


# Propagation of Uncertainty for Volunteered Geographic Information in Machine Learning

**Jin Xing**

Centre for Research in Geomatics, Laval University, Quebec City, Canada

jin.xing.1@ulaval.ca

 <https://orcid.org/0000-0001-5693-3414>

**Renee E. Sieber**

Department of Geography, McGill University, Montreal, Canada

renee.sieber@mcgill.ca

---

## Abstract

Although crowdsourcing drives much of the interest in Machine Learning (ML) in Geographic Information Science (GIScience), the impact of uncertainty of Volunteered Geographic Information (VGI) on ML has been insufficiently studied. This significantly hampers the application of ML in GIScience. In this paper, we briefly delineate five common stages of employing VGI in ML processes, introduce some examples, and then describe propagation of uncertainty of VGI.

**2012 ACM Subject Classification** Information systems → Uncertainty

**Keywords and phrases** Uncertainty, Machine Learning, Volunteered Geographic Information, Uncertainty Propagation

**Digital Object Identifier** 10.4230/LIPIcs.GIScience.2018.66

**Category** Short Paper

## 1 Background of VGI in Machine Learning

Machine Learning (ML) represents a set of methods that automatically learn from “experience” or training data with respect to given tasks. The learning can be implemented via a large body of models and algorithms, such as heuristic rules [32], decision trees [27], and cellular automata [31]. In Geographic Information Science (GIScience), ML has attracted considerable interest due to its wide applications in place recognition [34], ecology models [25], remote sensing image classification [33], transportation pattern discovery [22], and gazetteer analysis [9]. The rapid grow of ML has intensified due to the increasing ‘bigness’ of geospatial data, which describes the exaflood of geographic information at unprecedented volume, velocity, and variety, as well as challenges to veracity.

Among the diverse sources of big data, Volunteered Geographic Information (VGI) is considered a main provider of input data/services [12]. For example, OpenStreetMap OSM, in which individuals have crowdsourced editable web mapping services and content, has become a powerful platform for building, training, and evaluating ML algorithms and models in GIScience [15]. VGI describes the process of obtaining geographic data or services (e.g., rating accuracy of feature labels) from large groups of users in an open call that is self-organizing via the Internet [10]. Uncertainty is innate within VGI, which means data is noisy, containing redundancies, irrelevant content, errors and biases contributed by users, who are often non-experts [26]. VGI also is disorderly, in which data may be unstructured, incorrectly ordered, mis-formatted (e.g., lacking a header), and possibly poorly geo-registered. Finally, users may be unreliable in providing consistent input and inputting within the appropriate



© Jin Xing and Renee E. Sieber;

licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 66; pp. 66:1–66:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** Uncertainty Issues in Applying VGI for ML

ML Process	Uncertainty Type		Examples in VGI
Data Collection, Annotation, and Cleaning	Data Uncertainty		Inaccurate geolocation; spatial unevenness in data contributions; redundancies; gender, culture, and race bias in training data
Data Distribution	Operation	Uncertainty	Boundary Vagueness (e.g., artificial boundaries introduced by data splitting); aggregation errors (e.g., heaping error in determining the existence of a traffic jam, binning of VGI point data)
Feature/Topic Detection	Representation	Uncertainty	Interpreting location from place (from a well-defined to a poorly defined object)
Model/Algorithm Selection and Training	Decision Uncertainty		Simpler/alternate models than ML may be better like linear regression
Evaluation and Tuning	Service Uncertainty		Biased classification; Inconsistency in grading

time periods. Noisy, disordered, and unreliable data and service can significantly lower the value of VGI in ML.

Previous work in VGI's uncertainty largely concentrates on the data quality. Researchers focused, for example, on uncertainty regarding the non-expert (e.g., skill levels and motivation), the thematic diversity of input (scattered focus relative to analysis needs), and the spatial unevenness of contributions (e.g., popularity of places relative to others) [11]. In ML, VGI is viewed primarily for its ability to provide data for ML, either as training data or general input data. It also has been employed for result evaluation and tuning of ML [18]. A worrying trend in GIScience inquiry into ML is its treatment as a big black box, where issues of data uncertainty are treated as I/O problems. We break down the black box of ML into a collection of workflow processes to briefly identify uncertainty from VGI that can occur within the ML as well as in its parameterization and refinement.

