Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

An Automated Approach to Identifying Corporate Editing

Veniamin Veselovsky[1], Dipto Sarkar[2], Jennings Anderson[3], Robert Soden[4]

[1]EPFL, [2]Carleton University, [3]University of Colorado Boulder, [4]University of Toronto
veniamin.veselovsky@epfl.ch, diptosarkar@cunet.carleton.ca, jennings.anderson@colorado.edu, soden@cs.toronto.edu

Abstract

OpenStreetMap (OSM) is the world's largest peer-produced geospatial project. As a freely-editable open map of the world to which anyone may contribute or make use of, the dynamics and motivations of its contributors have been the object of significant scholarship. A growing phenomena in the OSM community is the increasing contributions of paid editing teams hired by tech corporations, such as, Microsoft, Apple, and Facebook. Though corporations have long supported OSM in various ways, the recent growth of teams of paid editors raises challenges to the community's norms and policies, which are historically oriented around contributions by individual volunteers, making it hard to track the contribution of paid editors. This research addresses a fundamental problem in approaching these concerns: understanding the scale and character of corporate editing in OSM. We use machine-learning to improve upon prior approaches to estimating this phenomena, contributing both a novel methodology as well as a more robust understanding of the latest corporate editing behavior in OSM.

Introduction

OpenStreetMap (OSM) is the world's largest peer-produced geospatial project. OSM started in 2004 as a volunteer effort and in those initial years, people organized in-person 'mapping parties' to collect local spatial data and introduce people to the project (Solis 2017). This tradition has continued and even today many OSM groups organize mapping parties to recruit new volunteers. Over time, OSM garnered significant momentum resulting in exponential growth in the amount of spatial data contributed as well as the size of the community of data contributors. At the same time, OSM data quality has vastly improved across a wide spectrum of quality metrics (Girres and Touya 2010; Barron, Neis, and Zipf 2014; Mooney, Corcoran, and Winstanley 2010; Pourabdollah et al. 2013) and is comparable, if not better across some dimensions than authoritative data in many places (Cipeluch et al. 2010; Girres and Touya 2010; Haklay 2010).

OSM data have been notably useful in post disaster recovery scenarios in the Global South, offering a viable alternative where there is, as is often the case, a lack of authoritative geospatial data. (Soden and Palen 2014; Anderson et al. 2018). Coordinating efforts to create large volumes of data in a short time has also led to the creation of organizations such as the Humanitarian OpenStreetMap (HOT) (Palen et al. 2015). As the value of OSM grew, government agencies also started contributing (Johnson 2017). Such organized editing efforts differ in motives from the purely hobbyist notion around which many initial contributors self-identified (Budhathoki and Haythornthwaite 2013) and have the capacity to create significant amounts of data in specific geographies over a short duration. In order to distinguish the provenance of individual voluntarily contributed data from organized efforts, the OSM Foundation (OSMF) came up with the Organized Editing Guidelines (OEG) in 2018 which asks such activities to self-report on the OSM wiki (Chapman 2018) .

In addition to volunteers, governments, and nonprofits, private sector corporations also rely on OSM data and support the OSM project. These include tech giants such as Apple, Amazon, Facebook, Microsoft, and Uber who are hiring editing teams to contribute data to OSM (Anderson, Sarkar, and Palen 2019). This list also includes a growing number of smaller companies, such as Digital Egypt, LightCyphers, Bolt, and more, who contribute regionally-specific data, such as building addresses in Cairo, Egypt (Anderson and Sarkar 2020). As an organized effort, OSM policy dictates these corporate editing teams comply with the OEG. This includes publishing a summary of their activity and a list of their mappers on publicly available sites. At present, there is no central website or format for these lists and corporation routinely use a number of different public outlets, such as the OSM wiki or GitHub repositories to publish this information. However, as paid editing activity grows, it is becoming increasingly difficult for researchers to track and collate these lists across a variety of websites. These lists of known paid editors are the starting point of characterizing the impacts of paid editing activity on OSM. There are also concerns brewing within the community about paid activity overshadowing volunteer efforts and the consequent future of the OSM project (Foundation 2021). The reasons for these concerns stem from the rapid rate at which paid editing teams can create data (e.g. 1000s of Kms of road data per year), as well as the power imbalances engendered due to the difference in resources available between the corporations and other mappers.

At this time, quantifying and mapping paid contributions is a vital step to broker informed dialogue within the larger OSM community about the impact of corporate editing. Another method of tracking corporate contributions is through hashtags found in the OSM changeset record (e.g. #adt, #kaart used by the Apple Data Team and Kaart respectively). This method also lacks scalability as there is no definitive nor comprehensive list of hashtags used by the various teams, nor do all teams use specific hashtags. The OEG include mention of the team for which an individual is mapping for in the their user OSM user page bio (Chapman 2018). However, in the absence of a machine-readable standard for how this information is disclosed, tracking paid editors through these bios is also not scalable. Thus, under the present scenario, there exists no method of tracking paid editors which does not rely on collating lists from multiple sources. Here, we propose an automated technique for identifying paid editors based on analyzing patterns of contributions of known paid editors. Further, we map the global scale and intensity of corporate-sponsored mapping in OSM between 2018 and 2021, expanding on previous work in this space.

Background & Related Work: Analyzing Editing Behavior in OpenStreetMap

OpenStreetMap is one of the largest online peer-production communities and the most extensive repository of open spatial data in existence. It hosts more than 1.7M unique contributors and more than 800M geospatial objects mapped to date, made up of over 7B individual points. Significant research has been conducted on various aspects of the platform and the social dynamics of the community, including investigations into the quality of OSM data (Haklay 2010), the motivations of its participants (Budhathoki and Haythornthwaite 2013), its use during humanitarian emergencies (Soden and Palen 2014), and, more recently, its adoption by large corporations in the tech sector (Anderson, Sarkar, and Palen 2019). However, important aspects of how the community collaboratively produces the underlying spatial data that powers the platform remains unclear. This is partly due to the sheer size of the dataset and challenges presented by the format of the data (Anderson et al. 2016). Significant work is under way by OSM researchers to build novel tools and approaches for analyzing and studying contributions to the platform. Two example tools produced by the larger OSM community include OSMCha, which allows users to review individual changesets and OHSOME, the OSM History Explorer that lets a user examine the full history of OSM in a given region. These tools are optimized for qualitative inspection of changes to a region, with OHSOME allowing for regional or temporal aggregation. This remains an active research space, however, with many tools developed each year to better understand the social and collaborative processes that produce the map.

