On the Role of Spatial Data Science for Federated Learning

Anita Graser[1][0000−0001−5361−2885], Clemens Heistracher[1], and Viktorija Pruckovskaja[1,2]

[1] AIT Austrian Institute of Technology, 1210 Vienna, Austria
anita.graser@ait.ac.at
[2] Technical University Vienna, 1040 Vienna, Austria

Abstract. Federated learning (FL) has the potential to mitigate privacy risks and communication costs associated with classical machine learning and data science approaches. Given the distributed nature of FL, many of its use cases face challenges related to spatiotemporal data, geographical analysis, and spatial statistics. However, so far, FL has received little attention by the GIScience community. In this paper, we provide a first overview of the key challenges in FL and how they relate to spatial data science. This paper thus aims to provide the basis for future contributions to federated learning practices by the (geo)spatial research community.

Keywords: Machine Learning · Federated Learning · GeoAI.

DOI: https://doi.org/10.25436/E24K5T

1 Introduction

Federated learning (FL) is a machine learning (ML) setting where many clients (such as mobile/edge/Internet-of-Things devices or whole organizations) collaboratively train a model under the orchestration of a central server, while keeping the training data decentralized [6]. This means that, instead of sending their data to the central server, the clients build their own local models and only share model updates with the central server. The central server aggregates these updates into the global model and transfers the global model back to the clients for further training. Since only model updates are shared, FL can mitigate privacy risks and reduce communication costs resulting from traditional, centralized machine learning and data science approaches [6]. FL applications include, for example, keyboard apps and vocal classifier in iOS, finance risk prediction for reinsurance, electronic health records mining, and smart manufacturing [6].

GIScience or Geoinformatics is not new to ML and ML is arguably central to Geographic or Spatial Data Science [1, 21] and GeoAI [4, 5]. Early uses of neural networks in GIScience include, for example, spatial interaction modeling [16] and hydrological modeling of rainfall runoff [3]. More recently, neural networks and deep learning have, for example, enabled object recognition in georeferenced

images [18, 2] as well as travel time [22, 8] and flow predictions [23] in transport networks. Generative adversarial networks (GANs) have also found their application in GIScience, for example, [26] demonstrate how GANs can generate road maps from aerial images and vice versa, and [25] generate artificial digital elevation models. The integration of spatial data challenges into machine learning (also known as spatially-explicit machine learning), for example, through location encodings for GeoAI [13], is an ongoing area of research.

The term federated learning (FL) was introduced in 2016 [14]. Given the distributed nature of FL settings, it stands to reason that a significant subset of FL use cases have to deal with questions where the geographic location or context of the sensor and/or observed phenomena is of relevance. A closely related early example is privacy-centred selective cloud computing for location-based services [15]. Other works dealing with spatiotemporal data include privacy-preserving fault diagnosis for the Internet-of-Ships [24], detecting anomalous vehicle trajectories at intersections [7], or privacy-preserving location recommendations [19] and location predictions [9]. However, overall, FL has received very little attention by the GIScience community so far, as illustrated by the low number of related articles (see Table 1).

Table 1. Number of articles including the search terms "machine learning" (ML) and "federated learning" (FL) in selected journals from GIScience and related fields (in alphabetical order). Searches were performed on the respective journal websites on 2022-07-26 using full text search and without temporal restriction.

Journal	ML articles	FL articles
Applied Geography	105	0
Cartography and Geographic Information Science	45	0
Geographical Analysis	39	0
International Journal of Geographical Information Science	270	0
ISPRS International Journal of Geo-Information	227	0
Journal of Location Based Services	37	0
Journal of Transport Geography	74	0
Remote Sensing	32	0
Transactions in GIS	206	1 [19]

In the following section, we introduce key challenges of FL and how they relate to spatial data science.

2 Spatially-explicit Federated Learning

Spatially-explicit machine learning research is necessary since conventional machine learning (ML) ignores most of the challenges related to spatial data, such as spatial autocorrelation [17]. For example, convolutional neural networks (CNNs), which are generally regarded as appropriate for any problem involving pixels or spatial representations, have been shown to fail even for the seemingly trivial

coordinate transform problem, which requires learning a mapping between coordinates in Cartesian space and coordinates in one-hot pixel space [12].

