Chapter 20
Mitigating Spatial Bias in Volunteered Geographic Information for Spatial Modeling and Prediction

Guiming Zhang

Abstract VGI (volunteered geographic information) observations are often spatially biased, which degrades the quality of inferences drawn from field sample sets consisting of VGI observations. This chapter presents a novel representativeness-directed approach to mitigating spatial bias in VGI for spatial modeling and prediction. The approach, based on the Third Law of Geography (the similarity principle), defines the representativeness of a field sample set as the degree to which the field sample locations capture the spatial variability of environmental covariates in the study area. Sample representativeness is then quantified as the overlap between the probability distribution of covariate values over sample locations and the distribution over the whole study area. Adjusting the weights for individual sample locations towards increasing the overlap thus mitigates spatial bias in the sample locations and improves sample representativeness. Applications of the approach to species habitat suitability mapping and digital soil mapping demonstrate its effectiveness in mitigating spatial bias to improve the accuracy of spatial modeling and prediction.

Keywords Volunteered geographic information (VGI) · Sample representativeness · Spatial bias mitigation · Modeling and prediction · Predictive mapping

20.1 Introduction

Volunteered geographic information (VGI) refers to geographic information created by citizen volunteers (Goodchild, 2007). It has proliferated in recent years as advancements in geospatial and communication technologies enable the general public to contribute geographic data. With ubiquitous access to the Internet, ordinary citizens can now easily create and share geographic observations of the world, for example, by sharing geo-referenced photos of species observations in citizen science communities or on social media through location-aware smartphones. VGI broadly

G. Zhang (✉)
Department of Geography & the Environment, University of Denver, Denver 80208, USA
e-mail: guiming.zhang@du.edu

B. Li et al. (eds.), *New Thinking in GIScience*,
https://doi.org/10.1007/978-981-19-3816-0_20

encompasses geographic data generated by volunteer participants in citizen science, crowdsourcing, social media, etc. as they all involve voluntary and non-expert geographic data creation (Zhang, 2021). VGI is useful in many domains such as emergency response, environmental monitoring, land cover map validation, and biodiversity modeling (Yan et al., 2020). Exemplary VGI projects include OpenStreetMap (Haklay & Weber, 2008) that compiles an open and free geographic databases for the world, and iNaturalist (Unger et al., 2020) and eBird (Wood et al., 2011) which document species observations across the globe on a daily basis. VGI represents a paradigm shift in how geographic data is created and shared and in its content and characteristics (Elwood, 2008). In a broader context, VGI is an important source of geospatial big data (Yang, 2017) which is propelling geographic research towards emerging paradigms such as "data-driven geography" (Miller & Goodchild, 2014) and "data-intensive science" (Kelling et al., 2009).

VGI has become a supplementary or even alternative mechanism for acquiring geographic data due to its unique advantages. First, VGI contains rich local information because citizens as local experts and sensors (Goodchild, 2007) have long been accumulating knowledge of their local environments (Zhang et al., 2018; Zhu et al., 2015b). Second, VGI makes it feasible to collect geographic data over large areas given that potential VGI contributors are all over the world. Third, VGI can provide timely updated data that are difficult to obtain through other means. Lastly, VGI is much less expensive than traditional spatial data collection protocols (e.g., survey). As such, VGI has a great potential to reveal the spatiotemporal dynamics of geographic phenomena at high spatiotemporal resolutions over large areas.

Such potential can be realized through spatial modeling and prediction based on VGI observations. Nonetheless, VGI observations still represent only a set of sample observations regarding the phenomenon under concern, despite its seemingly extensive coverages (Zhang & Zhu, 2018). For instance, occurrence locations of a bird species reported by volunteers is a sample set from the population consisting of all possible species occurrence locations. In this respect, VGI observations are similar to field sample data collected through traditional geographic sampling. One of the significant differences, though, is that locations for designed geographic sampling are carefully chosen (e.g., following statistical sampling design) so that the sampled locations are representative of the spatial variabilities in the study area (Jensen & Shumway, 2010). In contrast, VGI contributors decide where (and when) to conduct observations at their own discretion without following a coordinated sampling scheme. This characteristic of voluntary data creation often results in spatial bias in VGI data, which has profound implications on drawing inferences about the target phenomenon (i.e., population) from VGI observations (i.e., sample). This chapter focuses on this issue and presents a novel representativeness-directed approach to mitigating spatial bias in VGI for spatial modeling and prediction.