Contribution analysis in OSM has largely focused on who edits the data, what is edited, and the artifacts resulting from the contribution. In keeping with OSM's roots as a Volunteered Geographic Information (VGI) platform, the majority of contributors are volunteers. However, not all contributors participate equally. Like other online platforms, OSM follows the 90-9-1% rule where the majority of the data (87% in the case of OSM) has been created by 1% of users (Anderson, Sarkar, and Palen 2019). The disparity of contributions has been partially attributed to online mirroring of offline inequalities in the form of exclusion of women, non-urbanites, and people in the global south (Stephens 2013; Quinn 2017; Gilbert 2010; Shelton et al. 2014; Graham et al. 2014; Burns and Meek 2015) due to structural barriers, such as, internet access, free time to participate, geo-politics, and gatekeeping (Sui, Goodchild, and Elwood 2013; Stephens 2013). Unequal participation manifested artefacts in the data in terms of over-representation of data of interest to the parties who have the privilege to be active contributors. Thus, even though OSM data passes several quality thresholds, equity in terms of feature representation and global data coverage remains an issue.

In recent years, OSM has garnered significant interest from major corporations including Apple, Amazon, Facebook, Microsoft, and Uber (Anderson and Sarkar 2020; Anderson, Sarkar, and Palen 2019; Sarkar and Anderson 2021, 2022). International organizations and non-profits working in the humanitarian sector such as the Red Cross and United Nations also use OSM data and contribute to mapping efforts. Editing teams hired by these organizations are often professional mappers tasked with contributing data to OSM. Consequently, these editing teams have produced great volumes of data. In addition to data volume, the involvement of paid editors introduces new dynamics into a community that has historically consisted primarily of volunteers. Many organizations are editing in developing countries which has the potential to reduce data inequalities in global data coverage in OSM, but risks prioritizing the goals of the companies and organizations doing the mapping over those of the residents of such countries.

The growth of corporate editing raises concerns over unequal influence on the future of the platform. The current state of the art method to track corporate editing is to collate a list of corporate editors from various sources and then extract their edits from the changeset record (Anderson, Sarkar, and Palen 2019). However, in the absence of a central list of all paid editing teams, these criteria are insufficient for tracking all paid mapping activity across OSM. As a result, neither researchers nor the OSM community itself has a complete understanding of the current extent and impact of paid editing activities.

One way to identify paid editors is to look at the associated salient features. The biographies, temporality of edits, spatiality of edits, software used for editing, and similarity to known paid editors can be used to find a probability score for an editor being a member of a paid editing team. Similar approaches have been developed in Wikipedia, where, for example, time series analysis was used to find specific patterns in user editing behaviour across the world (Yasseri, Sumi, and Kertész 2012). In OSM, contributor focused data assessment showed that senior mappers tend to map more complex features and use more advanced OSM editing software (Jacobs and Mitchell 2020), produce quality data (Yang, Fan, and Jing 2016), and also tend to interact with other members of the community more often (Mooney and Corcoran 2012). On the other hand, new mappers are often attracted by HOT and consequently map developing countries (Madubedube, Coetzee, and Rautenbach 2021). The aforementioned studies look at contribution patterns of editors and thus mostly revert to seniority in terms of time since creation or number of edits performed. Characterizing paid editing is more complex as most editor profiles are relatively new and the heterogeneity of the group means that simple features are not sufficient to identify the wide variety of contribution patterns of paid editors.

Identifying Corporate Editing in OSM

Constructing a Training Dataset

We first extracted a list of known corporate mappers. To do this we scraped the user bios (osm.org/user/<username>) of all OSM mappers with more than 50 unique changesets (129,557 users) and then queried their user profiles for mentions of publicly disclosed corporations mapping on OSM. The publicly disclosed list was found on the OSM Organised Editing Page, and after analzying the profiles we found 14,901 users with non-empty user profiles, and 3,512 that were self-disclosed corporate mappers.

After acquiring the list of known corporate mappers, we extracted their editing histories through a query of OSM changesets between 2018 through June 2021, which included 53,921,921 unique changesets representing over 4.9B map changes by 852K distinct users. We then limited to users with more than 50 unique changesets during this period to temporally restrict our analysis leaving us with 35,479 users, of which 2,393 are corporate.

A changeset is a logical grouping of up to 10,000 edits made by a single user in one editing session. When submitting a changeset to OSM, the mapper is required to add a comment that is recorded in the metadata, along with the time, user information, and geographic bounds of the edits. Of particular interest to this work are the geographic bounds, timestamp, tags, and unique user ID. Tags at minimum feature a comment, but may also include hashtags, data source, and editing software information. Provided as free-entry text, comments typically include descriptions of what was mapped, and perhaps why. Additionally, hashtags are a way for users to denote changesets as belonging to a particular mapping activity or project. For example, all changesets submitted during a mapathon—an event in which many editors may map together—may include the same user-defined hashtag. This makes it easier to identify edits performed during the event and can be fed into edit-tracking tools such as the MissingMaps Leaderboard, which allows users to see the top mappers for given hashtags.[1]

We only consider changesets after 2018 because that was the first year when corporate editing began to grow dramatically, as well as the year that the OEG were implemented, resulting in the disclosure of mapping activities by corporate mappers on their bios (Chapman 2018). Comparing 2017 to 2018 the number of corporate mappers increases almost two-fold (Anderson, Sarkar, and Palen 2019).

Developing User Features

In this paper we adopt a feature engineering approach to predicting whether a user is corporate. This requires developing features likely to differentiate the corporate mappers from the rest of the OSM community. OSM's editing history contains rich metadata about each edit including when it was made, how it was made, and more. This simplifies the problem to designing features whose variance is explained by the association of the editor; in other words, finding dimensions which are most correlated with corporate association. As was shown in related work, prior work exists on contributor level analysis and these features aided us in the design of our user features. Simply put, our features were based on a few questions: Where does a user edit? When does a user edit? How does a user edit? What does a user edit? Some of these questions were easier to answer than others, whereas others required more feature engineering. Table 1 includes a summary of the features we developed and the observation each feature meant to capture.

Time series metrics. We constructed two temporal features based on two separate observations. The first observation is that corporate mappers typically map five days in a row, and then take a break during the weekend. To capture this, we take the mode number of consecutive days a user maps for over their editing history. This observation is strongly supported by comparing the mode of edits between known corporate and volunteer mappers. In fact, the mode for consecutive days for corporate mappers is five, where as for volunteer mappers it's one. The second temporal feature is based on the observation that corporate mappers map during the workday.

