




A representativeness-directed approach to mitigate spatial bias in VGI for the predictive mapping of geographic phenomena

Guiming Zhang & A-Xing Zhu


To cite this article: Guiming Zhang & A-Xing Zhu (2019) A representativeness-directed approach to mitigate spatial bias in VGI for the predictive mapping of geographic phenomena, International Journal of Geographical Information Science, 33:9, 1873-1893, DOI: [10.1080/13658816.2019.1615071](https://doi.org/10.1080/13658816.2019.1615071)

To link to this article: <https://doi.org/10.1080/13658816.2019.1615071>

 View supplementary material 

 Published online: 10 May 2019.

 Submit your article to this journal 

 Article views: 75

 View Crossmark data 

RESEARCH ARTICLE



A representativeness-directed approach to mitigate spatial bias in VGI for the predictive mapping of geographic phenomena

Guiming Zhang^a and A-Xing Zhu^{b,c,d,e,f}

^aDepartment of Geography & the Environment, University of Denver, Denver, CO, USA; ^bJiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China; ^cKey Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing, China; ^dState Key Laboratory Cultivation Base of Geographical Environment Evolution, Nanjing, China; ^eState Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; ^fDepartment of Geography, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

Volunteered geographic information (VGI) contains valuable field observations that represent the spatial distribution of geographic phenomena. As such, it has the potential to provide regularly updated low-cost field samples for predictively mapping the spatial variations of geographic phenomena. The predictive mapping of geographic phenomena often requires representative samples for high mapping accuracy, but samples consisting of VGI observations are often not representative as they concentrate on specific geographic areas (i.e. spatial bias) due to the opportunistic nature of voluntary observation efforts. In this article, we propose a representativeness-directed approach to mitigate spatial bias in VGI for predictive mapping. The proposed approach defines and quantifies sample representativeness by comparing the probability distributions of sample locations and the mapping area in the environmental covariate space. Spatial bias is mitigated by weighting the sample locations to maximize their representativeness. The approach is evaluated using species habit suitability mapping as a case study. The results show that the accuracy of predictive mapping using weighted sample locations is higher than using unweighted sample locations. A positive relationship between sample representativeness and mapping accuracy is also observed, suggesting that sample representativeness is a valid indicator of predictive mapping accuracy. This approach mitigates spatial bias in VGI to improve predictive mapping accuracy.

ARTICLE HISTORY

Received 8 October 2018
Accepted 30 April 2019

KEYWORDS

Volunteered geographic information (VGI); spatial bias; sample representativeness; predictive mapping; habitat suitability mapping

Introduction

Volunteered geographic information (VGI) refers to geographic information created by citizen volunteers (Goodchild 2007). VGI has proliferated in recent years, mainly due to the technological advancements enabling the public to contribute geospatial data. For

CONTACT Guiming Zhang  guiming.zhang@du.edu
 supplemental data for this article can be accessed [here](#).

© 2019 Informa UK Limited, trading as Taylor & Francis Group

example, with ubiquitous access to the Internet and positioning technologies, the average citizen can now easily create and share geo-referenced observations using smartphones, personal computers, and other portable devices. Collectively, these networked, volunteering individual sensors are producing rich information, revealing the spatiotemporal patterns of geographic phenomena (Goodchild 2007, Graham *et al.* 2011, Zhang and Zhu 2018).

VGI has several advantages as a mechanism of acquiring and compiling geographic data to reflect the spatial distribution of geographic phenomena. First, VGI contains rich local information that spans a broad temporal spectrum because the citizens, as local experts and sensors, sense and accumulate knowledge of their respective areas over long time periods (Goodchild 2007). As such, VGI also has the potential to provide geographic data over large areas, as billions of networked human sensors exist worldwide (e.g. eBird; Sullivan *et al.* 2009). Second, VGI can provide regularly updated geographic information that is difficult to obtain through remote sensing but can easily be collected by citizens on the ground (Goodchild 2007, Kelling *et al.* 2013). Third, VGI is far less expensive than traditional geographic data collection protocols (e.g. geographic sampling, biological survey) (Goodchild 2007, Coleman *et al.* 2009). The low cost of VGI is thus supporting real-world applications, such as wildlife conservation programs in poor and remote areas (Anadón *et al.* 2009, Zhu *et al.* 2015a; Zhang *et al.* 2017).

Since VGI contains valuable field observations regarding the spatial distribution of geographic phenomena, which may be the only data reflecting the geographic distribution of the phenomena of interest in some cases, it also has the potential to provide data for mapping these geographic phenomena. Information on the spatial variation of geographic phenomena is essential to many environmental modeling and geographic decision-making efforts (Goodchild *et al.* 1993, Zhu and Mackay 2001, Zhu *et al.* 2015, Zhang *et al.* 2018a). For example, maps revealing the spatial variation of soil, vegetation, and temperature are necessary inputs for hydrological modeling (Zhu and Mackay 2001). Additionally, species habitat suitability maps are needed to support decision-making for conservation prioritization and systematic reserve design (Margules and Pressey 2000, Wilson *et al.* 2005). In this context, *predictive mapping* is a commonly used framework for mapping the spatial variation of geographic phenomena (Zhu *et al.* 1997, McBratney *et al.* 2003), which assumes that geographic phenomena are influenced by other environmental factors and, as a result, their spatial variation is usually correlated with the spatial variation of their environmental determinants. Specifically, predictive mapping maps the spatial variation of a target geographic phenomenon (e.g. soil) based on the spatial variation of its environmental covariates (e.g. parent material, terrain relief, and vegetation) (Guisan and Zimmerman 2000, Scull *et al.* 2003, Franklin and Miller 2009), as conceptualized in the following equation:

$$T = f(E) \quad (1)$$

where T is the target geographic phenomenon, E a set of environmental covariates, and f the covariation relationship between T and E . The functional form and/or coefficient estimates of f are often obtained from field samples (McBratney *et al.* 2003, Franklin and Miller 2009).