Other taxonomies tend to focus on classifying ML methods (e.g., supervised, unsupervised, and reinforcement learning) and application areas (e.g., computer vision, natural language processing, and speech recognition)[16]. The importance of uncertainty and its propagation have not been highlighted. We view the interaction between VGI and ML as five stages throughout the processing of VGI: data collection and cleaning, data distribution, feature/topic detection, model/algorithm selection and training, and evaluation and tuning.

## 2 A General Framework for Integrating Geospatial Crowdsourcing and ML

Our framework (Table 1) follows the standard ML workflow (data collection and cleaning, splitting of training from testing data, model training, evaluation, parameter tuning) [28] and adds components from big data handling [21] and ML computation [4] for de-/re-composition. Since the five stages may occur iteratively (e.g., the evaluation result could be fed back to the training process to improve accuracy), uncertainty also can propagate if we fail to attend to the origin of the uncertainty.

## 2.1 Data Collection and Cleaning

The primary utility of VGI in ML is for training and, more generally, input data. Training refers to data used by ML to calculate its parameters/weights so that input data generates expected outputs. Geospatial content is available across a wide range of VGI. It can be raster (landscape photographs) and vector (social checkins, binned aggregations of points); structured (Twitter metadata) and unstructured (Twitter text), explicit (x,y's, placenames in hashtags) and implicit (colloquial names for neighborhood), absolute (latitude/longitude) and relative (concepts of home), passive (geo-fencing) and active (Amazon Mechanical Turk-AMT). It can be static or dynamic (harvesting of Flickr geotags at point in time or movement data), compensated or voluntary (AMT or VGI) [19]. Considerable research has been conducted to assess uncertainty with various VGI (cf., [14]).

Like other crowdsourced content, VGI data contains considerable error, vagueness, and ambiguity, and is vulnerable to malicious contributions (e.g., via GPS spoofing). As suggested above, this is the richest area of current research so this section is admittedly brief. Most research on the negative impact of ML focuses on the issue of algorithmic bias due to input data [26]. Location often serves as a proxy for race so one needs to debias on the basis of primary variable as well as data which functions as its surrogate [1]. Often debiasing requires human intervention (cf., gendered word2vec example in [2]) so this stage also can utilize crowdsourcing. Geographic unevenness in data contributions can further distort ML output, for example the low OSM participation in Africa or the differential accuracy of OSM in urban areas versus rural regions [29]. Privacy protections, like the EU's General Data Protection Regulation, will increase distortions in VGI as whole swaths of data are removed or masked [6]. Lastly, much of VGI is streamed, which requires new sampling techniques (e.g., reservoir sampling) to normalize temporal spikes or redundancies.

## 2.2 Data Distribution

The attraction of VGI to ML is both in its source (geosocial media) and its potential as big data. The latter likely requires de-/re-composition to distribute the computing. Data distribution may suffer from disorder in VGI because geographic data has its own internal topology and geometry that can be destroyed by arbitrary decomposition or splitting. For example, rectangular decomposition can distort the boundary of geographic objects and increase output uncertainty [5]. Most VGI is point-based and may need to be binned. A more sophisticated feature type, a polygon like a hexagon, does not easily alleviate the problem and any aggregation is subject to modifiable areal unit problems [24] that can alter ML output.

ML can be employed to reduce uncertainty in data distribution. Felzenszwalb et al. [7] employed latent support vector machine to decompose the original raster data into multiple object-based rectangles to lower boundary distortions. Temporal disorder in VGI, such as burstiness of reporting of natural disasters, could be addressed by decomposition with parallel processing.