Federated learning (FL) distinguishes between horizontal and vertical federated learning. Horizontal federated learning (HFL) is applicable when clients share a similar feature space, but the observed phenomena/objects are different. Vertical federated learning (VFL) is employed when the feature spaces among the clients are rather disjoint, but these clients store features or characteristics of the same phenomena/objects. For example, operators of different vehicle fleets may be collecting data from their vehicles to feed a traffic model. Due to privacy concerns, they may not want to share raw vehicle data with each other but with HFL, they could all contribute to and profit from the better spatial and temporal coverage of a joint global model. HFL could also happen directly on board of the moving vehicles, for example, to reduce communication costs (assuming that the model updates are smaller than the data collected by the vehicle). On the other hand, a VFL approach may be used, for example, to build a model that can leverage different sensors (such as cameras, lidar, and radar) at in intersection to detect anomalous and potentially dangerous behaviour.

Key challenges in federated learning (FL) relate to model optimization under statistical and systems heterogeneity, considering communication limitations and privacy. Model optimization essentially encompasses two aspects: selecting an appropriate ML algorithm for local training and selecting an appropriate model aggregation method for obtaining the global model.

FL needs to account for *statistical heterogeneity* since dataset sizes may vary from client to client and data originating from different clients usually cannot be considered as independent identically distributed (i.i.d.) [6]. Indeed, there may not just be temporal violations of independence (as acknowledged by [6][3]) but also spatial violations of independence (such as spatial autocorrelation) regarding both the observed phenomena as well as the FL clients. This calls for spatiotemporally-explicit ML models for local training as well as for model aggregation.

The baseline method for *model aggregation* in FL is FedAvg [14], a federated optimization method, where a global model is derived by taking a weighted average of the local models' parameters. Numerous improvements have been proposed to address different shortcomings, including FedProx [10] (to restrict strong drifts of the global model towards any of the local ones), FedYogi [20] (to integrate the knowledge from previous communication rounds), and qFedAvg [11] (to improve fairness by giving clients with poor performance higher relative weights). Spatial thinking could enhance the definition of fairness, for example, by ensuring that the global model performs equally well in all geographic regions, irrespective of potential skews in client and data distribution.

[3] For example, temporal violations of independence may occur "in cross-device FL, where devices typically need to meet eligibility requirements in order to participate in training, devices typically meet those requirements at night local time (when they are more likely to be charging, on free wi-fi, and idle)" [6].

Geographic theory may also help explain or interpret drifts towards certain local models or to provide additional knowledge to inform model aggregation.

Spatiotemporally-explicit FL models could also improve *client-side model personalization* (aiming to surpass the performance of the best fixed global model) by addressing spatial heterogeneity[9]. For example, a client may use different personalization settings while staying at the countryside than after moving to an urban environment.

Even though FL "embodies the principles of focused collection and data minimization" [6], sharing model updates can still pose a risk of revealing some sensitive information. Since spatiotemporal data presents very specific *privacy challenges* [15], appropriate measures have to be taken to avoid leakage of sensitive information.

Communication may become a critical bottleneck in federated approaches, especially when a great number of clients have to frequently transfer model updates. Particularly, in the case of mobile clients, the communication bandwidth may be fluctuating with potential connection losses. Spatiotemporal information may help foresee these connection issues, for example, if a ship is leaving coastal water and moving towards the edge of the radio communications service area.

So far, most FL approaches deal with "supervised learning tasks where labels are naturally available on each client" [6]. Spatial data science approaches, including spatial visualization, may be instrumental in extending FL to other ML paradigms, such as active learning and explainable AI. Active learning with humans in the loop, for example, provides alternative ways to label data in environments where labels are not naturally available on each client. These labeling tasks require appropriate visualizations that enable the human analyst to interpret the spatiotemporal data. Similarly, explainable AI (XAI) solutions require visualizations to make the AI's decisions understandable. For example, XAI for classic object detection in images often employ saliency heat maps on top of images. To extend XAI concepts to GeoAI, GeoXAI requires appropriate cartographic visualizations of big spatiotemporal data, particularly when dealing with GeoAI approaches for Graph neural networks (GNNs) instead of raster-based Convolutional neural network (CNN) approaches where many XAI methods can be readily transferred from computer vision.

3 Conclusion and Outlook

This paper provides a first overview of the key challenges in FL and how they relate to spatial data science. However, the FL challenges discussed here are certainly not comprehensive. There are diverse non-learning problems which need to be solved in the course of a practical machine learning project that attempts to use decentralized data. These include simple problems such as computing basic descriptive statistics since existing algorithms for solving such problems often do not have an obvious federated version [6]. In addition, legal and business issues may motivate or constrain the use of FL.

Spatial data science – in its role as a bridge between geography, GIScience, and data science – has the potential and maybe even responsibility to bring geographic thinking to data science and machine learning. This short paper represents work in progress which we aim to deepen in upcoming projects that will focus on matters of federated learning for mobility data science. It thus presents a starting point for discussion of the potential role of spatial data science in advancing federated learning practices. At the same time, advances in federated learning need to be reflected in GIScience for spatial data science and GeoAI to remain relevant.