To achieve a high mapping accuracy, predictive mapping requires that the field sample is representative to capture the relationship between the spatial variation of the covariates and of the geographic phenomenon over the area to be mapped (McBratney *et al.* 2003, Qi and Zhu 2003, Franklin and Miller 2009). Representative field samples are often

collected by following well-designed geographic sampling schemes. Commonly used geographic sampling schemes include probabilistic sampling methods (e.g. simple random, stratified random, systematic sampling) (Gregoire and Valentine 2007, Jensen and Shumway 2010) and purposive sampling (Yang *et al.* 2013), where sampling locations are allocated so that the geographic and/or the covariate space is well covered by the collected field samples (e.g. sample observations are collected across the complete gradient of the covariate space) (Minasny and McBratney 2006, Gregoire and Valentine 2007, Jensen and Shumway 2010). However, obtaining representative field samples through geographic sampling is costly, labor-intensive, and time-consuming.

Although the observations contributed by volunteers can provide timely updated field samples at a relatively low cost for the predictive mapping of geographic phenomena, VGI observations suffer from spatial bias and, thus, they may not be representative. That is, the VGI observations are often concentrated in some geographic areas over others, as the observations made by volunteers are mostly opportunistic (Zhu *et al.* 2015). As such, the spatial distribution of the observation efforts of volunteers would be considered neither random nor regular in the sense of geographic sampling design, as the individual volunteers decide where to conduct observations and there is no coordination among these observation efforts. As a result, VGI observations are typically spatially biased towards areas with denser population or higher route accessibility (Kadmon *et al.* 2004). Due to this spatial bias, field samples consisting of VGI observations (VGI-based samples, hereafter) might not be 'representative' of the mapping area. Spatial bias in VGI, if not appropriately accounted for, would thus adversely affect the accuracy of predictive mapping using VGI-based samples (Leitão *et al.* 2011, Pardo *et al.* 2013, Zhu *et al.* 2015).

Zhang and Zhu (2018), among others, offered a comprehensive review of the state-of-the-art methods developed to mitigate spatial bias in VGI and correct sample selection bias for predictive mapping. Fink *et al.* (2010) proposed an *AdaSTEM* approach to accommodate spatial bias for broad-scale VGI data (e.g. continent-scale eBird data). Specifically, the mapping area is partitioned into smaller sub-areas, and local predictive models are trained with VGI observations in each sub-area. However, this approach does not address the potential spatial bias of VGI observations within each sub-area. In addition, it requires very large sample sizes to ensure sufficient sample size for each sub-area. Filtering sample locations in the geographic or environmental space is also applied to reduce spatial bias (Boria *et al.* 2014, Varela *et al.* 2014). This method assumes that removing localities within a certain distance from each other would cancel out the unequal sampling effort.

Nonetheless, objectively setting distance thresholds is challenging and removing sample locations reduces sample size and discards useful information. Zhu *et al.* (2015) proposed to compensate for VGI spatial bias by weighting VGI observations with weights inversely proportional to the cumulative visibility at observation sites, which is applicable only in cases where cumulative visibility is a reasonable approximation of the observation efforts.

The *FactorBiasOut* method was developed to correct for spatial bias in species occurrence data when using MAXENT for species distribution modeling (Phillips *et al.* 2009). In this approach, background data are used that have the same spatial bias as the species occurrence data. However, information on the observation effort is required to generate background data, and such information is not always available in VGI genesis.

Various methods have been developed to correct for sample selection bias in general. One approach explicitly models the selection processes (Heckman 1979, Vella 1998),

which requires a clear understanding of the underlying sample selection processes. However, it is difficult to adopt this approach to correct the spatial bias in VGI, as detailed information on the selection processes underlying VGI genesis is often missing. Another approach is importance weighting, where the training sample is weighted using an importance weighting function to correct for sample selection bias (Shimodaira 2000). The optimal weighting function is the ratio of the probability density functions of test data features and training data (Cortes *et al.* 2008). In any case, to estimate the optimal weighting function requires a sufficiently large sample size and density estimation in high-dimensional cases is difficult (Shimodaira 2000). Therefore, this method is not applicable for correcting spatial bias in VGI, where the predictive mapping may involve many environmental covariates (i.e. high-dimensional) and a smaller number of VGI-based sample observations.

In summary, each bias correction method has its own data requirements, and a particular VGI application may not meet these requirements. Additionally, many existing bias mitigation methods require information on the underlying sampling or observation process (e.g. selection probabilities, sampling effort). However, such information is not always available during VGI genesis, because volunteers are not committed or unable to report such information. Thus, how to mitigate spatial bias in VGI to improve the accuracy of predictive mapping using VGI-based samples remains a challenge.

In this article, we propose a novel representativeness-directed approach for mitigating spatial bias in VGI for predictive mapping. The main idea and implementation details of the proposed approach are presented in Section 2. A case study of species habitat suitability mapping using VGI data was also conducted to demonstrate the approach and is presented in Section 3. The discussion and conclusions of the study are presented in Sections 4 and 5, respectively.

Methodology

Basic idea

Assessing sample representativeness with respect to the target geographic phenomenon is challenging because the spatial variation of the target phenomenon is unknown (i.e. to be predicted). Nevertheless, it is feasible to assess sample representativeness with respect to the environmental covariates used in modeling. Provided that the spatial variations of the target phenomenon and the covariates are correlated (i.e. the fundamental assumption of predictive mapping), it is reasonable to expect that sample representativeness with respect to the environmental covariates would approximate that of the target phenomenon (Hijmans *et al.* 2000, Minasny and McBratney 2006, Yang *et al.* 2008, 2013).

Sample representativeness is measured here as the ‘goodness-of-coverage’ of the field sample locations over the covariate space, which is in turn quantified by the similarity between the probability density distributions of the sample locations over the covariate space (i.e. sample distribution) and all spatial mapping units in the area (e.g. cells within the mapping area) over the covariate space (i.e. population distribution) (Figure 1). Stronger spatial bias in the field sample would thus lead to lower sample representativeness.

