Data Quality of Points of Interest in Selected Mapping and Social Media Platforms

Hartwig H. Hochmair, Levente Juhász and Sreten Cvetojevic

Abstract A variety of location based services, including navigation, geo-gaming, advertising, and vacation planning, rely on Point of Interest (POI) data. Mapping platforms and social media apps oftentimes host their own geo-datasets which leads to a plethora of data sources from which POIs can be extracted. Therefore it is crucial for an analyst to understand the nature of the data that are available on the different platforms, their purpose, their characteristics, and their data quality. This study extracts POIs for seven urban regions from seven mapping and social media platforms (Facebook, Foursquare, Google, Instagram, OSM, Twitter, and Yelp). It analyzes the POI data quality regarding coverage, point density, content classification, and positioning accuracy, and also examines the spatial relationship (e.g. segregation) between POIs from different platforms.

Keywords VGI · Crowd-sourcing · Point of interest · Data quality

1 Introduction

Location based services (LBS) play an important part in our everyday life. For many tasks LBS use an inventory of points of interest (POI), also often called places, venues, or businesses, which can originate from commercial (e.g. Google) or crowd-sourced platforms (e.g. OpenStreetMap (OSM)). POIs can be used to geo-reference social activities, such as posting a picture on Instagram, sending a geo-tagged tweet, or checking into a Foursquare/Swarm location (Rösler and Liebig

H. H. Hochmair (✉) · L. Juhász · S. Cvetojevic
Geomatics Program, University of Florida, Davie, FL 33314, USA
e-mail: hhhochmair@ufl.edu

L. Juhász
e-mail: levente.juhasz@ufl.edu

S. Cvetojevic
e-mail: scvetojevic@ufl.edu

P. Kiefer et al. (eds.), *Progress in Location Based Services 2018*, Lecture Notes in Geoinformation and Cartography, https://doi.org/10.1007/978-3-319-71470-7_15

2013). POIs are also relevant for navigation solutions, e.g. when providing reference points in travel directions (Nothegger et al. 2004; Duckham et al. 2010) or for suggesting venues along a route as part of personalized route recommendations (Lim et al. 2015). Due to the importance of POIs in geo-applications including LBS a solid understanding of their nature and data quality is essential to determine their fitness for purpose. Through successful conflation of POIs from different sources, attributes could be complemented, the number of objects increased, and the data quality of POIs improved (Hastings 2008). Successful conflation necessitates, however, to handle the challenging problem of data integration and achieving data interoperability. It requires also an understanding of the nature of POIs provided on the different platforms in order to identify promising candidates for conflation and integration in the first place. This study will tackle the latter task by a joint comparison of various quality aspects of POIs from seven commercial and Volunteered Geographic Information (VGI) crowd-sourcing platforms. This joint analysis is the novel aspect of this contribution, which builds on quality measures (e.g. richness in POI categories) that have been applied in other similar studies before.

From among the various data quality elements that are commonly used to determine how well a geo-spatial dataset meets its specified criteria, this research will closer examine relative completeness (abundance, categorization) and positional accuracy, as well as location bias by analyzing attraction and repulsion between marked point sets from different data sources through Cross-K functions. Most analyses in this study do not use ground truth data since perfect knowledge about POI locations in the different cities is difficult to obtain. Instead it applies comparative measures.

Current studies of POI quality assessment focus primarily on single data sources. They often use intrinsic quality measures (Barron et al. 2014; Gröchenig et al. 2014) or compare the data source in question to a proprietary or governmental reference data set (Senaratne et al. 2017). Especially OSM received considerable attention in these aspects (Jackson et al. 2013; Fan et al. 2014). Using a set of OSM POIs, Mülligann et al. (2011) use geo-ontologies to determine the plausibility of a POI type within a given neighborhood, which can be used for tag recommendation, data cleaning, and coverage recommendation. Several studies address also quality and conflation aspects of multiple POI sources. For example, a comparison of POIs from proprietary (TomTom, NAVTEQ, ESRI), governmental (TIGER/Line, USGS GNIS) and crowd-sourced (OSM) data sources finds that categorization schemes in the different platforms change over time, and that no single data source outperforms another in all aspects (Hochmair and Zielstra 2013). To address the challenges of integrating heterogeneous POI data sets from different sources, McKenzie et al. (2014) developed a weighted multi-attribute method which matches POIs from different sources and applies a variety of similarity measures, such as the Levenshtein distance on feature names, or category alignment based on WordNet. Similarly, Li et al. (2016) allocate Entropy based weights to POI attributes (e.g. distance between objects, name and sound similarity, category similarity) to improve POI matching from different sources.

The digital divide determines in which geographic regions users have the technical and economic means to participate in VGI and social media activities (Heipke 2010), which has a direct effect on VGI data quality. Furthermore, in OSM data coverage is affected by data imports, including a 2009 import of POIs from GNIS (Hochmair and Zielstra 2013), and a 2007 import of TIGER/Line road data in the US (Zielstra et al. 2013). Such an import will affect not only data coverage but also the range of OSM categories found in local datasets.

2 Data Collection

Data for the study were collected from sub-areas of the following seven cities: Albuquerque (New Mexico), Cairns (Australia), Gainesville (Florida), London (England), Nairobi (Kenya), Qingdao (China), and Salzburg (Austria). POI data with geographic coordinates were downloaded through Application Programming Interfaces (APIs) from seven selected data sources (Facebook, Foursquare, Google, Instagram, OSM, Twitter, and Yelp) and inserted into a PostgreSQL database. Twitter places at the different hierarchical levels (POI, neighborhood, city, administration) were extracted from worldwide tweets posted between 20 September 2016 and 20 October 2016 rather than from the Twitter REST API because of faster data access. The location of all Twitter place types except for POIs are defined through a bounding box. Tweets themselves were downloaded in JavaScript Object Notation (JSON) format through the Twitter Streaming API using the Tweepy python library.

For data download of other sources, requests were made in a Python environment using existing API wrappers, where available (Instagram, Facebook, Yelp, Foursquare), or using custom solutions (OSM, Google). The typical approach was to search places within a given radius around a center point (Facebook, Instagram, Google) or within a rectangular area (Foursquare, Yelp), which was moved along in a grid like pattern to cover the area to be analyzed. Since APIs often limit the returned data volume, locally refined rectangles or circles were inserted to ensure the capture of all POIs within an area whenever this threshold was met. An illustration of this refinement process for Yelp data retrieval is provided in (Juhász and Hochmair 2017).

For OSM, a different approach was chosen since large areas can be queried via the OverpassAPI. The query extracted all nodes with names, all ways with names (except for waterways and routes), all ways that are bridges and have names, and all relations (an ordered lists of nodes or ways) with a name and type = multipolygon tag. Since in OSM certain map features, such as parks and buildings, are often mapped as ways or relations, these were represented by their centroids in the final dataset. A set of working code examples that illustrate the different methods and libraries to use APIs for selected VGI and social media services, including Twitter, Instagram, Foursquare, and OSM, are provided in the literature (Juhász et al. 2016).

