DOI: 10.1111/tgis.12454

WILEY Transactions in GIS

RESEARCH ARTICLE

Modeling aggregated expertise of user contributions to assess the credibility of OpenStreetMap features

Bani Idham Muttaqien[1] | Frank O. Ostermann[2] | Rob L. G. Lemmens[2]

[1] Jakarta 12920, Indonesia

[2] Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, The Netherlands

Correspondence

Frank O. Ostermann, Faculty of Geo-Information Science and Earth Observation, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
Email: f.o.ostermann@utwente.nl

Funding information
Graduate scholarship from Indonesia Endowment Fund for Education (LPDP), Republic of Indonesia

[Correction added on 17 October 2018, after first online publication: the affiliation address of author Bani Idham Muttaqien was updated]

Abstract

The emergence of volunteered geographic information (VGI) during the past decade has fueled a wide range of research and applications. The assessment of VGI quality and fitness-of-use is still a challenge because of the non-standardized and crowdsourced data collection process, as well as the unknown skill and motivation of the contributors. However, the frequent approach of assessing VGI quality against external data sources using ISO quality standard measures is problematic because of a frequent lack of available external (reference) data, and because for certain types of features, VGI might be more up-to-date than the reference data. Therefore, a VGI-intrinsic measure of quality is highly desirable. This study proposes such an intrinsic measure of quality by developing the concept of aggregated expertise based on the characteristics of a feature's contributors. The article further operationalizes this concept and examines its feasibility through a case study using OpenStreetMap (OSM). The comparison of model OSM feature quality with information from a field survey demonstrates the successful implementation of this novel approach.

1 | INTRODUCTION

During the past decade, the role of a typical internet user has expanded from only consuming information to also contributing content. This development is driven by user preferences and interests, and is enabled by recent technologies that are commonly referred to as Web 2.0. Much of this user-generated content includes or refers to geographic features, and has been termed "volunteered geographic information" (VGI) (Goodchild, 2007).

Presently, VGI already serves as a data source for a broad range of applications, such as disaster management, environmental monitoring, and urban planning, with OpenStreetMap (OSM) being a prime example of VGI (Capineri et al., 2016). Assessing the quality and fitness-for-purpose of VGI is still an ongoing area of research, as there are no standards and specifications for a crowdsourced and volunteered data creation process. Several studies have evaluated different quality elements of VGI. For example, Haklay (2010) calculated the overlap of motorway features between OSM and Ordnance Survey to assess the positional accuracy and completeness of OSM, while Neis, Zielstra, and Zipf (2011) examined the completeness and topological error of the OSM street network by comparing it with commercial datasets.

Those methods require reference datasets for comparison purposes that should ideally have the same geographical coverage and a similar data model. However, such extrinsic comparisons have limitations because of the absence of a comparable global authoritative dataset, prohibitive procurement costs, and licensing restrictions (Barron, Neis, & Zipf, 2014). Antoniou and Skopeliti (2015) have argued that VGI and authoritative data are not comparable due to their different data acquisition processes. Additionally, an authoritative dataset does not necessarily indicate better data; for example, OSM now has better coverage in some areas than authoritative datasets (Vandecasteele & Devillers, 2015).

Other studies explored VGI quality without comparing it to reference data, so-called "intrinsic quality analysis" (Mooney & Corcoran, 2012; Keßler & de Groot, 2013; Yang, Fan, & Jing, 2016). However, current methods for assessing the quality of VGI do not identify and combine the parameters from VGI metadata and VGI contributors. Whereas contributor expertise is an important factor for producing high-quality authoritative datasets, crowdsourced projects also rely on contributors cross-validating and correcting each other's work. Duly, we aim to create an intrinsic measure for VGI quality that combines contributor expertise in a new metric we call "aggregated expertise." The contributor profile characteristics serve as a proxy for contributor expertise, while a feature's edit history serves to aggregate the expertise. OSM is used for this case study because it is the most utilized, analyzed, and cited VGI platform (Neis & Zielstra, 2014). This article develops and implements a model based on aggregated expertise to assess the credibility of OSM features. We analyze and compare the model outputs with the results of a field survey in a neighborhood in Jakarta, Indonesia. The article addresses therefore two main research questions: How can we operationalize the concept of aggregated expertise for the purpose of assessing the credibility of collaborative contributions? How well can aggregated expertise predict the quality of OSM features in our case study area?

Section 2 of this article provides a review of related work. Section 3 introduces the concept of aggregated expertise and its operationalization, the field survey setup, and the analysis methods. In Section 4, we present and discuss the results and reflect on the limitations of model implementation in the study area. Section 5 summarizes the key outcomes of the study and provides recommendations for future work.

2 | RELATED WORK ON INTRINSIC VGI QUALITY MEASURES

2.1 | Credibility

Credibility as a quality measure has been studied in various disciplines such as communication, social sciences, and computer sciences, and its definition varies accordingly. However, credibility can be understood in general as *a perceived quality made up of multiple dimensions such as trustworthiness and expertise* (Fogg & Tseng, 1999).

The concept of credibility for VGI proposed by Flanagin and Metzger (2008) defines credibility as the believability of the source or message. In the absence of direct measures, the authors argue that determining the credibility of information becomes critical as people involved in mapping processes have different knowledge, motivations, and skills. VGI credibility comprises two primary dimensions: the contributor's expertise on the information content, and the trustworthiness of the source of information (which can be the contributor of some other source). Thus, credibility is usually considered to combine at least some degree of both expertise and trustworthiness. However, we need to consider that these two elements have both objective and subjective components.

2.2 | Trustworthiness

Artz and Gil (2007) argued that provenance information is a key factor to measure trust on the web. Keßler, Trame, and Kauppinen (2011) propose a model to assess VGI quality based on its provenance. They created a specific vocabulary of provenance information of OSM and used it to evaluate a trust dimension. Mooney and Corcoran (2012) focused their analyses on OSM features which have been edited over 15 times and argued that the edit history of OSM data is the baseline for evaluating its quality. Finally, Keßler and de Groot (2013) investigated indicators influencing the trust derived from data provenance. However, this approach could not predict the trustworthiness of features with fewer than six versions after editing. Further, the authors state that their implemented approach is data-oriented only and suggest addressing user reputation for future work.

Bishr and Kuhn (2007) argued that *interpersonal trust* is a subjective measure but cannot directly be adopted in a geospatial context since trust in social networks has an unknown relation with the spatial and temporal dimensions of trust. In terms of VGI, informational trust can be related to the concept of *people–object transitivity*, where the trust between VGI consumer and VGI creator is mediated by the information produced (Bishr & Janowicz, 2010). We argue that this highly subjective and situational nature of personal trust puts it outside the scope of this study: we trust a person if we think we know his/her motivation well and are sure that they do not want to cause us any harm. Further, trust is transitive (i.e. if a trusted friend relays an information request to a third party, we are more likely to trust that third party). On the other hand, a statement from someone we do not explicitly trust might still be credible if that person has much expertise, and other potentially credible sources agree or endorse that information, or at the least if no one disagrees. The difficulty in modeling these complex interactions lets us focus on expertise instead.