Spatial bias in VGI-based samples can thus be mitigated by improving sample representativeness. This is achieved by increasing the sample distribution’s similarity

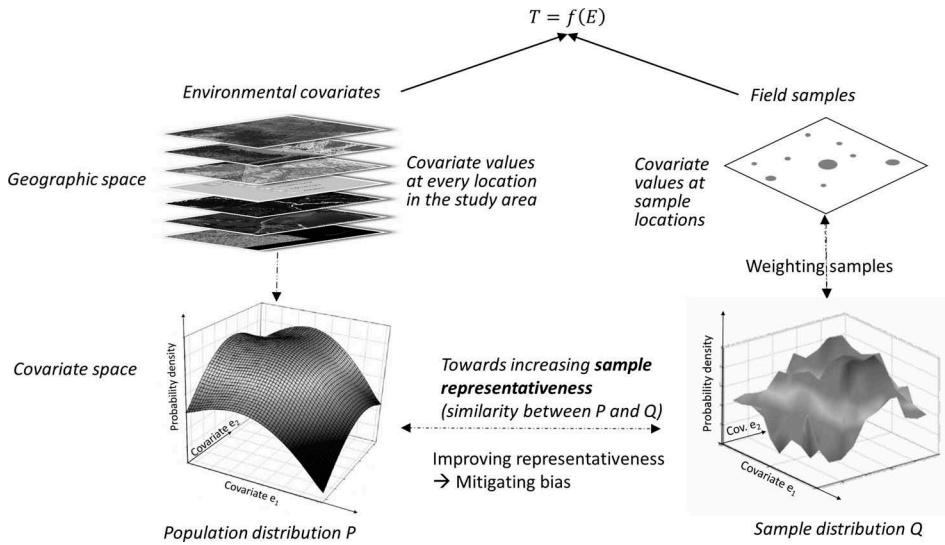


Figure 1. The basic principle of representativeness-directed spatial bias mitigation.

to the population distribution by weighting the VGI-based sample. That is, sample observations in an under-represented area would receive larger weights and be treated as more important in training predictive models. Weighting the sample in this way is expected to mitigate spatial bias and improve sample representativeness.

Overview

An overview of the proposed methodology is shown in Figure 2 and a detailed workflow chart in Figure 3. First, the raw VGI-based sample locations and a set of environmental

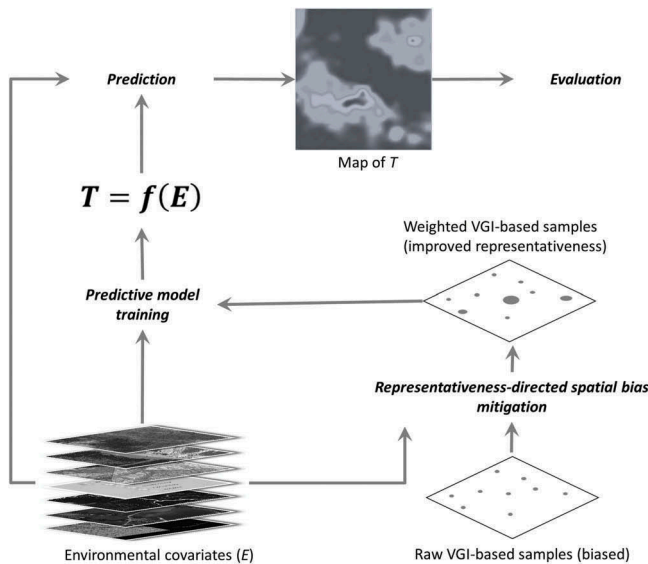


Figure 2. Methodology overview.

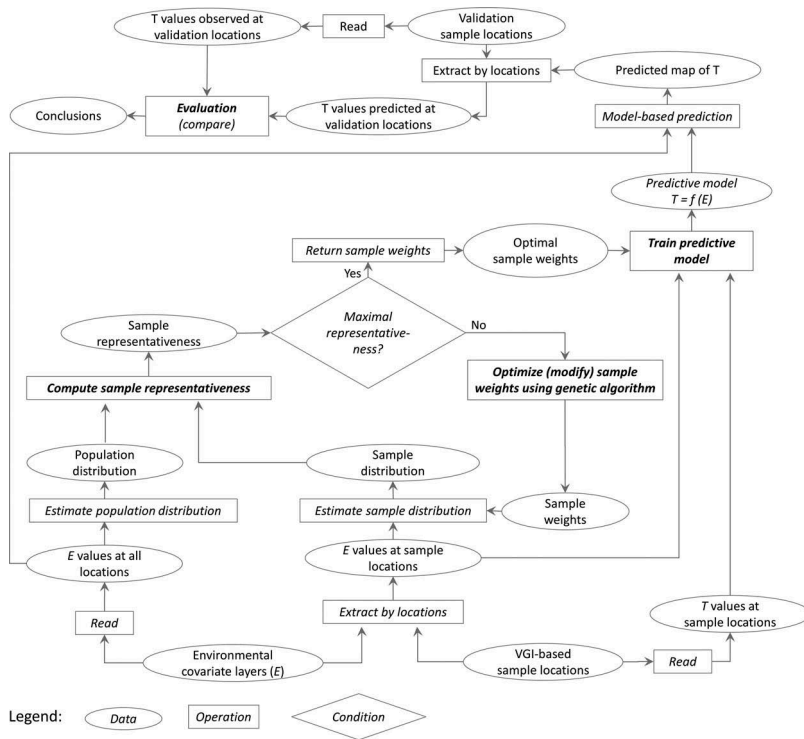


Figure 3. Detailed workflow of the methodology.

covariate layers are taken as inputs for the representativeness-directed spatial bias mitigation approach to obtain the optimal sample weights that maximize the representativeness of the VGI-based sample (Section 2.3). Second, the VGI-based sample weighted by the optimal weights is used to train predictive models. Third, the trained predictive models, which encode the covariation relationships, are used in combination with environmental covariate data to predict the spatial variation of the target geographic phenomenon (Section 2.4). Finally, the accuracy of the predicted map of the target geographic phenomenon is validated, and the effectiveness of the representativeness-directed approach evaluated to improve prediction accuracy (Section 2.5).

Representativeness-directed spatial bias mitigation

Measuring sample representativeness

Sample representativeness is measured as the similarity between the sample and population distributions over the covariate space. Principal component analysis (PCA) (Jolliffe 2002) was adopted to reduce the effects of high-dimensionality and multicollinearity (Shimodaira 2000). Only the first few principal components that explained 80% of the variance were used as new environmental covariates for predictive mapping.

