits Involving Amenities

Table 6. Pearson Correlation Coefficients Between Overall City Rankings and Selected Metrics

6. Conclusions and Recommendations

6.1. Conclusions

There have been and continue to be definitive differences in the activities and resulting data of US and European OpenStreetMap communities. Historically, US user activity has not only lagged behind Europe in raw numbers, it has also grown at a slower rate. The same has generally been true of event frequency, at least at the local level. Both of these trends have begun to reverse in the last two years, with both growth and number of events now on par with Europe. Events have gained much traction as a result of the Maptime chapters that came out of the 2013 SOTM US conference. Actual participation in the US is still far behind, however.

US cities tend to rely more heavily on imported data, and much less on physical surveys in order to collect data. Data quality can suffer as a result, if imports are not done very carefully, and cleanup of past mistakes still require much laborious effort. This may also discourage further efforts to add metadata not

present in the import data, as opposed to creation through physical surveying or other means when it is most easily gathered concurrently.

The maturity of data in the US is generally lower across the most advanced metrics, but not substantially so except in business hours for POIs and speed limits for roads. More significantly, its completeness of data other than roads is far behind Europe. Given the very high correlation between user activity and completeness, this result seems very logical. The US has fewer mappers, so completeness has suffered.

In order to identify which characteristics make up successful communities, four primary categories were chosen as potential metrics of success. High User Activity, Local Event Frequency, Maturity of Data, and Completeness of Data were indeed common for each successful community. Maturity does not need to be sacrificed in order to achieve completeness, for example. These communities also tend to map by hand or GPS rather than import massive amounts of data, and they focus more of their efforts on complex features and attributes not visualized on maps. A city is likely to more successful by making efforts to improve in each of these categories.

6.2. Recommendations to US OpenStreetMap Community

Based on the conclusions of this study, several recommendations can be made to US communities in order to increase participation and data quality. First, communities should set up consistent and engaging local events. This is already starting to happen through the establishment of Maptime and their associated rebranding of OpenStreetMap communities, but it is crucial that the momentum continues and event frequency does not peak as it did in 2009.

Regardless of whether frequent events are a primary driver of user participation, increased outreach efforts drive event frequency, or some combination of the two, prioritizing growth in the number of mappers is crucial to success. With user activity being most closely correlated to data completeness, and significantly correlated to all the other metrics of success, one cannot lose sight of the importance of raw numbers.

As part of these events and increased collaboration, communities should set goals and strategies unique to their needs. Continuing to clean up TIGER data may be necessary, but this should be balanced with the editing of more complex and potentially more interesting features. All of this must be done alongside a strong educational effort that teaches the skills necessary to map accurately and completely, in addition to clearly laying out expectations for quality standards within the OpenStreetMap community. Increasing maturity of attributes requires understanding which attributes should be included, and being disciplined in doing so. This study showed that data completeness does not have to come at the cost of decreased attribute maturity.

In order to support the above mentioned efforts, future research could focus on better identifying the gaps between the ideal OpenStreetMap standards of attribute data and actual community data across more features and attributes. Using recently available tools, such as OSM-Epic or potential future services derived from it, would allow even non-experts to conduct detailed analysis of attribute data at a much lower time cost than methods used in this study (“OSM History”, 2015). These steps will arm communities with the tools necessary to precisely understand their needs, and then craft the most effective strategies to address them. Understanding the current state of data and taking deliberate steps to improve it will pay dividends as the communities grow.

7. References

"About OpenStreetMap." (n.d.). OpenStreetMap Wiki. Retrieved March 18, 2014, from http://wiki.openstreetmap.org/wiki/About

"American FactFinder." (2014). United States Census Bureau. Retrieved August 1, 2014, from http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml

"City population by sex, city and city type." (2014). United Nations Statistics Division. Retrieved August 1, 2014, from http://data.un.org/Data.aspx?d=POP&f=tableCode%3A240

Corcoran, P. and Mooney, P. (2013). "Characterising the metric and topological evolution of OpenStreetMap network representations." The European Physical Journal Special Topics 215(1), 109-122.

