DOI: 10.1111/tgis.12679

RESEARCH ARTICLE

Transactions in GIS WILEY

Sample size and spatial configuration of volunteered geographic information affect effectiveness of spatial bias mitigation

Guiming Zhang[1] | **A-Xing Zhu**[2,3,4,5]

[1]Department of Geography and the Environment, University of Denver, Denver, CO, USA

[2]Department of Geography, University of Wisconsin-Madison, Madison, WI, USA

[3]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China

[4]School of Geography, Nanjing Normal University, Nanjing, China

[5]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

Correspondence
Guiming Zhang, Department of Geography and the Environment, University of Denver, Denver, CO 80208, USA.
Email: guiming.zhang@du.edu

Abstract

Volunteered geographic information (VGI) can provide field samples for predictively mapping geographic phenomena. Yet the biased spatial coverage of VGI observations often undermines the fitness of use of VGI samples for predictive mapping. Although methods have been developed to mitigate spatial bias in VGI samples to improve predictive model performance, there exist limited investigations into the impacts of VGI sample size and spatial distribution characteristics on the effectiveness of the methods. This article presents an empirical evaluation on how the two factors affect the effectiveness of bias mitigation methods with a case study of mapping habitat suitability of the red-tailed hawk (*Buteo jamaicensis*) using eBird data. Results reveal positive correlations between model performance improvement and sample size, given samples of similar spatial configuration. VGI samples with more spread-out spatial coverage (i.e., more representative) are more amenable to bias mitigation. However, performance improvement plateaued beyond a certain sample size and sample representativeness thresholds.

1 | INTRODUCTION

Predictive mapping is a widely used framework for producing information about the spatial variation of geographic variables (e.g., species habitat suitability, soil, landslide susceptibility), which are essential to support environmental modeling and decision-making efforts (Franklin, 2013; McBratney, Mendonça Santos, & Minasny, 2003; Zhu et al., 2014). The basic premise behind predictive mapping is that there is a relationship between the variable to be predicted (referred to as target geographic variable) and other geographic variables (referred to as covariates), about which we already have information on their spatial variation. Then the spatial variation of the target variable is predicted by coupling the relationship with the spatial variation of the covariates. Clearly, the relationship is key and is often obtained from field samples. To make it applicable to the entire geographic area of interest (i.e., mapping area), the relationship must be representative, which in turn requires the field samples used to derive this relationship to be representative and in sufficient number. Collecting a sufficient number of representative samples is no small undertaking in geography (De Gruijter, Brus, Bierkens, & Knotters, 2006; Zhang et al., 2016).

Volunteered geographic information (VGI) (Goodchild, 2007) has emerged as a major source of geographic data that can provide field samples for predictive mapping (Connors, Lei, & Kelly, 2012; Gao, Barbier, & Goolsby, 2011; Jokar Arsanjani, Helbich, Bakillah, Hagenauer, & Zipf, 2013; Mozas-Calvache, 2016; See et al., 2016; Sun, Fan, Helbich, & Zipf, 2013; Zhu et al., 2015). Although the strengths of VGI for predictive mapping can never be overstated (e.g., extensive spatiotemporal coverage, large quantity of data, cost-effectiveness, timeliness), data quality issues of VGI are under constant scrutiny (see comprehensive reviews in Flanagin & Metzger, 2008; Goodchild & Li, 2012; Haklay, 2010; Hung, Kalantari, & Rajabifard, 2016).

The spatial bias of VGI is among the top issues concerning the application of VGI for predictive mapping (Beck, Böller, Erhardt, & Schwanghart, 2014; Zhang & Zhu, 2018). Different from field samples collected at locations designed following rigorous geographic sampling schemes (e.g., stratified random sampling) (Jensen & Shumway, 2010; Wang, Stein, Gao, & Ge, 2012), most VGI observations are conducted by individual volunteers at locations selected in an ad-hoc or opportunistic manner (Zhu et al., 2015). As a result, samples compiled from VGI observations (VGI-based samples hereafter) are often non-probability samples. Moreover, VGI observations are often concentrated more in certain geographic areas (e.g., populous urban areas or areas with better accessibility) than in other areas (e.g., remote or less accessible areas). Such imbalanced spatial coverage of observations is referred to as spatial bias. Spatial bias usually renders VGI observations less representative of the spatial variation of the geographic variable of interest, which further impedes the accuracy of predictively mapping the target geographic variable based on VGI-based samples (Beck et al., 2014; Kadmon, Farber, & Danin, 2004; Zhu et al., 2015).

Methods could be adopted to mitigate spatial bias in VGI-based samples (see Zhang & Zhu, 2018 for a review) to improve predictive model performance. First, spatial bias in geographic samples is one particular type of sample selection bias. Thus, methods for accommodating sample selection bias are applicable for correcting spatial bias in VGI-based samples, for example, the methods of modeling sample selection process (Bethlehem, 2010) and importance weighting (Cortes, Mohri, Riley, & Rostamizadeh, 2008; Shimodaira, 2000). Second, a few methods have been developed to accommodate spatial bias in geographic samples and they can be applied to mitigate spatial bias in VGI-based samples, such as training local instead of global models (Fink et al., 2010), filtering sample locations in the geographic space or in the feature space (Boria, Olson, Goodman, & Anderson, 2014; Varela, Anderson, García-Valdés, & Fernández-González, 2014), weighting sample locations by effort information (Zhu et al., 2015), factoring bias out (Dudik, Schapire, & Phillips, 2005), and, most recently, representativeness-directed weighting (Zhang & Zhu, 2019a, 2019b). Different spatial bias mitigation methods have their respective data requirements which a specific VGI application scenario may or may not meet. For example, modeling sample selection process, weighting sample locations by effort information, and factoring out bias need information on the observation process (e.g., selection probabilities, observation effort) for bias mitigation, whilst importance weighting and representativeness-directed weighting require no additional information besides the basic inputs for predictive mapping (i.e., field samples, environmental covariates). The fitness of use of each method for spatial bias mitigation in a particular VGI application needs to be determined on a case-by-case basis (Zhang & Zhu, 2018).

The size of a VGI-based sample (i.e., number of VGI observations) and its spatial configuration (i.e., spatial distribution) are amongst the key factors that affect the effectiveness of spatial bias mitigation methods (Zhang & Zhu, 2019a, 2019b). A large sample size could provide more flexibility for bias mitigation. For example, if spatial bias were to be reduced by filtering sample locations (Boria et al., 2014; Varela et al., 2014), a larger sample size allows more variety of choices regarding which sample locations to remove. Similarly, if spatial bias were to be alleviated by importance weighting (Cortes et al., 2008) or representativeness-directed weighting—where sample locations are weighted differentially in training predictive models such that in feature space the weighted frequency distribution of covariate values at sample locations (sample distribution) approximates the covariate frequency distribution in the mapping area (population distribution) (Zhang & Zhu, 2019a)—a larger sample size allows examining more variety of weight combinations. It needs to be stressed here that the weights assigned to sample locations in the weighting methods are different from sampling weights, which is related to the inclusion probability in probability sampling (most VGI-based samples are non-probability ones).

However, large sample size by itself does not warrant effective bias mitigation. The spatial configuration of sample locations also plays an important role. For instance, if all sample locations are clustered in a small area covering a narrow niche of the environmental gradients, there is little room to adjust the sample distribution towards resembling the population distribution through weighting sample locations. In contrast, if the sample locations spread across the full range of environmental gradients, more variety of environmental conditions is present in the sample and thus there is more flexibility for differentially weighting the sample locations so that the sample distribution closely resembles the population distribution.

There exists only limited investigation on the impacts of sample size and sample spatial distribution characteristics on the effectiveness of the spatial bias mitigation methods. Varela et al. (2014) observed that ecological niche models built with fewer but environmentally filtered species occurrence locations outperformed models built with many unfiltered biased occurrence locations. Yet they did not examine how sample size affects the effectiveness of the bias mitigation method. Zhang and Zhu (2019b) briefly explored how soil sample size would affect the effectiveness of the representativeness-directed weighting approach for bias mitigation. They found that the approach brought less accuracy improvement for soil mapping on soil samples of larger sample size. To the best of our knowledge, there are few to no studies examining the impact of the spatial distribution characteristics of VGI-based samples on the effectiveness of the spatial bias mitigation methods.

This study presents an empirical evaluation of how VGI sample size and spatial distribution characteristics impact the effectiveness of spatial bias mitigation methods through a habitat suitability mapping case study. Compared to Zhang and Zhu (2019a), who developed the representativeness-directed weighting method for bias mitigation, this work makes new contributions to the community by seeking answers to the research question of how sample size and sample spatial distribution characteristics impact the effectiveness of spatial bias mitigation methods in general. In addition to the representativeness weighting method, the importance weighting method for sample bias correction was also examined (see Section 2.2). Even though the dataset used in this study (Section 2.1) has been used in previous publications, an entirely new set of experiments were designed and conducted (Section 2.4) to answer the research question.

The remainder of this article is organized as follows. Section 2 presents the data and methods used in this study and the experiment design for evaluation. Results are presented in Section 3, and discussed in Section 4. Conclusions are drawn in Section 5.

2 | DATA AND METHODOLOGY

2.1 | Study area and data

A case study of mapping habitat suitability of the red-tailed hawk (*Buteo jamaicensis*) using VGI data from the eBird citizen science project (Sullivan et al., 2014) and environmental covariate data was conducted in Wisconsin, USA, to examine the impacts of VGI sample size and spatial distribution characteristics on the effectiveness of two bias mitigation methods.

2.1.1 | Covariate data

Many environmental factors influence the habitat suitability of *B. jamaicensis* (Preston, 2000). A set of 71 environmental covariates representing spatial variation of the human population (housing density, population density, etc.), terrain (elevation), climatic conditions (temperature, precipitation, etc.), landscape level, and land cover class level indices and statistics that reflect habitat configuration (edge density, patch index, patch density, etc.) were compiled. Principal component analysis was applied on the covariates. Only the first 11 principal components that retain more than 80% of the total variance of the original covariates were used as covariates for habitat suitability mapping in this study (Zhang & Zhu, 2019a).

2.1.2 | VGI data

eBird data (Munson et al., 2012) were used for mapping *B. jamaicensis* habitat suitability. eBird checklist locations indicate bird watchers' observation efforts. A set of 655 geographically unique eBird checklist locations reported in June 2012 in the study area were extracted (Figure 1) (Zhang & Zhu, 2019a). These locations tend to spatially bias toward populous urban areas. This set of checklist locations, regardless of whether *B. jamaicensis* was observed, was treated as a biased VGI sample. Spatial bias mitigation methods (Section 2.2) were applied to determine the weights for these checklist locations.

B. jamaicensis occurrences were reported at 75 of the 655 checklist locations (Figure 1). Weights associated with these 75 occurrence locations were then extracted and used to weight the occurrence locations in training predictive models for suitability mapping (Section 2.3.1).

2.2 | Spatial bias mitigation methods

Some methods for mitigating spatial bias in VGI require information on the observation process, which is often not available in many VGI applications. The representativeness-directed weighting (Zhang & Zhu, 2019a) and importance weighting (Cortes et al., 2008; Shimodaira, 2000) methods need no additional data besides the basic inputs for predictive mapping. Thus, these two methods were adopted in this study to mitigate spatial bias in VGI-based samples.