Kernel density estimation (KDE) (Silverman 1986), with the commonly used Gaussian kernel, was used to estimate probability density distributions over the covariate space consisting of selected principal components. Based on the KDE method, sample

representativeness was computed using the following steps. First, sample distribution and population distribution with respect to the i^{th} selected principal component were estimated as per Equations 2 and 3, respectively:

$$Q_i(v_i) = \sum_{i=1}^n w_i \frac{1}{h_{iQ}} K\left(\frac{v_i - V_{li}}{h_{iQ}}\right) \quad (2)$$

and

$$P_i(v_i) = \sum_{j=1}^m \frac{1}{h_{iP}} K\left(\frac{v_i - V_{lj}}{h_{iP}}\right). \quad (3)$$

In the above equations, n is the number of sample locations and m the number of locations (cells) in the study area to be mapped. Q_i and P_i are the estimated sample and population distributions of the i^{th} principal component, respectively. v_i is a variable corresponding to the i^{th} principal component. V_{li} is the value of the i^{th} principal component at the i^{th} sample location and w_i a normalized sample weight (i.e. $\sum_{i=1}^n w_i = 1$) associated with this sample location. V_{lj} is the value of the i^{th} principal component at the j^{th} cell in the study area. h_{iQ} and h_{iP} are the bandwidths. Here, h_{iP} (i.e. bandwidth for estimating the population distribution) is determined using the 'rule-of-thumb' algorithm (Silverman 1986) given the large number of cells in the study area (Silverman 1986). h_{iQ} (i.e. the bandwidth for estimating the sample distribution) is determined more diligently using the 'golden section search optimization procedure' based on a maximum likelihood criterion through cross-validation of the sample data (Brunsdon 1995).

Second, the similarity between Q_i and P_i , that is, was computed as the overlapping area between the two distributions (Zhu 1999) (Equation 4):

$$SIM_i = \frac{2 \times A_{Q_i} \cap A_{P_i}}{A_{Q_i} + A_{P_i}}, \quad (4)$$

where A_{Q_i} and A_{P_i} are the areas under the sample and population distribution curves, respectively, and $A_{Q_i} \cap A_{P_i}$ is the overlapping area. SIM_i , ranging from 0 to 1, reflects the goodness-of-coverage of the sample regarding the i^{th} principal component (Figure 5 is a schematic example, illustrating sample and population distribution on one principal component and the overlapping area). The similarities between the sample and population distributions with respect to each of the L selected principal components were computed using Equations 2–4.

Finally, sample representativeness was computed as the overall similarity between the sample and population distributions with respect to all selected principal components. The overall similarity is a weighted average of the similarities with respect to each of the L selected principal components, with the weight proportional to the proportion of the variance each principal component retains (Equation 5):

$$R = SIM_{overall} = \sum_{i=1}^L \frac{\lambda_i}{\sum_{j=1}^L \lambda_j} SIM_i, \quad (5)$$

where R is sample representativeness, $SIM_{overall}$ the overall similarity, SIM_i the similarity between the two distributions with respect to the i^{th} principal component, and λ_i the eigenvalue of the i^{th} principal component indicating the amount of variance it retains.

The proportionality of the weight to the eigenvalue is a desirable property because a principal component retaining larger variance is more important for a sample to capture covariate variations. R ranges from 0 to 1, with a larger value indicating higher sample representativeness.

By computing sample representativeness this way, it is assumed that the principal component retaining larger variance is more important for measuring sample representativeness. This should hold in general given the underlying assumption of predictive mapping. That is, the variation of the target geographic phenomenon is correlated with the variation of the covariates. However, it does not account for other types of knowledge that may be available in specific applications. For example, it may be known in an application that some covariates are better correlated with the target than others (e.g. according to expert opinions, pilot studies). Such knowledge can be incorporated by computing sample representativeness directly using covariates and assigning more weight to covariates that are better correlated with the target.

Representativeness-directed spatial bias mitigation

Spatial bias mitigation was accomplished by improving sample representativeness (i.e. increasing the similarity between the sample and population distributions) by weighting the VGI-based sample. VGI observations in over-observed (over-represented) areas would have smaller weights, while observations in under-observed (under-represented) areas would have larger weights. The key is to determine these sample weights. This is conceived as an optimization problem, where the objective is to find a set of optimal sample weights that maximize the representativeness of the VGI-based sample.

The genetic algorithm (GA) (Davis 1991, Mitchell 1998) was adopted to determine the optimal sample weights where each weight is within $[1.0, W_{max}]$ (W_{max} is the maximum possible sample weight; $W_{max} = 10$ by default). The workflow of GA is shown in Figure 4

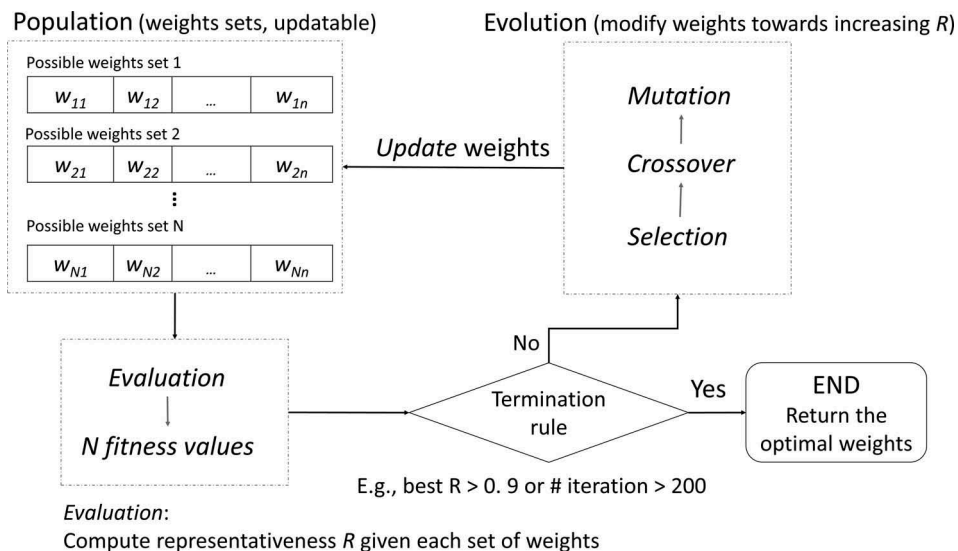


Figure 4. Workflow of the genetic algorithm to determine optimal sample weights.

and detailed descriptions can be found in Supplemental Materials S2. Figure 5 is a schematic example illustrating the effects of GA.