2.3 | Expertise

Another dimension of credibility is defined by the perceived knowledge and skill of the information provider (i.e. expertise) (Fogg & Tseng, 1999), assuming that expertise implies a motivation for providing objective information.

Bégin, Devillers, and Roche (2013) studied VGI contributors' mapping behavior to assess the completeness of VGI. They concluded that analyzing the mapping processes of only a few contributors could represent most datasets. Budhathoki and Haythornthwaite (2013) introduced as measures the number of features edited, contributing days, and longevity of contributions as indicators of contributors' expertise, using questionnaires to investigate the motivations and characteristics of OSM contributions. Van Exel, Dias, and Fruijtier (2010) discuss that local knowledge helps to determine quality, where local familiarity correlates with quality of contribution. Local knowledge helps the user to identify incorrect information. Research on geographic crowdsourcing processes also considered the information contributed by local Wikipedia editors to be more credible (Sen et al., 2015). Yang et al. (2016) identified evidence for expertise using a behavior-based approach, using three indicators to assess the expertise of major contributors: practice, skill, and motivations. Practice represents the number of user efforts to dedicate their time to the OSM project. Skill indicates how well a user is making his/her contribution. Motivation shows the willingness and persistence of the contributor. However, they only distinguished two classes of expertise: professional and amateur, and assessed only the major contributors using descriptive statistics.

To summarize, existing studies show a research gap in combining all available information. Thus, this study suggests a more systematic approach: the extension of existing research on VGI contributors by examining the contributing behaviors together with the quality of information being produced.

3 | MEASURES AND METHODS

3.1 | Aggregated expertise

In a collaborative environment with multiple contributors per feature, we argue that expertise cannot be modeled independently from the feature and introduce the concept of "aggregated expertise." In a nutshell, it models that the more expert contributors acknowledge a certain piece of geographic information, the more credible it is perceived to be.

Figure 1 illustrates how we aggregate the expertise level of each contributor by taking advantage of a versioning system for the stored data. Aggregated expertise uses the basic parameters of collective edits: number of versions created and number of users involved, adjusting those parameters using the properties of the contributors to model their skill, experience, and local knowledge.

Every contributor *u* has a different level of expertise *E(u)*, indicated by parameters *p(i)* such as mapping days, longevity, number of edits, software skill, and distance to the edited feature. We normalize those parameters, combining them in a weighted sum as the integrated expertise score *E(u)* of the VGI contributors, with *w(i)* being the weight for expertise parameters:

$$E(u)=\sum_{i=1}^{N} w(i)\cdot p(i) \tag{1}$$

Each contributor's edits to a feature result in a version *f(v)* of that feature. Types of editing are creation, confirmation, geometric correction, tag correction, main tag correction, rollback, tag addition, tag removals, or a combination. These editing types are indicated by *et(u)*. Consequently, the aggregation of different expertise levels

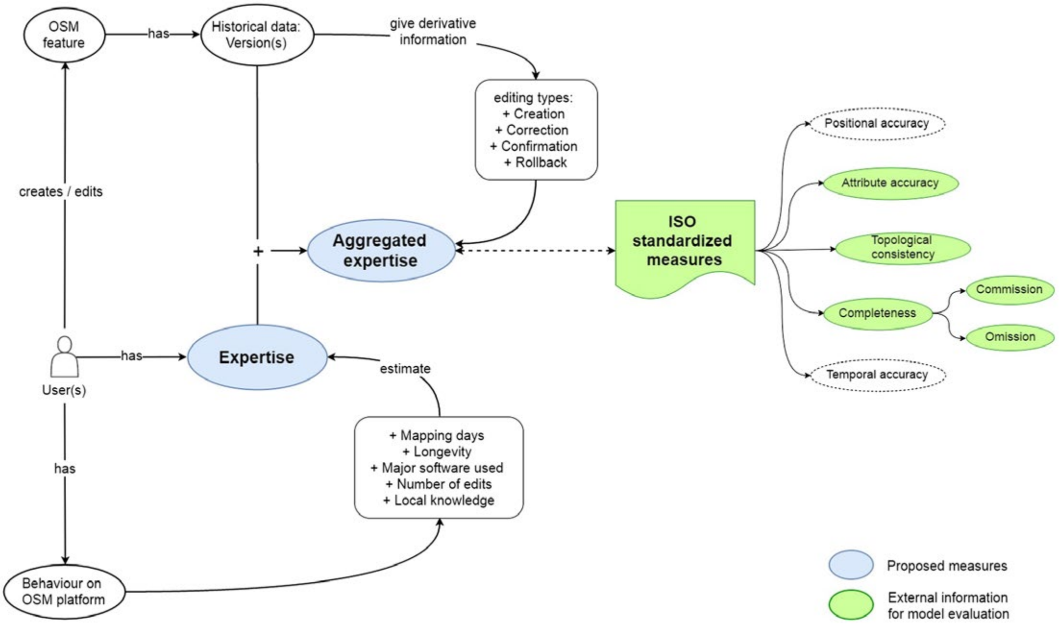

FIGURE 1 Aggregated expertise to assess the credibility of VGI

with the editing pattern of the contributors determines the credibility of the feature, since it accumulates the proficiency of the involved contributors. This aggregated expertise is denoted as *AE* and defined as the sum of contributor expertise weighted by edit type:

$$AE\ (f\,i\,) = \sum_{i=1}^{v} et(u)\ E(u) \tag{2}$$

where *et(u)* is the weight for the type of editing contributed by user*u*, and *v* is the number of versions for feature*fi*.

3.2 | Acquiring the OSM data

Weekly updated OSM data is available on the website of Planet OSM (https://planet.openstreetmap.org/) for the entire world, with a current compressed size of 38 GB under a Creative Commons Attribution-ShareAlike 2.0 license. It contains not only up-to-date data, but also historical data as well as deleted features. However, its uncompressed size makes processing problematic on desktop workstations. Smaller extracts of specific regions are offered by Geofabrik (https://download.geofabrik.de/) in OSM XML and Shapefile formats. Another alternative for a particular area of interest is to use the JOSM editor to retrieve an OSM XML file. Unfortunately, these OSM XML files contain none of the historical data and only contain information on the last edit. Additionally, they could include errors as a result of a bug in the Potlatch 1 OSM editor (up to 2011) that led to an invalid increase of the feature's version number (Barron et al., 2014).

To address these issues, we developed a semi-automated data acquisition workflow to derive all required information. Figure 2 shows the workflow of obtaining all OSM historical data and involved contributors as well as their profiles.

First, the dataset of the study area in Jakarta was obtained on October 3, 2016 in *.osm XML format using JOSM. The complete OSM XML file contains an extensive collection of nodes, ways, relations, as well as their associated tags and common attributes, but only user, version, timestamp, and changeset for the last version of the feature.