2.2.1 | Representativeness-directed weighting

The representativeness-directed approach to mitigating spatial bias in VGI for predictive mapping (Zhang & Zhu, 2019a) is based on the idea of the third law of geography, which contemplates that the representativeness of a sample location should be measured using its degree of closeness to other locations in the feature space constructed by a set of covariates (Zhu, Lu, Liu, Qin, & Zhou, 2018). Under this notation, the representativeness of a VGI-based sample is evaluated as the degree of how well the covariate domain occupied by the locations in the mapping area (population distribution) is represented by the set of sample locations from VGI (sample distribution). This evaluation can be used not only to reveal the spatial bias of the sample distribution, but also to mitigate the spatial bias by weighting the sample locations over the less represented area heavier in predictive mapping, such that the sample distribution more closely resembles the population distribution. The process of bias mitigation using the representativeness-directed approach consists of three major steps (Zhang & Zhu, 2019a).

The first step is to select a set of covariates which are relevant to the target variable and for which data reflecting their spatial variation are available. To reduce the dimensionality of the feature space and eliminate collinearity

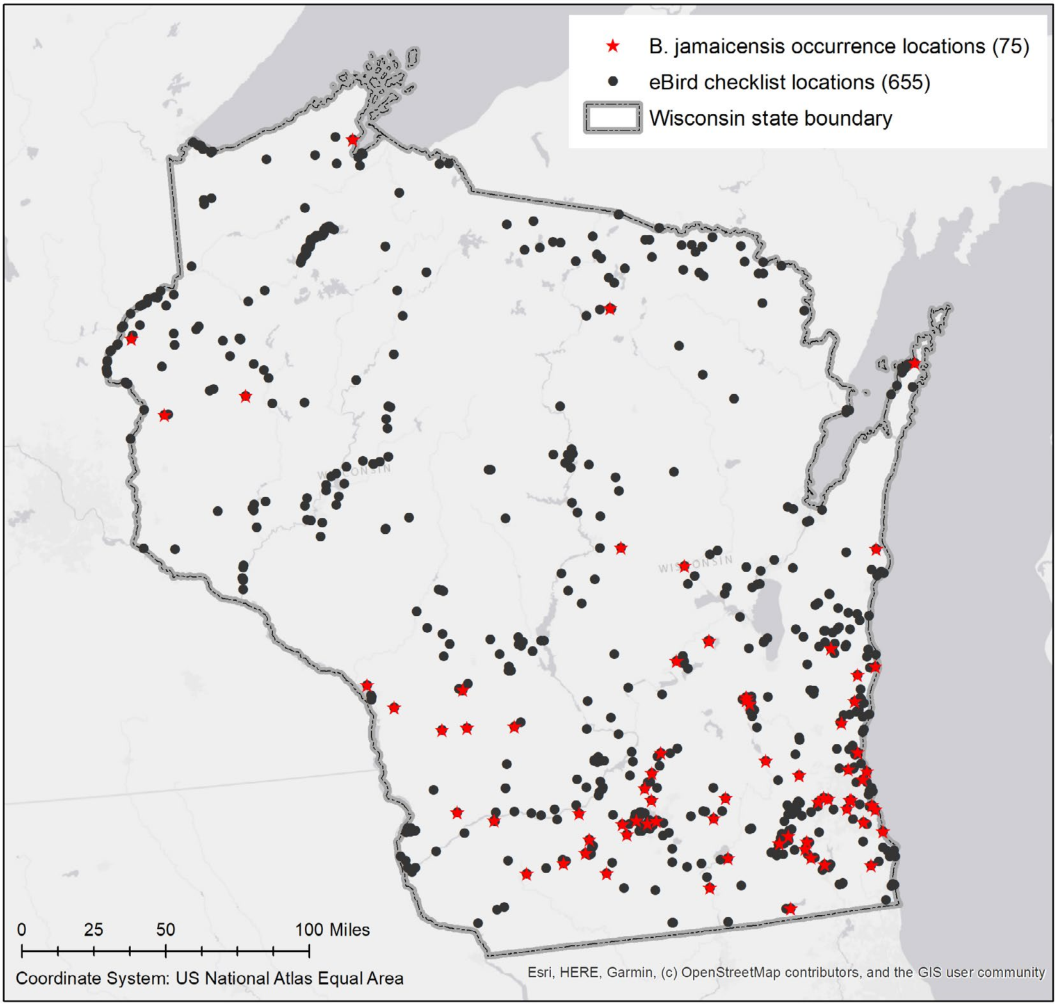
★ B. jamaicensis occurrence locations (75)
● eBird checklist locations (655)
Wisconsin state boundary
WISCONSIN
0 25 50 100 Miles
Coordinate System: US National Atlas Equal Area
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

FIGURE 1 eBird checklist locations and *B. jamaicensis* occurrence locations in June 2012

among the covariates, principal component analysis was applied to transform the original covariates into a set of principal components, and only the first few components retaining most of the variance in the original covariates are used as the new set of covariates (Section 2.1.1).

The second step is to quantify sample representativeness by measuring the similarity between the probability distribution of covariate values at sample locations (sample distribution) and the covariate distribution in the mapping area (population distribution). Given the sample locations and the covariate data layers, values of each covariate at the sample locations can be extracted to estimate the sample distribution using univariate kernel density estimation (Silverman, 1986):

$$\%f(v)=\sum_{i=1}^{n} w_i \frac{1}{\%h} K\left(\frac{v-V_i}{\%h}\right) \tag{1}$$

where $\tilde{f}(v)$ is the estimated sample distribution regarding covariate *v*, V_i is the value of *v* at sample location *i*, w_i is the weight for location *i*, and *n* is the total number of sample locations. The Gaussian kernel was adopted for the kernel function *K*:

$$K\left(\frac{v-V_i}{\%\tilde{h}}\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(v-V_i)^2}{2\tilde{h}^2}} \tag{2}$$

where $\tilde{h}$ is the bandwidth and was determined using the rule-of-thumb algorithm (Silverman, 1986):

$$\tilde{h}=1.06\times\tilde{\sigma}_v\times n^{-1/5} \tag{3}$$

in which $\tilde{\sigma}_v$ is the standard deviation of the values of v at all n sample locations.

The probability distribution of covariate v in the mapping area $f(v)$ (population distribution regarding v) was estimated similarly:

$$f(v)=\sum_{j=1}^{m}\frac{1}{h}K\left(\frac{v-V_j}{h}\right) \tag{4}$$

where V_j is the value of v at cell j and m is the total number of cells in the mapping area. The Gaussian kernel was used for K and the bandwidth h was determined using the rule-of-thumb algorithm (based on the standard deviation of the values of v at all m cells).

The similarity between the sample distributions $\tilde{f}(v)$ and the population distribution $f(v)$ was then computed as:

$$S^v=\frac{2\times A_{\tilde{f}(v)}\cap A_{f(v)}}{A_{\tilde{f}(v)}+A_{f(v)}} \tag{5}$$

where S^v is the similarity between the two distributions regarding covariate v, $A_{\tilde{f}(v)}$ and $A_{f(v)}$ are the areas under the two probability distribution curves, respectively, and $A_{\tilde{f}(v)}\cap A_{f(v)}$ is the overlapping area under the two curves (Figure 2).

The similarity between sample distribution and population distribution regarding each of the covariates was computed following Equations (1) through (5).

Finally, the sample representativeness R considering all covariates was computed as a weighted average of the similarities regarding individual covariates using the eigenvalues from principal component analysis as weights:

$$R=\sum_{v=1}^{L}\frac{\lambda^v}{\sum_{j=1}^{L}\lambda^j}S^v \tag{6}$$

where L is the number of covariates, λ^v is the eigenvalue of covariate (principal component) v, which is proportional to the percentage of variance it retains. The sample representativeness R has a value range of [0, 1.0], with a higher value indicating better sample representativeness.

The third step is to allocate weights to sample locations such that sample representativeness R is maximized. When estimating the sample distribution [Equation (1)], unequal weights can be assigned to sample locations. Different weight assignments will result in different sample distributions and thus different similarities to the population distribution. That is, it is possible to adjust the weights such that the sample distribution matches as closely as possible the population distribution (i.e., maximizing sample representativeness). Weights for the sample locations were determined using an optimization procedure (genetic algorithm) where the objective is to find a set of optimal weights for the sample locations that maximize sample representativeness. Under this representativeness-directed weighting scheme, sample locations over a less represented area are weighted heavier.

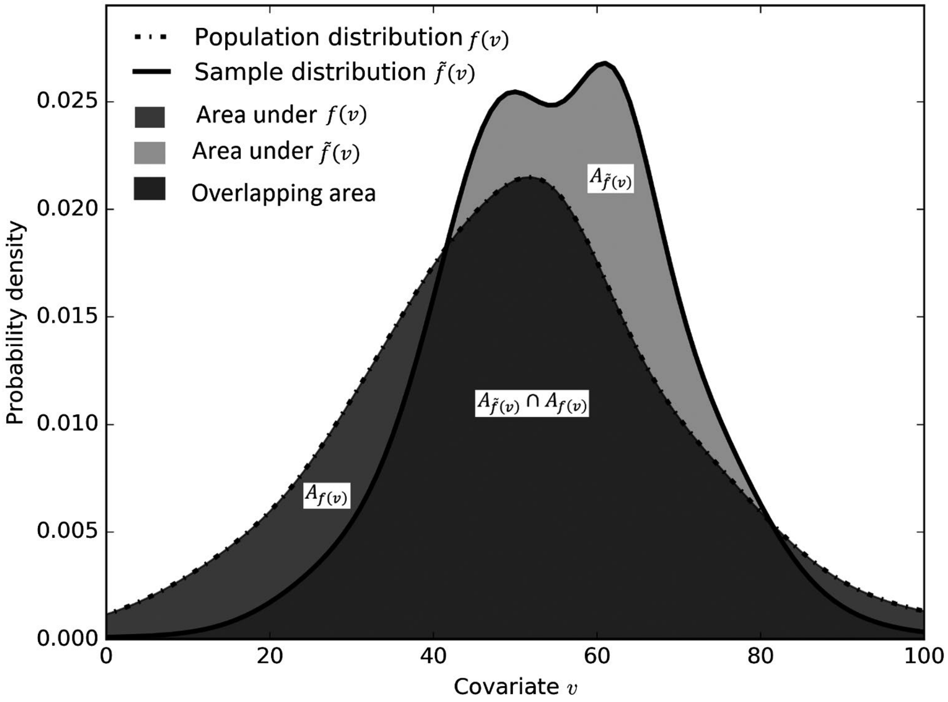

FIGURE 2 Similarity between the sample distribution and the population distribution regarding one covariate

The determined weights are used to weight sample locations in training models for predictive mapping. For example, when using weighted sample locations to train a regression model, the weights can be used to weight individual residual terms in determining regression coefficients based on ordinary least squares.

