Exploring the Potential of Machine and Deep Learning Models for OpenStreetMap Data Quality Assessment and Improvement

Salim Miloudi[1]

Spatial Reference Information Systems Department, Space Techniques Center, Oran, Algeria

Bouhadjar Meguenni

Spatial Reference Information Systems Department, Space Techniques Center, Oran, Algeria

Abstract

The OpenStreetMap (OSM) project is a widely-used crowdsourced geographic data platform that allows users to contribute, edit, and access geographic information. However, the quality of the data in OSM is often uncertain, and assessing the quality of OSM data is crucial for ensuring its reliability and usability. Recently, the use of machine and deep learning models has shown to be promising in assessing and improving the quality of OSM data. In this paper, we explore the current state-of-the-art machine learning models for OSM data quality assessment and improvement as an attempt to discuss and classify the underlying methods into different categories depending on (1) the associated learning paradigm (supervised or unsupervised learning-based methods), (2) the usage of extrinsic or intrinsic-based metrics (i.e., assessing OSM data by comparing it against authoritative external datasets or via computing some internal quality indicators), and (3) the use of traditional or deep learning-based models for predicting and evaluating OSM features. We then identify the main trends and challenges in this field and provide recommendations for future research aiming at improving the quality of OSM data in terms of completeness, accuracy, and consistency.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases OpenStreetMap (OSM), Volunteered Geographic Information (VGI), Machine Learning (ML), Deep Learning (DL), Quality Assessment (QA), Building Footprint Detection, Semantic Segmentation

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.53

Category Short Paper

Acknowledgements We would like to thank the anonymous reviewers for their valuable comments.

1 Introduction

The OpenStreetMap (OSM) project [2] is a collaborative effort to create a free, editable map of the world. The OSM database is built and maintained by a community of volunteers who contribute data on various geographical features such as roads, buildings, and points of interest. For this purpose, there are various editors that can be used to edit OSM data, including web-based editors such as iD and Potlatch [3], and desktop editors such as JOSM [4] and Merkaartor [5]. Each editor has its own set of features and tools, making them suitable for different types of mapping tasks. For example, JOSM is a powerful editor that has a wide

[1] Corresponding author
[2] `https://www.openstreetmap.org`
[3] `https://www.systemed.net/potlatch/`
[4] `https://josm.openstreetmap.de/`
[5] `http://merkaartor.be/`

12th International Conference on Geographic Information Science (GIScience 2023).
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 53; pp. 53:1–53:6
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

range of advanced features and is suitable for experienced mappers, while iD is a web-based editor that is easy to use and is suitable for beginners. Additionally, some editors like JOSM have plugins that can automate certain tasks, such as checking for errors in the data.

The importance of OSM data lies in its wide range of applications. It is essentially used in many fields such as navigation, emergency response, transportation and urban planning. It is also used to create custom maps for specific needs, such as hiking and biking maps, and is used as a base map for many other applications. Besides, OSM data can be used to create map tiles and other map products that can be used on websites and mobile applications. Despite its usefulness and its reliability, the quality of OSM data is strongly dependent on the accuracy and completeness of the contributions (edits or changests) made by volunteers. In fact, as the OSM database continues to grow, the need for automated methods to assess and improve its data quality becomes increasingly important.

On the other hand, numerous machine and deep learning models have been applied to various different tasks in the area of GIS (geographic information systems) and web-mapping, including map digitization using features generated by artificial intelligence (AI) predictions. For instance, one could extract building footprint binary masks from drone imagery via different deep learning segmentation models, transform those masks into georeferenced polygons, and then overlay those geometries on OSM base-map for quality assessment. This will allow us to build AI tools capable to assist the mappers detect incomplete regions and vandalism cases when there is a mismatching between the predicted features and the existing annotations created by the contributors within a certain area of interest (AOI) in OSM.

In this paper, we review some of the state-of-the-art machine learning models for OSM data quality assessment while describing the proposed approach and the important findings for each work.