OSM records the time of a changeset in UTC, without including the local time or timezone of the user who conducted the edits. This means someone editing at 8am in Lausanne, Switzerland and another user editing at 8pm in Honolulu, Hawaii would in fact appear to be editing at the same time. This standardization renders it difficult to know if an edit was done during the regular workday or even on a weekday, as Monday or Friday could appear to have been done on Sunday and Saturday respectively. This is evident from figure 1 where we plot the user time signatures for six known corporations. This plot shows that even individual organizations have editors who are located in different regions of the world. Longitude and latitude data are not an effective method of extracting the mappers timezone, since editing on OSM is primarily done remotely, through "armchair mapping," meaning an editor may be physically located anywhere in the world while editing anywhere on the map.

To account for this time zone difference, we propose a novel method to extract the underlying user timezone. This metric is motivated by the observation that if a paid editor is located x hours ahead of UTC and works the traditional nine-to-five work day, then it would appear that they start editing at $(9 - x)$ am and edit until $(17 - x)$ pm. So, by

[1] missingmaps.org/leaderboards/

Question	Feature	Observation
When	Consecutive days mapped	Corporate mappers usually map Monday through Friday
	Time series score	Corporate mappers map during the work day
	First edit day	Corporate mappers represent proportionally more new users
	Rate of edits	Corporate mappers edit a lot throughout the day
	Longevity	Volunteer mappers can map for a long period uninterrupted
Where	Geographic dispersion	Corporate mapping done remotely so a corporate signature more likely to be dispersed
	Edits: ”home country”	Corporate mappers often map in unmapped regions representing a large portion of the edits in that area
How	Editor used	Corporate mappers tend to use more professional mappers like JOSM
What	Objects edited	Past work suggests corporate mappers disproportionately map roads, driveways, and services

Table 1: User feature description broken down by motivating question and observation.

translating this editing pattern x hours back we can uncover the original time zone. To adjust for this we define a “corporate time signature” based on known corporate editors and then realign all user weekly time signatures to minimize the distance between the corporate signature and their weekly signature. Explicitly, the calculation was done as follows

$$adj(ts_i, cs) = min_j[ts_i - cs] \tag{1}$$

Here ts_i represents the weekly editing signature for editor i, and cs represents the “corporate signature”. This realignment recovers a paid mapper’s local time zone on the assumption that they map during normal working hours. Figure 2 shows the same users as Figure 1, with the realignment implemented. As is clear there is less variance in the corporate signatures than by using the changeset data alone. We then calculate the distance between the adjusted time series and the organized signature to measure how similar an editing pattern is to an expected organized signature. We denote this distance as the ts-score. Examining the top five-hundred users by the ts-score we see that almost all of them are known corporate mappers (95.4%).

Geometric dispersion. Corporations map remotely all around the world. Kaart’s OSM Wiki page lists mapping endeavours in Africa, Asia, Europe, the Middle East, North America, and South America. Similarly, Apple maps over the Global South, Grab maps all over Asia. In general, we observe that individuals who map for these organizations have a wider global footprint than volunteer mappers, who, while still mapping across borders, center most of their mapping activity at home. To quantify this observation we develop a set of geometric features. The first feature captures the global distribution of a user’s mapping. The second feature instead focuses on the nuanced observation that while volunteer mappers may also map around the world, their mapping is concentrated close to home.

To measure the geographic dispersion of a user editing, we project the (max) latitude and (max) longitude data from each changesets for each user to the Earth’s

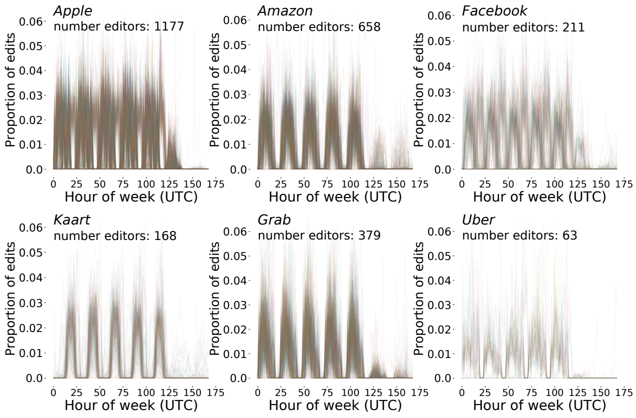

Figure 1: Raw corporate time editing signatures for mappers from six known corporations.

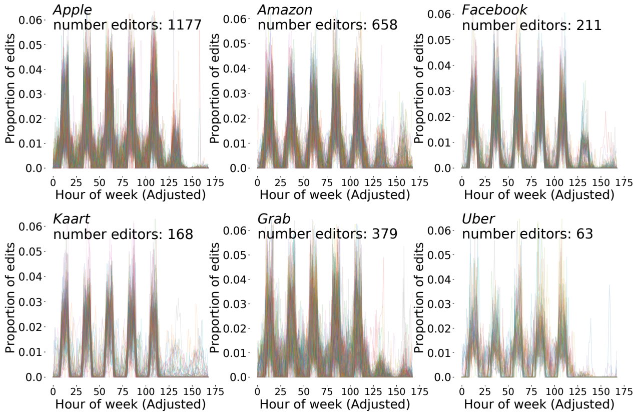

Figure 2: Time series editing signature for mappers from six known corporations, after realigning with the corporate time signature.

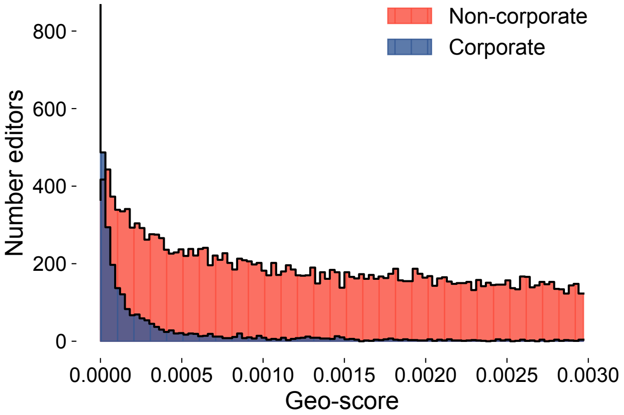

Figure 3: Plot shows how geometric dispersion differs between known corporate and non-corporate actors.