The OSM feature IDs found in the OSM XML file form the input to extract the historical data using a Python script (see Supporting Information) that downloads all versions that ever existed for all input feature IDs directly, relying on OSM history API v0.6 (OpenStreetMap Wiki, 2016), where <objtype> could be "node" or "way" and <id> is the feature id of a particular object (https://api.openstreetmap.org/api/0.6/<objtype>/<id>/history). The script checks the recurring changeset id, indicating errors resulting from the editor tool's bug, and deriving the true number of versions. Also, a list of contributors, timestamps, and changesets for all versions is included in the output file as comma-separated values.

Information about contributors' profiles is available from the OSM platform using third-party web-based tools provided by Neis (2012), which show in detail when, where, and how long the OSM contributors participated in the project as well as the tools they used (https://hdyc.neis-one.org/?Steve). To extract this data automatically for a given list of contributors, we wrote another Python script. The output file contains all essential information, indicating the profile for all input users.

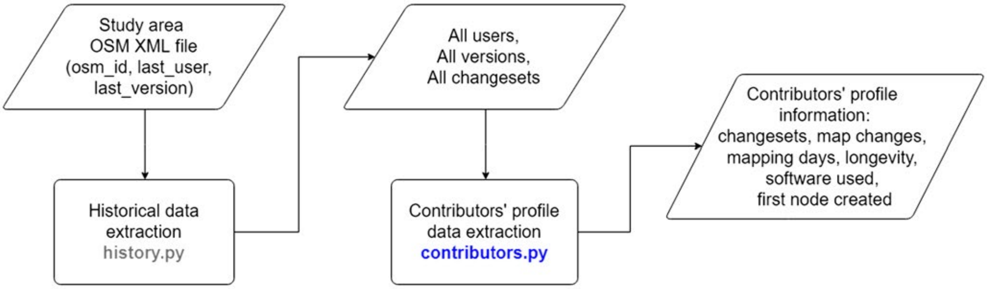

FIGURE 2 Workflow to obtain historical OSM data and OSM contributors' profile information

3.3 | Checking the ground truth

Although our aim is to develop a method for assessing VGI quality without extrinsic reference datasets, to determine the performance of the proposed approach we need to compare the model outputs with a ground truth. To collect such ground-truth data, we conducted a field survey to determine whether the OSM features correspond with reality. The field study is a 1km^2 area of a touristic neighborhood in Jakarta that has a high density of points of interest (POIs), roads, and buildings. As a result, the data produced from this field survey is expected to be of high quality and can be used as the reference data to validate the model output. Overall, it took 3 person-days (from October 12 to 14, 2016) to collect information on 398 OSM features in the field. We have provided the entire field survey dataset as supplementary material (see Supporting Information).

Not all common ISO spatial data quality measures are feasible to use for our model validation. It needs considerable effort to measure the positional accuracy of OSM data in the field, and a low positional accuracy does not make the object itself necessarily less credible. It was not practical to measure temporal accuracy, and it seems most relevant and useful to check current VGI. The remaining three ISO quality measures of attribute accuracy, completeness, and topological consistency are suitable and feasible to measure in the field.

To summarize the quality of OSM features, we calculate a feature quality score (for details see Table 1) using up to three input variables for each of the three quality measures. The three quality measures are weighted individually to calculate the feature quality score. The weights reflect our aim to create a useful classification that does not suggest wrong precision through too many classes. They were chosen after careful consideration of this particular case study, and might need adaptation for different contexts, as we do not assume that there is a universally acceptable threshold for "credible enough."

TABLE 1 Methods to assign the feature quality score

Quality measures	Variables	Score	Weight
Attribute accuracy	correctness of feature type (Q_1)	if feature type is correct: $Q_1 = 1$ if not: $Q_1 = 0$	0.35
	correctness of feature name (Q_2)	if feature name is correct: $Q_2 = 1$ if not: $Q_2 = 0$	0.35
	correctness of other tags information (Q_3)	Q_3 = number of correct tags/ number of tags	0.3
Completeness	tag completeness (Q_4)	Q_4 = number of tags/number of total available tags	0.5
	commission (Q_5)	if feature exist in reality: $Q_5 = 1$ if not: $Q_5 = 0$	0.5
Topological consistency	topological relationship between data (Q_6)	if building overlapped: $Q_6 = 0$ if line segment contains overshoot or undershoot: $Q_6 = 0$ if POIs do not fall inside the associated object: $Q_6 = 0.5$ Otherwise: $Q_6 = 1$	1

3.3.1 | Attribute accuracy

The field survey mainly checked the correctness of the main tag value for each defined tag key. Additional tag values were also investigated, such as the name of the road, the name of buildings, building levels, and so on. During the field survey, we used our in-field local knowledge to assess these data. Consequently, the attributes' correctness was calculated as a ratio of the correct information to the total information provided:

$$\mathit{Attribute\ accuracy} = 0.35 \times Q_1 + 0.35 \times Q_2 + 0.3 \times Q_3 \tag{3}$$

where the feature type and feature name have 70% weight as these form the main attribute information.

3.3.2 | Completeness

To measure any data absent from a dataset (*error of omission*), attribute completeness was defined as the ratio of number of tags provided to the total available tags for each map feature type. The highest score for attribute completeness is 0.5. Similarly, excess data present in a dataset (*error of commission*) was checked in the field (e.g. informal buildings that have been demolished due to building infractions). If the existence of an OSM feature in reality corresponds to the OSM database, then an additional score of 0.5 is assigned. The completeness for each OSM feature is therefore calculated as follows:

$$\mathit{Completeness} = 0.5 \times Q_4 + 0.5 \times Q_5 \tag{4}$$

3.3.3 | Topological consistency

While the positional accuracy of the OSM feature is less important for most use cases, the topological relationship between data must be correct. Point features in the dataset should be located within the footprint of the features they refer to. The analysis for this feature type was done by visually interpreting the data. The topological inconsistency of polyline feature such as roads can be seen at the road junctions. For polygon features, such as buildings, one can be categorized as topologically inconsistent when there is an overlap between the features. The JOSM validator tool was employed to automatically check the polyline and polygon data type for suspected mistakes. Topological consistency scores were given following the criteria shown in Table 1.

3.4 | Statistical analysis

We first examined the association between the variables that contribute to the aggregated expertise and the feature quality score derived from the field survey. Because these variables are not normally distributed, we used a non-parametric test for which all the variables were converted to ranks (where high ranks mean high scores and low ranks mean low scores). We chose Kendall's tau correlation (τ) that measures the statistical dependence between the ranks of two variables, since the dataset has a large number of tied ranks and many scores have the same rank.

Further, we attempted to assess feature credibility by predicting the observed feature quality from the aggregated expertise (see Section 3.1). Given the non-linearity of the feature quality score (see Section 3.3), we categorized the features into three quality classes based on the attribute accuracy, completeness, and topological consistency scores, and used a multinomial regression model to predict the quality class. The classification rules are as follows.