## 2.3 Feature/Topic Detection

ML is designed in large part to recognize patterns, generate rules, approximate functions, and classify data sets. An important use of VGI in ML can be for feature or topic detection (e.g., forest, alternate route to avoid traffic jam). We lack explicit control over the feature representation in VGI. Users may not provide feature identification as planned or neural networks may fail to extract useful features from noisy VGI. For example, uncertainty in

placename makes it difficult to infer locations; “downtown nearby” could be interpreted as multiple locations [8]. Although iterative feature/object detection in ML can reduce uncertainty, there is no easy way to clean data to better disambiguate place to a location and location to a place. This resembles the challenge of NLP regarding semantic modeling to disambiguate slang (e.g., “bad”, “hot”, “sick”) in ML. Aggregation (pattern detection) is a likely outcome of ML that is based on VGI and therefore is subject to Sorites paradox and modifiable areal unit problems here as well (e.g., how many cars constitute a jam; how many trees constitute a forest).

The temptation for users new to ML is to treat it as a blackbox, an algorithm amongst many in a software library. Treating ML as a black box means that ML cannot necessarily accommodate the geography of VGI. For example, max pooling, which is a widely used method to pass features from one layer of neural network to another, is considered problematic in convolutional neural network by Sabour et al. [30] because max pooling lacks topology. In another example, a word embedding algorithm may produce very different vectors to represent “pub” and “bar” due to the surrounding content, which may then require multiple detection iterations.

## 2.4 Model/Algorithm Selection and Training

Which ML model or algorithm achieves the highest accuracy with a given input dataset and features? What is the best way to calculate the weights or parameters of the ML model/algorithm? Should we rely on a single ML model/algorithm or combine several ones together? These questions are difficult in ML and there are no clear answers. VGI can potentially assist this selection process with existing knowledge about model/algorithm selection and training strategies (think a wiki of appropriate ML) [23]. However, knowledge contributed via VGI may be unreliable because of a “follow the crowd” mentality with little investigation into alternate approaches [17]. Deep neural network is increasingly popular in ML research but a linear regression may be more appropriate, considering the quality of the data at hand and the ease of an ML implementation.

## 2.5 Evaluation and Tuning

Performance of ML algorithms needs to be evaluated with datasets different from the training process. VGI plays a pivotal role in collecting evaluation datasets and crowdsourcing can play a role in the evaluation process. To avoid overfitting (i.e., model is too closely fitted to the training data), ML scientists usually employ cross-validation, which can reduce the influence of uncertainty from VGI training data. Evaluation can be conducted with crowdsourcing services, such as the translation validation within the Google Translate Community [20] or Captcha [3]. Here, issues similar to data collection re-emerge, with potential biases introduced by the evaluators, who may be drawn from a particular gender, race, class, or skill level. These issues resemble the social approach to assessing spatial data accuracy in [13], in which the focus shifts from the uncertainty of the contribution to that of the contributor. One may wish to implement ranking or rating systems to improve confidence in the validators.

## 3 Propagation of Uncertainty in ML and Conclusion

In this paper, we propose a general framework to explore VGI uncertainty in ML. This includes the concrete importance of VGI for training data as well as the use of crowdsourcing for model/algorithm selection and performance evaluation in ML.

Uncertainty also can propagate across the ML workflow. Uncertainty in data collection can make data distribution more difficult because we do not know the appropriate aggregation size or scale. Without adequate cleaning, noisy data can generate messy features or false positives that will invalidate the chosen ML models and algorithms. Crowdsourcers bring their own bias to the evaluation of ML, which can influence the training of ML for parameter tuning. Disagreements during the cross validations may generate inconsistency in iterations of ML and force us to re-run the process. Where possible, it is critical to identify uncertainty at each stage to minimize the propagation of uncertainty. However, the cost (e.g., human intervention) of reducing the uncertainty in the early stages of ML (e.g., data collection and cleaning) is generally less than later stages (e.g., evaluation and tuning), so it is useful for us to consider at which stages it is appropriate to insert geographic crowdsourcing and crowdsourcers.