2 Machine Learning Models for OSM Data Quality Assessment

Mapping systems are crucial for navigation, transportation and other applications, but they can be costly to maintain due to the need for regular updates. Traditional maps, also known as authoritative maps, may not be updated as frequently due to budget constraints and may result in inaccuracies in terms of temporal, spatial and completeness. An alternative solution is Volunteered Geographic Information (VGI) [9], which relies on the contributions of individuals to create and update maps. One of the most popular VGI projects is OpenStreetMap (OSM), which was launched in 2004 and currently has over 10 million users from around the world. OpenStreetMap (OSM) is a VGI project which serves as an alternative to traditional map sources and is open to the public for retrieving, adding and editing spatial features. While OSM data is constantly being improved, the completeness and quality of the data may vary depending on the number of contributors and their mapping skills [19]. For instance, OSM coverage is more or less complete in urban areas compared to rural areas [10]. Additionally, it is not uncommon to encounter missing roads and inaccuracies in terms of positional accuracy [28] and semantic tags [6, 14].

Despite these limitations (i.e., issues related to data completeness and its quality), OSM has been widely used in a variety of applications, including land cover mapping and classification [5, 25, 3, 26], navigation (e.g., traffic estimation) [16], 3D city modeling and location-based services [22], building footprint detection using aerial imagery [27, 18], location-based map services [30] and indoor mapping [8].

To evaluate and improve the quality of OSM data, researchers have proposed various methods to tackle issues related to completeness [15], positional accuracy [4], semantic tag accuracy [7] and topological consistency [21]. Other works approached OSM data quality assessment [24, 13] by performing OSM meta-analysis, such as examining the activities of the contributors [20, 2].

In recent years, there has been an increasing interest in automating tasks related to OSM data. In fact, numerous works have used machine learning and remote sensing techniques to improve OSM data, while deep learning [29] has been used to extract information from OSM data to train image recognition models. Overall, the combination of machine learning, earth observation and OSM data has the potential to address global challenges in new ways. Several supervised machine learning based models have been trained on properties of OSM objects to find potential annotation errors. The authors in [1] have proposed three different machine learning based approaches to identify errors (inconsistent tags) in OSM object annotations. The first approach, *consistency checking*, involves applying a classifier while the user is editing and assigning tags to OSM objects. In this case, the editing tool can inform the volunteer if the assigned tag value is inconsistent with what the classifier predicted. Usually, geometrical, topological, and contextual properties (e.g., the object area) are used to train the supervised learning classifier. The second approach, *manual checking*, consists of applying a supervised classifier on a selected set of objects from OSM and then having OSM users to manually validate the objects whose tags present inconsistencies with the predictions of the classifier. The third approach, *automatic checking*, involves using a classifier to automatically correct annotations based on its predictions without human verification.

Conventional methods typically compare Volunteered Geographic Information (VGI) against an *authoritative* dataset to evaluate the quality of VGI data such as OSM data. While authoritative data is generated by official organizations, VGI is contributed voluntarily by individuals or communities. Also, VGI can be less reliable and accurate due to varying quality and expertise, while authoritative data is more trusted. In addition, VGI is more dynamic but lacks consistent quality control, can have biases, legal concerns, and sustainability challenges. Despite the previous limitations, combining VGI with authoritative or reference data is recommended to tackle the aforementioned shortcomings.

In cases where reference data is unavailable to assess the quality of OSM data, intrinsic methods that evaluate the data itself and its metadata can be employed. The study described in [17] utilizes unsupervised machine learning (k-means clustering algorithm) to analyze OSM history data in Mozambique, aiming to gain insights into the contributors, their timing, and their contributions. The results obtained from the analysis showed that a majority of the data in Mozambique (93%) was contributed by a small group of active contributors (25%). The study also identified a new category of contributors who were newcomers to the area, likely attracted by HOT mapping events during disaster relief operations in Mozambique in 2019. While intrinsic methods cannot serve as a substitute for ground truthing or extrinsic methods, they offer alternative means of gaining insights into data quality and can contribute to efforts aimed at enhancing it.

The study presented in [12] takes a similar approach by examining the OSM database in Ottawa-Gatineau. The focus of the investigation is on historical map features and contributor data to understand how accurately users contribute to the OSM database. To classify the changesets and OSM contributors, two unsupervised machine learning models, namely K-means and Principal Component Analysis (PCA), are utilized. The findings reveal a cluster of skilled contributors identified as OSM experts, based on their strong contribution loadings related to the use of advanced OSM editors, and weaker loadings associated with feature creation and frequency of contributions resulting in further correction. Therefore, attributing data quality is done by identifying experienced contributors who are likely to make further corrections and improvements to the OSM database.