- Correct: Attribute accuracy = 1 AND Completeness = 1 AND Topological consistency = 1
- Partially correct: 1 > Attribute accuracy >= 0.7 AND 1> Completeness >= 0.5 AND Topological consistency >= 0.5
- Incorrect: Attribute accuracy < 0.7 OR Completeness < 0.5 OR Topological consistency < 0.5

4 | RESULTS

4.1 | Data analysis

4.1.1 | Versions

Our script identifies any duplications introduced through editor software bugs and returns the true version number, which is crucial to judge how many times a particular feature has been modified. In total, the collaborative edits on the OSM platform in the study area resulted in 737 unique versions for 398 OSM features. These feature versions are further analyzed to reveal information about the contributors, kind of edits, and recency of edits.

4.1.2 | Users

Features with only one version have only one user involved. For features with more than one version, the number of users participating in the editing process often differs from the number of versions created. The same user could edit the same feature several times for different versions. This case indicates that a specific user might have a special concern in the development of that particular OSM feature. Further analysis of user metadata is described in Section 4.2.

4.1.3 | Recency of data (temporal effect)

In principle, in a collaborative effort that relies on Linus's law for spotting errors, the longer a feature has remained unchanged over time, the higher the probability that this particular feature corresponds to reality. However, since we cannot count how many times other contributors have looked at that feature and tacitly validated it by not changing it, we approximate this information with the type of feature, assuming that some features have a longer life cycle than others (e.g. religious buildings only rarely change, while informal settlements or restaurants might). Instead of solely looking at the general type (tag key), one should consider the specific type of OSM features by examining the tag value. However, this specific type categorization has limitations due to the small sample size for each. Thus we combine them, assuming similar behavior regarding a feature's constancy over time as shown in Table 2.

This feature type will be used for further correlation analysis to expose the specific behavior of feature type and persistence conditions over time. For example, we would expect that a long-lived object such as a place of worship or historical building has an extended number of days since the last edit compared with informal buildings.

4.2 | Characteristics of the contributors

In total, there are 52 contributors involved in the editing process in the study area. About 75% of the data creation processes of OSM features in the study area are done by only 8% of all contributors, while 77% of data modifications are performed by only 21% of contributors. About 50% of all contributors are the creators of the first version of all features in the study area, while 88% of the contributors are involved in modification processes resulting in a newer version.

More detailed profile information about those OSM contributors, regarding their activity and experience on the OSM platform, is outlined in Table 3. The table reviews the descriptive statistics for the derived expertise indicators in terms of practice, skill, and local knowledge. For all parameters, the median and mean values differ considerably, indicating that the data is not normally distributed.

4.2.1 | Number of mapping days

The number of contributing days of each contributor varies from only 1 day to 986 days. About 31% of contributors have contributed to OSM on less than 10 days, indicating less practice and motivation. By contrast, 52% of

TABLE 2 Feature type generalization based on the longevity of its characteristics

Map features category	Feature type	Number of features in sample	Minimum of days	Maximum of days	Average of days
Commercial and informal settlement	café, marketplace, restaurant, night club, convenience store, commercial building, and informal settlement	130	55	1,366	559
General building and land use	house, residential, general building, and land use	93	55	2,702	1,034
Highway, railway, waterway	primary, secondary, tertiary, residential, service, pedestrian, a general road, footway, path, bus stop, rail, platform, station, river, and riverbank	122	10	2,859	627
Long-lived object	fire station, kindergarten, public building, attraction, museum, place of worship, post office, police, school, and historic building	50	21	2,822	652

TABLE 3 Descriptive statistics of contributors' characteristics

Parameters	Min	Q_1	Median	Mean	Q_3	Max
Number of mapping days	1	8	34.5	164.9	243	986
Longevity of activity (days)	0	114.5	735.0	1,059.0	1,887	3,113
Number of changesets	1	43.75	413	7,660	2,212	111,166
Number of map changes	24	2,836	13,670	333,600	276,200	3,205,766
Distance to study area (km)	0.15	10.60	55.10	5,384	9,989	39,330

contributors have contributed to OSM on more than 30 days, indicating that these contributors have a sustained interest in this geographic area and/or OSM.

4.2.2 | Longevity of activity

A short time span between the first contribution and the last one suggests that this user is a rather hit-and-run user. About 11% of contributors could be categorized as this type of user. However, 73% of contributors contribute over a longer period, and likely pay closer attention to OSM development.

4.2.3 | Number of changesets and number of map changes

Both are an indicator of user motivation to contribute to the platform. However, the number of map changes, or *number of edits*, was used for further analysis instead of the number of changesets (because the number of map changes gives more detailed information about the edits done by a user). A large number of edits can hardly occur for amateurs. About 17% of contributors have less than 1,000 edits.

4.2.4 | Distance to study area

We approximate local knowledge by calculating the distance from the first created node by each user to the study area. This serves as a proxy to assess local knowledge, assuming that most users will have their first attempts at editing close to the place they are familiar with. Figure 3 illustrates all the locations of all the contributors and the line indicates the distance to the study area.

Figure 4 shows histograms for all the above-mentioned expertise parameters. The plots show that the data is not normally distributed and presents a positive skew as the mean value is higher than the median value.

We converted distances into three classes, since there is no linear relationship between distance and local knowledge. A distance of the first edit to the study area of 0–100, 100–5,000, and >5000km, respectively, represents high, low, and no local knowledge. These intervals were chosen based on visible breaks in a QQ plot showing the contributor distance distribution. We argued that the classification based on distance is better than that based on country of origin of the contributors. Table 4 shows the results of the local knowledge classification.

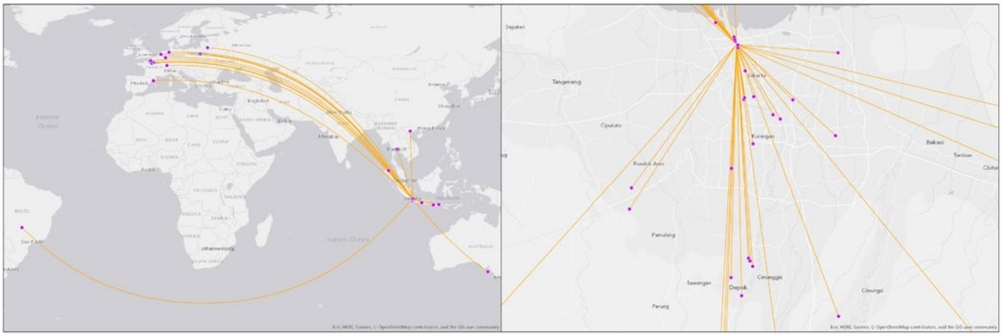

FIGURE 3 Distribution of origin of contributors (left) and distribution of contributors located near study area (right)