---

### References

- 1 Julia Angwin. Make algorithms accountable. *The New York Times*, 1, 2016.
- 2 Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.
- 3 Kumar Chellapilla and Patrice Y Simard. Using machine learning to break visual human interaction proofs (hips). In *Advances in neural information processing systems*, pages 265–272, 2005.
- 4 Cheng-Tao Chu, Sang K Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Kunle Olukotun, and Andrew Y Ng. Map-reduce for machine learning on multicore. In *Advances in neural information processing systems*, pages 281–288, 2007.
- 5 Joao Porto De Albuquerque, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4):667–689, 2015.
- 6 Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.
- 7 Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.
- 8 Andrew J Flanagin and Miriam J Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72(3-4):137–148, 2008.
- 9 Noah W Garfinkle, Lucas Selig, Timothy K Perkins, and George W Calfas. Geoparsing text for characterizing urban operational environments through machine learning techniques. In *Geospatial Informatics, Fusion, and Motion Video Analytics VII*, volume 10199, page 101990C. International Society for Optics and Photonics, 2017.
- 10 Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
- 11 Michael F Goodchild. Commentary: whither vgi? *GeoJournal*, 72(3-4):239–244, 2008.
- 12 Michael F Goodchild and J Alan Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.
- 13 Michael F Goodchild and Linna Li. Assuring the quality of volunteered geographic information. *Spatial statistics*, 1:110–120, 2012.
- 14 Joel Grira, Yvan Bédard, and Stéphane Roche. Spatial data uncertainty in the vgi world: Going from consumer to producer. *Geomatica*, 64(1):61–72, 2010.

- 15 Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.
- 16 Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
- 17 Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 467–474. International Foundation for Autonomous Agents and Multiagent Systems, 2012.
- 18 M Kanevski, A Pozdnukhov, and V Timonin. Machine learning algorithms for geospatial data. applications and software tools. In *Proceedings of the 4th International Congress on Environmental Modelling and Software*, 2008.
- 19 Leyla Kazemi and Cyrus Shahabi. Geocrowd: enabling query answering with spatial crowdsourcing. In *Proceedings of the 20th international conference on advances in geographic information systems*, pages 189–198. ACM, 2012.
- 20 Somesh Kumar. *Methods for community participation: a complete guide for practitioners*. New Delhi (India) Vistaar Pub., 2002.
- 21 Jae-Gil Lee and Minseo Kang. Geospatial big data: challenges and opportunities. *Big Data Research*, 2(2):74–81, 2015.
- 22 Jin Liu, Xiao Yu, Zheng Xu, Kim-Kwang Raymond Choo, Liang Hong, and Xiaohui Cui. A cloud-based taxi trace mining framework for smart city. *Software: Practice and Experience*, 47(8):1081–1094, 2017.
- 23 Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
- 24 Amelia McNamara and Aran Lunzer. Exploring the effects of spatial aggregation, 2016.
- 25 Julian D Olden, Joshua J Lawler, and N LeRoy Poff. Machine learning methods without tears: a primer for ecologists. *The Quarterly review of biology*, 83(2):171–193, 2008.
- 26 Cathy O’Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.
- 27 Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis, and Constantine D Spyropoulos. Learning decision trees for named-entity recognition and classification. In *ECAI Workshop on Machine Learning for Information Extraction*, 2000.
- 28 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- 29 Chris Perkins. Plotting practices and politics:(im) mutable narratives in openstreetmap. *Transactions of the Institute of British Geographers*, 39(2):304–317, 2014.
- 30 Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.
- 31 Hossein Shafizadeh-Moghadam, Ali Asghari, Mohammad Taleai, Marco Helbich, and Amin Tayyebi. Sensitivity analysis and accuracy assessment of the land transformation model using cellular automata. *GIScience & Remote Sensing*, 54(5):639–656, 2017.
- 32 Anna L Swan, Dov J Stekel, Charlie Hodgman, David Allaway, Mohammed H Alqahtani, Ali Mobasher, and Jaume Bacardit. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC genomics*, 16(1):S2, 2015.
- 33 Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.
- 34 Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.