20.2 Spatial Bias in VGI

Data quality of VGI is under constant scrutiny (Goodchild & Li, 2012), and spatial bias is a prominent concern when using VGI for mapping, modeling, and prediction (Zhang & Zhu, 2018). VGI observations in spatial distribution tend to concentrate in certain geographic areas as observations made by volunteers are opportunistic in nature, which results in spatial bias in sampling. Spatial distribution of the observation effort can be considered neither random nor regular in the sense of geographic sampling design, but 'ad hoc' (Zhu et al., 2015b). As a result, VGI observations are often of higher density in specific areas, for example, populous and accessible areas (Kadmon et al., 2004; Zhang, 2020).

Due to spatial bias, a field sample set consisting of VGI observations may not be representative of the spatial variabilities of the phenomena in the study area. Spatial bias, if not appropriately accounted for, would adversely affect the quality of inferences drawn from VGI observations (Leitão et al., 2011). Spatial bias is one form of sample selection bias (Zhang & Zhu, 2018). Many methods rely on information of the underlying observation process (e.g., selection probabilities) to correct for sample selection bias, but such information is often unavailable in VGI data.

Here a novel representativeness-directed approach was developed to mitigate spatial bias in VGI to improve the quality of spatial modeling and prediction (Zhang, 2018; Zhang & Zhu, 2019a, 2019b). Specifically, it is for mitigating spatial bias in field sample sets to improve the accuracy of predictive mapping, a framework for predicting the spatial variation of a target variable based on environmental covariate data and a model capturing the covariation relationship (f) between the target variable (T) and the covariates (E) (Fig. 20.1).

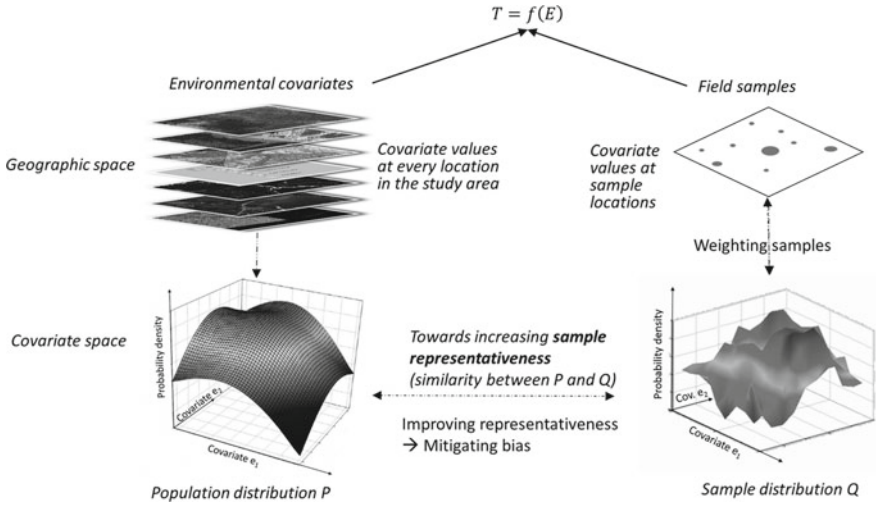

Fig. 20.1 Basic idea of representativeness directed spatial bias mitigation. Reused from Zhang and Zhu (2019b) with permission