Technical details about the geo-tagging process in Twitter and Instagram, and its effect on positioning accuracy, are discussed in another study (Cvetojevic et al. 2016). Instagram positions were obtained in October 2015 where the API could still be used without going through an approval process, which changed effective June 1, 2016.[1] All other place data, except for those from tweets, were download in May 2017.

In several data sources many POIs were stacked on top of each other at the exact same point location. Such a case is plausible if various parties reside at a single building. Examples are hospitals (with doctor's offices as individual POIs geo-coded at the same geographic location), or commercial buildings that host several business offices. In many other cases, however, the stacking of POIs appears incorrect, especially if there are no major buildings in the vicinity. In some cases, stacked POIs aggregate places (e.g. businesses, plazas, parks) from across a whole city district. Several possible explanations can be found, such as (1) different locations being aggregated to a single point location by the platform, or (2) users uploading information of different POIs from one physical location (e.g. after a wireless network connection became available), and being unaware of the app attaching that same location to all POIs. Large POI stacks often contain POIs with made up place names (e.g., My Bed, Smoker's Paradise, Hell, Mi Casa, LETS OPEN Our BIBLE), people's names or unlikely business names (e.g. Herstyle, Happy Healthy Life) with no actual business found nearby. These kind of stacked POIs could be the outcome of location spoofing, i.e., the intentional falsifying of one's locational information (Zhao and Sui 2017). To avoid massive, incorrect POI stacks biasing subsequent cluster and density analyses, all point locations with 15 or more stacked POIs were manually reviewed for plausibility. If no building or market plaza of appropriate size was found at the posted position, the stacked POIs were removed from further analysis. The POI stack size of 15 is arbitrarily chosen. Although this stack size does not capture small clusters, it reduces the workload for manual cluster checking to a manageable amount.

Table 1 lists the size of identified POI clusters in descending order that were removed before further data analysis. Most removed clusters are found in Facebook and Instagram, which suggests that user-added places on these platforms undergo only little review and quality control. Incorrect clusters occur in all analyzed cities, but mostly in Nairobi, which is possibly indicative of poor positioning accuracy in that city. In fact, one point had 1195 different Facebook places stacked on top of each other.

All seven analyzed platforms operate worldwide, however, with some differences in data coverage between countries. Yelp and Twitter POIs, for example, are available in only five of the seven analyzed cities.

[1] http://developers.instagram.com/post/133424514006/instagram-platform-update.

Table 1 Size of removed clusters

Platform	City	Sizes of removed clusters
Facebook	Albuquerque	16
	Cairns	445, 246, 118, 37, 25, 24
	Gainesville	279, 54, 37, 21, 20
	London	–
	Nairobi	1195, 813, 245, 121, 64, 37, 37, 31, 26, 24, 23, 20, 18, 18, 17
	Qingdao	357
	Salzburg	113
Instagram	Albuquerque	105, 37, 28, 28, 21, 18
	Cairns	33, 20, 19
	Gainesville	58, 34, 30, 29, 28, 18
	London	33, 24, 23, 17
	Nairobi	33, 33, 33, 33, 33, 31, 26, 26, 20, 20, 18, 18, 17, 17, 16
	Qingdao	33, 33
	Salzburg	33, 31
Google	Nairobi	15, 15

3 POI Classification

3.1 Platform Comparison

Facebook, Google, OSM, Yelp, and Foursquare provide a categorization of their POIs into different hierarchical levels. New places added by users need to follow the provided POI categorization for these platforms. As opposed to this, the content of Twitter and Instagram POIs can only be characterized by their name since these platforms do not provide POI categories. For an application there can be differences in categories between POIs shown on a Web map and POIs downloaded through an API. The latter categorization is typically limited in detail. One example is a youth hostel that in the Google Places API Web Service is classified as type "lodging" whereas on Google Maps (in the Web browser) the same feature is classified as a more detailed "2-star hotel". Another example is "school" (API) versus "high school" (browser map).

The list of Google place types cannot be directly retrieved from the API. Alternative methods include extraction of place types from downloaded POIs (similar to how it was done for Twitter), or using classifications from third parties, such as Blumenthals.[2] That Website lists Google place categories for different language-country combinations. POI category numbers vary strongly between languages, e.g., US English (N = 2465), British English (N = 847), German

[2]http://blumenthals.com/google-lbc-categories/search.php?q=&val=hl-gl%3Den-US%28PfB%29%26ottype%3D1.

(DE) (N = 997), French (FR) (N = 1183), or Spanish (ES) (N = 2365). The place types returned in Google Maps (browser map) depend on the Google domain used (i.e., google.com, google.fr, google.it, etc.). Similarly, if users suggest a new place to be added to Google Maps, the available categories with their languages depend on the chosen Google domain. However, when downloading Google places through the Places API Web Service, only English place categories are returned, independent of the language setting. The language setting does also not affect the number of features returned. Returned place features often contain a list of place categories, where the first category appears to be the most specific one. Examples for POI categories include {doctor, health, point of interest, establishment} or {car dealer, store, point of interest, establishment}.

OSM offers a free tagging system for nodes, ways, and relations, which allows a user to add an unlimited number of attributes to features. However, the OSM community agrees to certain key-value combinations for commonly used features. Features are divided into 23 primary feature categories[3] (amenity, building, highway, etc.) which represent the key of a feature. Each key can take many different values to further specify the sub-type of the mapped feature. Key-value examples include highway = motorway, or amenity = restaurant.

Yelp places are structured in a four-tiered hierarchy which is customized for 32 countries.[4] POI classification schemes contain between about 900 and 1200 categories. A single venue can be assigned to several categories even from different hierarchical levels, e.g. Gyms (L3), Sports Clubs (L2), and Day Spas (L2). Setting a country determines which Yelp place categories can be added to Yelp. This is because many POI categories are associated with a whitelist (which specifies in which countries a category can be added) and/or a blacklist (a list of countries where the category cannot be added). For example, category "Bird Shops" has the following whitelist: NO, NL, DE, IT, SG, BE, ES, US, DK, SE. Thus with Sweden (SE) as chosen country bird shops are recognized by autocomplete during manual data entry (Fig. 1a), whereas, for example, in the UK they are not (Fig. 1b).

POI categories returned through the Yelp API depend on the local setting of the app/request as well. Yelp uses an "alias" for each category, similar to a unique category ID, but the name of the category returned by the API depends on the country setting. For example, the "landmarks" alias returns a "Landmarks & Historical Buildings" feature category for the US, and a category "Sehenswürdigkeiten" for Germany.

The list of 785 Facebook place topics can be extracted from the Facebook Graph API Explorer, which is not hierarchically structured. This list is more comprehensive than what is offered when interactively creating a Facebook page for a new business, brand, product etc. on the Facebook Web site.

All 920 Foursquare place categories are organized in a five-tiered hierarchical structure, where the top hierarchy contains 10 entries including Arts &

[3] http://wiki.openstreetmap.org/wiki/Map_Features.

[4] https://www.yelp.com/developers/documentation/v2/category_list.