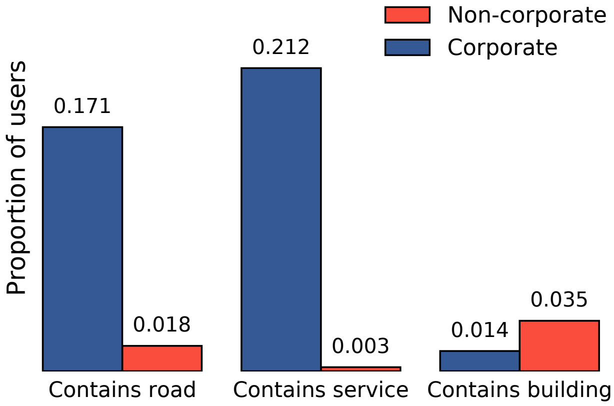

Figure 4: Difference in language use between corporate and non-corporate editors.

surface using the following transformation: $\pi(\phi, \lambda) = (R\cos\phi\cos\lambda, R\sin\phi\cos\lambda, Rsin\lambda)$. Then we define the users geographic dispersion as

$$\text{geo-score}(u) = \sum_{i=1}^{n} \frac{\pi(u_{\phi_i}, u_{\lambda_i})}{n} \tag{2}$$

$$\pi(\phi, \lambda) = (R\cos\phi\cos\lambda, R\sin\phi\cos\lambda, Rsin\lambda) \tag{3}$$

Here we let u_{ϕ_i} represents that max latitude and u_{λ_i} the max longitude for user u. Geometrically, this represents the location of the mean point inside of the sphere. If the points are broadly dispersed, the mean will be closer to the origin and if they are located in one portion of the earth then the mean will be located closer to the surface. Figure 3 shows a histogram of the geo-score for corporate versus non-corporate editors. A smaller score indicates that the user edits more broadly across the globe, whereas a higher score means that it is more specific. As is clear from the figure, corporate mappers are far more likely to edit broadly across the globe than volunteer mappers, who, while still editing across the globe, tend to have a specialized editing pattern.

Qualitatively analyzing non-corporate mappers revealed that while local mappers may also map broadly across the world, their editing is mostly centered in their home country, so the proportion of edits done at their home country greatly outnumbers the number of edits done in other countries. In contrast, with exception, corporate mappers will have a smaller fraction of their edits made up in one country. We measured the proportion of edits made in a user's "home country", where "home" represents the country where most of the users edits have taken place.

Tags data. A variety of metadata "tags" exist for each changeset. These can be any key-value pair denoting a particular attribute of the changeset. Common tags are "created_by" (denoting the editing software), "comment" (a mapper's comment about the changeset), and "source" (disclosing the source of the information the mapper used). Empirical investigation of tags shows that organizations have developed rules and practices for their teams as there often exists uniformity in the language used in comments, hashtags, and editing software. We make use of two of these tag features. First, the editing software the user is using. We find that corporate editors disproportionately use the more sophisticated JOSM editor (86%), whereas volunteer mappers mostly use the user-friendly iD (61%) and mobile editors like OSMAnd, MAPS.ME, and Vespucci. Second, is the type of object the user is editing. Comments often include keywords disclosing the type of features that the mapper edited, such as "road", "building", or "service". In Figure 4 we compare the proportion of users that use these terms based on corporate and non-corporate affiliation. The text in the comments differs widely between corporate and volunteer mappers providing critical insight into the difference in objects edited by the two groups. To measure this, we randomly sampled 100,000 comments from corporate and non-corporate mappers and calculated the tf-idf between the comments in the two groups. The 20 words most associated with each group were chosen, removing any words that relate to the identity of the mapper like the name of a corporation the user is mapping for. We then use these textual features as additional variables to capture the semantic features of an editor.

We combined the three datasets into one user feature space used for training the model to predict the likelihood that a user is editing on behalf of a corporation or not.

Model Training and Results

Our aim with this research was to develop a model that could predict whether or not a mapper is a corporate editor based on their editing history. This required identifying a corporate editing signature that was sufficiently distinct such that it could be algorithmically distinguished from other users. Given the many features used to embed each user we relied on Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of the space and locate clusters of similar users to see if there is some underlying order to how they edit. Figure 5 presents a two-

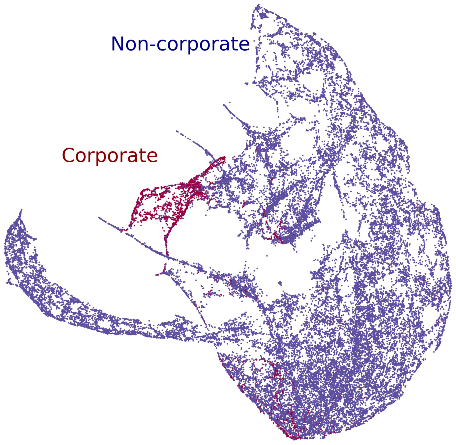

Figure 5: Two-dimensional UMAP of 35,479 mappers on OSM. Each user is represented as a point and the color indicates if they are corporate or not using the original scraping. The UMAP's parameters were the 15 nearest neighbours and minimum distance of 0.1.

dimensional UMAP of OSM editing behaviour coloured by corporate association, with all features scaled. Given the reduced dimensionality, this projection contains less information than the original embedding yet it was able to group the corporate editors together. The UMAP also reveals a clear imbalance in the dataset: corporate users make up only 7% of all 35,479 users with more than 50 changesets. Training a model on such a dataset is challenging since most models would fall back to a naive classification, classifying all users as non-corporate and achieving an accuracy of 93.5%. To adjust for this, we took steps to balance the datasets.

Creating the Model

During model selection, four main metrics were used to capture the effectiveness of a model: recall, precision, $F_{0.5}$ score, and the area under the curve (AUC) metric for the receiver operating characteristic curve (ROC). Recall measures how many of the known corporate samples the model was able to predict and all models shared a high recall of around 87%. In our context, precision can be interpreted as the number of additional corporate editors the model predicts, that were not present in the initial dataset. Identifying these additional mappers is critical to understanding the true extent of corporate editing on OSM. $F_{0.5}$ score, more generally known as the F_β score, is a measure of a test's accuracy, combining both the precision and recall into one metric. The β value determines class weighting, a lower score means that the model should value false positives more than false-negatives, and a higher score represents the opposite. Explicitly, the score is defined as $F_\beta = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$. Finally, the ROC curve plots the true positive rate by the false positive rate based on a probability threshold.

Given the large distinction between non-corporate and corporate features outlined in the feature creation section, we initially used simpler classification models like k-nearest neighbours, logistic regression, support vector machines, and a simple perceptron (single layer neural network). These models were used to provide an interpretable result to easily understand how the model works.