20.3 A Representativeness-Directed Approach to Bias Mitigation

Spatial bias has adverse effects on spatial modeling and prediction as it impedes the representativeness of VGI-based field sample sets. In the context of predictive mapping, sample representativeness essentially is the degree to which the observations made at sample locations capture the spatial variability of the relationship between the target variable (e.g., species habitat suitability) and the environmental covariates (e.g., elevation, land cover, precipitation) in the area. This is achieved by capturing the variability in the target variable and that in covariates. With covariate data (raster layers), it is feasible to assess sample representativeness. Sample representativeness with respect to the target variable is hard to assess as its spatial variation is unknown (to be predicted). Nonetheless, according to the Third Law of Geography (Zhu et al., 2018; Zhu & Turner, 2022), which states that similar values of the target variable can be expected at locations with similar geographic configurations (e.g., environmental conditions), it can be reasonably expected that the representativeness measured on the covariates would approximate the representativeness on the target variable because the target variable and the covariates should correlate (Zhu et al., 2015a). Based on this idea, sample representativeness can be defined and measured to guide spatial bias mitigation.

20.3.1 Measuring Sample Representativeness

Sample representativeness is defined as the "goodness-of-coverage" of the sample locations in the covariate space, which in turn is measured as the similarity between the probability density distribution of the sample locations in the covariates space (i.e., sample distribution Q) and the probability density distribution of all spatial units (e.g., raster cells) in the area (i.e., population distribution P) (Fig. 20.1). Stronger spatial bias in the sample locations would lead to poorer sample representativeness.

Sample representativeness is computed as the similarity between the sample and population distributions over the covariate space (Zhang & Zhu, 2019b). Kernel density estimation was used to estimate probability density distributions for computing sample representativeness. First, sample and population distributions with respect to individual covariate were estimated as per Eqs. (20.1) and (20.2), respectively.

$$Q_l(v_l) = \sum_{i=1}^{n} w_i \frac{1}{h_{lQ}} K\left(\frac{v_l - V_{li}}{h_{lQ}}\right) \tag{20.1}$$

$$P_l(v_l) = \sum_{j=1}^{m} \frac{1}{h_{lP}} K\left(\frac{v_l - V_{lj}}{h_{lP}}\right) \tag{20.2}$$

In the above equations, $K(\cdot)$ is the Gaussian kernel function, n is the number of sample locations and m is the number of locations (cells) in the study area. Q_l and P_l are the estimated sample and population distributions on covariate l (denoted as v_l), respectively. V_{li} is the value of v_l at sample location i and w_i is a normalized sample weight ($\sum_{i=1}^{n} w_i = 1$) associated with location i. V_{lj} is the value of v_l at cell j in the study area. h_{lQ} and h_{lP} are kernel bandwidths. Second, the similarity between Q_l and P_l (S_l) was computed as the overlapping area between the two distributions (Eq. 20.3) (Zhu, 1999):

$$S_l = \frac{2 \times A_{Q_l} \cap A_{P_l}}{A_{Q_l} + A_{P_l}} \tag{20.3}$$

where A_{Q_l} and A_{P_l} are the areas under the sample and population distribution curves, respectively and $A_{Q_l} \cap A_{P_l}$ is the overlapping area (Fig. 20.2). S_l reflects the goodness-of-coverage of the sample regarding this covariate. Finally, sample representativeness was computed as the overall similarity between the sample and population distributions with respect to all covariates. It is a weighted average of the similarities with respect to individual covariates (Eq. 20.4):

$$R = \sum_{i=1}^{L} \frac{\lambda_i}{\sum_{j=1}^{L} \lambda_j} S_i \tag{20.4}$$

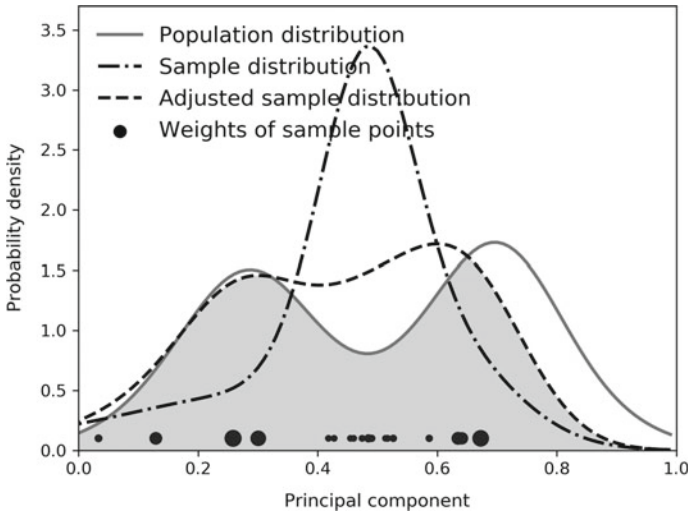

Fig. 20.2 An illustration of the effects of representativeness-directed spatial bias mitigation. Reused from Zhang and Zhu (2019b) with permission