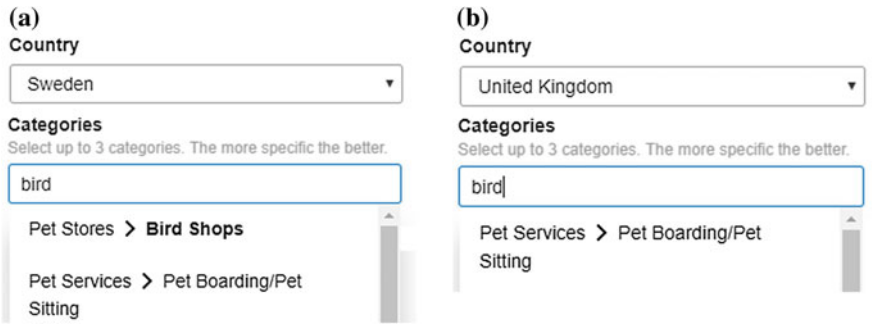

Fig. 1 Trying to add a place for a country that is listed (**a**) and not listed (**b**) on the white list of category "Bird Shops" in Yelp

Entertainment or Travel & Transport. Some venues are restricted to certain countries, as is specified in the Foursquare category documentation. For example, category "Anhui Restaurant" is available only in China and some other nearby countries.

3.2 Observed Distribution of POI Categories

Figure 2 shows the 10 most frequently used categories of POIs that were downloaded in Albuquerque, Gainesville, and London from the five social media platforms that provide POI categories. The analysis was limited to these three cities since they are the only ones that provide POI information (not necessarily category information though) for all seven platforms, and were therefore also used as common geographic areas for other data quality metrics. The top row in Fig. 2 shows that the most prominent OSM primary key is building, followed by amenity and shop, and that restaurant and hotel are the most prominent sub-types which were identified through querying OSM key-value pairs. Since shops in OSM are distributed across 73 potential shop types according to the wiki Map Features site, no shop type makes it to the top ten in that second chart. Figure 2 as a whole reveals that the most frequently used POI categories vary between analyzed platforms, comprising tourism facilities (hotels, restaurants, cafes in Yelp, OSM, and Foursquare), health infrastructure (doctors, hospitals, medical in Google, Facebook), and university buildings (Foursquare).

These differences indicate that the different platforms have strengths in different category groups and may thus complement each other in a meaningful way. It should be noted that charted category frequencies are based on only three cities in the US and Europe. They will differ from category frequencies found in other cities, especially those located in other parts of the world. Nairobi, for example, shows

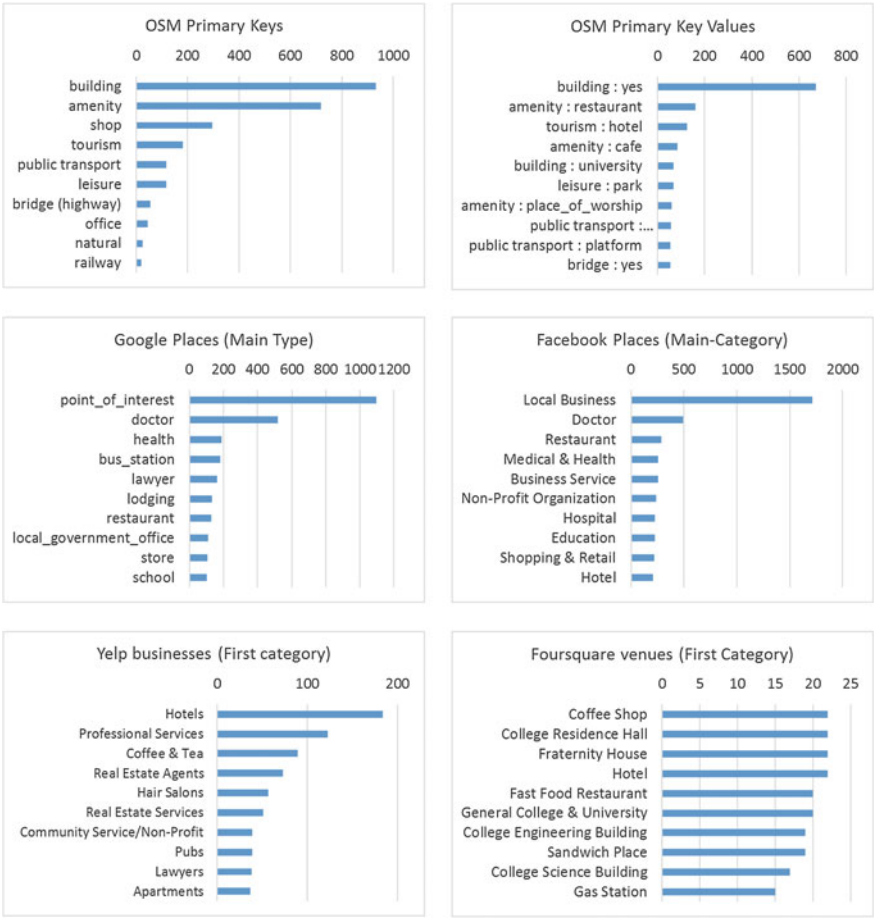

Fig. 2 Most frequent POI categories found in Albuquerque, Gainesville, and London

"sport" on 6th place in OSM primary keys, includes finance and banks in the top five Google categories, maps only a single doctor's office in Facebook, and misses fraternity housing in Foursquare altogether. These differences reflect local variations in the presence of venues between cities.

Twitter place types do not contain thematic categories but are organized in a spatial hierarchy instead. The coarsest place level is admin polygons which vary in size, density, and coverage between and within countries (Fig. 3a).

A more refined place level is cities which are found primarily in Europe, Canada, Mexico, Brazil, Japan, New Zealand, and a few other countries. Neighborhood places represent the sub-city level with a few clusters around the world, including the Netherlands, Australia, and New Zealand. Although metadata for city,