The rationale of setting the weight range as $[1.0, W_{max}]$ instead of $[0.0, 1.0]$ was as follows. Using a weight range of $[1.0, W_{max}]$ ensures that every sample observation would weight at least 1.0 and at most W_{max} . Given this sample weight range, all sample observations contribute to training predictive models (i.e. no sample observations are excluded). This also implies that a sample observation can be considered at most W_{max} times as more important than another sample observation (i.e. the ratio of the relative importance between two sample observations is bounded). Moreover, compared to $[0.0, 1.0]$, the weight range of $[1.0, W_{max}]$ is wider, which allows more flexibility for the genetic algorithm to explore optimal weights. Sample weights can be normalized as necessary in training predictive models using a weighted sample.

Predictive mapping using a weighted sample

The VGI-based sample weighted by optimal weights is used to train the predictive models (e.g. statistical, machine learning) to derive covariation relationships between the target geographic phenomenon and its environmental covariates. The mechanism of incorporating sample weights in the model training process depends on the specific predictive models in use. For example, for training a linear regression model from a weighted sample using ordinary least squares, sample weights can be used to weight the individual squared error terms (Pedregosa *et al.* 2012).

Effectiveness evaluation of the proposed approach

The effectiveness of the proposed approach in improving prediction accuracy was evaluated using two predictive maps of the target geographic phenomenon. One was generated using an unweighted VGI-based sample and the other a VGI-based sample weighted by

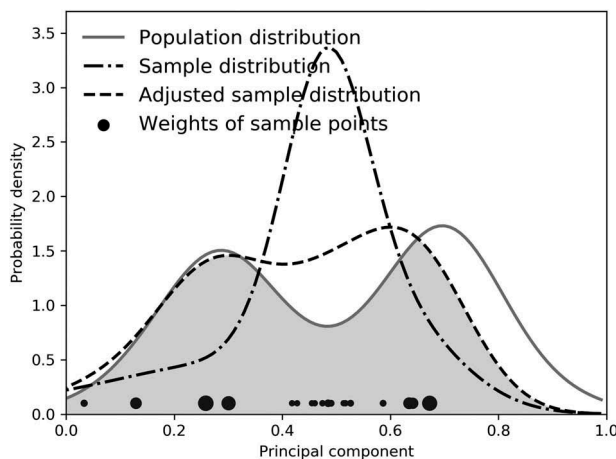


Figure 5. A schematic example of the effects of the genetic algorithm. Sample weights were returned after 25 iterations. Similarity, that is, the overlapping area between the adjusted sample and population distributions, increased from 0.579 to 0.835.

optimal sample weights. The prediction accuracies of the two maps were evaluated using independent validation data. The effectiveness of the approach was assessed using the change in prediction accuracy between the weighted and unweighted maps.

Case study

Study area and data

Study area and species

The case study area is the state of Wisconsin, located in the north-central area of the United States (in the Midwest and Great Lakes regions). Habitat suitability of the Red-tailed hawk (*Buteo jamaicensis*) was mapped over this study area. The Red-tailed hawk is a large bird species that is easy to identify, typically weighing from 690 to 1600 g and measuring 45–65 cm in length, with a wingspan from 110 to 145 cm. The Red-tailed hawk occupies a wide range of habitats and altitudes, including deserts, grasslands, coniferous and deciduous forests, agricultural fields, and urban areas (Preston 2000).

Biased VGI sample

VGI data from the eBird citizen science project (Munson *et al.* 2012) were used for habitat suitability mapping. eBird checklist locations indicating bird watchers' observation efforts in the study area were treated as a biased VGI sample in the representativeness-directed approach. The rationale is as follows. As species occurrence data resulted from observers' observation effort, it is reasonable to expect that better representativeness of the underlying observation effort implies better representativeness of the recorded species occurrences. It is thus the representativeness of the sample representing the observation effort (instead of species occurrence locations) that needs to be improved. In this case study, eBird checklist locations at which the observers carried out observations, regardless of the species observed, were treated as a biased sample.

A set of 655 eBird checklists with unique geographic locations reported in June 2012 (Figure 6) were extracted from eBird (Supplemental Materials S2). The checklist locations

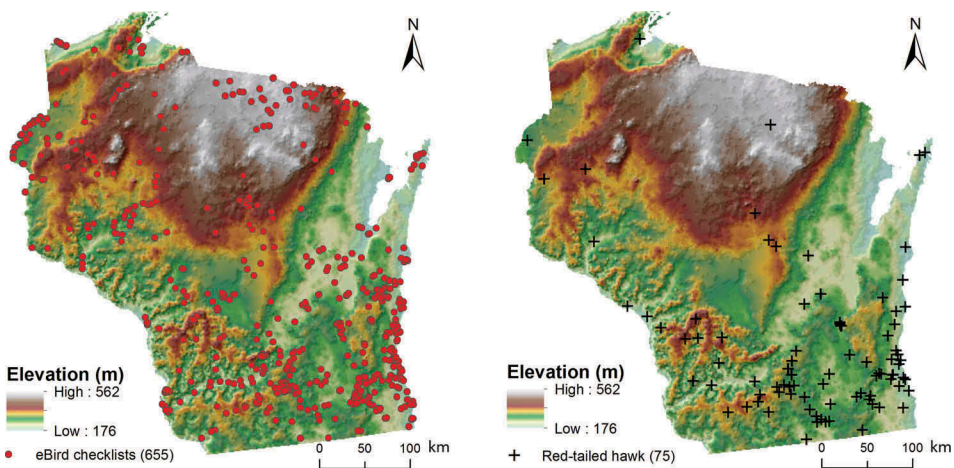


Figure 6. Selected eBird checklist locations and occurrence locations of the Red-tailed hawk in June 2012.

are spatially biased, that is, they tend to be clustered in areas with a denser human population (e.g. in the vicinity of large cities) and better accessibility (e.g. along roads). The proposed representativeness-directed approach was then applied to determine the optimal sample weights for these checklist locations, which were then used to mitigate spatial bias in training predictive models (see Section 3.4 for details).

Training sample

The Red-tailed hawk was reported at 75 of the 655 checklist locations (Figure 6). The occurrences were reviewed and approved by regional experts (Sullivan *et al.* 2009). A set of 1000 random locations were used as pseudo-absences (Franklin and Miller 2009), where each location/cell in the area had an equal probability of being selected. The occurrences and pseudo-absences were used to train predictive models for habitat suitability mapping (see Section 3.2 for details). Checklist locations where the species was not observed were not used as absences for modeling because not observing the species may simply be a failure of detection (Franklin and Miller 2009). Neither were these locations used as pseudo-absences because they could not be representative of the background environmental conditions due to spatial bias.