Feature and model selection. The first issue we came by is known broadly within the literature as the "curse of dimensionality," which essentially states that by adding more dimensions to a dataset you need exponentially more data to train a good classifier (Bellman and Kalaba 1959). To handle this issue, we adopted forward feature selection, a greedy approach that recursively adds the most predictive features until the model's accuracy declines. We then iterated across five models (support vector machines, k-nearest neighbours, logisitc regression, neural networks, aid XGBoost) and determined the optimal feature selection for each model. This feature selection was run on a upsampled subset of the data since most models would predict all users to be non-corporate and maintain a high accuracy on the full dataset.[2] We then split the data into three groups with 64% of the data for training, 16% for validation, and 20% held out for testing. The validation set was used to decide on the optimal set of features for each model.

XGBoost had the best results across the metrics, with a recall of 0.89, precision 0.78, f-0.5 score of 0.80 and AUC of 0.93 on the test dataset. The feature selection process for the XGBoost model is presented in Figure 6. It's suprising to note that even with just the ts-score, the model achieves a high AUC score of 0.907. We also show the feature importance based for the top eight features after inputting all features into the XGBoost model in Figure 7.

The other models provided similar results, performing a few points lower than the XGBoost model. They similarly agreed upon the most important features, which included the time-series score, the mean number of consecutive days a user maps, the geographic dispersion score, the choice of editor, and the proportion of edits in the home country.

Model Prediction and Validation

Rerunning the XGBoost model on the entire dataset of users since 2018 captured 2266 of the 2393 known corporate editors and predicted another 350. We manually went through each of the users to better understand the prediction. Thirty of the mappers explicitly declared their corporate affiliation in their user bio, which belonged to companies not listed on the organised editing Wiki or were not picked up on our initial data pull due to inconsistent formatting (for example, only listing their email as an affiliation). Careful analysis of the next 320 users revealed many mappers participating in organized editing activities that are not necessarily corporate, but are indistinguishable because they are mapping as part of their occupation. The United Nations Global Service Center (UNGSC), for example, employs a number of

[2] Note, this was repeated for an downsampled dataset which produced slightly worse results.

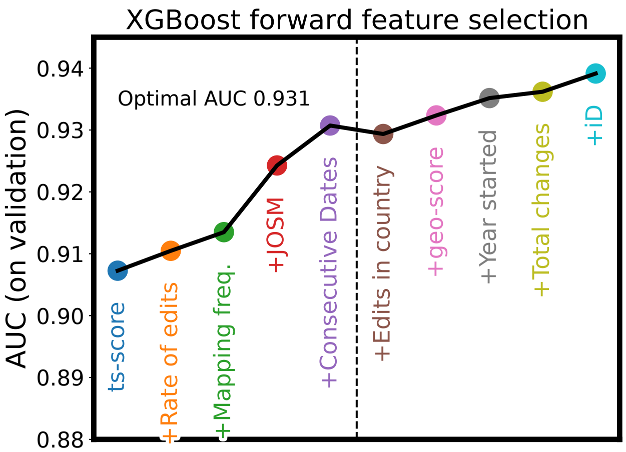

Figure 6: Forward feature selection for the XGBoost model. Features prior to the dashed vertical line were selected for the final model.

mappers. Additionally, mappers working on Humanitarian OpenStreetMap Team (HOT) tasks on a regular basis also match the corporate signature.

It is impossible to fully validate the employment status of each of these mappers as a majority of their user bios are empty. However, we did find mappers on this list that consistently submitted changesets with hashtags related to Apple, Digital Egypt, Grab, or Kaart. Some of these mappers are listed on the company's wiki page, but do not have employment information in their user bio. Additionally, this list of 320 users also included mappers from smaller companies whose presence in OSM was previously unknown to us.

Summary Statistics

OpenStreetMap is a massive repository of open geospatial data and an important study platform for social computing research into online collaboration. In February 2021, a member of the OSM community in Senegal contributed the 100 millionth edit to the OpenStreetMap project: 77 buildings and a water tower in Senegal. As the size of both the community and dataset continues to increase, so does its value[3]. In 2020, Accenture estimated the replacement value of OSM data to be $1.67 billion[4]. The number of private sector tech firms involved in the project and the diversity of products and services they develop using OSM data along with the contributions they make to the dataset are fast increasing. Therefore, it is important to understand these contributions in terms of their overall impact on the project, as well as for what they can tell us about the role of the private sector in online open source peer-production communities more generally. In this section, we draw on the predictions to present an updated assessment of patterns and trends in corporate contributions to OSM.

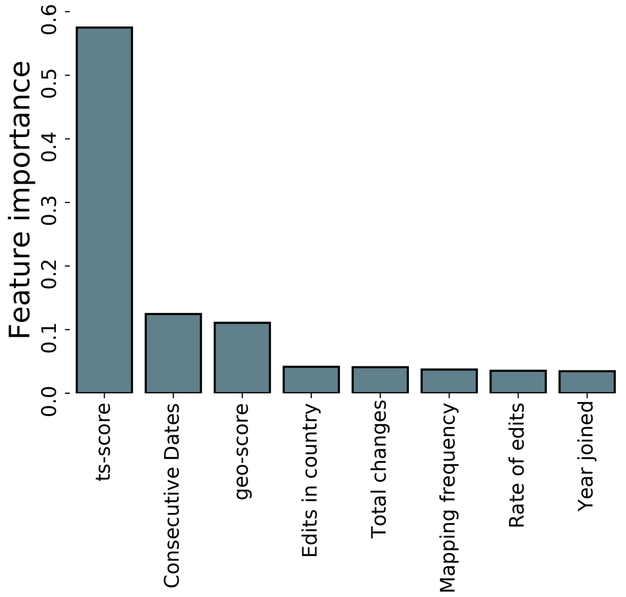

Figure 7: Absolute value of feature importance when training the XGBoost model.

	Any	Highways	Buildings	Amenities
Non-Corp	*0.867*	0.710	0.915	0.970
Corporate	*0.133*	0.290	0.085	0.030

Table 2: Proportion of specific features edited by corporate and non-corporate users.

Our results show that the corporate presence increased four-fold in the past few years, growing from representing about 5% of all edits in 2018 to representing nearly 20% of edits to date in 2021. The maps in Figure 8 show how the percentage of edits performed by corporate editors (out of all edits to OSM each year) has consistently increased each year. This growth has been driven by both a rapid escalation of corporate team sizes and the number of corporate teams.