where R is sample representativeness with a larger value indicating higher sample representativeness, and λ_i the weight associated with covariate i indicating covariate importance in measuring sample representativeness.

20.3.2 Representativeness-Directed Bias Mitigation

Spatial bias in field sample sets can then be mitigated by improving sample representativeness. Weights of the sample locations (Eq. 20.2) affect the estimated sample distributions and hence sample representativeness. Therefore, improving sample representativeness is achieved by adjusting the sample distribution towards increasing its similarity to the population distribution through weighting sample locations (Zhang & Zhu, 2019b). That is, sample locations in under-represented areas would receive larger weights and be treated as more important in training models. Weighting the sample locations in this way is expected to mitigate spatial bias and improve sample representativeness. The key is to determine the optimal weights. This can be conceived as an optimization problem, where the goal is to find a set of optimal weights associated with the sample locations that maximizes sample representativeness. A Genetic Algorithm was adopted to search for the optimal weights using sample representativeness as the objective function.

The weighted sample locations were used to train models to establish the relationships between the target variable and the covariates. Weights can be incorporated in the model training process by weighting the error term associated with each sample location (e.g., training a regression model using weighted ordinary least square) (Zhang & Zhu, 2019a, 2019b). The trained models can be applied to the covariate data layers (cell-by-cell) to predict spatial variation of the target variable.

20.4 Applications

The representativeness-directed approach to spatial bias mitigation was evaluated through case studies in two application domains: species habitat suitability mapping, and digital soil mapping.

Occurrence locations of the red-tailed hawk (*Buteo jamaicensis*) obtained from eBird were used to model and predict the species habitat suitability in Wisconsin, United States. The approach was applied to determine weights for species occurrence locations (Fig. 20.3) to train a habitat suitability model with logistic regression. Validation shows that the accuracy of predicted suitability map (Fig. 20.4) improved with weighted occurrence locations. Additionally, a positive relationship between sample representativeness and prediction accuracy was observed (Fig. 20.5), suggesting that sample representativeness is a valid indicator of suitability prediction accuracy (Zhang & Zhu, 2019b).

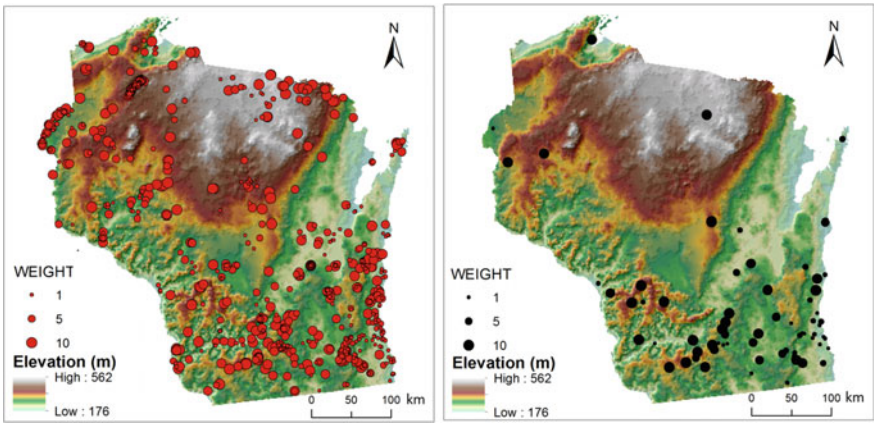

Fig. 20.3 Optimal weights determined through the representativeness-directed approach for all eBird observation locations (left) and for the red-tailed hawk occurrence locations (right). Reused from Zhang and Zhu (2019b) with permission