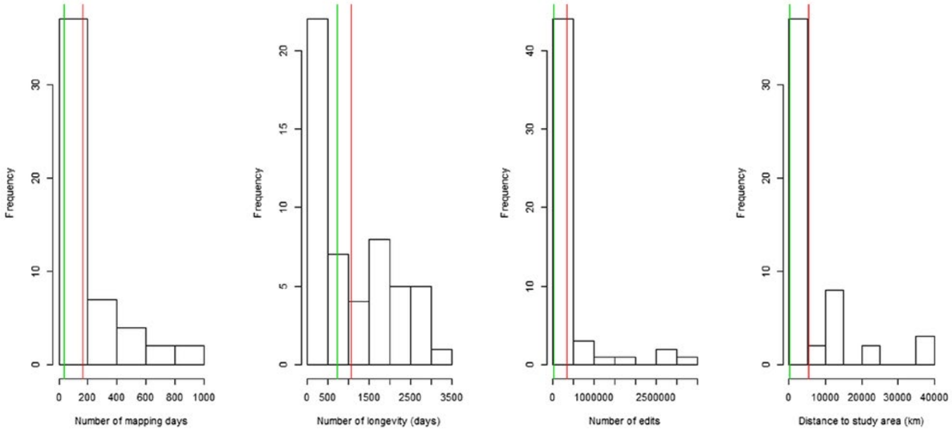

FIGURE 4 Histogram showing data distributions. The green and red lines indicate the median and mean of the data, respectively

TABLE 4 Local knowledge classification

Distance (km)	Number of contributors	% contributors	Given score
0–100	29	56%	1
100–5,000	8	15%	0.5
>5,000	15	29%	0

4.2.5 | Main software used

The ability to use different OSM editor tools shows the skill level of the contributors (Yang et al., 2016). The software used most for contributing OSM data is the software used to generate most map changes. JOSM and other powerful desktop applications require more training and arguably a higher level of skill. Potlatch and iD were categorized as software requiring moderate and low skill levels, respectively. Mobile editors such as OsmAnd, Maps.me, Vespucci, and so on require intermediate skill levels, where the user should also navigate through the field to collect the information.

As shown in Figure 5, JOSM, iD, Potlatch, and other tools are the most to least commonly used software by contributors in the study area. To bring this parameter into the model, scores were given for the different required skill levels as shown in Table 5.

4.3 | Feature quality scores obtained from the field survey

About 60% of OSM features have the maximum *attribute accuracy* score, which means that all the provided information was correct. However, only one OSM feature has wrong attribute information. The remainder of the OSM features have different attribute accuracy scores ranging from 0.225 to 0.925, where the features with a score of less than 0.7 contain wrong information about name and feature type information. Features with a score above 0.7 are of sufficient quality for map feature classification and POI search.

Regarding *completeness*, only 15% of the OSM features have the maximum completeness score. About 30% of the features have a score of less than 0.5, and therefore are errors of commission where this excess data is present

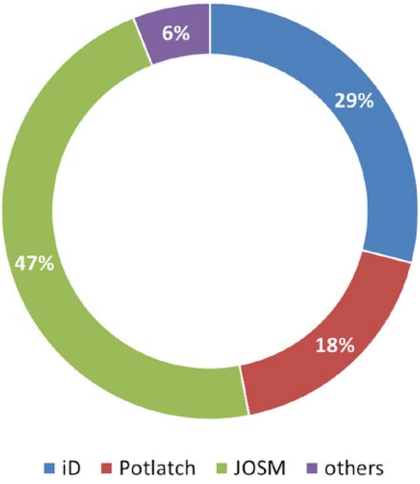

FIGURE 5 Skill of overall contributors based on the main software used

TABLE 5 Software scores

Software	Skill level	Given score
JOSM, Merkaartor, QGIS	High	3
Potlatch, Mobile editors	Moderate	2
iD	Low	1

in the dataset but does not exist in reality. The remainder of the features (55%) have different completeness scores ranging from 0.55 to 0.9, distinguished by number of tags provided for each map feature type.

We also found that the majority (96%) of OSM features in the study area are *topologically* consistent with their neighboring features. There are only six overlapping buildings and nine POIs that are not located inside the associated object. The minimum scores were given for the overlapping objects because they are topologically incorrect, while the average score was given to the objects located outside but still nearby the associated object. The frequencies of the three scores are shown in Figure 6.

Using the formulas and rules from Table 1, we calculated the feature quality scores. As Figure 7 shows, they are not normally distributed. For further statistical analysis (see the following sections), we normalized all the contributor characteristic and feature quality scores.

We then assigned each feature a quality class using the rules in Section 3.4. Figure 8 shows a map of the study area with the features colored according to their quality class. The lower-quality features appear clustered in the eastern part of the study area. Further identifications found that those incorrect features were created by the same user who has a low expertise score.

4.4 | Correlation analysis results

4.4.1 | Aggregated expertise vs. feature quality score

First, we examined the variables that contribute to the aggregated expertise for their potential utility in calculating the aggregated expertise. For each OSM feature, we aggregated for each single variable the normalized expertise

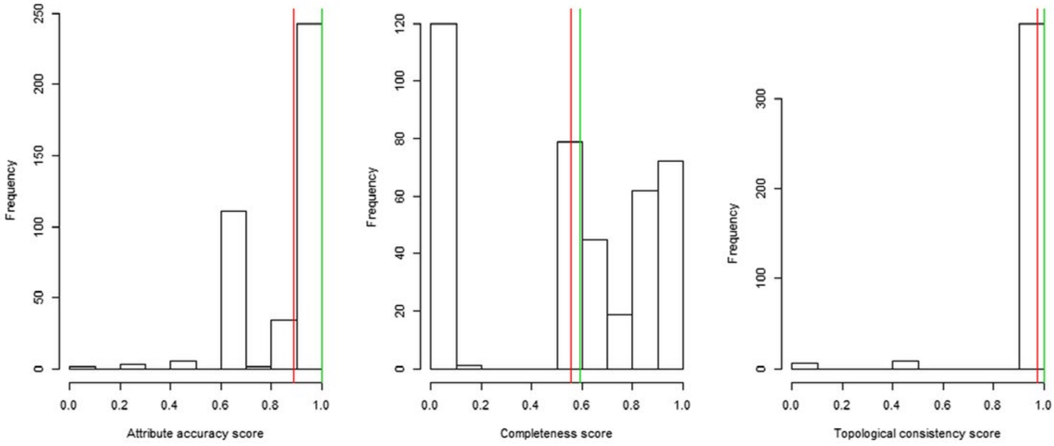

FIGURE 6 Histogram of feature's quality attributes. The green and red lines indicate the median and mean of the data, respectively