2.2.2 | Importance weighting

Importance weighting is a method developed in the machine learning community for correcting for sample selection bias, which is also referred to as covariate shift and transfer learning (Cortes et al., 2008; Pan & Wang, 2010; Zadrozny, 2004). Sample selection bias occurs when training data and test data are drawn from different underlying distributions. In the context of predictive mapping, due to spatial bias, the distribution of covariate values at the sample locations (training data) is different from the covariate distribution over the mapping area (from which test data are drawn).

To correct for sample selection bias, training examples are weighted by an importance weighting function to compute loss in learning classification and regression models (Shimodaira, 2000). For example, linear regression uses squared error as the loss function, which is a measure of the goodness-of-fit of the model. An importance weighting function can be used to weight individual residual terms in computing the sum of squared errors and therefore would affect the determination of model coefficients. It is proven that, asymptotically, the optimal weighting function is the ratio of the probability density function of features on the test data (test data distribution) and the density function on the training data (training data distribution) (Cortes et al., 2008; Zadrozny, 2004). The optimal weighting function can be computed based on empirical estimates of the two density functions.

Applying importance weighting to mitigate spatial bias in VGI-based samples for predictive mapping, the test data distribution is the multivariate probability density function of covariate values in the mapping area, as test data are assumed to be drawn from the mapping area. The training data distribution is the multivariate probability density function of covariate values at sample locations. These two density functions were estimated using multivariate kernel density estimation (Scott, 2015), implemented in the SciPy Python library (Jones, Oliphant, & Peterson, 2001). The ratios of the two density functions (at data points corresponding to sample locations in feature space) were then used to weight sample locations in training predictive models. Under this importance weighting scheme, sample locations in less represented areas receive heavier weights.

The representativeness-directed weighting method and the importance weighting methods both involve estimating a probability distribution of covariate values at sample locations, which is then compared to the covariate distribution over the mapping area to determine weights for bias mitigation. A large number of sample locations that spread across the mapping area, covering the full extent of covariate gradients, would be ideal for estimating robust probability distributions and examining more variety of weight assignments. Thus, sample size and spatial configuration of sample locations are supposed to affect the effectiveness of the two bias mitigation methods.

2.3 | Habitat suitability mapping

2.3.1 | Suitability modeling and mapping

Logistic regression (LR) was adopted for modeling and mapping species habitat suitability (Zhang & Zhu, 2019a). LR models can be trained using a training sample consisting of the *B. jamaicensis* occurrence locations and background locations, along with the in-situ covariate values. A set of 1,000 locations uniformly randomly selected from the study area were used as backgrounds. The *B. jamaicensis* occurrence locations from eBird and these background locations were used to train LR models. Species occurrence locations may carry different weights (as determined using the two bias mitigation methods; Section 2.2) when training the models, whilst the background locations had an equal weight of 1.

LR models were trained using procedures implemented in the scikit-learn package (Pedregosa et al., 2012), which is capable of accounting for weights of the training sample. A trained LR model was applied to the covariate values at every location (cell) in the study area to predict the in-situ habitat suitability and thereby produce a suitability map.

2.3.2 | Model performance evaluation

B. jamaicensis occurrence locations obtained from the North American Breeding Bird Survey (BBS) (https://www.pwrc.usgs.gov/BBS/RawData/) were used to evaluate the performance of the predictive LR model. BBS routes have roughly uniform spatial coverage and sample habitats representative of the entire region (Sauer et al., 2017). *B. jamaicensis* was observed at 73 stops on the active BBS routes surveyed in Wisconsin in the breeding season (May or June) of 2012 (Figure 3) (Zhang & Zhu, 2019a).

The AUC (area under the receiver operating characteristic curve) was adopted as a threshold-independent measure of predictive model performance (Phillips & Dudík, 2008). It can be computed based on a predicted suitability map and a set of validation data consisting of *B. jamaicensis* occurrence locations from BBS and background locations. A set of 1,000 locations were randomly chosen from the study area as backgrounds for validation. AUC is the probability that the predicted suitability at a randomly chosen species presence location is higher than that at a randomly chosen background location (Phillips, Anderson, & Schapire, 2006). The AUC value ranges from 0 to 1. AUC = 0.5 indicates random predictions; AUC < 0.5 indicates predictions contradicting the validation data; AUC > 0.5 indicates predictions agreeing with the validation data (AUC = 1 indicates perfect prediction).

Due to the lack of reliable absence locations of the species in this study, habitat suitability modeling was trained using species presence locations and randomly sampled background locations (Phillips et al., 2006). The model thus distinguishes environmental conditions that are suitable for the species from the background

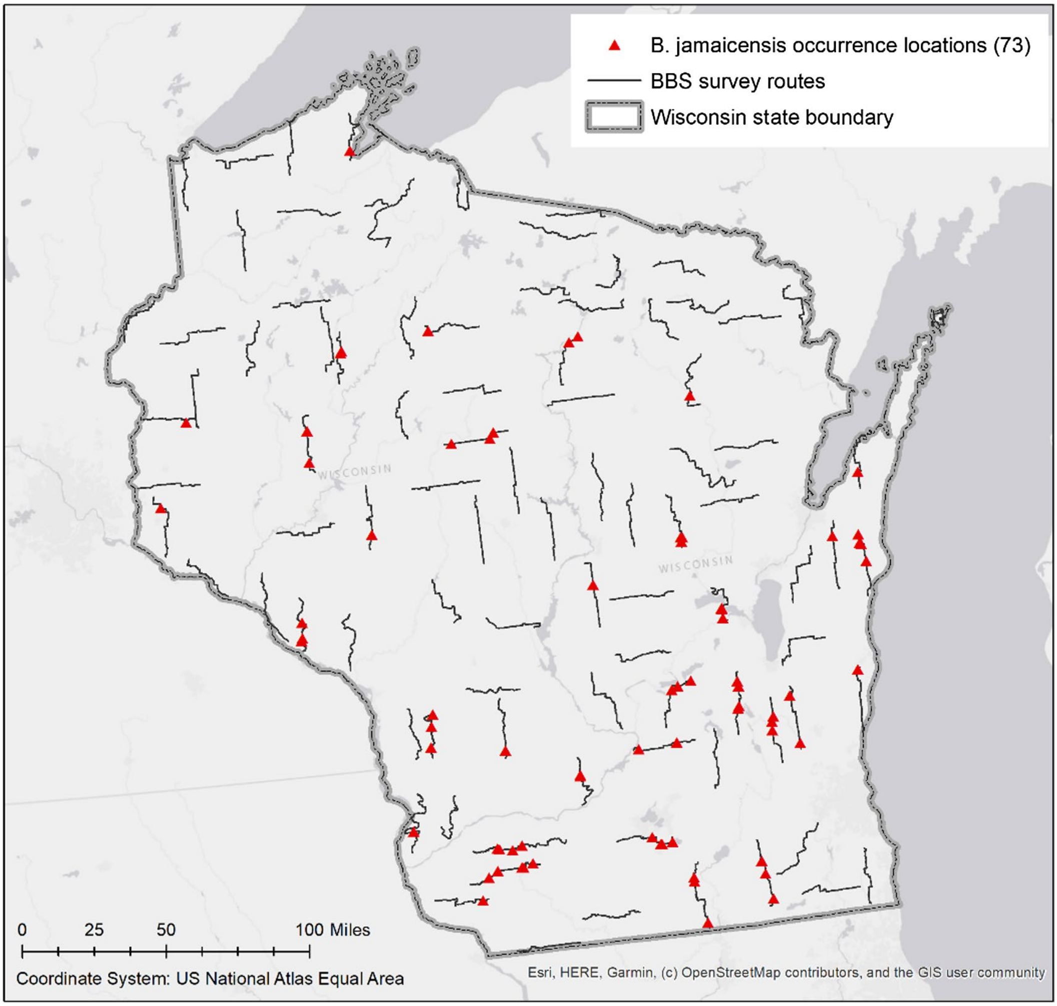
▲ B. jamaicensis occurrence locations (73)
BBS survey routes
Wisconsin state boundary
WISCONSIN
WISCONSIN
0
25
50
100 Miles
Coordinate System: US National Atlas Equal Area
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

FIGURE 3 *B. jamaicensis* occurrence locations obtained from 2012 BBS data

environmental conditions in the study area. The modeled habitat suitability is a relative index, not an estimate of species occurrence probability. It would be arbitrary to choose a suitability threshold to turn the suitability map into a presence/absence map for the purpose of evaluating the accuracy of the binary map (e.g., confusion matrix, overall accuracy, kappa index, etc.) (Hirzel, Le Lay, Helfer, Randin, & Guisan, 2006). In contrast, AUC was also computed based on species presence and background locations, and it is independent of any particular choice of suitability threshold. It has been widely used for evaluating the performance of species distribution and habitat suitability models (Dudik et al., 2005; Elith et al., 2006; Phillips et al., 2006, 2009; Phillips & Dudík, 2008; Zhang, Zhu, Huang, & Xiao, 2018; Zhang, Zhu, Windels, & Qin, 2018).

2.4 | Experiment designs

VGI-based samples with diverse characteristics in terms of sample size, spatial configuration, and number of species occurrences were constructed based on the VGI data. The spatial bias mitigation methods were then applied to mitigate spatial bias in the samples for habitat suitability mapping.

2.4.1 | Effort samples with varying sample sizes

eBird checklist locations were treated as a biased VGI-based effort sample, to which the spatial bias mitigation methods were applied to determine the weights (as discussed in Section 2.2). Effort sample size affects the determined weights and therefore would impact the bias mitigation methods' effectiveness in improving predictive model performance.

To investigate such impacts, effort samples were compiled at sample size ranging from 100 through 600 at an increment of 100. Specifically, a certain number of locations were randomly sampled at an equal selection probability from the 580 checklist locations where no *B. jamaicensis* occurrences were reported, such that an effort sample containing the sampled locations and the species occurrence locations would have the desired sample size. For example, at sample size 100, only 25 locations were sampled from the 580 non-occurrence checklist locations. The 25 locations were combined with the 75 *B. jamaicensis* occurrence locations to form an effort sample. This was repeated 20 times to account for randomness in the sampling at each sample size. As a result, there were 20 effort samples at each sample size.

The 75 species occurrence locations were always kept in the effort samples. This allowed examining the sole impact of effort sample size with an equal number of species occurrence locations used to train predictive models. Moreover, as the non-occurrence locations were randomly sampled from the 580 locations, the effort samples of different sizes were expected to have similar spatial configurations. This allowed assessing the impact of effort sample size while controlling for sample spatial configuration.

2.4.2 | Effort samples with varying spatial configurations

The spatial distribution characteristics of a VGI-based effort sample could also affect the effectiveness of the bias mitigation methods (as argued in Section 1). To examine such effects, effort samples with various spatial configurations were constructed, as described below.

The study area was first divided into four spatial zones, namely north west (NW), north east (NE), south east (SE), and south west (SW), by bisecting its rectangular bounding box along the north-south and the east-west directions (Figure 4). Accordingly, the 580 non-occurrence checklist locations were divided into the four zones. Checklist locations were dense in the SE zone, followed by the NW zone, whilst those in the NE and SW zones were relatively sparse.