Environmental covariates

A suite of 71 environmental covariates representing the human population (housing density, housing percent vacant, and population density), terrain (elevation), climatic conditions (average, minimum, maximum temperature, and total precipitation), landscape level and land cover class level indices and statistics reflecting habitat configuration (edge density, largest patch index, and patch density) were used for predictive habitat suitability mapping. The first 11 principal components, retaining 80.1% of the total variance of the original covariate layers, were used in the analyses (Supplemental Materials S3).

Habitat mapping method

Logistic regression (LR) was adopted for modeling and mapping habitat suitability. An LR model was calibrated using the training sample (Section 3.1.3) using procedures implemented in the scikit-learn package (Pedregosa *et al.* 2012) (Supplemental Materials S4). The model was then applied to every location (cell) in the study area to generate a habitat suitability map.

Here LR was adopted primarily as a prediction model for predictive mapping. An LR model based on principal components is difficult to interpret because there is a limited capability of revealing the causal relationship between species presence/absence and covariates or differentiating relative importance of covariates. In cases where the emphasis is on model interpretability instead of prediction, more structured model selection methods could be applied to conduct the logistic regression analysis for examining causal mechanisms.

Evaluation

Validation data

Red-tailed hawk occurrences were obtained from the North American Breeding Bird Survey (BBS) (Pardieck *et al.* 2016). BBS routes, each having 50 evenly spaced stops, are distributed following a stratified random design to ensure roughly uniform spatial

coverage and to sample habitats representative of the entire region, although it may not be a strictly equal probability sampling design (Robbins *et al.* 1986, Sauer *et al.* 2013, Pardieck *et al.* 2016). Red-tailed hawks were observed at 73 stops on the active BBS routes surveyed in Wisconsin in June 2012 (Figure 7) (Supplemental Materials S5). A set of 1000 random locations were chosen from the study area as pseudo-absences using a uniform random distribution. This set of pseudo-absences was only used for validation; they are different from the set of pseudo-absences in the training sample.

Evaluation metric

The area under the ROC (receiver operating characteristic) curve (AUC) (Fielding and Bell 1997, Phillips and Dudík 2008), which can be computed for a suitability map using the validation data, was adopted as an accuracy measure of the predicted suitability map. AUC has an intuitive interpretation, namely, the probability that the predicted suitability at a randomly chosen species presence location is higher than that at a randomly chosen background location (Phillips *et al.* 2006). AUC provides an accuracy measure independent of any particular choice of suitability threshold. It has been widely adopted in species habitat suitability mapping (Dudík *et al.* 2005, Elith *et al.* 2006, Phillips *et al.* 2006, 2009, Phillips and Dudík 2008, Zhang *et al.* 2018b, 2018c). AUC ranges from 0.5 to 1.0. A value of 0.5 indicates that the prediction is no better than random predictions, while a value of 1.0 indicates perfect model performance.

Experiment design

The proposed approach was applied to determine optimal weights for the 655 checklist locations (biased VGI sample). The sample distribution in the covariate space was computed from covariate values at these locations. The population distribution was computed from covariate values for all raster cells. Sample representativeness was the similarity between these two distributions (Section 2.3). The default parameter settings

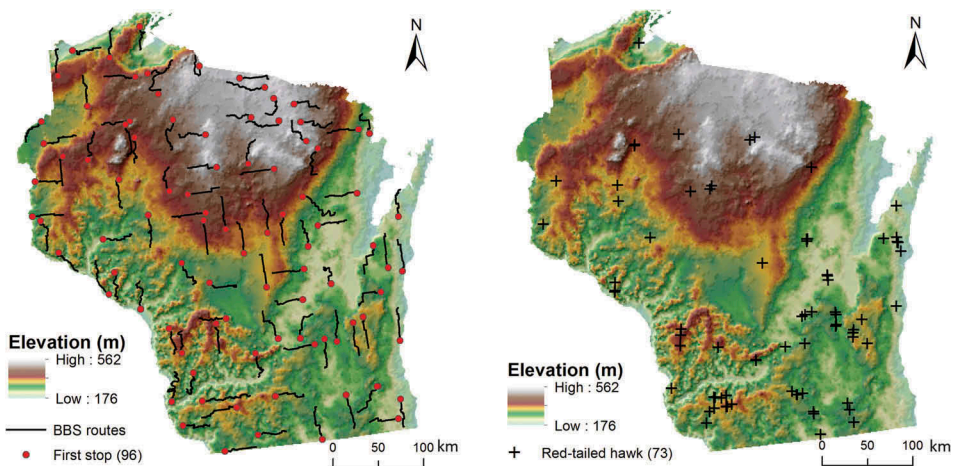


Figure 7. Active BBS routes in Wisconsin (left) and occurrences of Red-tailed hawks on these routes during June 2012 (right).

were: population size for the genetic algorithm = 500; number of generations for the genetic algorithm = 500; and an upper limit of sample weight $W_{max} = 10.0$.

The optimal weights for the checklist locations were then used to mitigate spatial bias in species occurrences. The optimal weights at the 75 species occurrence locations were used to weight the occurrences in training the predictive models (Section 3.2) for habitat suitability mapping.

Two habitat suitability maps were produced: one was predicted using the LR model trained using unweighted occurrences, and the other using the LR model trained using occurrences weighted by the optimal weights. The accuracies (i.e. AUC) of the predicted suitability maps were evaluated and compared.

The relationship between prediction accuracy and sample representativeness was also examined. Weights for the checklist locations and hence their representativeness evolve gradually over the generations of the genetic algorithm. At each generation, the weights corresponding to the best representativeness were recorded. The weights were used to weight the occurrences in training an LR model. AUC of the predicted suitability map was computed. A scatter plot was then created by plotting the AUC against the best representativeness over the generations.

To examine the impact of W_{max} the approach was repeated using W_{max} values of 5, 10, 20, 50, and 100. AUCs of the predicted maps were computed and compared.

Filtering sample locations is a common approach to reducing spatial sample bias (Boria *et al.* 2014, Varela *et al.* 2014). The optimal sample weights determined by the proposed approach are expected to be informative of the importance of individual sample locations. To test this hypothesis, the optimal weights of the 75 Red-tailed hawk occurrences ($W_{max}=10$) were used to filter the occurrence locations.