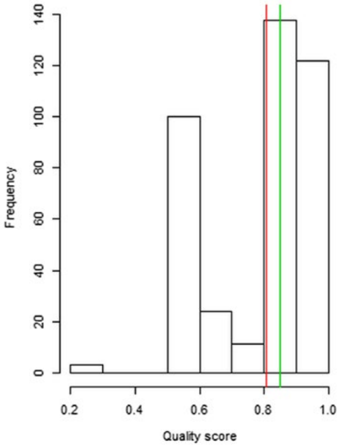

FIGURE 7 Histogram of feature quality scores. The green and red lines indicate the median and mean of the data, respectively

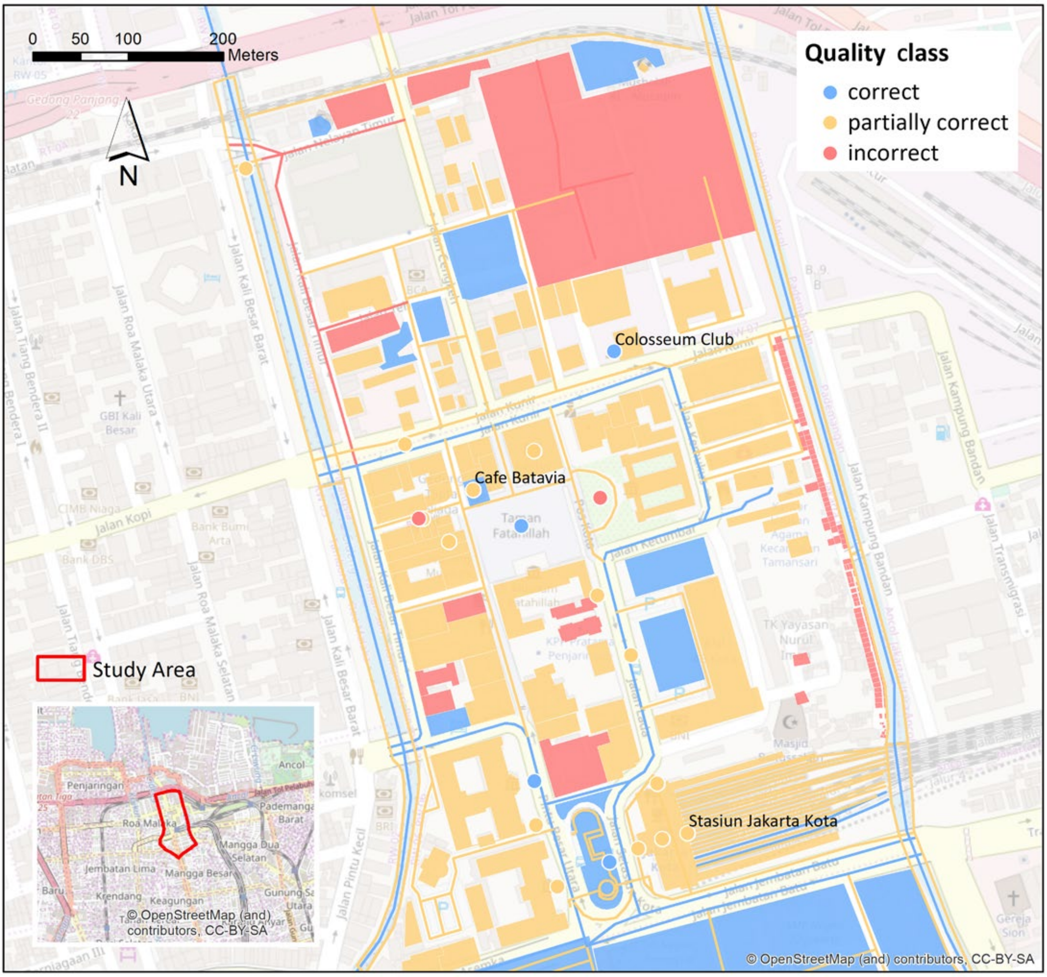

FIGURE 8 Map of the study area showing the feature quality class of OSM features

scores of all contributors to that OSM feature. These were ranked and then correlated with the ranked feature quality scores, with the results shown in Table 6.

The correlation between the expertise parameters and the feature quality scores is significantly different from zero ($p \ll 0.001$), but of varying positive strength. The distribution of the aggregated expertise scores of the 398 OSM features is illustrated as a QQ plot in Figure 9.

The aggregated expertise scores were transformed into ranks and correlated with the feature quality ranks. The correlation coefficient of 0.518031 (p value$\ll$0.001) supports the general hypothesis that the quality of OSM features could be estimated using aggregated expertise to indicate feature credibility, as described in Section 3.1.

4.4.2 | Temporal effect vs. feature quality score

Table 7 displays the results of investigating the relationship between feature quality scores and the recency of the data for the different feature types. Commercial and informal settlement features have the highest significant negative correlation (−0.38304, $p \ll 0.001$), indicating that the more recently a feature has been added, the higher the probability that the feature corresponds to reality. It supports the assumption that these kinds of features are the most unstable features. The other three map feature types have only a weak association between the two variables.

4.5 | Multinomial logistic regression

Before running the model, we chose quality class = incorrect as the baseline outcome. For this reason, the model conveys the effect of predictors on the probability in comparison to this reference class.

The logit coefficients relative to the reference class, along with the standard errors and odds ratios, are shown in Table 8. The final probability model is:

$$\ln\left(\frac{P\ (correct)}{P\ (incorrect)}\right) = -3.407 + 0.850\ (AE) \tag{5}$$

$$\ln\left(\frac{P\ (partially\ correct)}{P\ (incorrect)}\right) = -1.310 + 0.754\ (AE) \tag{6}$$

As shown, aggregated expertise has a positive impact on the log odds of the probability of an OSM feature being correct vs. the probability of an OSM feature being incorrect of 0.850*aggregated expertise. In other words, for an increase in aggregated expertise score of 1, the logit coefficient for "correct" relative to "incorrect" will increase by 0.850. Similarly, aggregated expertise has a positive impact on the log odds of the probability of an OSM feature being partially correct vs. the probability of an OSM feature being incorrect in the amount of 0.754*aggregated expertise (i.e. an increase of the aggregated expertise score by 1 increases the logit coefficient for "partially correct" relative to "incorrect" by 0.754).