One observation is that, in general, corporate editors tend to map different objects than non-corporate editors. For example, in Table 2 we show that while between 2018 and 2021 corporate mappers made only 13% of all edits to OSM, they were responsible for almost 30% of the 'highway' edits. In contrast, they represented only 8% of the buildings and 4% of the amenities mapped. However, there are some corporations with specific aims. For example, Digital Egypt aims to improve geocoding service in Egypt and has consequently edited more than 1.7M addresses in Egypt (Anderson and Sarkar 2020). When considering the growth in corporate mapping alongside the difference in what corporations map, we should expect the map to continue to evolve over the next several years with more and more roads being mapped and potentially fewer amenities and buildings. Prior work suggests that if the seeds by mapping roads are sown, then volunteer mappers will complete the map and fill in the details (Nagaraj 2021). Thus, corporate editing activ-

[3] blog.openstreetmap.org/2021/02/25/100-million-edits-to-openstreetmap/

[4] cupdf.com/document/how-to-make-the-most-of-openstreetmap-platform-accenture-how-to-make-the-most.html

ity in places with hitherto sparse information may spurn new community mapping activities in these locations.

Discussion

Features

This paper presents a novel methodology for studying user editing patterns on OSM. Using a public dataset of OSM changesets, we derived several metrics that help quantitatively differentiate corporate and non-corporate editing on the platform. First, corporate editors tend to follow a distinct and very uniform editing pattern, mapping on weekdays during business hours, between 9am and 5pm. Since OSM records editing timestamps in UTC, without including the timezone in which the editor is located, we developed a new method for recovering a user's timezone, which we then used to measure each user's distance to the corporate time signature. There exist new research directions that this approach opens up, like examining the often-stated assumption that corporate mappers map across time zones. By connecting the time zone of the editor with where the editor maps, we can measure the "locality" of the OSM edits. Additionally, given the often remote nature of corporate mapping (mappers not physically located in the region in which they are mapping), users are likely to map widely across the world instead of locally. For this, we developed a series of metrics to capture geographic dispersion/localization. Finally, we used the metadata associated with the changesets, like the comments and editor profiles to provide some context for how the edit was made. This method found, for example, that corporate mappers were likely to edit roads and disproportionately use the more advanced JOSM editor. Together, these features were sufficient to recall the vast majority of previously known corporate editors and to identify new editors that edit in an occupational manner that were not tracked by previous lists.

A key finding of this work is the strength of the temporal and geographic dispersion features. We have successfully captured a machine-readable signature that describes the working and engagement pattern of a distinct group of users. Each model we tried consistently found these features to be the strongest signals. While a more complex model like XGBoost performs better, it is not by an order of magnitude. Instead, in our analysis simple models tended to also fare well. We believe that this reflects the strength of the metrics we identified to characterize corporate editors. Additionally, these signals are not unique to corporate editing or OSM. Similar sets of metrics may also help identify contributions made by local vs remote mappers (a recurring challenge in OSM data analysis), or predict how and when to send encouragement to new mappers to keep them engaged in the platform. Furthermore, we believe that a similar approach could be applied to other online projects, such as Wikipedia, to differentiate between different user groups.

Compliance with Organized Editing Guidelines

The majority of edits made by corporate mappers appear to be in accordance with the organized editing guidelines (OEG). Our model successfully identified at least 25 corporate mappers who updated their profiles to be in compliance with the OEG since we created our training dataset. Furthermore, of the few hundred occupational and organized editors that the model successfully identified, a handful of editors are likely corporate but have not yet listed their employer in their bio. As more corporations and government bodies commit teams to contributing and maintaining data on OSM, these tools can help not only track new editors, but also to verify whether they are adhering to the OEG. For example, we identified one corporate mapping team that used the company logo as a means to note their affiliation in their bios instead of text, making it inadvertently out of compliance with OEG. Trust is a foundational requirement for community health in open source projects (Stephany, Braesemann, and Graham 2020), and reaching out to the editors unaware or out of compliance with the guidelines might alleviate concerns from being escalated needlessly. Further qualitative research into the specific strategies that the OSM community has used to successfully encourage good corporate behavior could help inform a broader research agenda on how online peer production and open source communities can best accomplish healthy working relationships.

Limitations

Features of Corporate Editing

This paper presents results of our effort to identify patterns in corporate editing between 2018 and 2021, a period during which corporate editing in OSM experienced rapid growth. As discussed, the features we have identified are strongly predictive of corporate and organized editing during this time period. However, it is likely that corporate editing practices have evolved over time, and will continue to as the community continues to grow and change. As a result, further work is necessary to evaluate and adjust as needed the extent to which the features presented are applicable to corporate editing behavior. It is possible that earlier, more nascent, forms of corporate editing practice were more experimental and varied, thus offering interesting challenges to the methods introduced here. In addition, work to "future proof" this analysis to ensure that the approach presented here stays current with corporate and other forms of organized editing behavior as they continue to evolve will also be required. For example, recent research shows that features edited by a corporate mapper is often edited soon after by another corporate editor, probably reflecting internal workflows (Sarkar and Anderson 2022). Thus, looking at the edit history of OSM objects in addition to editing patterns might improve the model.

Mining Corporate Editors

The first task of this paper was to define a set of known corporate mappers to act as reference when designing features and as a supervised dataset for prediction. Our approach was based on the observation that many corporate mappers publicly disclosed their association on their public page. This provided an initial list of corporate mappers, but does not