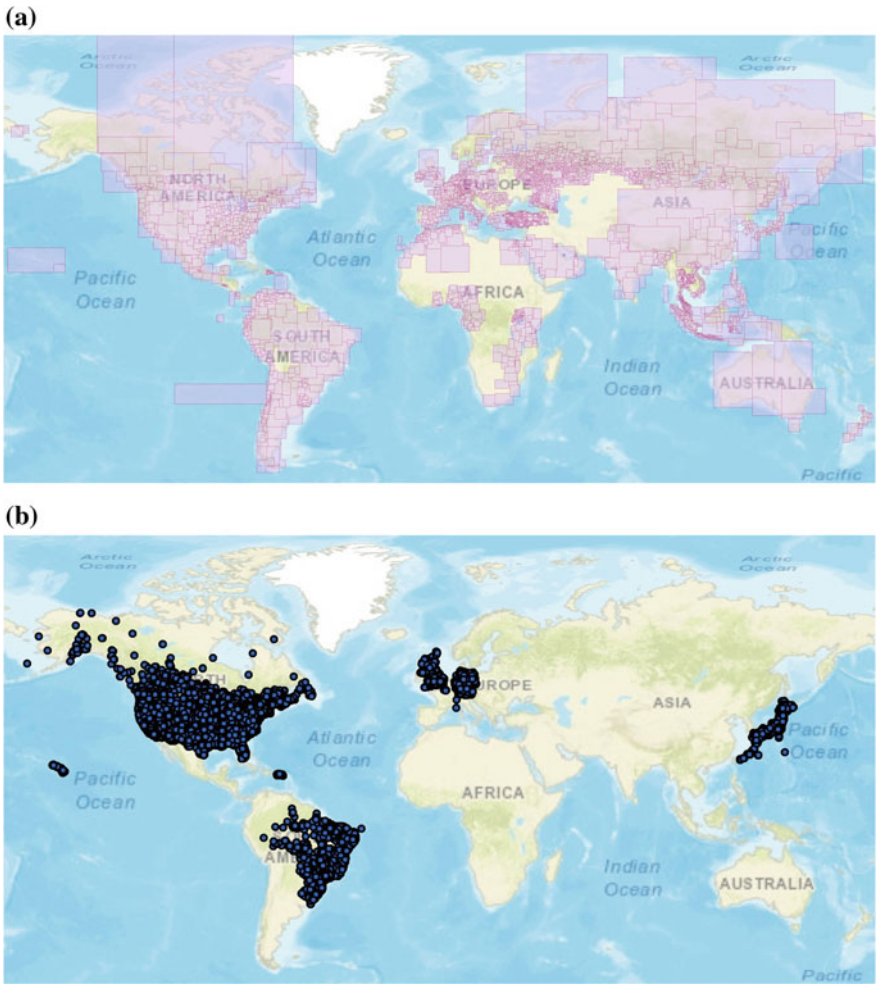

Fig. 3 Worldwide coverage of Twitter admin place type (Alaska and ocean polygons were removed from this map for clarity) (**a**), and place type Point of interest (POI) (**b**)

neighborhood, and administrative contain only a minimum bounding box as a feature geometry, Twitter uses accurate boundary polygons internally to determine which city, neighborhood, etc. a tweet needs to be assigned to during the geo-tagging process. The POI place type is the only one with point geometry and found in several countries, including Canada, the US, Brazil, Great Britain, Germany, and Japan (Fig. 3b). The POI level will be analyzed in more detail for selected cities.

4 POI Pattern Analysis

4.1 POI Density and Spatial Distribution

Table 2 summarizes descriptive statistics of POIs for the analyzed data sources in cities where they are available. Values in the upper row of each data source list the number of downloaded POIs per km^2, and values in the lower row denote the average nearest neighbor index (NNi), computed as the ratio of the observed over the mean nearest-neighbor distance (O'Sullivan and Unwin 2010). The NNi characterizes a point pattern relative to complete spatial randomness, i.e., a point pattern created by a homogenous Poisson process. An NNi < 1 indicates clustering, whereas an NNi > 1 indicates a tendency toward evenly spaced points (dispersion). A statistical test can be applied to check whether the NNi is significantly different from 1. The right-most column in Table 2 (M) reports for each data source the mean POI density and NNi from those three cities where POIs are available in all seven data sources (Albuquerque, Gainesville, London).

The highest mean POI density can be found for Instagram, which allowed users to add arbitrary place labels until 2015. As a consequence many POIs are incorrectly labeled, mislocated, duplicates, or stacked together due to positioning inaccuracies (Cvetojevic et al. 2016). Duplicate locations appeared to be at least partially cleaned out since October 2015, which is when Instagram was purchased by Facebook. Since then new Instagram places can be added through Facebook. We checked which of the Instagram places that we downloaded in October 2015 were still available in 2017, and stored these results in a newly added "available" POI attribute. Facebook POIs exhibit the second highest mean density. Also on that platform a frequent occurrence of stacked place labels at single locations poses a problem. Google shows the most consistent place density among all cities at approximately 100–200 POIs/km^2, suggesting that Google has access to high quality base data in different parts of the world. Twitter POIs demonstrate the lowest mean density, suggesting that the list of POIs is strictly controlled by the company. These POIs are only suitable for approximate geo-tagging of posted tweets, but less so for mapping and navigation purposes, which would require a more dense pattern of POIs. Foursquare venues reveal the second lowest place density, since they are limited to businesses (hotels, bars, bakeries), public buildings (city halls, university campuses and buildings, train stations), and public locations (parks, plazas) in the analyzed cities.

Yelp offers its service in five of the analyzed cities with a high variation of POI densities between cities, suggesting that different base datasets are used for this platform depending on the region.

The OSM POI density is higher in the two European than the three US cities. Compared to all other cities POI densities are lowest in Qingdao for the four data sources offered in that city, possibly due to a lower prominence of analyzed VGI and social media apps in that city. The analyzed area in London, which is a mixed business and residential district located between Hyde Park and Paddington

Table 2 Density and nearest neighbor index for POIs in analyzed cities

		Albuquerque	Cairns	Gainesville	London	Nairobi	Qingdao	Salzburg	$M_{A,G,L}$
Facebook	Dens.	244.5	530.8	247.9	1679.0	115.6	4.8	321.7	723.8
	NNi	0.45	0.57	0.51	0.58	0.44	0.66	0.53	0.51
Foursquare	Dens.	11.2	10.9	29.3	48.0	4.1	0.0	11.8	29.5
	NNi	0.71	0.68	0.64	0.54	0.72	*N/A*	0.62	0.63
Google	Dens.	117.6	121.2	114.7	165.5	136.6	99.0	185.8	132.6
	NNi	0.64	0.74	0.64	0.73	0.66	0.51	0.68	0.67
Instagram	Dens.	178.3	417.2	315.7	2753.6	260.1	8.1	431.6	1082.5
	NNi	0.62	0.59	0.60	0.72	0.59	0.57	0.57	0.65
OSM	Dens.	35.2	29.0	48.0	572.9	39.0	4.6	294.6	218.7
	NNi	0.70	0.68	0.74	0.75	0.55	0.68	0.72	0.73
Twitter	Dens.	5.3	0.0	7.1	25.4	0.0	0.0	0.0	12.6
	NNi	0.80	*N/A*	0.69	0.61	*N/A*	*N/A*	*N/A*	0.70
Yelp	Dens.	45.7	343.4	42.8	1160.4	0.0	0.0	708.8	416.3
	NNi	0.64	0.35	0.55	0.42	*N/A*	*N/A*	0.29	0.54

Railway station, shows the highest POI density in all data sources except for Google, which may have to do with frequent visitors in the area with its railway station and shopping streets.

NNi values < 1 in Table 2 indicate that POIs are clustered for all platforms and cities (where data is available). All clusters are significant at $p < 0.001$. Clustering can be expected since complete spatial randomness, though mathematically elegant, is often unrealistic in the physical world (O'Sullivan and Unwin 2010). Instead, point patterns typically display spatial dependence. It can either be modeled as first order effect (variation in the intensity of the process across space) or as second order effect (interaction of some kind between events). In the context of this work, a first order effect could be the tendency of restaurants and shops to be opened along selected roads of the built environment. Second order effects can result in clustering or dispersion. An example in the context of this study is the frequent co-occurrence of railway stations and restaurants (clustering), or the repulsion between public schools (dispersion).

