An automated approach to identifying corporate editing activity in OpenStreetMap

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

[1] Department of Computer Science, University of Toronto, Toronto, Canada; venia@cs.toronto.edu, soden@cs.toronto.edu
[2] Geography and Environmental Studies, Carleton University, Ottawa, Canada; diptosarkar@cunet.carleton.ca
[3] Department of Computer Science, University of Colorado Boulder, Boulder, USA; jennings.anderson@colorado.edu

* Author to whom correspondence should be addressed.

This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

In the past five years, the OSM community has seen a dramatic rise in organized editing, including corporate, humanitarian, and educational, on the platform. These new actors have continued the ongoing debate surrounding OSM's relationship with organized editing, with new rules and best-practices being implemented to align the interests of the organizations with those of the community.

We became interested to study how the editing habits of these new actors differed from the community as a whole, but were quickly confronted by the challenge of producing accurate measures of their activities. In this paper we aim to fill this gap by creating computational methods of understanding different editing behaviours on OSM to classify editors as being corporate or volunteer. Classifying individual editors has been done in the past, on a more local level, for example in the recent analysis on editing in Mozambique. [1]

Studying corporate editing behaviour, first requires a list of corporate editors. In the past, researchers have searched individual "organized editing team" webpages. Instead, our paper presents a novel method for classifying users on the platform, by scraping user profiles. There are two possible approaches to extract corporate mappers based on user profiles. The first approach uses a clustering of the keywords within the profiles. Though effective at uncovering relations between users (like students, programmers, Garmin editors, Colorado mappers), this method failed to properly capture all known corporate groups. Instead we did a keyword search for corporations listed on the Organized Editing List and classified similar users together. This included a list of 2,177 known corporate mappers with over 50 unique changesets.

Using this extracted list, we discern features that could act as "signals" for organized editors. Explicitly, which features from the changesets can point to an editor being corporate or volunteer. Do corporate editors edit specific types of items? Do their time series signatures differ?

Veselovsky, V., Sarkar, D., Anderson, J. & Soden, R. (2021). An automated approach to identifying corporate editing activity in OSM In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 31-33. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112248

For the creation of these features, we relied on Jennings Anderson's past work on corporate editing for inspiration [2]. The first set of features came from OSM changeset metadata which is rich with user descriptive data like the editor used, comments, and source. We find that most organizations use editors like JSOM and iD. Next, we attempted to model which objects corporations edit by finding descriptive words like "service", "road", and "building" in the comments of the changeset. We observed that most corporations focus on services and roads, as opposed to buildings which tend to be dominated by volunteer mappers.

The third feature was motivated by the observation that as the interests of a corporation change, the editing of its mapping team can also change. This has led to the documented phenomena of corporate mappers having a geographically dispersed editing pattern. This is markedly different from many volunteer mappers who often begin by mapping their local neighbourhoods. Using established metrics, we calculated the geographic dispersion for each user based on the latitude and longitude of their edits.

The metric we found most effective was the timeseries signature. Corporations have a traditional 9-5 mapping schedule, whereas non-corporate mappers tend to map far more haphazardly, including significant mapping on the weekend. When attempting to convert the time series signature into a usable metric, we came across a problem: time zones. All changesets in OSM are normalized to UTC time, this means that a user editing at 8am in Toronto, Canada and another user editing at 8pm in Beijing, China would in fact appear to be editing at the same time in OSM. Longitude and latitude data are not an effective method of extracting the mapper's time zone, since editing on OSM is increasingly done remotely, through "armchair mapping".

To utilize this strong signal, we developed a new method for normalizing a user's time signature, and it was based on the observation that individual corporations have several key editing patterns, depending on where their employees are located. For example, Facebook has two such patterns, each displaced by around 8 hours. This motivated us to create a "corporate editing signature" and translate user time signatures to find the minimal distance between the two. After using this method of adjustment, we were able to significantly improve the alignment of the time-series. In other words, we were able to recover the local time zone of most of these corporate editors. Figure 1 illustrates corporate mappers before and after adjustment.

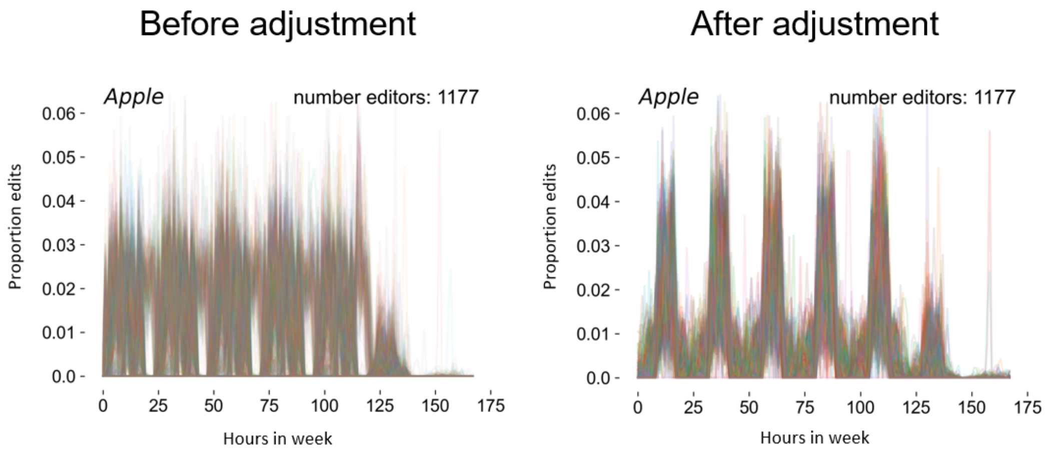

Figure 1. This plot shows how corporate time zones were recovered after minimizing distance between corporate actors and a "corporate mapping signature".

Once we realigned each user using this method, we calculated the distance between a user's adjusted time signature and the “corporate signature”. This feature ended up acting as a key determinant of the likelihood of a given editor being corporate. Out of the top 100 editors (who had the smallest distance to the corporate signature) all of them belonged to corporations.

Utilizing the user features we predict whether an editor is corporate or not. We experimented with several classification algorithms, including logistic regression, k-nearest neighbours, support vector machines, and neural networks. The four most important features in the prediction task, ordered by impact on model, were the geographic dispersion, time series score, first edit date, and the editor type. All models provided comparable results offering a high recall of 96%+ and predicting anywhere between 700 to 2,000 additional corporate mappers. Examining the newly predicted mappers reveals users that map for humanitarian groups like HOT, corporate mappers that the initial scrape didn’t pick up on, corporate mappers who reveal their association only in the hashtags, users who are likely corporate mappers with no ability to know for certain, and volunteers. After removing any “predicted mappers” who have known humanitarian associations from the most conservative model we arrived at a list of 500 newly identified corporate mappers. We are now entering the stage of further validating the different models based on a manually annotated set of users that any of the models predicted to be corporate.

References

[1] Madubedube, A., Coetzee, S., & 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.

[2] Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate Editors in the Evolving Landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 8(5), 232.