TABLE 6 Correlation test results for expertise parameters against feature quality scores

Expertise parameters	*p* value	Kendall's tau correlation coefficient (τ)
Number of edits	2.2e-16	0.30423
Number of mapping days	8.9e-16	0.29928
Longevity of activity	2.2e-16	0.33177
Local knowledge	0.00025	0.14538
Software skill	2.2e-16	0.53313

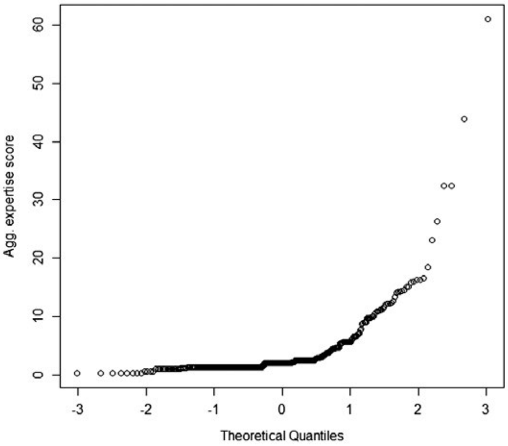

FIGURE 9 QQ plot of aggregated expertise score

TABLE 7 Correlation test results for recency of last version time effect parameter of trustworthiness against feature quality scores

Map feature type	Sample size	*p* value	Kendall's tau correlation coefficient (τ)
Highway, railway, waterway	122	0.04713	−0.13649
General building and land use	93	0.75360	0.02871
Commercial and informal settlement	130	2.7e-06	−0.38304
Long-lived object	53	0.09959	−0.16865

The results on the odds ratios indicate the ratio of the probability of an OSM feature being correct or partially correct over the probability of an OSM feature belonging to the baseline class (incorrect). The odds ratio for a one-unit increase in aggregated expertise score is 2.340 for an OSM feature being correct vs. incorrect, and 2.126 for an OSM feature being partially correct vs. incorrect.

From a 2-tailed *z* test, aggregated expertise as a predictor variable plays a significant role ($p \ll 0.001$). Further, the predicted probability was used to understand the model. This was calculated for each outcome level using a *fitted* function for all 398 observations. Figure 10 shows a plot of predicted probabilities across aggregated expertise scores.

The larger the aggregated expertise score, the higher the probability of OSM features being correct, as illustrated by the blue line. On the contrary, the red line indicates the larger aggregated expertise score, the lower the probability of OSM features being incorrect. The probability of an OSM feature being partially correct decreases

TABLE 8 Logit coefficient, standard error, and odds ratio results relative to the reference class

	Coefficients		Std. errors		Odds ratios	
	Intercept	AE	Intercept	AE	Intercept	AE
Correct	−3.407	0.850	0.335	0.130	0.033	2.340
Partially correct	−1.310	0.754	0.261	0.128	0.269	2.126

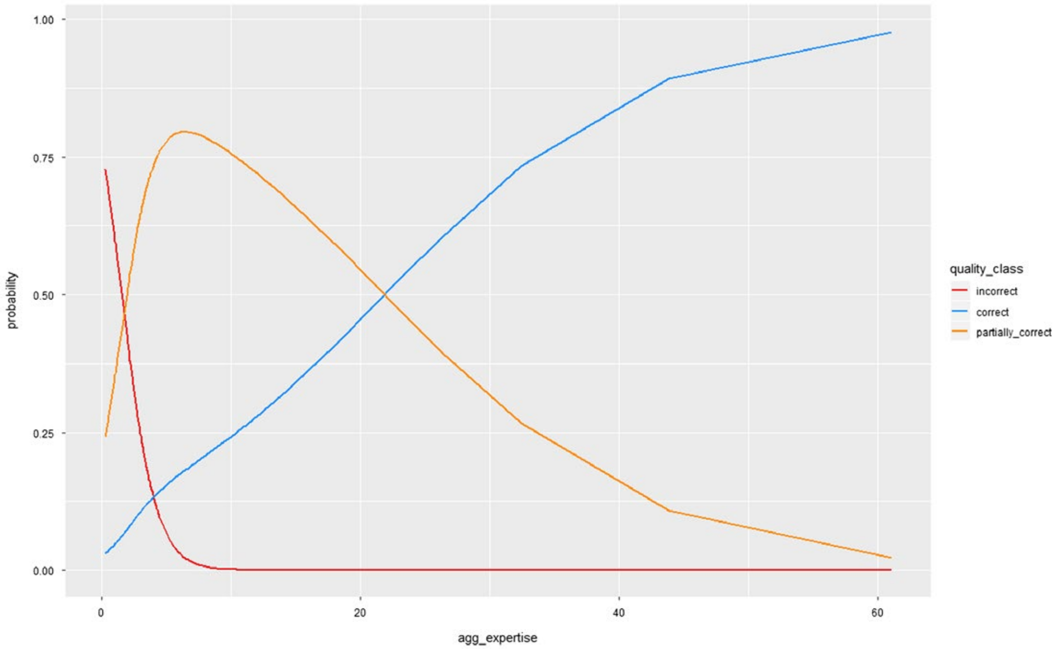

FIGURE 10 Plot of predicted probabilities across aggregated expertise scores for each quality class

for aggregated expertise scores above ~7. Overall, these results make sense: with increasing aggregated expertise, a feature is more likely to be at least partially correct, and from a certain level of aggregated expertise, it is most likely to be completely correct.

5 | DISCUSSION

Through a set of experiments, this study examines an intrinsic VGI quality assessment by looking at the properties of the contributors and the edit history of the features. We attempt to derive contributors' expertise from their editing history, using it as a proxy measure for their local knowledge, skill, and experience. We acknowledge that contributor expertise increases somewhat over time, but an examination of this factor was beyond the scope of this study.

Based on the contributor analysis discussed in Section 4.2, the majority of OSM data in the study area was created by a few contributors. Further examination confirmed that those creators are local users based near the study area (i.e. deemed within 100km of the study area, as per Table 4). In contrast, data modifications were not only from local users but also from remote contributors. Interestingly, this contribution pattern is different from the pattern of coordinated mapping initiatives, where remote volunteers first trace satellite imagery into OSM and then local mappers in the respective area add local details. Altogether, about 10% of the contributors provided most of the data within the study area. This finding corresponds closely to the phenomenon of participation inequality in online and crowdsourcing communities introduced by Nielsen (2006), where a small number of users account for most contributions.

A field survey provided the required external data quality metrics for model testing. The execution of the fieldwork provided a real insight in the sense that using ground-truth data to validate the quality of OSM is high in terms of effort and time-consuming.

Correlation analysis confirmed that the aggregated expertise measure allows the estimation of VGI quality by approximating a feature's credibility through the properties of the contributors. This shows that the quality is not solely determined by the number of versions created. In general, the quality improves when the number of contributors increases, with the sum of contributor expertise being more suitable than the average contributor expertise. Features which are edited twice collectively by skilled and unskilled users should have higher aggregated expertise than a feature with only a single edit by a skilled user as long as the edit type is not overriding or altering some existing information. Our implemented model has not yet taken into account the weight of the editing types, which could also influence the calculation score. By leaving the edit type information out of the model implementation, one could argue that this model is overestimating the credibility because it is not fit for a highly contested feature where edit "wars" occur. Further research might include this weighting variable in the aggregated expertise model.

Another significant finding was the strong relation between credibility and recency of last edit for commercial and informal settlement features. A more detailed investigation with a larger dataset for each feature type (especially for long-lived objects) would be useful to prove this assumption. Additionally, further research might include the editing of nearby features as a proxy for the implicit feature validation, as another parameter related to the temporal effect.