The locations in each zone were then combined with the 75 *B. jamaicensis* occurrence locations to form an effort sample. This resulted in four effort samples, which were denoted by NW, NE, SE, and SW, respectively. Besides, the locations in any two zones were combined with the occurrence locations to form an effort sample. This resulted in six effort samples, denoted by N, E, S, W, NW-SE, and NE-SW, respectively. Furthermore, the locations in any three zones (i.e., locations in the other zone were excluded) were combined with the occurrence locations to form an effort sample. This resulted in four effort samples, denoted by exNW, exNE, exSE, and exSW, respectively. Finally, the 580 non-occurrence locations were then combined with the 75 *B. jamaicensis* occurrence locations to form yet another effort sample, demoted by ALL.

These effort samples had varying spatial configurations. Nonetheless, the 75 species occurrence locations were always present in each effort sample. This allowed examining the impacts of effort sample spatial configuration with an equal number of species occurrence locations used to train predictive models. Although it is difficult to strictly control for effort sample size across various sample spatial configurations, results show that for this set of effort samples spatial configuration (not sample size) was the dominant factor determining sample representativeness (Section 3.2.1) and model performance improvement (Section 3.2.3). Therefore, the marginal effects of sample size can reasonably be assumed negligible.

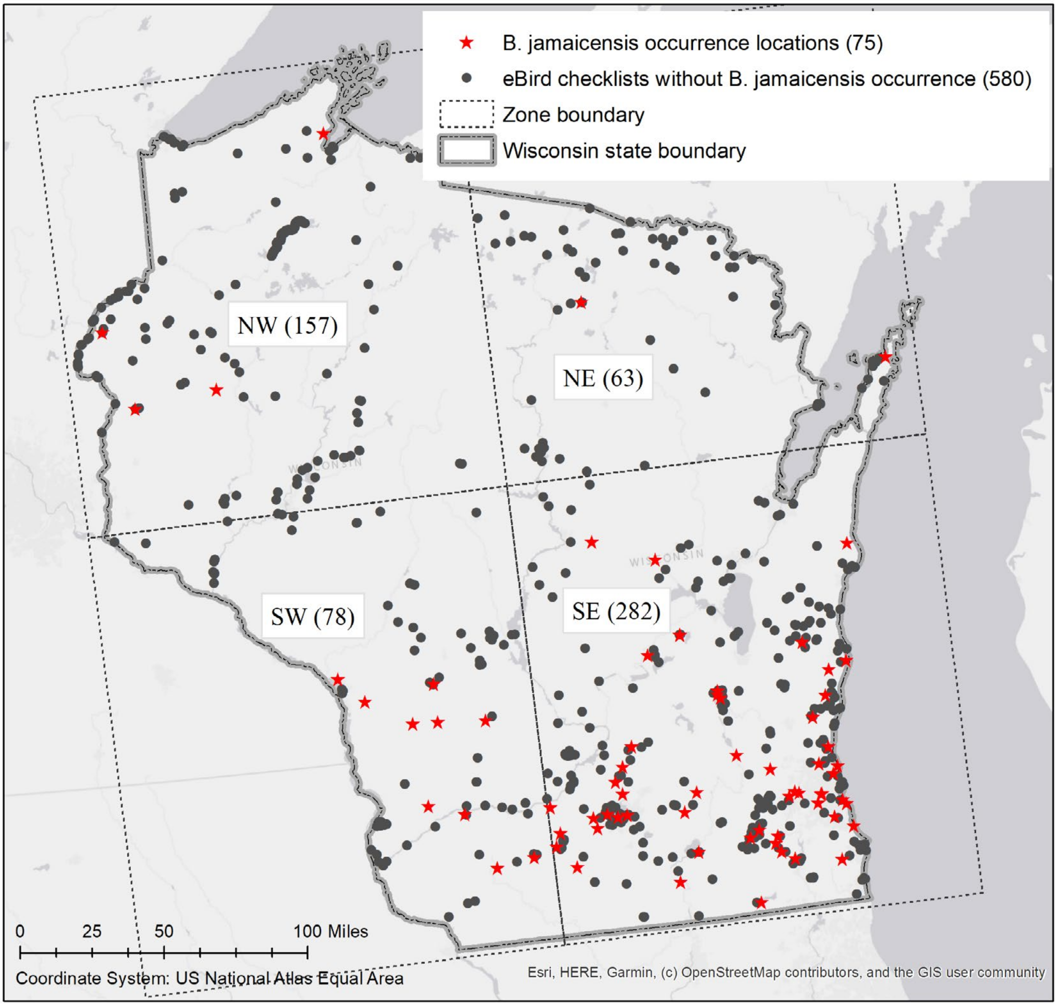
★ B. jamaicensis occurrence locations (75)
● eBird checklists without B. jamaicensis occurrence (580)
Zone boundary
Wisconsin state boundary
NW (157)
NE (63)
SW (78)
SE (282)
0 25 50 100 Miles
Coordinate System: US National Atlas Equal Area
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

FIGURE 4 Division of eBird checklist locations into four spatial zones

2.4.3 | Effort samples with varying number of species occurrences

Since only *B. jamaicensis* occurrence locations were used in training predictive models for habitat suitability mapping (Section 2.3.1), the number of species occurrence locations in effort samples could affect the effectiveness of the bias mitigation methods for improving model performance. To examine such effects, VGI-based effort samples containing varying number of species occurrence locations were compiled.

The number of species occurrences ranged from 10 through 70 at an increment of 5. For each number of species occurrence locations, a corresponding number of locations were randomly sampled at an equal selection probability from the 75 *B. jamaicensis* occurrence locations. The sampled locations were combined with the 580 non-occurrence checklist locations to form an effort sample. This was repeated 20 times to account for randomness in the sampling. As a result, there were 20 effort samples for each number of species occurrences.

The 580 non-occurrence locations were always present in each effort sample. Results indicate that variations in representativeness of this set of effort samples were very small due to the only slight differences in effort sample size (Section 3.3.1). This allowed examining the impacts of number of species occurrence locations with similar effort sample spatial configuration and close effort sample size.

2.4.4 | Spatial bias mitigation for suitability mapping

The two bias mitigation methods (Section 2.2) were applied to determine weights for each of the above VGI samples. Sample representativeness of the unweighted or weighted VGI samples (as defined in the representativeness-directed weighting approach; Section 2.2.1) was evaluated. Weights associated with *B. jamaicensis* occurrence locations were then used to weight the occurrence locations in training predictive models for suitability mapping.

AUC values of the suitability models were computed (Section 2.3.2) and compared using statistical analyses (i.e., *t* test, Spearman's rank correlation coefficient r_s, coefficient of determination r^2) to investigate the impacts of VGI sample size, effort sample spatial configuration, and number of species occurrences in the effort sample on the effectiveness of the bias mitigation methods. The statistical analyses were performed using the scipy.stats Python library (Jones et al., 2001).

3 | RESULTS

3.1 | Effort samples with varying sample sizes

3.1.1 | Representativeness versus effort sample size

Across various sample sizes, weighting effort samples with weights determined through the representativeness-directed weighting method consistently improved sample representativeness (Figure 5). The improvements were statistically significant based on paired-sample Student *t* tests (Table 1).

The representativeness of an effort sample generally increases as sample size increases, until reaching a plateau at a certain sample size (e.g., 400) (Figure 5). For unweighted and weighted effort samples, r_s between the mean effort sample representativeness and sample size were r_s = 1 (p = .000; n = 6) and r_s = .942 (p = .004; n = 6), respectively, suggesting a statistically significant strong positive correlation between effort sample size and effort sample representativeness. In addition, sample size explains about 76% (r^2 = .766; p = .000; n = 6) and 58% (r^2 = .586; p = .001; n = 6) of the variations in mean representativeness of unweighted and weighted effort samples, respectively.

3.1.2 | AUC versus sample size

Compared to the baseline performance of the suitability model trained using unweighted species occurrence locations (AUC = 0.716), the spatial bias mitigation methods effectively improved AUC by weighting species occurrence locations in training LR models (Figure 6). This observation holds across various effort sample sizes and the performance improvements were statistically significant (Table 2). Notably, the representativeness-directed weighting method performed better than the importance weighting method on improving AUC. At sample size 300 and beyond, the mean AUC achieved using representativeness-directed weighting was statistically significantly higher than that achieved using importance weighting (Table 2).

AUC achieved using the two weighting methods increases as effort sample size increases, before plateauing at sample size 300–400 (Figure 6). The representativeness weighting method performed better (as measured by AUC) on effort samples of larger sample sizes (r_s = .828; p = .041; n = 6) and sample size explains about 72% of the variations in mean AUC (r^2 = .724; p = .000; n = 6). Effort sample size did not statistically significantly correlate with performance of the importance weighting method (r_s = .771; p = .072; n = 6), but sample size still explains about 54% of the variations in mean AUC (r^2 = .547; p = .002; n = 6).

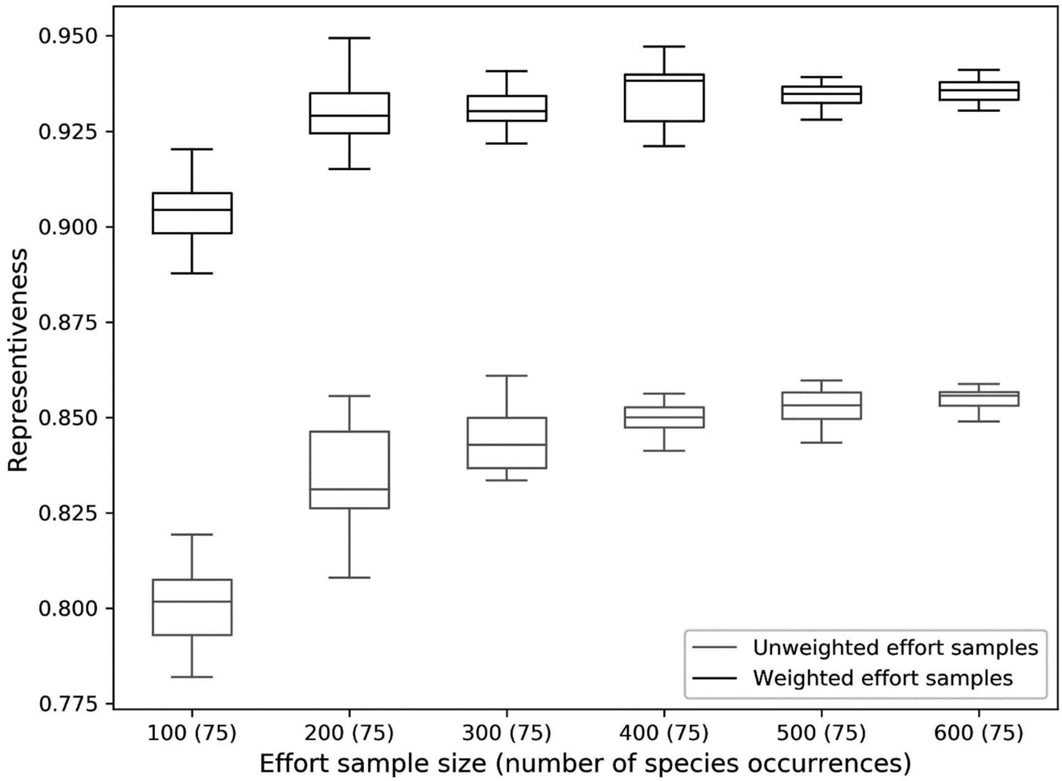

FIGURE 5 Sample representativeness improvement at various effort sample sizes

TABLE 1 Statistical significance tests on the difference between the mean representativeness of unweighted and weighted effort samples (paired-sample *t* tests; *df* = 19; one-tailed)

Sample size		100	200	300	400	500	600
Representativeness (unweighted sample)	Mean	0.802	0.834	0.844	0.850	0.852	0.855
	SD	0.011	0.012	0.008	0.006	0.005	0.003
Representativeness (weighted sample)	Mean	0.904	0.930	0.932	0.934	0.934	0.935
	SD	0.009	0.009	0.006	0.008	0.004	0.003
t Statistic		71.626	51.292	88.779	79.682	89.887	135.459
p Value		.000	.000	.000	.000	.000	.000

3.1.3 | AUC versus representativeness

Mean AUC achieved using the two weighting methods increases as the mean representativeness of effort samples increases, before plateauing around a mean representativeness of 0.85 (Figure 7). The representativeness weighting method was more effective in improving AUC on effort samples of higher representativeness (r_s = .828; p = .041; n = 6) and representativeness explains about 97% of the variations in mean AUC (r^2 = .978; p = .000; n = 6). Mean sample representativeness did not statistically significantly correlate with the performance of the importance weighting method (r_s = .771; p = .072; n = 6), but sample representativeness still explains about 94% of the variations in mean AUC (r^2 = .944; p = .000; n = 6).