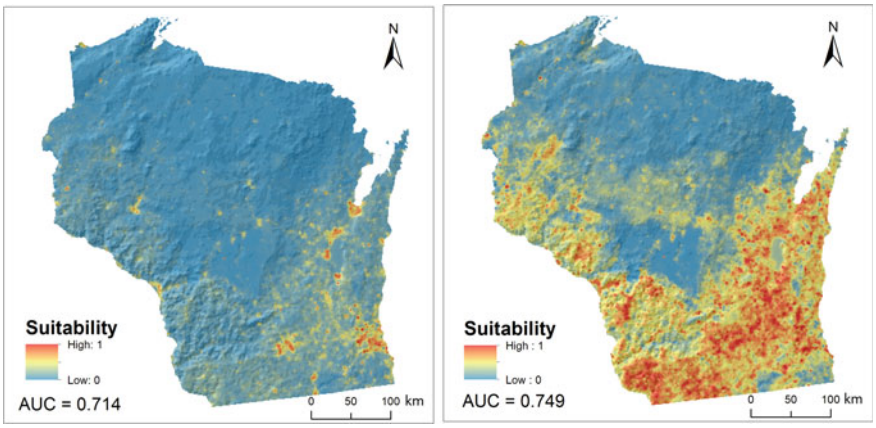

Fig. 20.4 Predicted habitat suitability maps based on unweighted species occurrence locations (left) and weighted occurrence locations (right). Higher AUC (area under the receiver operating characteristic curve) indicates higher prediction accuracy. Reused from Zhang and Zhu (2019b) with permission

The representativeness-directed approach was also applied to mitigate spatial bias in existing soil samples for digital soil mapping in Heshan study area, northeastern China. Existing soil samples in the study area were pooled from various sources and subject to spatial bias. Quantitative evaluations show that weighting soil samples using the weights determined from the approach (Fig. 20.6) improved A-horizon soil organic matter content prediction accuracy with either the iPSM method (Zhu et al., 2015a) or multiple linear regression (Fig. 20.7). A positive relationship between

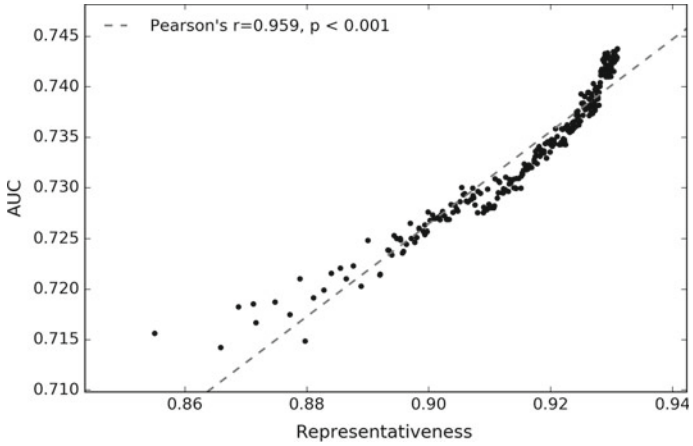

Fig. 20.5 The relationship between sample representativeness and prediction accuracy over the generations of the genetic algorithm. Reused from Zhang and Zhu (2019b) with permission

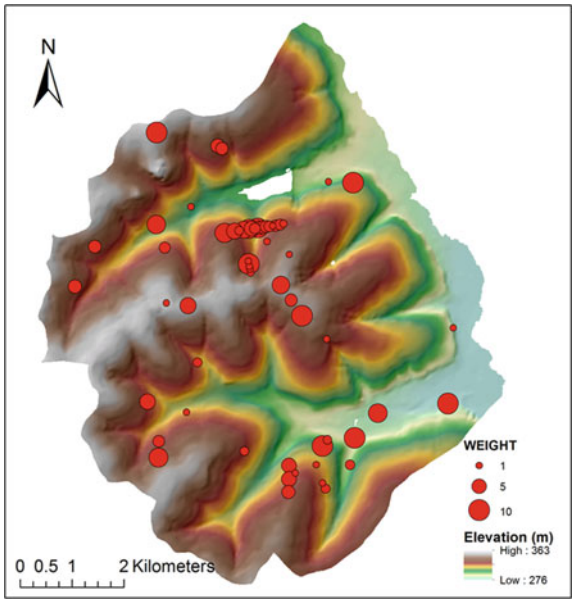

Fig. 20.6 Weights of the soil samples determined through the representativeness-directed approach. Reused from Zhang and Zhu (2019a) with permission