On the other hand, the authors in [23] introduced a deep learning approach to address the challenge of detecting buildings in areas with limited data. They achieved this by transferring a pre-trained building detection model on a well-mapped region in OSM to data-scarce areas. The transfer was accomplished through fine-tuning the model using a combination of training samples from the original and target areas. The effectiveness of the method was validated by applying a deep neural networks trained in Tanzania to a site in Cameroon. The fine-tuned model successfully identified numerous OSM buildings that were missing in a specific area of Cameroon. The results demonstrated a significant improvement in the f1-score, even with only 30 training examples from the target area.

Moreover, the paper in [11] presents a novel approach that combines deep learning and crowdsourcing using the MapSwipe [6] platform. The authors devised a strategy for assigning classification tasks to either deep learning or crowdsourcing based on the confidence level of the derived binary classification results. They conducted three case studies in Guatemala, Laos, and Malawi to assess the effectiveness of their proposed workflow. The findings indicated that both crowdsourcing and deep learning surpassed existing earth observation methods and products like the Global Urban Footprint in terms of performance and accuracy.

3 Conclusions and Perspectives

OpenStreetMap (OSM) is a collaborative, open-source project that aims to create a free and editable map of the world. The data in OSM is contributed and maintained by a global community of volunteer mappers, who use various tools to edit and update the map. However, the data in OSM can be inconsistent and contain errors, which can lead to inaccuracies in the map. This paper has discussed the contribution of machine and deep learning models to the assessment of the OSM data quality.

In fact, various traditional machine learning models have been used in several studies to automatically detect errors in OSM data, such as annotation errors, topological errors, and positional errors. Technically, classifiers have been trained to automatically detect errors in new data, and to recommend tag values for new objects being added to the map. Additionally, other works have deployed these models to extract rules from OSM data to help with the disambiguation of geographical objects.

On the other hand, deep learning models have been used largely to segment high-resolution satellite imagery for roads and building footprint detection. The extracted features could be used later on to assess and enrich OSM data quality.

Working on improving the existing machine learning models in terms of providing better training data quality and designing & optimizing larger models will certainly play an important role in making OSM data a more valuable and reliable data source for various real world applications.

References

1 Ahmed Loai Ali, Falko Schmid, Rami Al-Salman, and Tomi Kauppinen. Ambiguity and plausibility: Managing classification quality in volunteered geographic information. In *Pro-*

[6] https://mapswipe.org/en/index.html

ceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14, pages 143–152, New York, NY, USA, 2014. Association for Computing Machinery. `doi:10.1145/2666310.2666392`.

2 Jamal Jokar Arsanjani and Eric Vaz. An assessment of a collaborative mapping approach for exploring land use patterns for several european metropolises. *International Journal of Applied Earth Observation and Geoinformation*, 2015. `doi:10.1016/j.jag.2014.09.009`.

3 Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. *CoRR*, abs/1705.06057, 2017. `arXiv:1705.06057`.

4 Hongchao Fan, Alexander Zipf, Qing Fu, and Pascal Neis. Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science*, 2014. `doi:10.1080/13658816.2013.867495`.

5 Cidália Costa Fonte, Lucy Bastin, Linda See, Giles M. Foody, and Flavio Lupia. Usability of vgi for validation of land cover maps. *International Journal of Geographical Information Science*, 2015. `doi:10.1080/13658816.2015.1018266`.

6 Stefan Funke, Robin Schirrmeister, and Sabine Storandt. Automatic extrapolation of missing road network data in openstreetmap. In *Proceedings of the 2nd International Conference on Mining Urban Data - Volume 1392*, MUD'15, pages 27–35, Aachen, DEU, 2015. CEUR-WS.org.

7 Jean-François Girres and Guillaume Touya. Quality assessment of the french openstreetmap dataset. *Transactions in Gis*, 2010. `doi:10.1111/j.1467-9671.2010.01203.x`.

8 Marcus Goetz and Alexander Zipf. Using crowdsourced geodata for agent-based indoor evacuation simulations. *ISPRS international journal of geo-information*, 2012. `doi:10.3390/ijgi1020186`.

9 Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 2007. `doi:10.1007/s10708-007-9111-y`.

10 Mordechai Haklay. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets:. *Environment and Planning B-planning & Design*, 2010. `doi:10.1068/b35097`.

11 Benjamin Herfort, Hao Li, Sascha Fendrich, Sven Lautenbach, and Alexander Zipf. Mapping human settlements with higher accuracy and less volunteer efforts by combining crowdsourcing and deep learning. *Remote Sensing*, 11(15), 2019. `doi:10.3390/rs11151799`.