Aggregated expertise as a predictor in multinomial logistic regression has a significant impact on predicting outcome probabilities. However, this regression method would benefit from a larger sample size, since it uses a maximum likelihood estimation. Additionally, random sampling which is not geographically stratified would be better to avoid an internally homogeneous sample where few users produce the same quality of OSM data. Another predictor can be introduced such as feature type (point, line, and polygon) to see how the model predicts differently for each feature type.

6 | CONCLUSIONS AND FUTURE WORK

In this study, we operationalized the concept of aggregated experience based on feature edit history and contributor characteristics, with the aim of using it to assess the credibility of VGI features. We then successfully tested it in a case study against ground truth based on a recent field survey that used three ISO data quality measures to describe the feature quality of OSM data. The aggregated expertise allows us to predict the credibility of OSM features, and by extension the likelihood of a feature being of sufficiently high or low quality. This process can be fully automated and is computationally inexpensive. By helping to discover areas of low-quality scores, for example, it can help to prioritize mapping efforts.

However, there are still several limitations to our study. First, the exact implementation depends on the available metadata from the VGI platform. For example, other VGI platforms might not have detailed information on software used, hence the software skill parameter might have to be implemented differently or not at all. However, any VGI platform that supports feature versioning and records user profile/reputation information can adapt the approach. To the best of our knowledge, currently only Wikimapia (https://wikimapia.org) has a comparable versioning system in place. Therefore, including a versioning system is an important recommendation for any platform or project that plans to collect and manage VGI (e.g. citizen science projects). The use of linked data in relationship with a domain ontology could be another way to enrich the semantics of the contributors' profiles and the features they edit.

Second, our study had limited resources, resulting in a small pilot case study. Larger case studies in different geographic areas are required to demonstrate the geographic transferability of our approach. Carefully planned and executed ground surveys for model validation seem indispensable until we have sufficient information on the robustness of the model for different user communities. For larger case studies, it seems advisable to download and use the OSM Planet History files to reduce the load on OSM servers. To scale up the automated processing, a cloud computing infrastructure offers a flexible approach to host larger datasets. Future work could also explore additional combinations of variables and their weights for aggregated expertise scores, and test the validity of the approach for specific feature types. We hope that the provided supplementary material (see Supporting Information) will contribute to improving the overall state of reproducibility for VGI-based studies (Ostermann & Granell, 2017), and encourage follow-up research addressing the open challenges.

ACKNOWLEDGMENTS

We would like to thank Pascal Neis for allowing us to use his tools. Part of this work, including the field survey, has been funded by a Graduate Scholarship from the Indonesian Endownment Fund for Education.

ORCID

Frank O. Ostermann http://orcid.org/0000-0002-9317-8291

Rob L. G. Lemmens https://orcid.org/0000-0001-5269-6343

REFERENCES

Antoniou, V., & Skopeliti, A. (2015). Measures and indicators of VGI quality: An overview. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, *II-3/W5*, 345–351.

Artz, D., & Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services & Agents on the World Wide Web*, *5*(2), 58–71.

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

Bégin, D., Devillers, R., & Roche, S. (2013). Assessing volunteered geographic information (VGI) quality based on contributors' mapping behaviours. *ISPRS International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, *XL-2/W1*(2), 149–154.

Bishr, M., & Janowicz, K. (2010). Can we trust information? The case of volunteered geographic information. In *Proceedings of the Towards Digital Earth: Search, Discover and Share Geospatial Data Workshop at the Future Internet Symposium*. Berlin, Germany.

Bishr, M., & Kuhn, W. (2007). Geospatial information bottom-up: A matter of trust and semantics. In S. I. Fabrikant & M. Wachowicz (Eds.), *The European information society* (Lecture Notes in Geoinformation and Cartography, pp. 365–387). Berlin, Germany: Springer.

Budhathoki, N. R., & Haythornthwaite, C. (2013). Motivation for open collaboration: Crowd and community models and the case of OpenStreetMap. *American Behavioral Scientist*, *57*(5), 548–575.

Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., & Purves, R. (Eds.) (2016). *European handbook on crowdsourced geographic information*. London, UK: Ubiquity Press.

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3&4), 137–148.

Fogg, B. J., & Tseng, H. (1999). Elements of computer credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 80–87). Pittsburgh, PA: ACM.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment & Planning B*, *37*(4), 682–703.

Keßler, C., & de Groot, R. T. A. (2013). Trust as a proxy measure for the quality of volunteered geographic information in the case of OpenStreetMap. In D. Vandenbroucke, B. Bucher, & J. Crompvoets (Eds.), *Geographic information science at the heart of Europe* (Lecture Notes in Geoinformation and Cartography, pp. 21–37). Berlin, Germany: Springer.

Keßler, C., Trame, J., & Kauppinen, T. (2011). Tracking editing processes in volunteered geographic information: The case of OpenStreetMap. In *Proceedings of Workshop on Identifying Objects, Processes and Events in Spatio-Temporally Distributed Data*. Belfast, ME.

Mooney, P., & Corcoran, P. (2012). Characteristics of heavily edited objects in OpenStreetMap. *Future Internet*, *4*(1), 285–305.

Neis, P. (2012). *How did you contribute to OpenStreetMap?* Retrieved from http://neisone.org/2012/10/overhauling-hdyc/

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.

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

Nielsen, J. (2006). *Participation inequality: The 90-9-1 rule for social features*. Retrieved from https://www.nngroup.com/articles/participation-inequality/

OpenStreetMap Wiki. (2016). *History API*. Retrieved from http://wiki.openstreetmap.org/wiki/Contributors#Finding_contributors

Ostermann, F. O., & Granell, C. (2017). Advancing science with VGI: Reproducibility and replicability of recent studies using VGI. *Transactions in GIS*, *21*(2), 224–237.

Sen, S. W., Ford, H., Musicant, D. R., Graham, M., Keyes, O. S. B., & Hecht, B. (2015). Barriers to the localness of volunteered geographic information. In *Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems* (pp. 197–206). Seoul, South Korea: ACM.

van Exel, M., Dias, E., & Fruijtier, S. (2010). The impact of crowdsourcing on spatial data quality indicators. In *Proceedings of the Sixth International Conference on Geographic Information Science*. Zurich, Switzerland.

Vandecasteele, A., & Devillers, R. (2015). Improving volunteered geographic information quality using a tag recommender system: The case of OpenStreetMap. In J. Jokar Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.), *OpenStreetMap in GIScience: Experiences, research, and applications* (Lecture Notes in Geoinformation and Cartography, pp. 59–80). Berlin, Germany: Springer.

Yang, A., Fan, H., & Jing, N. (2016). Amateur or professional: Assessing the expertise of major contributors in OpenStreetMap based on contributing behaviors. *ISPRS International Journal of Geo-Information*, *5*(2), 21.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Muttaqien BI, Ostermann FO, Lemmens RLG. Modeling aggregated expertise of user contributions to assess the credibility of OpenStreetMap features. *Transactions in GIS*. 2018;22:823–841. https://doi.org/10.1111/tgis.12454