Current GIScience research questions that may benefit from FL revolve around learning of spatiotemporal patterns from privacy sensitive data and/or in settings with communication bandwidth bottlenecks. Research questions dealing with privacy preservation include but are not limited to the development of novel methods that enable learning of movement patterns derived from individual movement data (for example, to perform location prediction [9] and recommendation [19] or to detect anomalous movement behavior [7]) or learning spatiotemporal energy demand patterns (for example, from data generated by smart home thermostats that turn off when the houses are empty [6]). In these settings, information may be fragmented among different stakeholders so that every one only knows a fraction of the overall situation. Therefore, data security-preserving sharing of data with other stakeholders is likely to be beneficial [7]. Research questions dealing with communication bandwidth bottlenecks, on the other hand, include how to enable learning in remote locations and/or on moving sensor platforms, such as on board of ocean-going vessels which can only communicate via expensive satellite connections once they leave coastal waters. Solutions to these challenges may, for example, include spatial models that help predict communication loss and FL approaches that enable merging of local regional models into the global model once communication is reestablished.

Acknowledgements

This work was partly funded by the Austrian Research Promotion Agency (FFG) through the project INTERACTIVE (project number 883855).

References

1. Andrienko, G., Andrienko, N., Weibel, R.: Geographic Data Science. IEEE Computer Graphics and Applications **37**(5), 15–17 (2017). https://doi.org/10.1109/MCG.2017.3621219, http://ieeexplore.ieee.org/document/8047433/
2. Cermelli, F., Mancini, M., Rota Bulo, S., Ricci, E., Caputo, B.: Modeling the Background for Incremental Learning in Semantic Segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9230–9239. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00925, https://ieeexplore.ieee.org/document/9157089/

3. Dawson, C.W., Wilby, R.L.: Hydrological modelling using artificial neural networks. Progress in Physical Geography: Earth and Environment **25**(1), 80–108 (Mar 2001). https://doi.org/10.1177/030913330102500104, http://journals.sagepub.com/doi/10.1177/030913330102500104
4. Hu, Y., Gao, S., Lunga, D., Li, W., Newsam, S., Bhaduri, B.: GeoAI at ACM SIGSPATIAL: progress, challenges, and future directions. SIGSPATIAL Special **11**(2), 5–15 (Dec 2019). https://doi.org/10.1145/3377000.3377002, https://dl.acm.org/doi/10.1145/3377000.3377002
5. Janowicz, K., Gao, S., McKenzie, G., Hu, Y., Bhaduri, B.: GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. International Journal of Geographical Information Science **34**(4), 625–636 (Apr 2020). https://doi.org/10.1080/13658816.2019.1684500, https://www.tandfonline.com/doi/full/10.1080/13658816.2019.1684500
6. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R.G.L., Eichner, H., Rouayheb, S.E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S.U., Sun, Z., Suresh, A.T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F.X., Yu, H., Zhao, S.: Advances and Open Problems in Federated Learning. Tech. Rep. arXiv:1912.04977, arXiv (Mar 2021), http://arxiv.org/abs/1912.04977, arXiv:1912.04977 [cs, stat] type: article
7. Koetsier, C., Fiosina, J., Gremmel, J.N., Sester, M., Müller, J.P., Woisetschläger, D.: Federated cooperative detection of anomalous vehicle trajectories at intersections. In: Proceedings of the 4th ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities. pp. 13–22. ACM, Beijing China (Nov 2021). https://doi.org/10.1145/3486626.3493439, https://dl.acm.org/doi/10.1145/3486626.3493439
8. Kudinov, D.: Predicting travel times with artificial neural network and historical routes (Mar 2018), https://community.esri.com/t5/arcgis-pro-blog/predicting-travel-times-with-artificial-neural/ba-p/884972
9. Li, A., Wang, S., Li, W., Liu, S., Zhang, S.: Predicting Human Mobility with Federated Learning. In: Proceedings of the 28th International Conference on Advances in Geographic Information Systems. pp. 441–444. ACM, Seattle WA USA (Nov 2020). https://doi.org/10.1145/3397536.3422270, https://dl.acm.org/doi/10.1145/3397536.3422270
10. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated Optimization in Heterogeneous Networks. Tech. Rep. arXiv:1812.06127, arXiv (Apr 2020), http://arxiv.org/abs/1812.06127, arXiv:1812.06127 [cs, stat] type: article
11. Li, T., Sanjabi, M., Beirami, A., Smith, V.: Fair Resource Allocation in Federated Learning. Tech. Rep. arXiv:1905.10497, arXiv (Feb 2020), http://arxiv.org/abs/1905.10497, arXiv:1905.10497 [cs, stat] type: article
12. Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the CoordConv solution. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 9628–9639. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (Dec 2018)