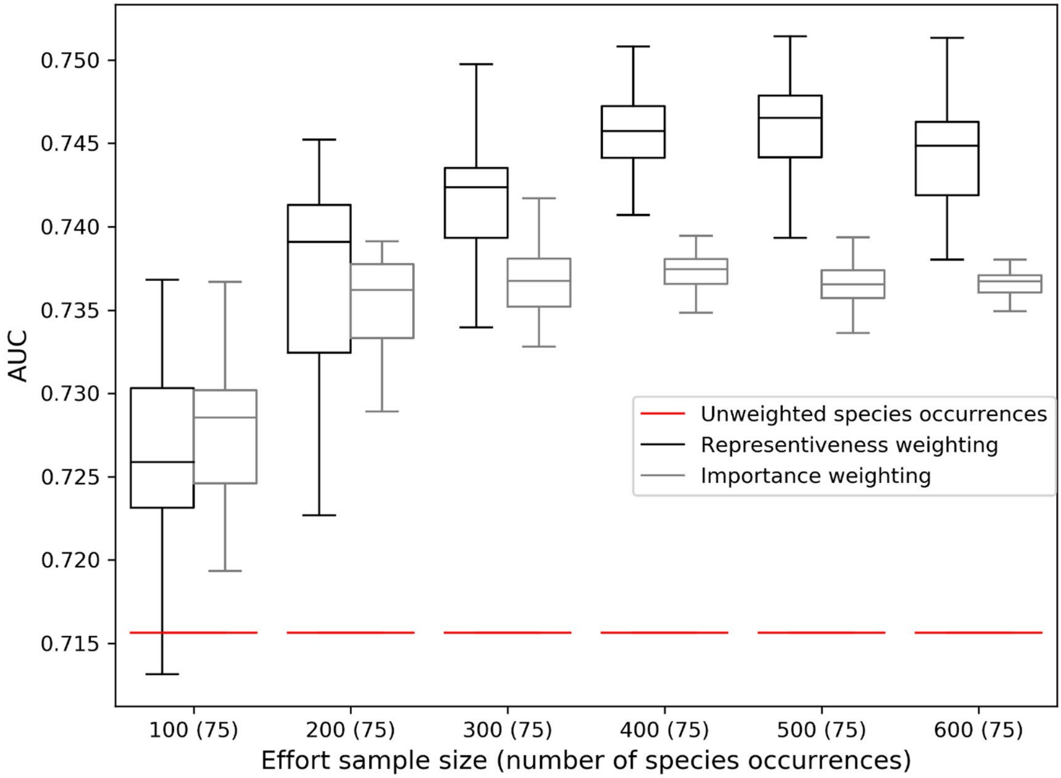

FIGURE 6 Performance of predictive suitability models trained using effort samples with various sample sizes

TABLE 2 Statistical significance tests on the differences between the mean AUC achieved using the two weighting methods and the baseline AUC = 0.716 using unweighted species occurrences (one-sample *t* tests; *df* = 19; one-tailed) and the differences between the mean AUC achieved using the two weighting methods (paired-sample *t* tests; *df* = 19; one-tailed)

Sample size		100	200	300	400	500	600
Representativeness weighting	Mean	0.726	0.737	0.742	0.746	0.746	0.744
	SD	0.007	0.006	0.004	0.003	0.003	0.004
	t Statistic	6.442	15.422	31.204	40.739	46.555	33.434
	p Value	.000	.000	.000	.000	.000	.000
Importance weighting	Mean	0.728	0.735	0.736	0.737	0.737	0.737
	SD	0.004	0.003	0.003	0.001	0.001	0.001
	t Statistic	12.463	26.125	29.317	67.336	61.536	88.595
	p Value	.000	.000	.000	.000	.000	.000
t Statistic		−1.435	1.384	4.680	10.261	12.825	9.818
p Value		.084	.091	.000	.000	.000	.000

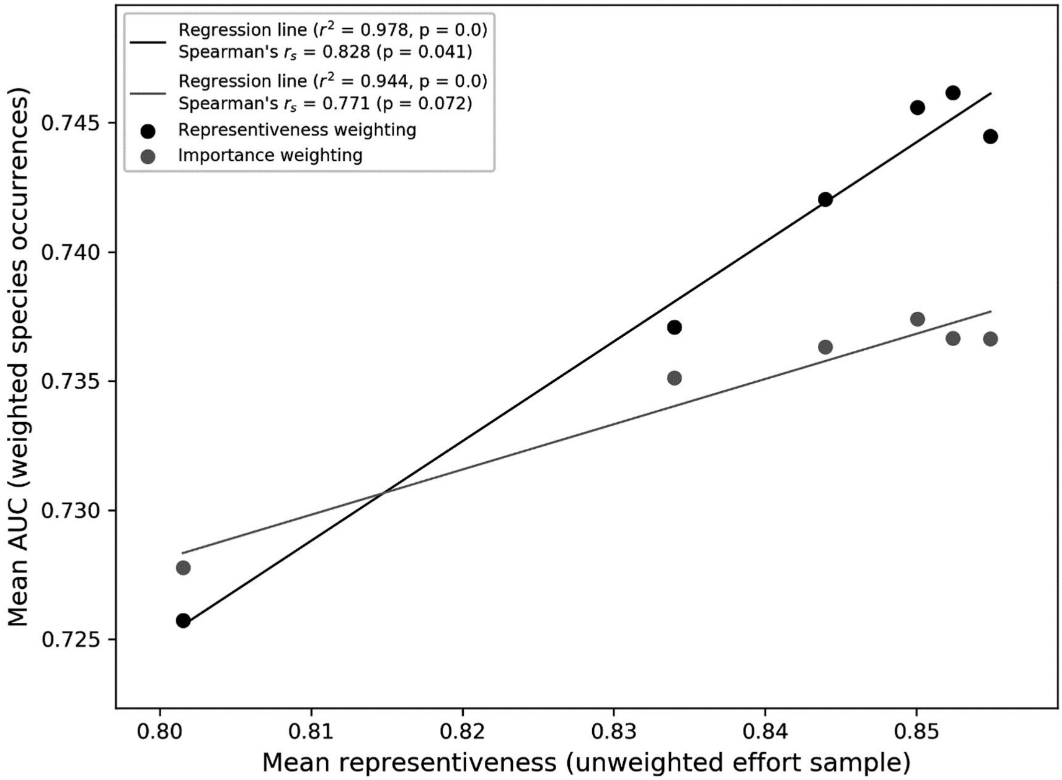

FIGURE 7 Relationship between mean AUC and mean representativeness of effort samples with varying sample size

3.2 | Effort samples with varying spatial configurations

The effectiveness of the two bias mitigation methods to improve suitability model performance on effort samples of varying spatial configurations is shown in Table 3. Generally, weighting species occurrences using the two methods helped improve suitability model performance (i.e., increased AUC) compared to the performance achieved based on unweighted occurrences (AUC = 0.716). The only exceptions were effort samples SE and E, where the importance weighting method degraded model performance (AUC = 0.701 and 0.708, respectively). This might be partially due to the two samples being of relatively low representativeness compared to other samples.

3.2.1 | Representativeness versus effort sample spatial configuration

Among the effort samples with varying spatial configurations, sample NE-SW has the highest representativeness of 0.860 at a relatively small sample size of 216, whilst sample SE has the lowest representativeness of 0.779 at a moderate sample size of 357 (Table 3). This disproportionality between sample representativeness and sample size was not surprising given the spatial distribution characteristics of the two effort samples (Figure 8). Sample NE-SW spreads widely and covers most parts of the study area, except the north-west. Sample SE, in contrast, is relatively clustered in the south-east with very limited coverage in the north-west, north-east, and south-west. Obviously, the representativeness of an effort sample is greatly affected by the spatial configuration

TABLE 3 Effectiveness of the two bias mitigation methods on effort samples with varying spatial configurations (all effort samples contain 75 species occurrences)

		Unweighted		Weighted		
Zone	Effort sample size	Representativeness	AUC	Representativeness	AUC (representativeness weighting)	AUC (importance weighting)
NW	232	0.817	0.716	0.928	0.742	0.742
NE	138	0.841	0.716	0.929	0.741	0.725
SE	357	0.779	0.716	0.863	0.738	0.701
SW	153	0.820	0.716	0.893	0.734	0.743
N	295	0.815	0.716	0.925	0.746	0.742
E	420	0.812	0.716	0.923	0.730	0.708
S	435	0.805	0.716	0.888	0.734	0.721
W	310	0.845	0.716	0.930	0.735	0.752
NW-SE	514	0.836	0.716	0.926	0.747	0.725
NE-SW	216	0.860	0.716	0.936	0.748	0.745
exNW	498	0.832	0.716	0.931	0.734	0.724
exNE	592	0.847	0.716	0.933	0.753	0.736
exSE	373	0.840	0.716	0.929	0.738	0.752
exSW	577	0.845	0.716	0.928	0.748	0.727
ALL	655	0.856	0.714	0.935	0.749	0.737

of the sample. Nevertheless, weighting effort samples using weights determined through the representativeness-directed approach consistently improved sample representativeness (Table 3).

3.2.2 | AUC versus representativeness

Across effort samples with various spatial configurations, weighting species occurrences in training suitability models using weights determined through the two bias mitigation methods also consistently increased the performance of the models compared to the baseline performance (AUC = 0.716). The mean AUC achieved by the representativeness weighting method was statistically significantly higher than that achieved by the importance weighting method (paired-sample *t* tests; *df* = 14; one-tailed; *t* statistic = 2.389; *p* = .016).