Subsets of the occurrences for specific sample sizes were randomly selected from the 75 occurrences, with selection probabilities proportional to the optimal weights squared (i.e. occurrence locations associated with larger weights have higher probabilities of being selected). These sets of occurrences were obtained at sample sizes ranging from 10% to 90% of the original sample size (75) in 10% increments. One hundred sets were drawn for each sample size. Each set of occurrences was used in the training sample to train an LR model (Section 3.1.3) (these training samples were denoted as *informative samples*). As a comparison, subsets of the occurrences were also selected from the 75 occurrences purely at random (i.e. an equal selection probability) and 100 sets of occurrences were drawn for each sample size. Each set was used in the training sample to train an LR model (these training samples were denoted as *random samples*). Both informative samples and random samples were not weighted in training LR models. AUC was computed for each predicted suitability map. At each sample size, the two-sample t-test was adopted to test if the mean AUC for the informative samples was statistically significantly higher than the mean AUC for the random samples.

Results

The effectiveness of the approach

The proposed approach allocated smaller weights to spatially clustered checklist locations than sparsely distributed locations (Figure 8 left). When using optimal weights, the overall representativeness of the checklist locations increases from 0.855 to 0.935 (for

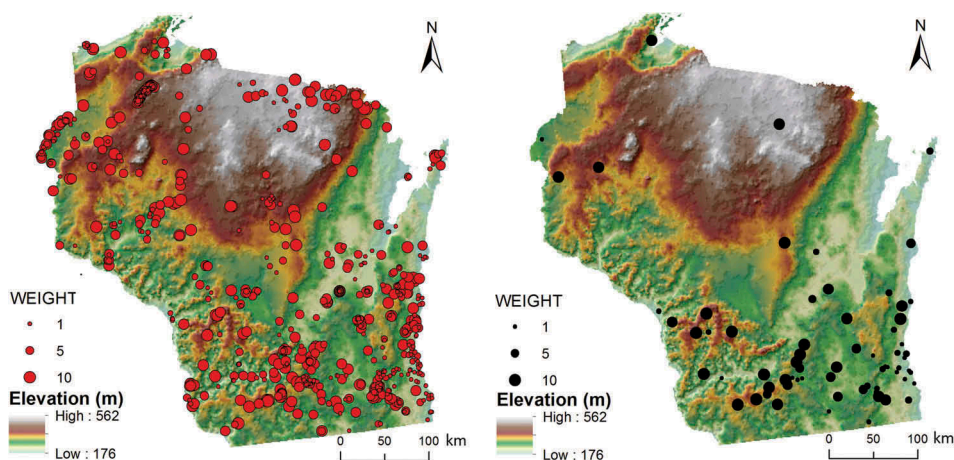


Figure 8. Optimal weights for eBird checklist locations (left) and the weights associated with the Red-tailed hawk occurrence locations (right).

comparison, the representativeness of the first stop locations of the BBS routes is 0.909). Similarly, densely distributed species occurrence locations (e.g. occurrences in the Milwaukee areas) tend to receive smaller weights than sparsely distributed occurrence locations (e.g. occurrences in northern areas) (Figure 8 right).

Using unweighted occurrences, the areas predicted to be of higher habitat suitability (e.g. suitability above 0.5) are limited to densely populated urban and suburban areas surrounding large cities such as Milwaukee, Madison, and Green Bay (Figure 9 left). This spatial pattern most likely reflects an artifact in the training sample rather than the ecological reality of the species. As there tend to be more bird watchers in these areas, the observations of the species are more frequent than other less observed areas (i.e. biased). The LR model overfits species occurrences in these areas. In contrast, weighting occurrences by optimal weights can reduce such overfitting and reveal the underlying ecological reality of the species better. The areas

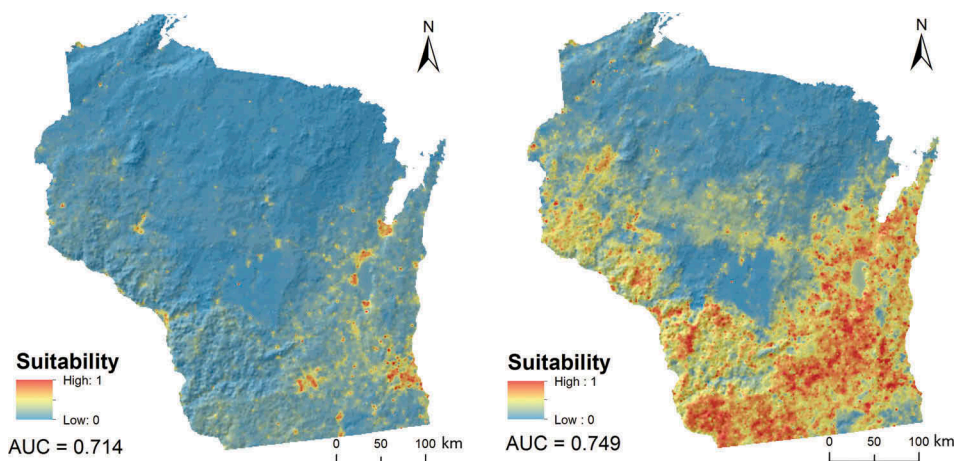


Figure 9. Habitat suitability maps predicted using unweighted species occurrence locations (left) and occurrence locations weighted by optimal weights (right).

predicted to be of higher habitat suitability have a much broader geographic range (Figure 9 right). Weighting species occurrences improved the accuracy of the predicted suitability map (i.e. AUC increased from 0.714 to 0.749).

Representativeness versus prediction accuracy

A clear positive relationship was observed between the representativeness of the checklist locations and AUC of suitability map predicted using weighted species occurrences (Figure 10). It suggests that sample representativeness can effectively indicate prediction accuracy.

Impact of w_{max}

Sample representativeness and prediction accuracy evolved differently under different W_{max} settings over the generations of the genetic algorithm (Figure 11). Under various W_{max} settings (Table 1), prediction accuracy using weighted species occurrences is higher than using unweighted occurrences (AUC = 0.714). The AUCs achieved under $W_{max} \leq 50$ are generally above 0.740. The AUC was highest when $W_{max} = 10$.