The degree of clustering varies strongly between the datasets. Figure 4 maps for the analyzed area in Gainesville (enclosing polygon) the POIs for the most clustered (Facebook, NNi = 0.51) and the least clustered (OSM, NNi = 0.74) point patterns. The primary POI categories mapped in each platform (Fig. 2) can help to explain some of the differences in NNi values. Facebook POIs depict primarily business data, including shops, restaurants, retail and fitness clubs. A high density of such venues can be found along University Avenue (running East-West to the north), and on Community Plaza to the east. As opposed to this the OSM POI pattern shows heavy mapping activities on the UF campus (south-west portion of the map). With university buildings being further apart than businesses on a plaza or a shopping street, the clustering is less pronounced than in Facebook. Though the magnitude of the NNi for a specific data source itself is not a quality criterion, it can reveal differences in POI distributions between compared data sources.

In the OSM POI dataset, when averaging (unweighted mean) the geometry proportions across the seven analyzed cities, the share of POIs with node

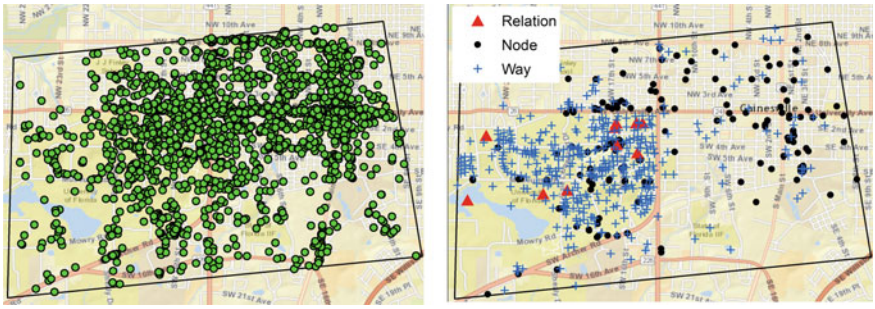

Fig. 4 POIs in Facebook (**a**), and OSM (**b**) for the Gainesville study area

Table 3 Number and proportion of OSM node features, way features, and relations

	Albuquerque	Cairns	Gainesville	London	Nairobi	Qingdao	Salzburg	Mean
Nodes	343	197	161	500	251	135	1743	
%	49.3	77.3	27.2	49.3	53.4	49.8	72.7	54.1
Ways	346	57	422	505	217	136	582	
%	49.7	22.4	71.2	49.8	46.2	50.2	24.3	44.8
Relations	7	1	10	9	2	0	74	
%	1.0	0.4	1.7	0.9	0.4	0.0	3.1	1.1

geometries is the highest (54.1%), followed by way features (44.8%) and relations (1.1%). However, considerable inter-urban variation in the use of geometries can be observed (Table 3). For example, Gainesville exhibits a high percentage of way features (71.2%) due to many UF campus buildings being mapped as closed polyline features (Fig. 4b). A high concentration of OSM way features on university campuses can also be observed for other cities, for example, around the University of New Mexico in Albuquerque.

Relation objects are generally more sparsely found in OSM maps. Salzburg stands out as an area with a relatively high proportion (3.1%) of relation objects. This can be attributed to the detailed mapping of plazas and historic buildings with court yards, which are mapped using inner and outer polygons as part of a relation (Fig. 5). In summary it can be stated that for OSM POI analysis both point and way objects should be considered since the ratio between those two geometry types varies significantly across cities.

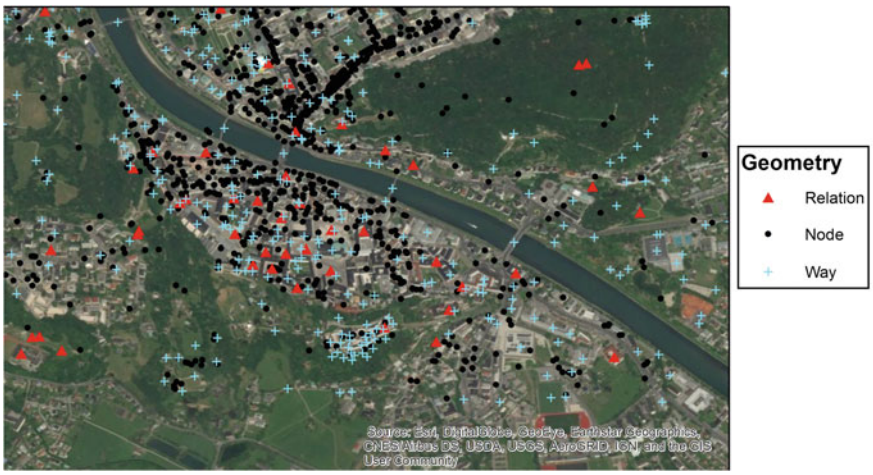

Fig. 5 OSM nodes, way centroids and relation centroids in the Salzburg study area

4.2 Relative Clustering of Point Patterns

While so far the level of clustering of POIs was discussed for individual platforms, statistical methods can be applied to analyze if and how POI locations from different platforms cluster relative to each other. For this purpose, a bivariate generalization of Ripley's K function, known as the Cross-K function (Dixon 2002), can be applied. The Cross-K function can be formulated as

$$K_{ij}(r) = \lambda_j^{-1} E[f(r)] \quad (1)$$

where E[f(r)] is the expected number of type j events within a distance r of a randomly chosen type i event, and λ_j is the density of j events per areal unit. If the two point patterns i and j are identical, the Cross-K function collapses to the self-K function K(r) which considers only locations of events but ignores information about the type of event.

Under random labeling, that is, assigning the n1 points from type 1 and n2 points from type 2 randomly to type 1 and type 2 events (keeping their original proportions), all four bivariate Cross-K functions should equal the K function, giving $K_{ii}(r) = K_{ij}(r) = K_{ji}(r) = K_{jj}(r) = K(r)$. Using place data from Gainesville as an example, this study analyzes the spatial relationship between all possible combinations of platform pairs, using events from different platforms as event types i and j in Eq. 1. Statistical inference of the difference between the observed Cross-K function and a Cross-K function generated by random labeling can be achieved through Monte Carlo simulation. Within each of the 99 completed permutations of the Monte Carlo simulation, the combined set of locations and the number of events of each type are held fixed. The labels (of the two platforms involved in the test) are randomly assigned to locations, which is followed by the computation of the Cross-K function. This establishes an upper and lower simulation envelope for random labeling at a 99% confidence level. If the observed Cross-K function falls within the simulation envelope, POIs from both platforms are similarly clustered around each other.