12 Kent T. Jacobs and Scott W. Mitchell. Openstreetmap quality assessment using unsupervised machine learning methods. *Transactions in GIS*, 24(5):1280–1298, 2020. `doi:10.1111/tgis.12680`.

13 Musfira Jilani, Michela Bertolotto, Padraig Corcoran, and Amerah Alghanim. Traditional vs. machine-learning techniques for OSM quality assessment. In Cláudio Elízio Calazans Campelo, Michela Bertolotto, and Padraig Corcoran, editors, *Volunteered Geographic Information and the Future of Geospatial Data*, pages 47–64. IGI Global, 2017. `doi:10.4018/978-1-5225-2446-5.ch003`.

14 Musfira Jilani, Padraig Corcoran, and Michela Bertolotto. Automated highway tag assessment of openstreetmap road networks. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, pages 449–452, New York, NY, USA, 2014. Association for Computing Machinery. `doi:10.1145/2666310.2666476`.

15 Thomas Koukoletsos, Mordechai Haklay, and Claire Ellul. Assessing data completeness of vgi through an automated matching procedure for linear data. *Transactions in Gis*, 2012. `doi:10.1111/j.1467-9671.2012.01304.x`.

16 Bill Y Lin, Frank F Xu, Eve Q Liao, and Kenny Q Zhu. Transfer learning for traffic speed prediction: A preliminary study. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

17 Aphiwe Madubedube, Serena Coetzee, and Victoria Rautenbach. A contributor-focused intrinsic quality assessment of openstreetmap in mozambique using unsupervised machine learning. *ISPRS International Journal of Geo-Information*, 10(3), 2021. doi:10.3390/ijgi10030156.

18 Volodymyr Mnih and Geoffrey Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 203–210, Madison, WI, USA, 2012. Omnipress.

19 Peter Mooney and Padraig Corcoran. The annotation process in openstreetmap. *Transactions in Gis*, 2012. doi:10.1111/j.1467-9671.2012.01306.x.

20 Pascal Neis and Dennis Zielstra. Recent developments and future trends in volunteered geographic information research: The case of openstreetmap. *Future Internet*, 2014. doi:10.3390/fi6010076.

21 Pascal Neis, Dennis Zielstra, and Alexander Zipf. The street network evolution of crowdsourced maps: Openstreetmap in germany 2007-2011. *Future Internet*, 2011. doi:10.3390/fi4010001.

22 Martin Over, Arne Schilling, S. Neubauer, and Alexander Zipf. Generating web-based 3d city models from openstreetmap: The current situation in germany. *Computers, Environment and Urban Systems*, 2010. doi:10.1016/j.compenvurbsys.2010.05.001.

23 J. Pisl, H. Li, S. Lautenbach, B. Herfort, and A. Zipf. Detecting openstreetmap missing buildings by transferring pre-trained deep neural networks. *AGILE: GIScience Series*, 2:39, 2021. doi:10.5194/agile-giss-2-39-2021.

24 Hansi Senaratne, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Mordechai Haklay. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 2017. doi:10.1080/13658816.2016.1189556.

25 Shivangi Srivastava, John E. Vargas Muñoz, Sylvain Lobry, and Devis Tuia. Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science*, 2018. doi:10.1080/13658816.2018.1542698.

26 Shivangi Srivastava, John E. Vargas-Munoz, and Devis Tuia. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sensing of Environment*, 2019. doi:10.1016/j.rse.2019.04.014.

27 John E. Vargas-Munoz, Sylvain Lobry, Alexandre X. Falcão, and Devis Tuia. Correcting rural building annotations in openstreetmap using convolutional neural networks. *Isprs Journal of Photogrammetry and Remote Sensing*, 2019. doi:10.1016/j.isprsjprs.2018.11.010.

28 Yongyang Xu, Zhanlong Chen, Zhong Xie, and Liang Wu. Quality assessment of building footprint data using a deep autoencoder network. *International Journal of Geographical Information Science*, 2017. doi:10.1080/13658816.2017.1341632.

29 Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 2017. doi:10.1109/mgrs.2017.2762307.

30 Alexander Zipf, Steffen Neubauer, G Walenciak, Martin Over, Pascal Neis, and Arne Schilling. Interoperable location based services for 3d cities on the web using user generated content from openstreetmap. *Urban and Regional Data Management*, 2009. doi:10.1201/9780203869352.ch7.