sample representativeness and prediction accuracy was again observed (Fig. 20.8). Moreover, the weights were informative of individual sample importance and thus can be used as guidance to filter soil samples to improve soil prediction accuracy (Zhang & Zhu, 2019a).

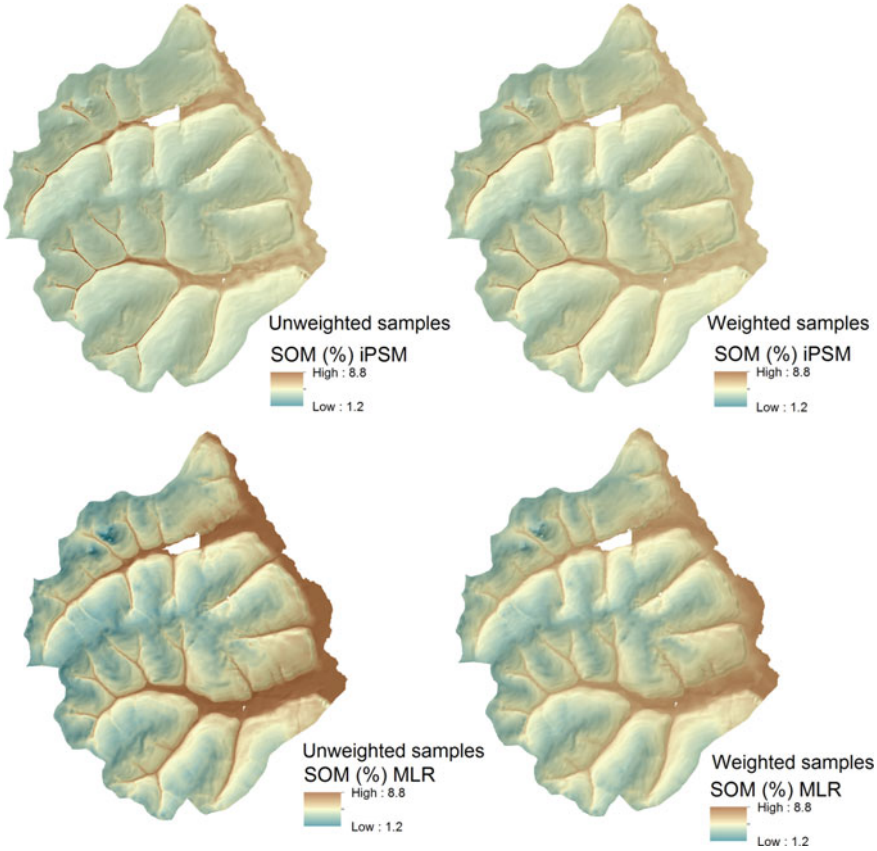

Fig. 20.7 A-horizon soil organic matter content predicted with iPSM (top row) and multiple linear regression (bottom row) based on unweighted soil samples (left column) and weighted soil samples (right column). Lower RMSE (root mean squared error) indicates higher prediction accuracy. Reused from Zhang and Zhu (2019a) with permission

20.5 Outlook on Future Research

Beyond the two application case studies, the idea of the representativeness-directed approach should apply to sample selection bias mitigation in general for spatial modeling and prediction. Specifically, beyond its applicability to global modeling methods, the approach can be extended to train localized models (e.g., modeling based on sample locations within a neighborhood of the prediction location) that account for spatial non-stationarity. It would also be interesting to examine the applicability of the approach for classification problems (e.g., soil class prediction) in addition to regression tasks explored. Lastly, spatial bias in a field sample set