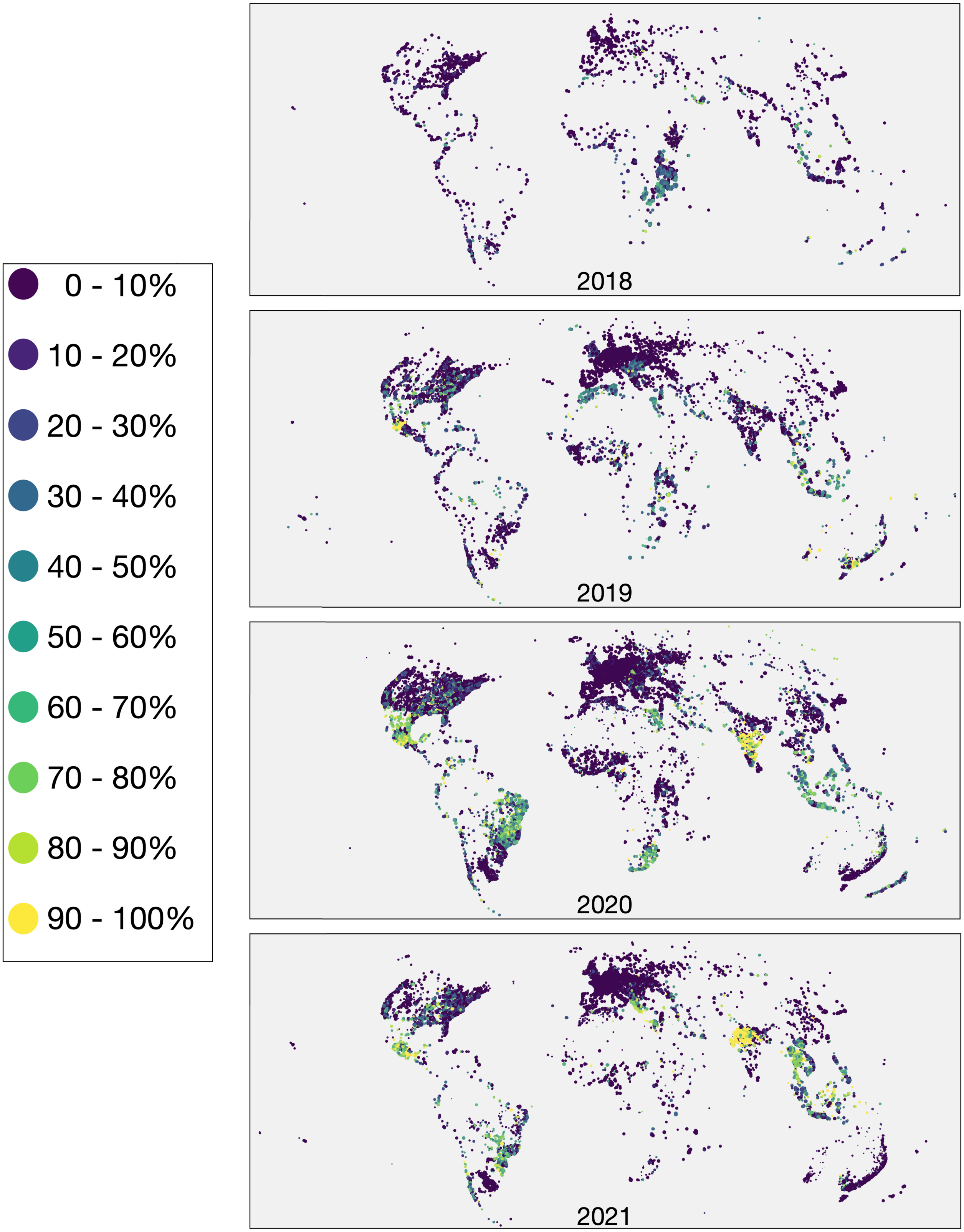

Figure 8: Percent of edits performed by corporate editors each year, aggregated at zoom-level 10 map tiles where radius is log-scaled by total number of edits to highlight areas with the most activity.

account for all possible ways mappers can signal their corproate affiliation, as we saw in the analysis of newly predicted corporate mappers. Future work can expand on our method by including hashtags and logos in their initial determination of corporate mappers. Moreover, a positive and unlabeled classification algorithm could also be adopted to account for the unlabelled corporate editors that may have been missed in this scrape. This method was not explored in this paper, but might provide a new approach for studying corporate mapping.

Corporate Contributions to OSM Besides Editing

Corporations have been involved in OSM since the early days of the project: sponsoring conferences, supporting local meetups, providing datasets, and developing tools for OSM (Anderson, Sarkar, and Palen 2019). The broad range of corporations taking an interest in OSM accelerates the evolution of the platform. Apple and Facebook, for example, list OSM as a data source for maps displayed in their products, creating a self-explanatory interest in OSM. As of 2020, ESRI, the geospatial software giant, has been releasing spatial datasets that users may easily upload to OSM. Additionally, Google has recently made a dataset of buildings extracted from satellite imagery available for import into OSM. These two geospatial giants are thereby involved in OSM, but not hiring teams to edit the data directly, highlighting the myriad ways of participation. Thus, corporate involvement in OSM is a vibrant and evolving dynamic requiring further enquiries from different perspectives.

Other Forms of Organized Editing

Historically, the OSM literature has highlighted the important role played by local volunteer communities and independent hobbyists. However, a significant and growing percentage of mapping is conducted by organized groups performing specific types of edits. In addition, to corporate mappers in this study, other research has tracked the growth of other organized editing communities, such as the Humanitarian OpenStreetMap Team (HOT) (Palen et al. 2015) that works on issues related to disaster or international development, or Youth Mappers, who organize local chapters of university students to contribute to OSM (Solís, Anderson, and Rajagopalan 2020). Still other research has noted the adoption of OSM by local governments, and the increasing contributions to the project made by government employees (Khan and Johnson 2020). ESRI, another technology corporation with significant interests in the project, is developing a set of tools that may further support these activities. As OSM adoption increases across these sectors, importance of organized editing, including by full-time and professional mappers will continue to grow. Further methods are needed to understand these forms of organized editing made by editing teams beyond those hired by corporations.

Conclusion

Tracking the contributions of corporate mappers to the OpenStreetMap is an important, though increasingly difficult task. As these contributions continue to grow, prior approaches based on manually managing lists of usernames reported by individual corporations (Anderson, Sarkar, and Palen 2019) are increasingly untenable. The approach we have outlined in this paper offers both researchers and the OSM community a more automated approach to identifying organized, occupational, and corporate editors. Based on our method, we have provided an updated understanding of the overall scope and contours of corporate mapping. Furthermore, the features we have identified will likely prove useful for answering other questions about OSM editing behavior such as evaluating the success of various approaches to recruiting new mappers or identifying contributions made through university coursework. Finally, this work contributes to ongoing debates in social computing research around algorithmic means of enforcing policy and on-boarding newcomers in peer production projects (Halfaker et al. 2013).

The current OEG are broad enough to encompass various forms of coordinated editing and thus not specifically designed for corporate editing. This makes it particularly challenging to track only edits performed by corporate mappers. Tracking these edits is important to understand the impact these activities can have. Our automated detection approach highlights that, within the current ambit of organized editing, corporations are largely adhering to the guidelines. The similarity in machine identifiable features in the editing pattern of corporations and other organized editors (e.g. HOT, government agencies) point to the similarities in volume, as well as temporal, object, and geographic patterns of editing amongst the groups. Thus, paid corporate editor behaviour is not dissimilar to other professional mappers or ardent hobbyists. At this juncture, OSM represents a critical data infrastructure. Geospatial data on the platform is of acceptable quality for a range of applications and in many places, it is the only form of geospatial data available. It is therefore unsurprising that there are many groups of organized editors, some of whom are professional editors — getting remunerated for their work (e.g. government employees), and some of the professional mappers are affiliated with corporations. If paid editing is a concern and if the larger OSM community is to develop mechanisms for tracking such paid editing, then a separate set subset of guidelines may be necessary to account for professional editors.

References

Anderson, J.; and Sarkar, D. 2020. Curious Cases of Corporations in OpenStreetMap. In Minghini, M.; Juhász, L.; Yeboah, G.; Mooney, P.; and Grinberger, A. Y., eds., *Proceedings of the Academic Track, State of the Map 2020*. Online Conference. ZSCC: 0000002 Library Catalog: Zenodo Publisher: Zenodo.

