What has machine learning ever done for us?

Peter Mooney[1,*] and Edgar Galvan[1]

[1] Naturally Inspired Computation Research Group, Department of Computer Science, Maynooth University, Maynooth, Ireland; peter.mooney@mu.ie, edgar.galvan@mu.ie

* 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.

Recently, machine learning (ML) based approaches have been applied frequently to many different types of problems in OpenStreetMap (OSM). Indeed, ML approaches have been used extensively by the research community for a plethora of applications and problems both related and unrelated to OSM. Wagstaff suggests ML offers "a cornucopia of useful ways to approach problems which defy manual solutions" [1]. In specific relation to the geospatial domain, ML approaches have been reported for at least the last two decades with the remote sensing community being particularly active in ML usage. The number of works appearing around ML in the geospatial domain began to noticeably increase around a decade ago with work by authors such as Werder et al. [2] on interpretation of buildings in settlements and detecting road intersections from GPS traces by Fathi and Krumm [3]. Around this time interest in the combination of ML and OSM began to emerge. Funke et al. argued that many aspects of OSM data might be suitable for "extrapolation or classification using ML" [4]. Many examples have emerged with ML approaches being used to consider problems such as: predicting or recommending tagging for objects, object classification based on contextual or proximity information, tag usage checking, automated mapping approaches etc. Anderson et al. showed that Facebook's recent mapping campaign in OSM used ML to detect road networks from satellite imagery which are then validated by OSM editors and the local OSM communities [5]. Examples also exist where OSM is used in ML approaches for other geospatial classification problems while authors such as Feldmeyer et al. used machine and deep learning algorithms with OSM for developing socio-economic indicators [6]. Audebert et al. [7] argued that OSM's richness means it can be used in difficult problems such as semantic labeling of aerial and satellite images.

In addition to the observations by Vargas-Munoz et al. [8] in their recent review of ML approaches in OSM we can usually observe ML and OSM interaction in one of two ways: (1) ML approaches are used to improve the quality and coverage of OSM layers by using GIS and Remote Sensing and (2) instances where OSM layers are used as a means of training ML models for some specific task such as building segmentation, population estimation, navigation or land use classification. In this research, we ask the following question: With all of the many applications and integration of ML with OSM over the past number of years, how many of these applications and approaches have been adopted or used by the OSM community? Furthermore, what are the benefits or impact of these efforts from the research community with ML approaches to the OSM project and OSM community?

Mooney, P. & Galván, E. (2021). What has machine learning ever done for us?
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, 9-12. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112219

We performed a systematic review of approximately 50 peer-reviewed academic journal and conference papers. These papers are selected on the following basis: the paper(s): (1) clearly outlines an ML approach using OSM data and (2) tackle a problem known in the OSM community such as tag prediction, contribution patterns, or geometry correction. We used the Google Scholar search engine to retrieve the papers for review. Paper metadata such as title, keywords, and abstract contents were used to select the papers. We processed the results in linear fashion, as returned by the Google Scholar search, and selected the first 50 papers. This included a manual check of all the papers to ensure that the content of each paper related to our selection criteria. From the initial search results, five of the papers were either literature review papers or used OSM as just a background map layer in visualisations. These were replaced by the next five qualifying papers in the search results. The GitHub repository at [9] contains the links to all papers and our classifications. We acknowledge that this sampling is far from representative of the whole field. With greater resources (search and analysis time, research availability), a much larger sample of papers could be assembled. These papers were analysed using the following set of questions for guidance:

- What are the most common ML approaches used by researchers for the three instances outlined above?
- What are the most common types of problems in OSM tackled by ML approaches?
- Are the approaches reproducible and replicable by others?
- What, if any, is the awareness and/or understanding of the OSM project or community outlined in the paper?

Using these questions, we now report a narrative on our findings on the benefits and impacts of these efforts to the OSM project and the OSM community. We do not evaluate the results of the papers such as analysis of the accuracy of ML approaches nor do we advocate a specific ML approach. Both tasks are outside the scope of this work.

- In Vargas-Munoz et al. [8] one interaction of ML and OSM is to improve the quality and coverage of OSM. We found that 23 papers (46%) could be classified as displaying this type of approach. Furthermore, here we attempted to group papers into different types of coverage and quality tasks for OSM. We used our own classification and allowed multiple selections. From these 23 papers we found that: 3 papers dealt with contribution patterns in OSM, 9 papers considered data quality issues in general, 14 papers were connected to annotation and tagging in OSM while the remaining 10 papers dealt with topological issues and geometric analysis (correction, alignment, creation, and other tasks)
- Also in Vargas-Munoz et al. [8] the other interaction of ML and OSM is to use OSM data to train models and serve applications. We found that 31 papers (62%) used this approach. We believe that 2 papers shared both classifications. Here, as before, we attempted to group papers into different types of application domains. We created our own classification as follows: Navigation and transportation (7 papers), use of OSM to generate ML training datasets (6 papers), Socio-economic analysis (6 papers), image analysis (3 papers) and miscellaneous (5 papers). We found great diversity in these applications with examples taken from areas such as multilane road extraction, electric vehicle routing, training datasets for urban areas, population estimation, air quality forecasting and image labelling using OSM.