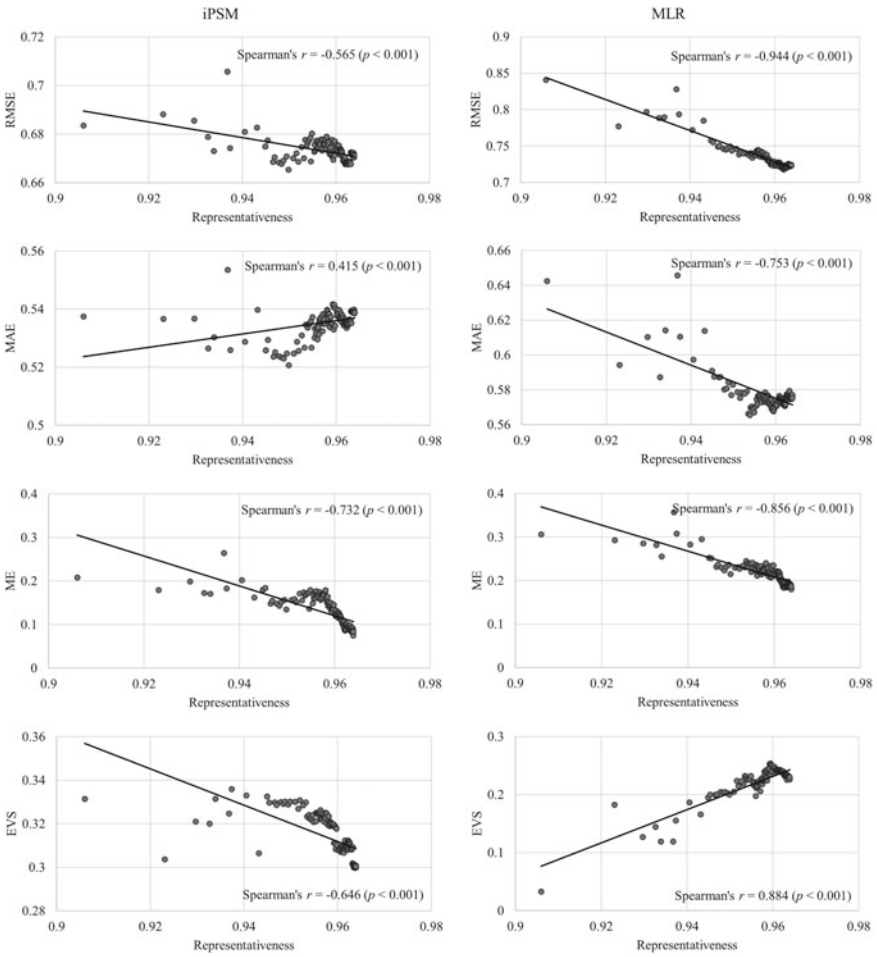

Fig. 20.8 Relationship between sample representativeness and prediction accuracy (measured by root mean squared error—RMSE, mean absolute error—MAE, mean error—ME, and explained variance score—EVS) over the generations of the genetic algorithm. Reused from Zhang and Zhu (2019a) with permission

may not be mitigated completely. It is thus worth exploring how to utilize information on sample representativeness to quantify modeling and prediction uncertainties, preferably in a spatially explicit manner.

At the core of the approach is defining and measuring sample representativeness in the covariate space. The idea can be translated to the social space. For example, it may be used to quantify demographic and socio-economic biases embedded within social media users to inform to what extent inferences drawn from social media data truly reflect the status of the population at large. In a broader sense, big data often suffers from biases. The concept of defining and quantifying representativeness

offers a novel perspective on how to appropriately deal with biases in big data so that more accurate insights can be gained from them.

Acknowledgements The author sincerely thanks the editors team Drs. Bin Li, A-Xing Zhu, Xun Shi, Cuizhen Wang and Hui Lin for their invitation to contribute a chapter and their tireless effort to organize and edit this book series.

References

Elwood, S. (2008). Volunteered geographic information: Key questions, concepts and methods to guide emerging research and practice. *GeoJournal, 72*(3), 133–135.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*, 110–120.