As an example for this analysis, Fig. 6 shows for all platform combinations in Gainesville the observed Cross-K function (black), the simulation mean from the Monte Carlo simulation (dashed red), and the 99% confidence envelope (gray area), for distances between 0 and 2000 meters. No significant attraction between two point patterns can be observed. While most pairwise platform point patterns are independent of each other, there are some platform combinations where the observed Cross-K function falls below the lower simulation envelope. This is clearly the case for some platform combinations that involve Foursquare and OSM. Whereas OSM is primarily contributed around the UF campus, business related contributions to Yelp and Facebook cluster around shopping strips and plazas (compare Fig. 4). Hence events from the platform pairs Foursquare-Yelp, OSM-Yelp, OSM-Facebook, and Foursquare-Facebook are spatially segregated (see Fig. 6). Google places are evenly distributed across the study area with no

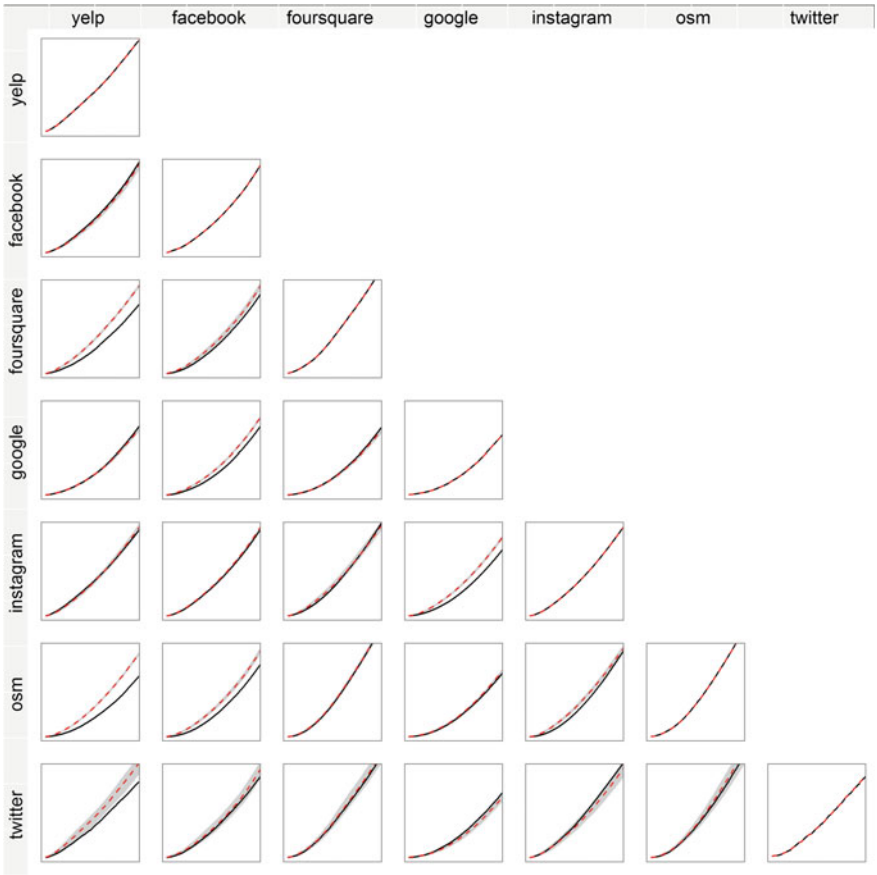

Fig. 6 Cross-K functions for Gainesville with 99% confidence envelopes

apparent clusters on the UF campus or in shopping areas. This makes it spatially segregated from Facebook (businesses) and Instagram POI locations (clustered around event places like student centers, campus food courts, hospitals, restaurant areas, market plazas). Overall, these findings suggest that the urban structure of the analyzed area is reflected in Cross-K functional patterns, and that a conflation of POIs from different sources can lead to improved data coverage for that city.

5 Positional Accuracy

All platforms analyzed in this study provide map interfaces and/or address search functions for manually adding new POIs. Provided that a user possesses basic map reading capabilities, such an approach supports accurate mapping of POIs. In

addition to this, mobile apps show a user's current position, which can also be used for adding a POI if the user is located directly at the new venue. It is also possible to add coordinates to a picture in different photo sharing platforms, e.g. in Flickr. In such cases it is encouraged to map the photographer's position, and not that of the photographed object (Zielstra and Hochmair 2013). However, guidelines for the analyzed platforms in this study encourage users to add the true POI position of a venue, and not that of the photographer's position (if any pictures are involved at all). An exception are older Instagram place locations (before October 2015) which could be added by users and were often placed at a photographer's position or biased by the physical location from which the image was uploaded to the platform (Cvetojevic et al. 2016).

Figure 7 provides a visual impression of the positional accuracy of POIs in analyzed platforms, using the Salzburg downtown area as an example. Twitter POIs are not shown since they are not available for Salzburg. No POIs should be located within the Salzach river, with a few potential exceptions, such as a ferry service, a river place label, or the city label. None of the POIs from Google, OSM, and Yelp are located in the river, indicating good positional accuracy. The Facebook map reveals a few POIs to be incorrectly placed in the river, which include a barber, a shopping strip, and a graphic design business. Foursquare uses a review of added locations through super users to verify locations and to increase data reliability. A Boolean attribute in the point data set ("verified") indicates whether a POI underwent such a check or not. Only a small percentage of Foursquare venues is actually marked as verified. In Salzburg this is true for 96 out of 2012 features (4.8%), and for all seven analyzed cities this rate is slightly higher with 910 out of 13040 (7.0%). This filter process is clearly discernible in Fig. 7, where only few verified points (light green) appear on top of non-verified points (orange). Whereas some POIs of the unfiltered Foursquare dataset are incorrectly placed in the Salzach river (e.g. old city hall, a snack stand, a person's name, a coffee shop), no point in the filtered dataset is. This indicates that the revision through super users has a positive effect on the positional accuracy of Foursquare POIs.

The Instagram POI file used for this study was downloaded in October 2015. Next, using the Instagram API it was verified if a POI was still available in 2017. This information was then coded in a Boolean "available" attribute. POIs not available any more were possibly cleaned due to inspection after Instagram has been purchased through Facebook. Using this attribute, yellow dots in the Instagram map indicate possibly reviewed (and retained) POIs, whereas the red dots show the locations of the remaining unverified POIs from the original dataset. The map suggests that a significant percentage of Instagram locations was removed since 2015, namely 27.3% of POI for Salzburg and 27.5% for all seven analyzed cities. Several POIs that are incorrectly placed in the river disappear when considering only "available" Instagram features, including a bus stop, the old city hall, a pub, and a road. However, remaining POIs in the river (yellow dots) include bars, restaurants, or shops, still revealing POI accuracy problems.