"Current Population Estimates." (2014). The City of New York. Retrieved August 1, 2014, from http://www.nyc.gov/html/dcp/html/census/popcur.shtml

Girres, JF. and Touya, G. (2010). "Quality Assessment of the French OpenStreetMap Dataset." Transactions in GIS, 14(4), 435-459.

Goodchild, M. (2007). "Citizens As Sensors: The World Of Volunteered Geography." GeoJournal, 69(4), 211-221.

Hochmair, H. H., Zielstra, D., & Neis, P. (2013, January). "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.

Hristova D., Quattrone G., Mashhadi A., and Capra L. (2013). "The life of the party: Impact of social mapping on openstreetmap." In Proceedings of the AAAI International Conference on Weblogs and Social Media(ICWSM2013). Association for the Advancement of Artificial Intelligence (AAAI).

"Import." (2014). OpenStreetMap Wiki. Retrieved August 10, 2014, from http://wiki.openstreetmap.org/wiki/Import

"Mapping L.A. Neighborhoods". (2014). Los Angeles Times. Retrieved August 5, 2014, from http://maps.latimes.com/neighborhoods/

Mashhadi, A., Quattrone, G., Capra, L., & Mooney, P. (2012). "On the accuracy of urban crowd-sourcing for maintaining large-scale geospatial databases." In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration. ACM.

"MaZderMind/osm-history-splitter." (n.d.). GitHub. Retrieved April 13, 2014, from https://github.com/MaZderMind/osm-history-splitter

Mondzech, J., and Sester, M. (2011). "Quality analysis of OpenStreetMap data based on application needs." Cartographica: The International Journal for Geographic Information and Geovisualization, 46(2), 115-125.

Mooney, P., and Corcoran, P. (2012). "Characteristics of heavily edited objects in OpenStreetMap." Future Internet, 4(1), 285-305.

"My City Mapps & Apps." (2014). City of Houston. Retrieved August 12, 2014, from http://mycity.houstontx.gov/home/

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

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(2), 146-165.

“OSM History.” (2015). The Regents of the University of Colorado. Retrieved April 20, 2015, from http://project-epic.github.io/epic-osm/

"Planet.osm/full." (2014). OpenStreetMap Wiki. Retrieved August 20, 2014, from http://wiki.openstreetmap.org/wiki/Planet.osm/full

"Population Estimates for UK, England and Wales, Scotland and Northern Ireland, Mid-2013." (2014). Office of National Statistics. Retrieved August 1, 2014, from http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-322718

Fan, H., Zipf, A., Fu, Q., and Neis, P. (2014). "Quality assessment for building footprints data on OpenStreetMap." International Journal of Geographical Information Science, (ahead-of-print), 1-20.

Neis, P., Zielstra, D., and Zipf, A. (2013). "Comparison of volunteered geographic information data contributions and community development for selected world regions." Future Internet, 5(2), 282-300.

Neis, P., & Zielstra, D. (2014). "Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap." Future Internet, 6(1), 76-106.

Stephens, M. (2013). "Gender and the geoweb: Divisions in the production of user-generated cartographic information." GeoJournal, 78(6), 981-996.

"State of the Map." (2014). OpenStreetMap Wiki. Retrieved December 1, 2014, from http://wiki.openstreetmap.org/wiki/State_Of_The_Map

"What is Maptime?" (2015). Maptime. Retrieved February 15, 2015, from http://maptime.io/about/

"Wiki History." (n.d.). Wiki History. Retrieved April 10, 2014, from http://c2.com/cgi/wiki?WikiHistory

Willis, N. (2007, October 11). "OpenStreetMap project imports US government maps." Linux.com. Retrieved March 27, 2014, from http://archive09.linux.com/feature/119493

Yasseri, T., Quattrone, G., and Mashhadi, A. (2013). "Temporal analysis of activity patterns of editors in collaborative mapping project of OpenStreetMap." In Proceedings of the 9th International Symposium on Open Collaboration (p. 13). ACM.

Zielstra, D., Hochmair, H. H., and Neis, P. (2013). "Assessing the Effect of Data Imports on the Completeness of OpenStreetMap–A United States Case Study." Transactions in GIS, 17(3), 315-334.

Zielstra, D., and Zipf, A. (2010). "A comparative study of proprietary geodata and volunteered geographic information for Germany." In 13th AGILE international conference on geographic information science (Vol. 2010).