AUCs of the suitability models trained using weighted species occurrences were statistically significantly positively correlated with the representativeness of effort samples (r_s = .675, *p* = .005, *n* = 15 and r_s = .521, *p* = .046, *n* = 15 for representativeness weighting and importance weighting, respectively) (Figure 9). Effort sample representativeness explains a moderate portion of variations in the AUCs (r^2 = .283, *p* = .000, *n* = 15 and r^2 = .362, *p* = .000, *n* = 15, respectively, for the two weighting methods).

3.2.3 | Effects of effort sample size

With varying spatial configurations, the correlation between sample representativeness and effort sample size was no longer statistically significant (r_s = .214; *p* = .443; *n* = 15). Sample size explains only less than 5% of the variations in sample representativeness (r^2 = .047; *p* = .044; *n* = 15). This implies that the representativeness of the effort samples was more affected by spatial configuration than sample size.

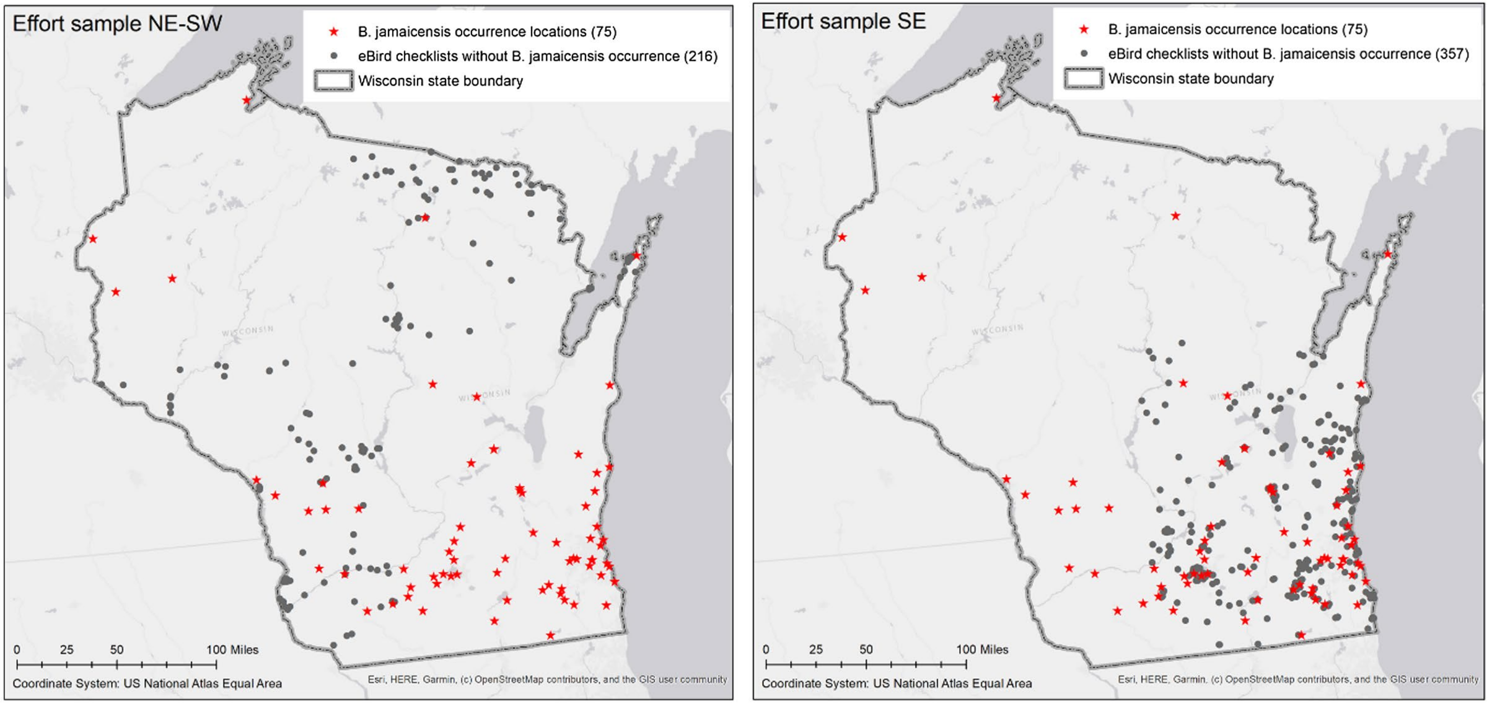
Effort sample NE-SW
★ B. jamaicensis occurrence locations (75)
● eBird checklists without B. jamaicensis occurrence (216)
Wisconsin state boundary
WISCONSIN
0 25 50 100 Miles
Coordinate System: US National Atlas Equal Area
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community
Effort sample SE
★ B. jamaicensis occurrence locations (75)
● eBird checklists without B. jamaicensis occurrence (357)
Wisconsin state boundary
WISCONSIN
0 25 50 100 Miles
Coordinate System: US National Atlas Equal Area
Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

FIGURE 8 Spatial distribution of effort samples NE-SW (left) and SE (right)

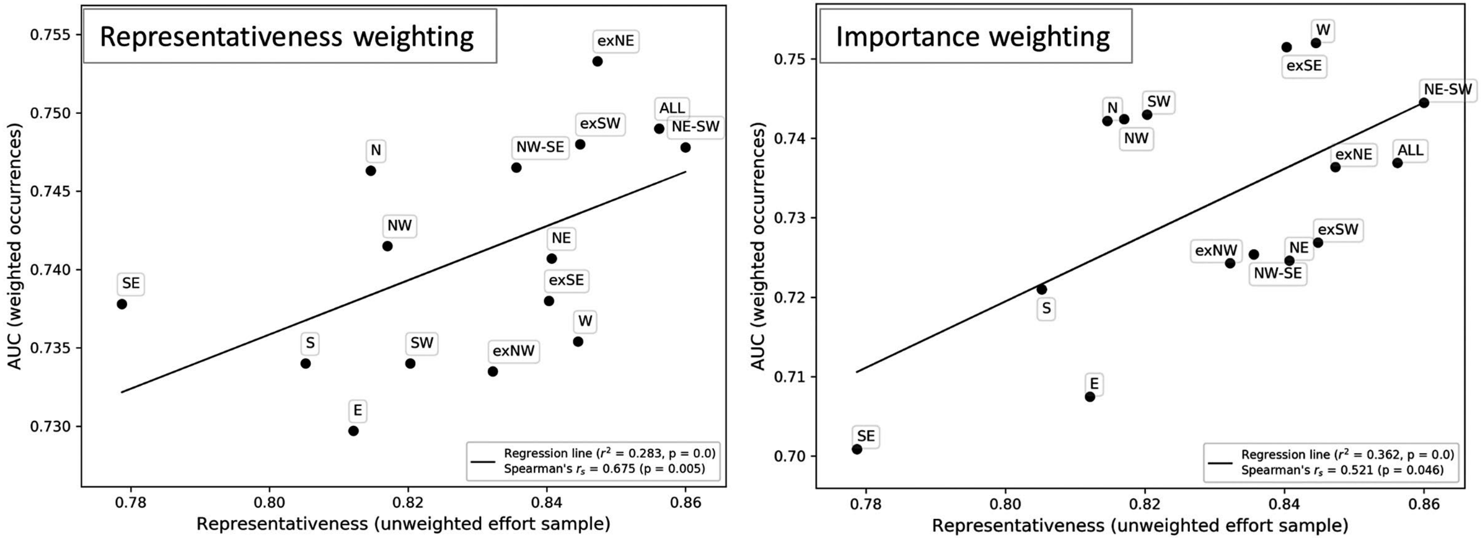

FIGURE 9 Relationship between AUC and representativeness of effort samples with varying spatial configurations

The correlation between AUC of the suitability model trained using weighted species occurrences and effort sample size was not statistically significantly either. For the representativeness weighting and importance weighting methods, r_s between AUC and effort sample size was $r_s = .328$ ($p = .231$; $n = 15$) and $r_s = -.31$ ($p = .259$; $n = 15$), respectively. Sample size explains a very small portion of the variations in AUC ($r^2 = .126$, $p = .002$, $n = 15$ and $r^2 = .055$, $p = .03$, $n = 15$ for the two methods, respectively). It again implies that, with varying spatial configurations, the effectiveness of the two methods in improving model performance was mostly affected by the spatial configuration of the effort samples rather than the sample size.

3.3 | Effort samples with varying number of species occurrences

3.3.1 | Representativeness versus number of species occurrences

Regardless of weighting the effort samples or not, variations in representativeness of the samples containing varying number of species occurrences were very small due to the only slight differences in effort sample size (Figure 10). Nonetheless, there is a strong positive correlation between effort sample size and mean representativeness for both unweighted and weighted effort samples ($r_s = 1$, $p = .000$, $n = 7$). Effort sample size explains over 97% of the variations in mean representativeness ($r^2 = .973$, $p = .000$, $n = 7$ and $r^2 = .992$, $p = .000$, $n = 7$, respectively, for unweighted and weighted effort samples). This observation is consistent with that in Section 3.1.1.

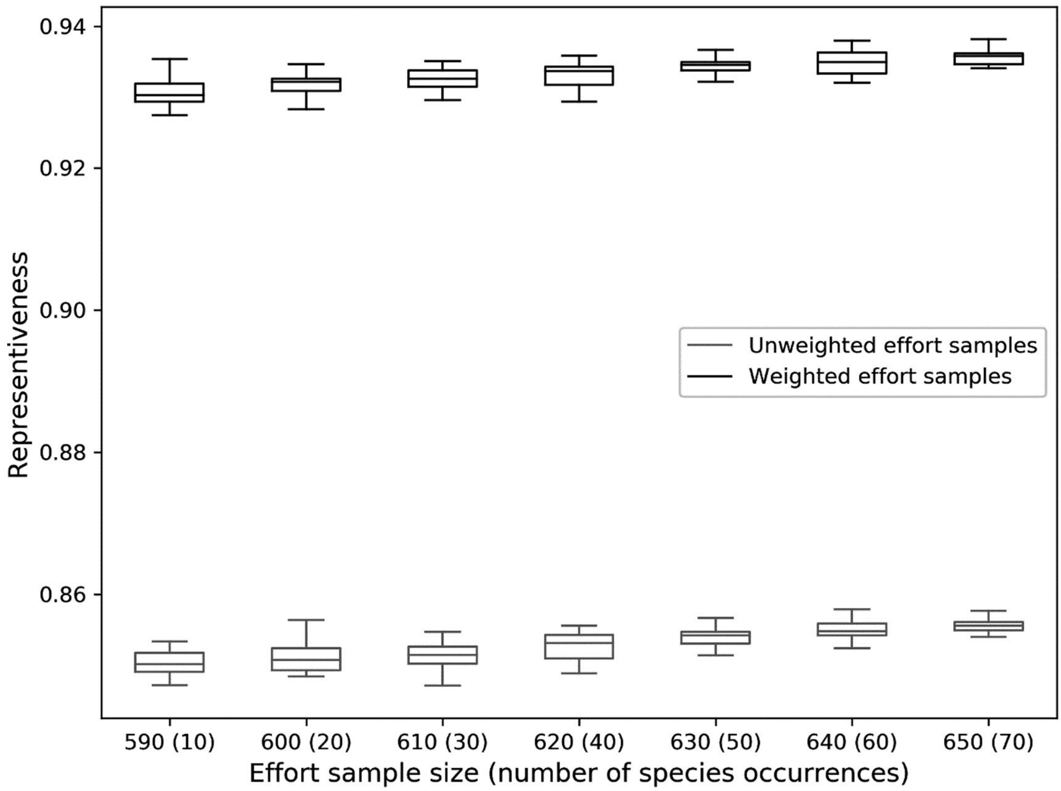

FIGURE 10 Representativeness of effort samples containing varying number of species occurrence locations

3.3.2 | AUC versus number of species occurrences

Compared to suitability models trained using unweighted species occurrences, models trained using species occurrences with representativeness weighting had statistically significant higher mean AUC across various number of species occurrences (Figure 11; Table 4). For importance weighting, the improvements were observed at 30 species occurrences and beyond. Overall, representativeness weighting improved AUCs statistically significantly more than importance weighting did (Table 4).