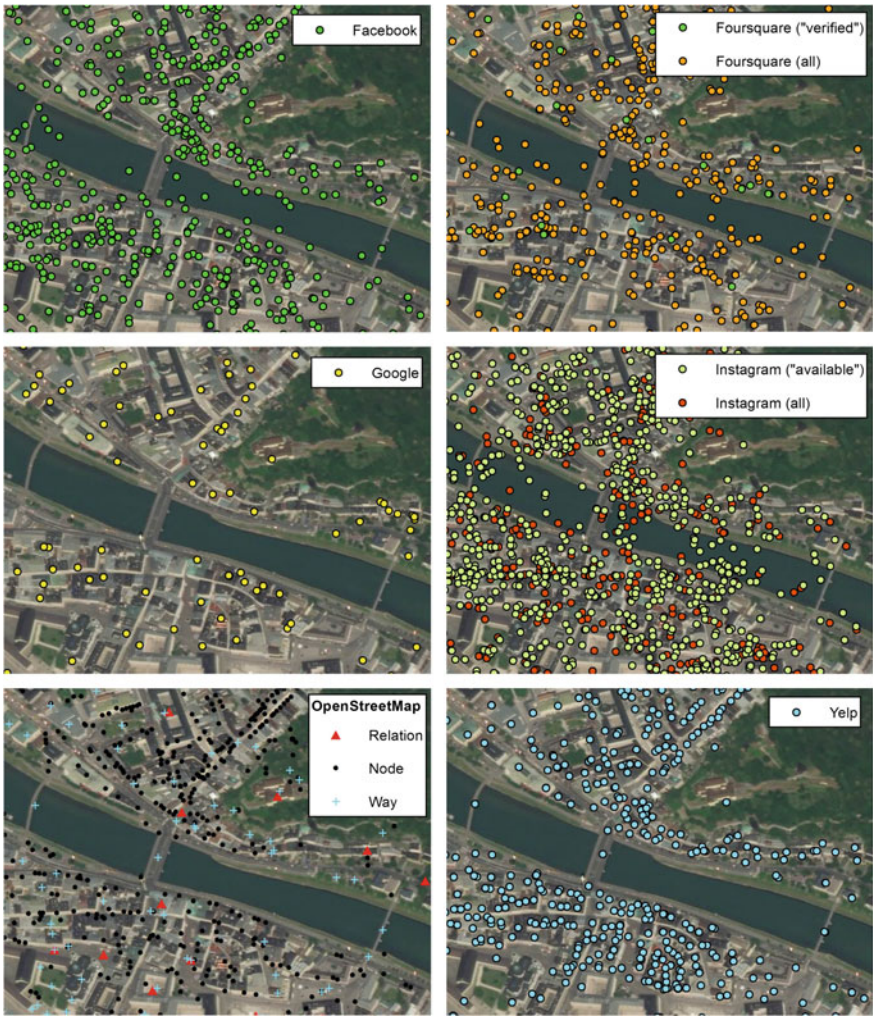

Fig. 7 Location of places features for different platforms in the Salzburg downtown area

To quantify the distance offsets between mapped POIs (based on coordinates provided on the platforms) and their true location, a sample of place points was selected for each of the available six data sources from the Salzburg downtown area. Using the name tag of a selected POI and the authors' local knowledge of the area, the corresponding true location was identified (if possible) and the offset computed. If a place was mapped on the street directly in front of the correct building, it was counted as correct as well and assigned an offset distance of 0. Seasonal POIs (e.g. a Christmas market) were also taken into account for the

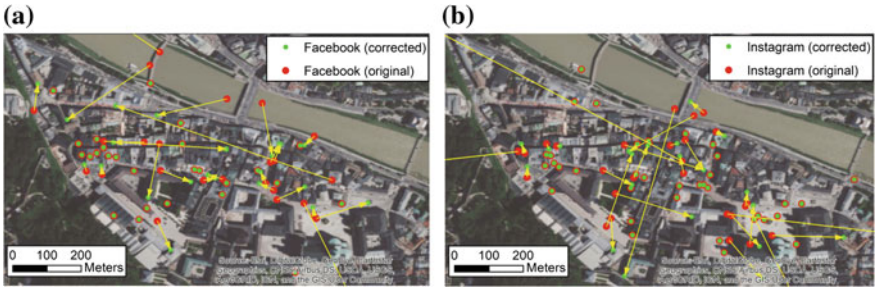

Fig. 8 Offset vectors for POIs in Salzburg for Facebook (**a**) and Instagram (**b**)

analysis. As an example, Fig. 8 shows for the samples of Facebook POIs (a) and Instagram POIs (b) their original positions (red) and corrected positions (green).

Yellow arrows denote the offset vectors from the published to the corrected position. For Facebook 24 POIs (43.6%) had to be corrected with offsets ranging up to 25,514 m. For Instagram this was the case for 26 POIs (40.6%) with a maximum offset of 715.4 km. Instagram has four outliers with an offset >10 km, whereas Facebook has only one.

Figure 9 plots the histograms of offset distances (in meters) for the six evaluated platforms in downtown Salzburg, using a logarithmic scale on the x-axis. The median distance offset is zero for all platforms, which means that at least half of all POIs in each platform is correctly placed. Google and OSM sample POIs do not reveal any positional errors, closely followed by Yelp which has moderate offsets in 12.7% of the cases. This finding suggests that, at least for the chosen test site, mapping platforms (Google, OSM) and business platforms with strong quality control (Yelp) provide most reliable POI positions. This finding is in-line with the distribution of point patterns observed in Fig. 7. Foursquare (verified POIs only) has the next smallest error rate (32.5%), followed by Instagram (40.6%) and Facebook (43.6%). Besides higher error rates, the latter two platforms are also the only ones with positional errors of over 10 km. Hence these two social media platforms together with an unverified (complete) Foursquare POI dataset achieve a lower POI reliability than other analyzed platforms. This indicates that social media platforms with little to no quality control through the governing company perform poorly in terms of positional accuracy when compared to mapping platforms or platforms that implemented stricter quality control measures (i.e. approval by moderators). To be able to provide more generalizable conclusions, however, offset measures would have to be expanded to other cities as well.

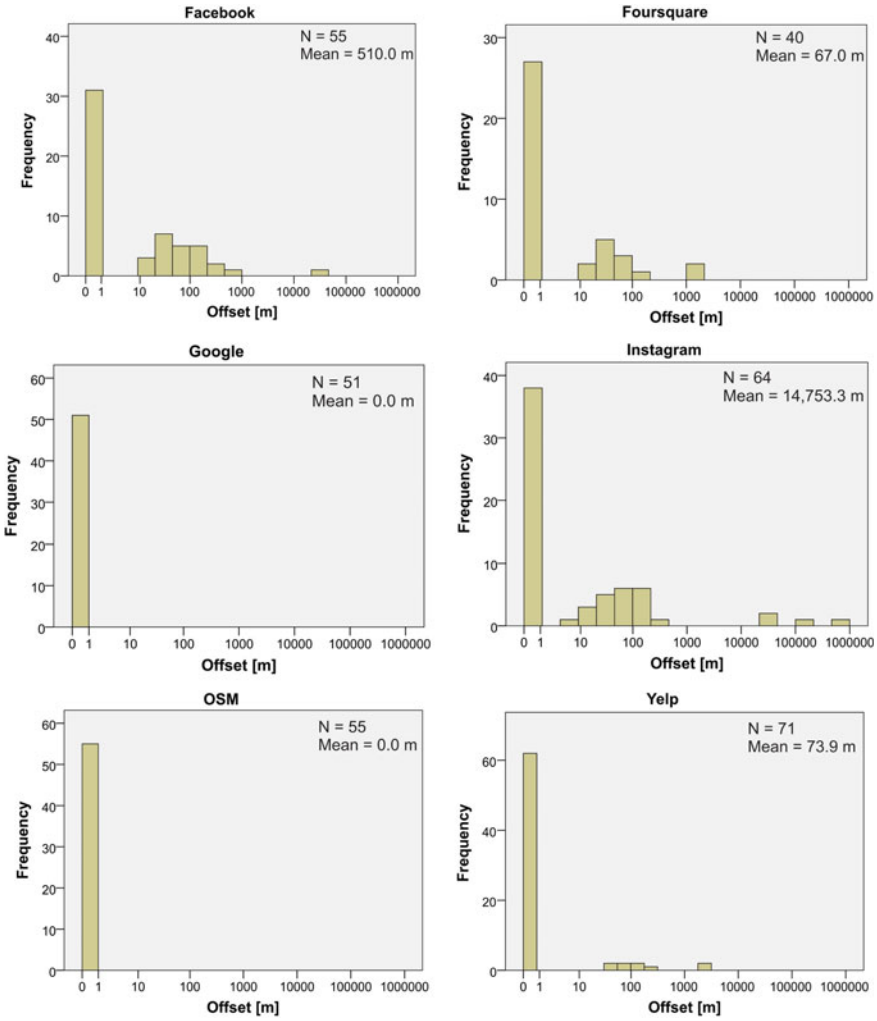

Fig. 9 Histograms of offset distances for evaluated features in downtown Salzburg