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-generated street maps. *Pervasive Computing, IEEE, 7*(4), 12–18.

Jensen, R. R., & Shumway, J. M. (2010). Sampling our world. In B. Gomez, & J. P. Jones III (Eds.), *Research methods in geography: A critical introduction* (pp. 77–90).

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications, 14*(2), 401–413.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience, 59*(7), 613–620.

Leitão, P. J., Moreira, F., & Osborne, P. E. (2011). Effects of geographical data sampling bias on habitat models of species distributions: A case study with steppe birds in southern Portugal. *International Journal of Geographical Information Science, 25*(3), 439–454.

Miller, H. J., & Goodchild, M. F. (2014). Data-driven geography. *GeoJournal, 80*(4), 449–461.

Unger, S., Rollins, M., Tietz, A., & Dumais, H. (2020). iNaturalist as an engaging tool for identifying organisms in outdoor activities. *Journal of Biological Education*, 1–11.

Wood, C., Sullivan, B., Iliff, M., Fink, D., & Kelling, S. (2011). eBird: Engaging birders in science and conservation. *PLoS Biology, 9*(12), e1001220.

Yan, Y., Feng, C., Huang, W., Fan, H., & Wang, Y. (2020). Volunteered geographic information research in the first decade: A narrative review of selected journal articles in GIScience. *International Journal of Geographical Information Science, 34*(9), 1765–1791.

Yang, C. P. (2017). Geospatial cloud computing and big data. *Computers, Environment and Urban Systems, 61*, 119.

Zhang, G. (2018). *A representativeness directed approach to spatial bias mitigation in VGI for predictive mapping*. University of Wisconsin-Madison.

Zhang, G. (2020). Spatial and temporal patterns in volunteer data contribution activities: A case study of eBird. *ISPRS International Journal of Geo-Information, 9*(10), 597.

Zhang, G. (2021). Volunteered geographic information. In J. P. Wilson (Ed.), *The geographic information science & technology body of knowledge* (1st Quarter 2021 Edition).

Zhang, G., & Zhu, A. X. (2018). The representativeness and spatial bias of volunteered geographic information: A review. *Annals of GIS, 24*(3), 151–162.

Zhang, G., & Zhu, A. X. (2019a). A representativeness directed approach to spatial bias mitigation in VGI for predictive mapping. *International Journal of Geographical Information Science, 33*(9), 1873–1893.

Zhang, G., & Zhu, A. X. (2019b). A representativeness heuristic for mitigating spatial bias in existing soil samples for digital soil mapping. *Geoderma, 351*, 130–143.

Zhang, G., Zhu, A. X., Huang, Z. P., Ren, G., Qin, C. Z., & Xiao, W. (2018). Validity of historical volunteered geographic information: Evaluating citizen data for mapping historical geographic phenomena. *Transactions in GIS, 22*(1), 149–164.

Zhu, A. X. (1999). A personal construct-based knowledge acquisition process for natural resource mapping. *International Journal of Geographical Information Science, 13*(2), 119–141.

Zhu, A. X., Liu, J., Du, F., Zhang, S., Qin, C. Z., Burt, J., Behrens, T., & Scholten, T. (2015a). Predictive soil mapping with limited sample data. *European Journal of Soil Science*, *66*(3), 535–547.

Zhu, A. X., Lu, G., Liu, J., Qin, C., & Zhou, C. (2018). Spatial prediction based on Third Law of Geography. *Annals of GIS, 24*(4), 225–240.

Zhu, A.-X., & Turner, M. (2022). How is the third law of geography different? *Annals of GIS, 28*(1), 57–67. https://doi.org/10.1080/19475683.2022.2026467.

Zhu, A. X., Zhang, G., Wang, W., Xiao, W., Huang, Z. P., Dunzhu, G. S., Ren, G., Qin, C. Z., Yang, L., Pei, T., & Yang, S. (2015b). A citizen data-based approach to predictive mapping of spatial variation of natural phenomena. *International Journal of Geographical Information Science, 29*(10), 1864–1886.