Anderson, J.; Sarkar, D.; and Palen, L. 2019. Corporate Editors in the Evolving Landscape of OpenStreetMap: A Close Investigation of the Impact to the Map & Community. In *Proceedings of the Academic Track at the State of the Map 2019*, 17–18. Heidelberg, Germany.

Anderson, J.; Soden, R.; Anderson, K. M.; Kogan, M.; and Palen, L. 2016. EPIC-OSM: A software framework for OpenStreetMap data analytics. In *2016 49th Hawaii In-*

ternational Conference on System Sciences (HICSS), 5468–5477. IEEE.

Anderson, J.; Soden, R.; Keegan, B.; Palen, L.; and Anderson, K. M. 2018. The Crowd is the Territory: Assessing Quality in Peer-Produced Spatial Data During Disasters. *International Journal of Human-Computer Interaction*, 34(4): 295–310. Publisher: Taylor & Francis.

Barron, C.; Neis, P.; and Zipf, A. 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, 18(6): 877–895. ArXiv: cs/9605103 ISBN: 0-7803-3213-X.

Bellman, R.; and Kalaba, R. 1959. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2): 1–9.

Budhathoki, N. R.; and Haythornthwaite, C. 2013. Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist*, 57(5): 548–575. ISBN: 0002-7642.

Burns, R.; and Meek, D. 2015. The politics of knowledge production in the geoweb. *ACME: An International Journal for Critical Geographies*, 14(3): 786–790.

Chapman, K. 2018. Organised Editing Guidelines. *OpenStreetMap Foundation*, (November): 1–4.

Cipeluch, B.; Jacob, R.; Winstanley, A.; and Mooney, P. 2010. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences*.

Foundation, O. 2021. 2021 Survey Results - OpenStreetMap Foundation.

Gilbert, M. 2010. Theorizing digital and urban inequalities: Critical geographies of 'race', gender and technological capital. *Information, communication & society*, 13(7): 1000–1018.

Girres, J. F.; and Touya, G. 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4): 435–459. ISBN: 1361-1682.

Graham, M.; Hogan, B.; Straumann, R. K.; and Medhat, A. 2014. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4): 746–764.

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

Halfaker, A.; Geiger, R. S.; Morgan, J. T.; and Riedl, J. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5): 664–688.

Jacobs, K. T.; and Mitchell, S. W. 2020. OpenStreetMap quality assessment using unsupervised machine learning methods. *Transactions in GIS*, 24(5): 1280–1298. ZSCC: 0000001 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12680.

Johnson, P. A. 2017. Models of direct editing of government spatial data: challenges and constraints to the acceptance of contributed data. *Cartography and Geographic Information Science*, 44(2): 128–138.

Khan, Z. T.; and Johnson, P. A. 2020. Citizen and government co-production of data: Analyzing the challenges to government adoption of VGI. *The Canadian Geographer / Le Géographe canadien*, 64(3): 374–387.

Madubedube, A.; Coetzee, S.; and Rautenbach, V. 2021. A Contributor-Focused Intrinsic Quality Assessment of OpenStreetMap in Mozambique Using Unsupervised Machine Learning. *ISPRS International Journal of Geo-Information*, 10(3): 156. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Mooney, P.; and Corcoran, P. 2012. How social is OpenStreetMap. *Proceedings of the 15th Association of AGILE 2012 International Conference on Geographic Information Science*, 24–27. ISBN: 9789081696005.

Mooney, P.; Corcoran, P.; and Winstanley, A. C. 2010. Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10*, 514. ISBN 978-1-4503-0428-3. ISSN: 00219258.

Nagaraj, A. 2021. Information seeding and knowledge production in online communities: Evidence from openstreetmap. *Management Science*.

Palen, L.; Soden, R.; Anderson, J.; and Barrenechea, M. 2015. Success & Scale in a Data-Producing Organization. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 4113–4122. New York, New York, USA: ACM Press. ISBN 978-1-4503-3145-6.

Pourabdollah, A.; Morley, J.; Feldman, S.; and Jackson, M. 2013. Towards an Authoritative OpenStreetMap: Conflating OSM and OS OpenData National Maps' Road Network. *ISPRS International Journal of Geo-Information*, 2(3): 704–728.

Quinn, S. 2017. Using small cities to understand the crowd behind OpenStreetMap. *GeoJournal*, 82(3): 455–473. Publisher: Springer Netherlands ISBN: 1572-9893.

Sarkar, D.; and Anderson, J. 2021. Community interactions in OSM editing. Conference Name: State of the Map 2021 (SotM 2021) Publisher: Zenodo.

Sarkar, D.; and Anderson, J. T. 2022. Corporate editors in OpenStreetMap: Investigating co-editing patterns. *Transactions in GIS*, n/a(n/a).

Shelton, T.; Poorthuis, A.; Graham, M.; and Zook, M. 2014. Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum*, 52: 167–179.

Soden, R.; and Palen, L. 2014. From Crowdsourced Mapping to Community Mapping: The Post-earthquake Work of OpenStreetMap Haiti. In *COOP 2014 - Proceedings of the 11th International Conference on the Design of Cooperative Systems, 27-30 May 2014, Nice (France)*, 311–326. ISBN 978-3-319-06497-0. ISSN: 0007-6813.

Solis, P. 2017. Building mappers not just maps: challenges and opportunities from YouthMappers on scaling up the crowd in crowd-sourced open mapping for development. In *Annual Meeting of the Association of American Geographers*. Boston, Massachusetts.

Solís, P.; Anderson, J.; and Rajagopalan, S. 2020. Open geospatial tools for humanitarian data creation, analysis, and learning through the global lens of YouthMappers. *J Geogr Syst*.

Stephany, F.; Braesemann, F.; and Graham, M. 2020. Coding together–coding alone: the role of trust in collaborative programming. *Information, Communication & Society*, 1–18.

Stephens, M. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal*, 78(6): 981–996.

Sui, D.; Goodchild, M.; and Elwood, S. 2013. Volunteered geographic information, the exaflood, and the growing digital divide. In *Crowdsourcing geographic knowledge*, 1–12. Springer.

Yang, A.; Fan, H.; and Jing, N. 2016. Amateur or Professional: Assessing the Expertise of Major Contributors in OpenStreetMap Based on Contributing Behaviors. *IJGI*, 5(2): 21.

Yasseri, T.; Sumi, R.; and Kertész, J. 2012. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one*, 7(1): e30091.