AUC achieved using the two weighting methods increases as the number of species occurrences increases, before reaching a plateau around 60 (Figure 11). A strong positive correlation exists between the number of species occurrences and the mean AUC achieved by representativeness weighting and importance weighting (r_s = .964, p = .000, n = 7 and r_s = 1, p = .000, n = 7, respectively). The number of species occurrences explains over 84% of the variations in the mean AUC achieved by the two weighting methods (r^2 = .844, p = .000, n = 7 and r^2 = .916, p = .000, n = 7, respectively).

3.3.3 | AUC versus representativeness

Mean AUC achieved using representativeness weighting and importance weighting increases as the effort sample representativeness increases, before the representativeness weighting method reaches a plateau around a representativeness of 0.855. The weighting methods were more effective in improving AUC on effort samples of higher

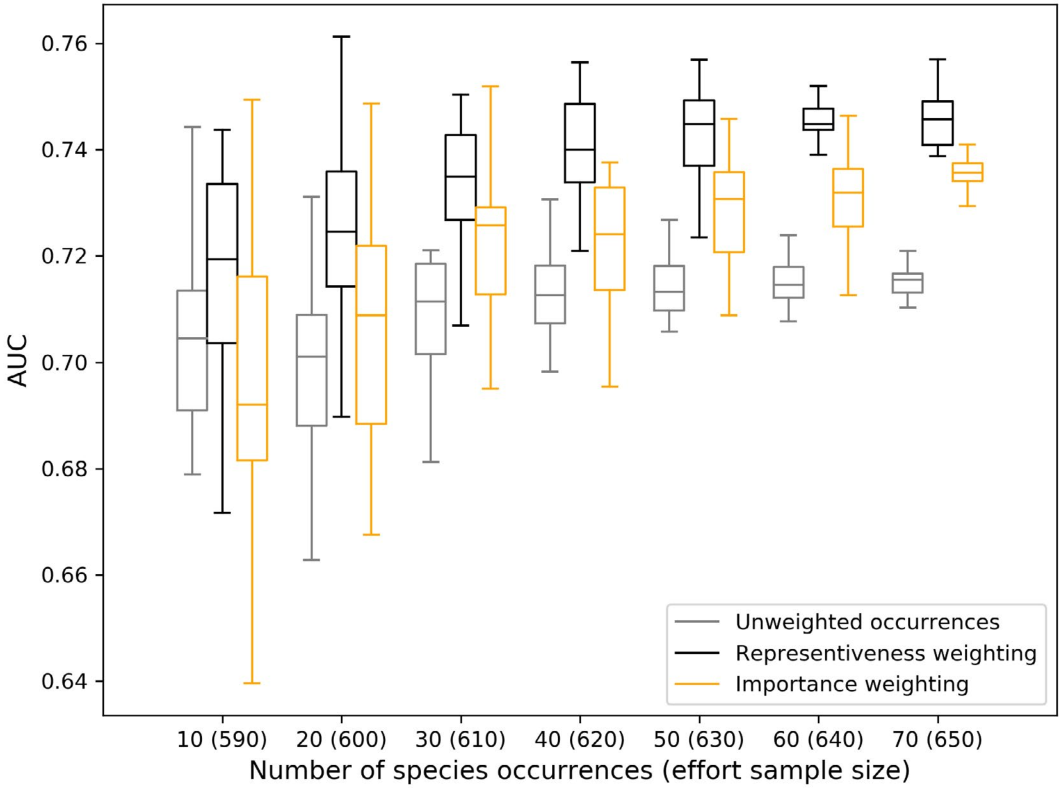

FIGURE 11 Performance of predictive suitability models trained using various number of species occurrence locations

TABLE 4 Statistical significance tests on the differences between the mean AUC achieved using the two weighting methods and using unweighted species occurrences (paired-sample *t* tests; *df* = 19; one-tailed) and the differences between the mean AUC achieved using the two weighting methods (paired-sample *t* tests; *df* = 19; one-tailed)

Number of species occurrences	10	20	30	40	50	60	70	
Unweighted	Mean	0.701	0.698	0.708	0.712	0.714	0.715	0.715
	SD	0.027	0.019	0.011	0.010	0.008	0.004	0.004
Representativeness weighting	Mean	0.715	0.722	0.734	0.741	0.742	0.744	0.744
	SD	0.026	0.022	0.012	0.010	0.009	0.006	0.009
	t Statistic	2.432	5.425	12.553	14.016	13.817	20.273	17.019
	p Value	.013	.000	.000	.000	.000	.000	.000
Importance weighting	Mean	0.695	0.707	0.720	0.720	0.729	0.730	0.735
	SD	0.029	0.020	0.019	0.017	0.010	0.010	0.004
	t Statistic	−1.004	1.479	2.885	2.420	5.429	6.927	20.065
	p Value	.164	.078	.005	.013	.000	.000	.000
t Statistic		3.767	3.219	3.357	7.129	5.226	7.614	5.367
p Value		.000	.003	.002	.000	.000	.000	.000

representativeness (r_s = .964, p = .000, n = 7 and r_s = 1, p = .000, n = 7, respectively). Effort sample representativeness explains over 76% of the variations in the mean AUC achieved by the two weighting methods (r^2 = .764, p = .000, n = 7 and r^2 = .831, p = .000, n = 7, respectively).

4 | DISCUSSION

As revealed in Section 3, across the VGI-based samples with diverse characteristics, the two spatial bias mitigation methods were effective in improving the performance of the predictive suitability model, although the representativeness weighting method consistently outperformed the importance weighting method. Nonetheless, there are commonalities regarding how sample size, spatial configuration, and representativeness of a VGI sample, and the number of species occurrences in the VGI sample, impacted the effectiveness of the two spatial bias mitigation methods.

4.1 | Impact of VGI sample size

A positive correlation between suitability model performance (i.e., AUC) and VGI sample size was observed, regardless of whether there were equal number of species occurrences in the VGI samples, provided that the samples have similar spatial configurations. For example, samples examined in Section 3.1 (which contained an equal number of species occurrences) and Section 3.3 (which contained a varying number of species occurrences) all had similar spatial configurations as they were randomly selected from the original eBird checklist locations. Across these samples, AUCs of the suitability models trained using the species occurrences weighted using the two spatial bias mitigation methods initially increase as the sample size increases, before reaching a plateau beyond a certain sample size (Figures 6 and 11). The plateau indicates a ceiling effect of sample size on effectiveness of the bias mitigation methods.

In general, as the size of the samples increases, the range of sample representativeness and AUC values decreases, because the variety in samples of larger size is not as high as that in samples of smaller size (e.g., there are only 10 possible different combinations of selecting 9 sample locations out of a total of 10 locations, while there are 252 combinations of selecting 5 locations from the 10).

4.2 | Impact of VGI sample spatial configuration

VGI samples examined in Section 3.2 had drastically different spatial configurations. Across the samples, a correlation between AUC and sample size was absent (Section 3.2.3). For such samples, it was the spatial configuration of the samples, not the sample size, that affected the effectiveness of the spatial bias mitigation methods in improving suitability model performance (Section 3.2). VGI samples with spatial configurations that are more spread out tend to be more amenable to bias mitigation for improving model performance, whilst samples with limited spatial coverage leave little space for bias mitigation.

4.3 | Impact of number of species occurrences in VGI samples

The number of species occurrences contained in VGI samples contributes to VGI (effort) sample size. It would thus impact the effectiveness of the bias mitigation methods in a way similar to effort sample size (as discussed in Section 4.1). Moreover, it should have a more direct impact on the performance of predictive suitability models as only species occurrence locations, not the non-occurrence locations, were used to train models. For example, one needs to add in a much larger number of non-occurrence locations than species occurrence locations to achieve a comparable amount of AUC improvement using the representativeness weighting method (comparing Figures 6 and 11).

4.4 | Impact of VGI sample representativeness

The representativeness of a VGI sample is an overall reflection of the intertwined effects of effort sample size, spatial configuration, and number of species occurrences. An observation that was rather consistent across the diverse VGI samples was that suitability model performance (AUC) was positively correlated with VGI sample representativeness, suggesting that bias mitigation methods were more effective on VGI samples of higher representativeness. Also, AUC plateaued beyond a certain representativeness threshold (see Figures 7 and 12), which may indicate a ceiling effect of sample representativeness on the effectiveness of the bias mitigation methods.

4.5 | Sampling designs in VGI

In most cases, volunteers conduct observations in an ad-hoc, opportunistic, and uncoordinated manner without following any (probability) sampling designs. As a result, the inclusion probability of sample locations is unknown and VGI-based samples are often non-probability ones (e.g., eBird data used in this study). In rare cases, volunteers could be directed to sample locations designed following probability sampling protocols to collect probability samples, and the inclusion probability of sample locations is known (see Section 3.3 in Stehman, Fonte, Foody, & See, 2018 for examples).