- OSM community understanding. Grinberger et al. [10] explored interactions between the academic and mapping communities in OSM. We have based our classifications on this work with 5 classes representing the understanding of OSM as outlined in the academic paper. Multiple selections were allowed for each paper. The five classes used are: OSM as a data source (25 papers, 50%), OSM as a data source produced by contributors (25 papers, 50%), OSM as a social data product (14 papers, 28%), or no understanding or perception of OSM specified (25 papers, 50%). We acknowledge that this particular aspect of the review is very subjective and reflects our personal interpretation. We also understand that authors may discuss their understanding of the OSM community in different ways with the constraints of the paper.
- There was great variation in the types of machine learning approaches used in these works. Random forests are by far the most popular but many other approaches are also implemented: ensemble ML, convolutional neural networks, logistic regression, supervised and unsupervised clustering, support vector machines, AdaBoost, Boosting, Latent Dirichlet Allocation and others.
- Finally, we considered the reproducibility and replicability of the studies in each of the papers. Would it be possible for other scientists in the academic community or members of the OSM community to reproduce or replicate the work in a given paper? We decided to classify reproducibility and replicability in three ways: red (appears very difficult to reproduce or replicate for reasons related to proprietary software usage, lack of description, etc); amber (appears that reproduction and replication is possible but access to certain APIs or datasets is required) and green (reproduction and replication appears to be a priority in the work). After our review, we found: red (8 papers, 16%), amber (29 papers, 58%) and green (13 papers, 26%). Grippa et al [11] and Zurbaran et al [12] are two excellent examples of papers with reproducibility and replicability as priority issues.

As stated previously, we acknowledge the subjective nature of these results and our Github repository will allow others to consider our classifications. There is incredible diversity, even within this small sample, of applications of ML with OSM. We believe that the ability to integrate OSM data with other data sources greatly adds to these opportunities. At this stage in this work, it is difficult to ascertain the level of adoption or indeed opportunities for adoption of published ML approaches by the OSM community. We were disappointed to see that in our evaluation 50% of the papers surveyed did not clearly indicate any understanding or connection to the OSM community. OSM provides incredible value to ML research. Indeed, we believe it is impossible to estimate the value of OSM as a data source to ML research and studies. In a similar way, and as shown above, many ML approaches do lend themselves very well as potential candidates for implementation by the OSM community. We conclude that it still remains difficult to understand how all of this ML could contribute effectively to the OSM database and OSM community without improved interactions between both communities.

We must not get carried away with the combination of ML and OSM purely for the sake of it. OSM, as a massive open geospatial database, is a very attractive source of (geo-)data for researchers and practitioners looking to train, benchmark and test ML approaches. Consequently, we can confidently state that, after well over a decade of reported results in this domain, researchers have produced many excellent research and knowledge outputs using the ML and OSM combination. Now is a good time to ask, as we

have done in this paper, how could all of this ML knowledge contribute effectively to the OSM database and OSM community. Grinberger et al. [10] argue that efforts to establish and strengthen interaction between the research community interested in working with or in OSM and the OSM community itself have generally been positive. However, opportunities exist to enhance interactions between these two communities and perhaps ML could be the catalyst for a new interaction. Based on this the scientific contribution of this work is multi-faceted. Firstly, this paper will stimulate debate about the contribution of these ML approaches to the improvement of OSM data and enhancement of the OSM community. Secondly, this work will highlight situations where these ML approaches have delivered genuinely new and novel outputs of interest to OSM in general. Finally, this work will issue the challenge to the academic community to apply ML to several interesting and open problems which are of mutual interest to both the academic and OSM community, returning in kind the considerable impact OSM has had on this academic field.

References

[1] Wagstaff, K. (2012). Machine learning that matters. *arXiv preprint*, arXiv:1206.4656.
[2] Werder, S., Kieler, B., & Sester, M. (2010). Semi-automatic interpretation of buildings and settlement areas in user-generated spatial data. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 330-339.
[3] Fathi, A., & Krumm, J. (2010). Detecting road intersections from GPS traces. In: *International conference on geographic information science*, 56-69. Berlin, Heidelberg: Springer.
[4] Funke, S., Schirrmeister, R., & Storandt, S. (2015). Automatic extrapolation of missing road network data in OpenStreetMap. In: *Proceedings of the 2nd International Conference on Mining Urban Data*, 1392, 27-35.
[5] Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate editors in the evolving landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 8(5), 232.
[6] Feldmeyer, D., Meisch, C., Sauter, H., & Birkmann, J. (2020). Using OpenStreetMap Data and Machine Learning to Generate Socio-Economic Indicators. *ISPRS International Journal of Geo-Information*, 9(9), 498.
[7] Audebert, N., Le Saux, B., & Lefèvre, S. (2017). Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 67-75.
[8] Vargas-Munoz, J. E., Srivastava, S., Tuia, D., & Falcao, A. X. (2020). OpenStreetMap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 9(1), 184-199.
[9] Mooney, P., & Galvan, E. (2021). What has machine learning ever done for us?. Retrieved from https://github.com/petermooney/sotm2021
[10] Grinberger, A. Y., Minghini, M., Juhász, L., Mooney, P., & Yeboah, G. (2019). Bridging the map? Exploring interactions between the academic and mapping communities in OpenStreetMap. In: Minghini, M., Grinberger, A. Y., Juhász, L., Yeboah, G., & Mooney, P. (Eds.) *Proceedings of the Academic Track at the State of the Map 2019*, 1-2.
[11] Grippa, T., Georganos, S., Zarougui, S., Bognounou, P., Diboulo, E., Forget, Y., Lennert, M., Vanhuysse, S., Mboga, N., & Wolff, E. (2018). Mapping Urban Land Use at Street Block Level Using OpenStreetMap, Remote Sensing Data, and Spatial Metrics. *ISPRS International Journal of Geo-Information*, 7(7), 246.
[12] Zurbaran, M. A., Wightman, P., & Brovelli, M. A. (2019). A machine learning pipeline articulating satellite imagery and OpenStreetMap for road detection. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W14, 255-260.