6 Discussion and Conclusions

The study examined various aspects of data quality and POI clustering for seven mapping and social media platforms in seven study sites across the world. The findings provide information to help determine the suitability of a given POI source for an intended geo-application, such as an LBS. Various quality metrics (e.g. nearest neighbor index, density, bulk uploads, spatial offsets) were compared between different platforms in the absence of ground truth data. Even with a correct a POI reference dataset available, POIs from the different platforms would first have

to be matched to reference features in order to determine certain quality measures (e.g. categorization). An example for such detailed analysis that considers both feature location and attributes for manual feature matching (e.g. schools) in order to determine the relative completeness of data sources is presented in (Jackson et al. 2013).

The findings of this study can be summarized as follows:

- POIs are more abundant in selected social media platforms (Facebook, Foursquare, Instagram) than in mapping and business oriented platforms with strict quality control (Google, OSM, Yelp). Twitter is an exception with the lowest POIs density in all areas (where present), and a lack of POI categories.
- Mapped POIs of the three social media platforms (Facebook, Foursquare, Instagram) show higher mean offsets from their true locations than the three map/business related platforms, based on a Salzburg sample analysis.
- Presence of erroneous POI stacks uploaded to the same point location is primarily a problem of Facebook and Instagram and was observed in all cities.
- Different platforms map different POI categories as the most prominent ones. Therefore conflation of POIs from different platforms could improve POI completeness.
- The level of POI clustering, as determined by the nearest neighbor index, differs between platforms, reflecting a different topical focus of platforms.
- Cross-K functions for marked point patterns showed that point patterns cluster sometimes differently between pairs of platforms, which means that POIs are spatially segregated. This reflects also different types of POIs mapped in compared platforms.

Aspects of future work include consideration of other quality measures, including errors of omission and commission in selected test areas, and a closer examination of POI contribution patterns of users across different crowd-sourced platforms (Juhász and Hochmair 2016).

References

Barron C, Neis P, Zipf A (2014) A comprehensive framework for intrinsic OpenStreetMap quality analysis. Trans GIS 18(6):877–895

Cvetojevic S, Juhász L, Hochmair HH (2016) Positional accuracy of twitter and instagram images in urban environments. GI_Forum 1:191–203

Dixon PM (2002) Ripley's K function. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics. Wiley, Chichester

Duckham M, Winter S, Robinson M (2010) Including landmarks in routing instructions. J Location Based Serv 4(1):28–52

Fan H, Zipf A, Fu Q, Neis P (2014) Quality assessment for building footprints data on OpenStreetMap. Int J Geogr Inf Sci 28(14):700–719

Gröchenig S, Brunauer R, Rehrl K (2014) Estimating completeness of VGI datasets by analyzing community activity over time periods. In: Huerta J, Schade S, Granell C (eds) Connecting a

digital Europe through location and place. Lecture notes in geoinformation and cartography. Springer, Berlin, pp 3–18
Hastings JT (2008) Automated conflation of digital gazetteer data. Int J Geogr Inf Sci 22 (10):1109–1127
Heipke C (2010) Crowdsourcing geospatial data. ISPRS J Photogramm Remote Sens 65:550–557
Hochmair HH, Zielstra D (2013) Development and completeness of points of interest in free and proprietary data sets: a Florida case study. In: Jekel T, Car A, Strobl J, Griesebner G (eds), GI_Forum 2013. Creating the GISociety. Wichmann, Berlin, pp 39–48
Jackson SP, Mullen W, Agouris P, Crooks A, Croitoru A, Stefanidis A (2013) Assessing completeness and spatial error of features in volunteered geographic information. ISPRS Int J Geo-Inf 2:507–530
Juhász L, Hochmair HH (2016) Cross-linkage between Mapillary street level photos and OSM edits. In: Sarjakoski T, Santos MY, Sarjakoski T (eds) Geospatial data in a changing world: selected papers of the 19th AGILE conference on geographic information science. Lecture notes in geoinformation and cartography. Springer, Berlin, pp 141–156
Juhász L, Hochmair HH (2017) Where to catch 'em all?'—a geographic analysis of Pokémon Go locations. Geo-spat Inf Sci 20(3):241–251
Juhász L, Rousell A, Arsanjani JJ (2016) Technical guidelines to extract and analyze VGI from different platforms. Data 1(3):15
Li L, Xing X, Xia H, Huang X (2016) Entropy-weighted instance matching between different sourcing points of interest. Entropy 18(2):45
Lim KH, Chan J, Leckie C, Karunasekera S (2015) Personalized tour recommendation based on user interests and points of interest visit durations. In: 24th international joint conference on artificial intelligence (IJCAI 2015), Buenos Aires, Brazil
McKenzie G, Janowicz K, Adams B (2014) A weighted multi-attribute method for matching user-generated points of interest. Cartogr Geogr Inf Sci 41(2):125–137
Mülligann C, Janowicz K, Ye M, Lee W-C (2011) Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In: Egenhofer MJ, Giudice NA, Moratz R, Worboys MF (eds) Conference on spatial information theory (COSIT 2011). LNCS 6899. Springer, Berlin, pp 350–370
Nothegger C, Winter S, Raubal M (2004) Computation of the salience of features. Spat Cogn Comput 4:113–136
O'Sullivan D, Unwin DJ (2010) Geographic information analysis, 2nd edn. Wiley, Hoboken, New Jersey
Rösler R, Liebig T (2013) Using data from location based social networks for urban activity clustering. In: Vandenbroucke D, Bucher B, Crompvoets J (eds) Geographic information science at the heart of Europe. Lecture notes in geoinformation and cartography. Springer, Berlin
Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay M (2017) A review of volunteered geographic information quality assessment methods. Int J Geogr Inf Sci 31(1):138–167
Zhao B, Sui DZ (2017) True lies in geospatial big data: detecting location spoofing in social media. Ann GIS 23(1):1–14
Zielstra D, Hochmair HH (2013) Positional accuracy analysis of Flickr and Panoramio images for selected world regions. J Spat Sci 58(2):251–273
Zielstra D, Hochmair HH, Neis P (2013) Assessing the effect of data imports on the completeness of OpenStreetMap—A United States case study. Trans GIS 17(3):315–334