The representativeness-directed weighting and importance weighting methods fall under the model-based inference framework (Gregoire, 1998). They do not rely on inclusion probability of sample locations to mitigate

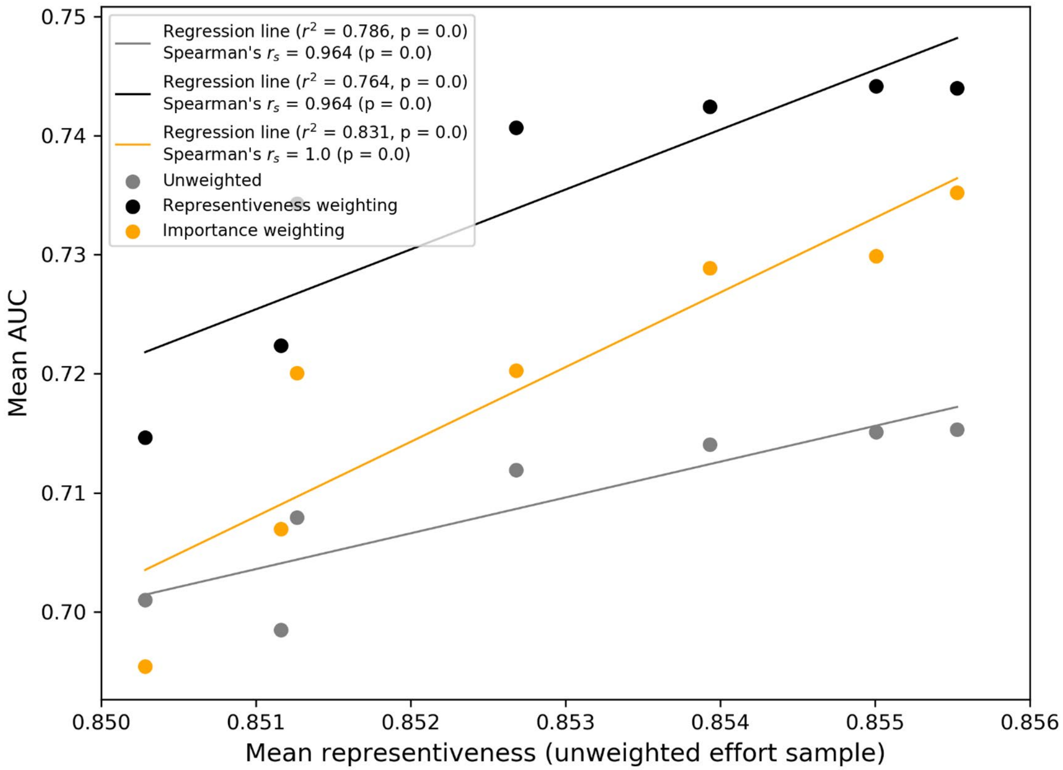

FIGURE 12 Relationship between mean AUC and mean representativeness of effort samples with varying number of species occurrences

spatial bias and thus are supposed to be applicable to both non-probability and probability VGI-based samples. However, when the methods are applied to probability samples, information on inclusion probability is simply discarded (samples are treated as if they are non-probability ones). For mitigating bias in probability samples, other methods under the design-based statistical inference framework that make use of inclusion probability (e.g., Bethlehem, 2010; Stehman et al., 2018) may be more beneficial and efficient. How sample size and spatial configuration affect the effectiveness of such bias mitigation methods is beyond the scope of this article and deserves separate treatment.

4.6 | Limitations

There are limitations in this study. Samples of varying spatial configurations were obtained simply by dividing the sample locations into quadrants of the study area based on cardinal directions. In general, the spatial configuration of a point pattern can be more complex, and the clustering or sparseness of sample locations does not always align with the division of quadrants.

Ideally, when investigating the impacts of sample size and sample spatial configuration, the condition of one factor should be kept constant while varying the condition of the other factor (i.e., controlled experiments). For instance, when examining the impact of sample size, the spatial configuration of the samples should be held changeless or changes be negligible across different sample sizes. When investigating the effect of sample spatial configuration, sample size should be held constant. Yet strict controls are difficult to implement when working with real-world data. In this study's experiment designs, such controls were maintained as much as possible and experiment results support that the controls are reasonable (see detailed discussion in Sections 2.4.1–2.4.3, respectively).

5 | CONCLUSIONS

This study presents an empirical evaluation of the impacts of VGI sample size and spatial configuration on the effectiveness of spatial bias mitigation methods in improving the performance of predictive models through a habitat suitability mapping case study. Evaluation results showed that: (a) predictive model performance improvement brought by the bias mitigation methods was positively correlated with sample size when samples have similar spatial configurations; (b) VGI samples with more spread-out spatial coverages were more amenable to bias mitigation for improving model performance; (c) the bias mitigation methods were more effective on VGI samples of higher representativeness, where sample representativeness is a measure encapsulating the intertwined effects of sample size and sample spatial configuration; and (d) model performance improvement plateaued beyond a certain sample size and sample representativeness thresholds, indicating ceiling effects on the effectiveness of the bias mitigation methods.

Spatial bias and other forms of bias are a prominent VGI data quality issue when using VGI for spatial analyses and geographic or environmental modeling (Zhang & Zhu, 2018). The findings of this study, though derived from a habitat suitability modeling case study, are expected to inform assessing the fitness and effectiveness of VGI spatial bias mitigation methods for improving the performance of predictive models in general. Many predictive modeling methods used for regression or classification problems (multivariate regression, decision trees, random forests, etc.) fall in the same paradigm as the logistic regression method used for suitability modeling in this study, wherein a target variable (continuous or discrete) is modeled as a function of a set of covariates such that the values of that target variable can be inferred from observations of the covariates. The function is fitted on a training sample that contains observed values of the target variable and the covariates at a series of field sample locations by minimizing the differences between observed values of the target variable and those predicted from the function. Thus, spatial bias in the field sample locations would affect the modeling methods in similar ways (Zhang & Zhu, 2018). As a result, the findings of this study have broader implications for coping with spatial bias in VGI-based samples for many application domains beyond habitat mapping—for example, species distribution modeling (Zhang, 2019), digital soil mapping (Rossiter, Liu, Carlisle, & Zhu, 2015), and land cover classification (Comber et al., 2013).

ACKNOWLEDGMENTS

Supports to Guiming Zhang through the Faculty Start-up Funds and the Faculty Research Fund at the University of Denver are greatly appreciated. Supports to A-Xing Zhu from the National Natural Science Foundation of China (Grant No. 41871300) and through the Vilas Associate Award, the Hammel Faculty Fellow Award, and the Manasse Chair Professorship from the University of Wisconsin-Madison are acknowledged.

ORCID

Guiming Zhang https://orcid.org/0000-0001-7064-2138

REFERENCES

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species geographic distributions. *Ecological Informatics*, *19*, 10–15.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, *78*, 161–188.

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., & Foody, G. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation & Geoinformation*, *23*, 37–48.

Connors, J. P., Lei, S., & Kelly, M. (2012). Citizen science in the age of neogeography: Utilizing volunteered geographic information for environmental monitoring. *Annals of the Association of American Geographers*, *102*, 1267–1289.

Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample selection bias correction theory. In Y. Freund, L. Györfi, G. Turán, & T. Zeugmann (Eds.), *Algorithmic learning theory: ALT 2008* (Lecture Notes in Computer Science, Vol. *5254*, pp. 38–53). Berlin, Germany: Springer.

De Gruijter, J., Brus, D. J., Bierkens, M. F. P., & Knotters, M. (2006). *Sampling for natural resource monitoring*. Berlin, Germany: Springer.

Dudik, M., Schapire, R. E., & Phillips, S. J. (2005). Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems*, *18*, 323–330.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151.

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., … Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, *20*, 2131–2147.

Flanagin, A., & Metzger, M. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*, 137–148.

Franklin, J. (2013). Species distribution models in conservation biogeography: Developments and challenges. *Diversity & Distributions*, *19*(10), 1217–1223.

Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, *26*, 10–14.

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

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

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Canadian Journal of Forest Research*, *28*, 1429–1447.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment & Planning, B*, *37*, 682–703.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, *199*, 142–152.

Hung, K.-C., Kalantari, M., & Rajabifard, A. (2016). Methods for assessing the credibility of volunteered geographic information in flood response: A case study in Brisbane, Australia. *Applied Geography*, *68*, 37–47.

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

Jokar Arsanjani, J., Helbich, M., Bakillah, M., Hagenauer, J., & Zipf, A. (2013). Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science*, *27*, 2264–2278.

Jones, E., Oliphant, T., & Peterson, P. (2001). *SciPy: Open source scientific tools for Python*. Retrieved from http://www.scipy.org/

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

McBratney, A., Mendonça Santos, M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*, 3–52.

Mozas-Calvache, A. T. (2016). Analysis of behaviour of vehicles using VGI data. *International Journal of Geographical Information Science*, *30*(12), 2486–2505.

Munson, A. M., Webb, K., Sheldon, D., Fink, D., Hochachka, W. M., Iliff, M., … Kelling, S. (2012). *The eBird reference dataset, version 4.0*. Ithaca, NY: Cornell Lab of Ornithology and the National Audubon Society.

Pan, S. J., & Wang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge Data Engineering*, *22*, 1345–1359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*, 161–175.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*, 181–197.

Preston, C. R. (2000). *Red-tailed hawk* (1st ed.). Mechanicsburg, PA: Stackpole Books.

Rossiter, D. G., Liu, J., Carlisle, S., & Zhu, A.-X. (2015). Can citizen science assist digital soil mapping? *Geoderma*, *259–260*, 71–80.

Sauer, J. R., Niven, D. K., Hines, J. E., Ziolkowski, Jr, D. J., Pardieck, K. L., Fallon, J. E., & Link, W. A. (2017). *The North American breeding bird survey, results and analysis 1966-2015 (Version 2.07.2017)*. Laurel, MD: USGS Patuxent Wildlife Research Center.

Scott, D. W. (2015). *Multivariate density estimation: Theory, practice, and visualization* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., … Rutzinger, M. (2016). Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, *5*, 55.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log likelihood function. *Journal of Statistical Planning & Inference*, *90*, 227–244.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London, UK: Chapman & Hall.

Stehman, S. V., Fonte, C. C., Foody, G. M., & See, L. (2018). Using volunteered geographic information (VGI) in design-based statistical inference for area estimation and accuracy assessment of land cover. *Remote Sensing of Environment*, *212*, 47–59.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40.

Sun, Y., Fan, H., Helbich, M., & Zipf, A. (2013). Analyzing human activities through volunteered geographic information: Using Flickr to analyze spatial and temporal pattern of tourist accommodation. In M. J. Krisp (Ed.), *Progress in location-based services* (pp. 57–69). Berlin, Germany: Springer.

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*, 1084–1091.

Wang, J.-F., Stein, A., Gao, B.-B., & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, *2*, 1–14.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada (pp. 1–8). New York, NY: ACM.

Zhang, G. (2019). Enhancing VGI application semantics by accounting for spatial bias. *Big Earth Data*, *3*, 255–268.

Zhang, G., & Zhu, A.-X. (2018). The representativeness and spatial bias of volunteered geographic information: A review. *Annals of GIS*, *24*, 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*, 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., & Xiao, W. (2018). A heuristic-based approach to mitigating positional errors in patrol data for species distribution modeling. *Transactions in GIS*, *22*, 202–216.

Zhang, G., Zhu, A.-X., Windels, S. K., & Qin, C.-Z. (2018). Modelling species habitat suitability from presence-only data using kernel density estimation. *Ecological Indicators*, *93*, 387–396.

Zhang, S., Zhu, A.-X., Liu, J., Yang, L., Qin, C.-Z., & An, Y.-M. (2016). An heuristic uncertainty directed field sampling design for digital soil mapping. *Geoderma*, *267*, 123–136.

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

Zhu, A.-X., Wang, R., Qiao, J., Qin, C.-Z., Chen, Y., Liu, J., … Zhu, T. (2014). An expert knowledge-based approach to landslide susceptibility mapping using GIS and fuzzy logic. *Geomorphology*, *214*, 128–138.

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

How to cite this article: Zhang G, Zhu A-X. Sample size and spatial configuration of volunteered geographic information affect effectiveness of spatial bias mitigation. *Transactions in GIS*. 2020;24: 1315–1340. https://doi.org/10.1111/tgis.12679