13. Mai, G., Janowicz, K., Hu, Y., Gao, S., Yan, B., Zhu, R., Cai, L., Lao, N.: A review of location encoding for GeoAI: methods and applications. International Journal of Geographical Information Science **36**(4), 639–673 (Apr 2022). https://doi.org/10.1080/13658816.2021.2004602, https://www.tandfonline.com/doi/full/10.1080/13658816.2021.2004602
14. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-Efficient Learning of Deep Networks from Decentralized Data. Tech. Rep. arXiv:1602.05629, arXiv (Feb 2017), http://arxiv.org/abs/1602.05629, arXiv:1602.05629 [cs] type: article
15. Meier, S.: Personal Big Data: A Privacy-centred Selective Cloud Computing Approach to Progressive User Modelling on Mobile Devices. PhD Thesis, Universität Potsdam, Mathematisch-Naturwissenschaftliche Fakultät (2017)
16. Openshaw, S.: Neural Network, Genetic, and Fuzzy Logic Models of Spatial Interaction. Environment and Planning A: Economy and Space **30**(10), 1857–1872 (Oct 1998). https://doi.org/10.1068/a301857, http://journals.sagepub.com/doi/10.1068/a301857
17. Openshaw, S., Turton, I.: A parallel Kohonen algorithm for the classification of large spatial datasets. Computers & Geosciences **22**(9), 1019–1026 (Nov 1996). https://doi.org/10.1016/S0098-3004(96)00040-4, https://linkinghub.elsevier.com/retrieve/pii/S0098300496000404
18. Porzi, L., Bulo, S.R., Colovic, A., Kontschieder, P.: Seamless Scene Segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8269–8278. IEEE, Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/CVPR.2019.00847, https://ieeexplore.ieee.org/document/8954334/
19. Rao, J., Gao, S., Li, M., Huang, Q.: A privacy-preserving framework for location recommendation using decentralized collaborative machine learning. Transactions in GIS **25**(3), 1153–1175 (Jun 2021). https://doi.org/10.1111/tgis.12769, https://onlinelibrary.wiley.com/doi/10.1111/tgis.12769
20. Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B.: Adaptive Federated Optimization. Tech. Rep. arXiv:2003.00295, arXiv (Sep 2021), http://arxiv.org/abs/2003.00295, arXiv:2003.00295 [cs, math, stat] type: article
21. Singleton, A., Arribas-Bel, D.: Geographic Data Science. Geographical Analysis **53**(1), 61–75 (Jan 2021). https://doi.org/10.1111/gean.12194, https://onlinelibrary.wiley.com/doi/10.1111/gean.12194
22. Wang, D., Zhang, J., Cao, W., Li, J., Zheng, Y.: When Will You Arrive? Estimating Travel Time Based on Deep Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018). https://doi.org/10.1609/aaai.v32i1.11877, https://ojs.aaai.org/index.php/AAAI/article/view/11877
23. Zhang, J., Zheng, Y., Sun, J., Qi, D.: Flow Prediction in Spatio-Temporal Networks Based on Multitask Deep Learning. IEEE Transactions on Knowledge and Data Engineering **32**(3), 468–478 (Mar 2020). https://doi.org/10.1109/TKDE.2019.2891537, https://ieeexplore.ieee.org/document/8606218/
24. Zhang, Z., Guan, C., Chen, H., Yang, X., Gong, W., Yang, A.: Adaptive Privacy-Preserving Federated Learning for Fault Diagnosis in Internet of Ships. IEEE Internet of Things Journal **9**(9), 6844–6854 (May 2022). https://doi.org/10.1109/JIOT.2021.3115817, https://ieeexplore.ieee.org/document/9548946/

25. Zhu, D., Cheng, X., Zhang, F., Yao, X., Gao, Y., Liu, Y.: Spatial interpolation using conditional generative adversarial neural networks. International Journal of Geographical Information Science **34**(4), 735–758 (Apr 2020). https://doi.org/10.1080/13658816.2019.1599122, https://www.tandfonline.com/doi/full/10.1080/13658816.2019.1599122
26. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2242–2251. IEEE, Venice (Oct 2017). https://doi.org/10.1109/ICCV.2017.244, http://ieeexplore.ieee.org/document/8237506/