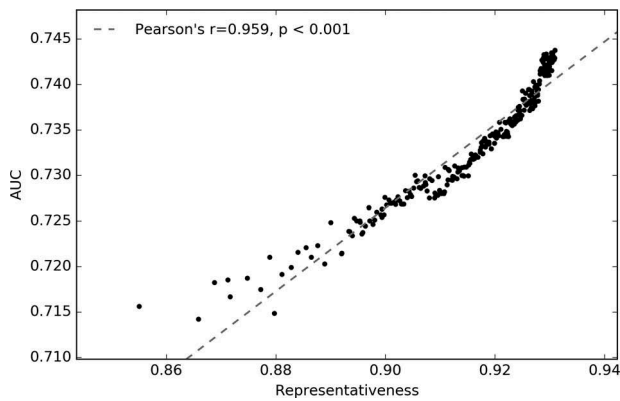


Figure 10. The relationship between sample representativeness and prediction accuracy over the generations of the genetic algorithm.

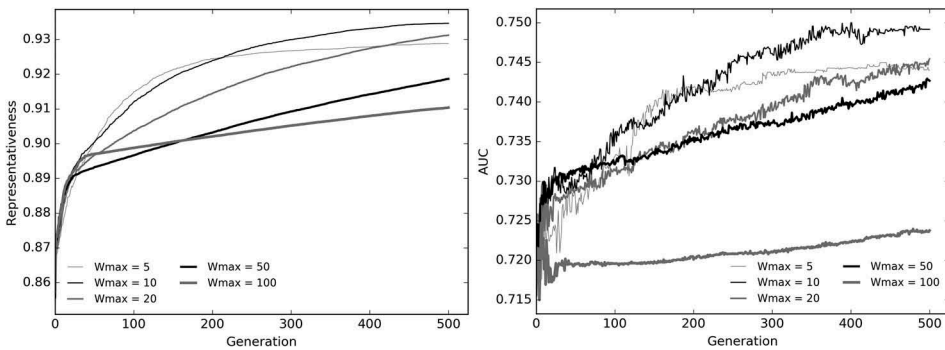


Figure 11. Evolution of sample representativeness (left) and prediction accuracy (right) over the generations of the genetic algorithm.

Table 1. The accuracy of suitability maps predicted from species occurrence locations, weighted by optimal weights under different W_{\max} settings.

W_{\max}	5	10	20	50	100
AUC	0.744	0.749	0.745	0.743	0.724

Effects of filtering sample locations

The mean AUC of the suitability maps predicted using *informative samples* is statistically significantly higher than that using *random samples* (Table 2). When the number of selected occurrences is 37 (approximately 50% of the original sample size), the mean AUC reaches the highest point of 0.748, which is higher than the AUC achieved for unweighted occurrences (AUC = 0.714) and comparable to the AUC achieved by weighting the occurrences (AUC = 0.749). In practice, if independent validation data were not available, cross-validation might be used to determine the optimal number of occurrences to filter. Nonetheless, utilizing the optimal weights to filter species occurrences can be as effective as weighting the occurrences to improve predictive mapping accuracy. The optimal weights are informative for differentiating the importance of individual occurrence locations which can thus be useful guidance for filtering species occurrences.

Discussion

The effectiveness of the approach

The prediction accuracy of the predictive models trained using species occurrences weighted by optimal weights is higher than that of predictive models trained using unweighted species occurrences. Moreover, a strong positive relationship between sample representativeness and prediction accuracy was observed, suggesting that representativeness is a valid indicator of prediction accuracy. In addition, the optimal weights are informative for differentiating the importance of individual species occurrence locations and can be used to filter sample locations to improve predictive mapping accuracy. Overall, the proposed approach can effectively mitigate spatial bias in VGI-based samples to improve predictive mapping accuracy.

Table 2. The accuracy of habitat suitability maps predicted using species occurrences selected under the guidance of optimal weights (informative samples) and selected purely at random (random samples).

Number of occurrence locations	Informative samples		Random samples		t-test
	Mean AUC	Std. AUC	Mean AUC	Std. AUC	t statistic
7	0.710	0.027	0.677	0.047	-6.111***
15	0.727	0.017	0.695	0.021	-11.719***
22	0.732	0.014	0.704	0.018	-12.220***
30	0.744	0.012	0.707	0.016	-18.307***
37	0.748	0.009	0.711	0.012	-25.242***
45	0.746	0.006	0.710	0.009	-33.834***
52	0.735	0.005	0.713	0.007	-24.676***
60	0.725	0.004	0.715	0.005	-15.334***
67	0.719	0.003	0.714	0.004	-10.801***

Parameter settings

The population size and number of generations for the genetic algorithm were both set to 500 for the experiments in this case study. These settings were mostly determined in accordance with the optimization problem size and the related computational cost. That is, the population size and the number of generations should be large enough relative to the problem size (e.g. 655 sample weights to optimize). At the same time, they should be kept as small as possible to save computation time if only limited computing resources are available for running the genetic algorithm.

W_{max} is a key parameter in the proposed approach. The physical meaning of W_{max} is that a species occurrence with weight W_{max} will be treated as W_{max} times as more important than a species occurrence with weight 1.0 in training predictive models. The case study shows that weighting species occurrences by optimal weights obtained under $W_{max} = 10.0$ indeed achieved the largest prediction accuracy improvement. In cases where data availability allows, W_{max} may be determined through more robust data-driven procedures such as cross-validation, in addition to considering its physical meaning.

Comparison against other weighting methods

The proposed approach was compared against two other sample weighting methods. The first method simply used sample weights proportional to the area of the Thiessen polygon associated with each sample location. Using this weighting method, geographically sparse sample locations received larger weights (Supplemental Materials S6). The second was the importance weighting method (Cortes *et al.* 2008). The multidimensional probability density function of covariate values in the study area and the probability density function of covariate values at the sample locations were estimated using multivariate kernel density estimation (Scott 2015). The ratios of the two density functions (at sample locations in the covariate space) were then used as sample weights. Using this weighting method, sample locations at under-represented covariate gradients received larger weights (Supplemental Materials S7).

Species occurrences were weighted using the two methods in training LR models. The AUCs for the predicted suitability maps are 0.711 and 0.735, respectively. Our approach (AUC = 0.749) outperformed the Thiessen polygon-based weighting method, suggesting such a simple scheme to define sample weights should be discouraged. The performance of our approach was slightly better than the importance of weighting method. Nonetheless, one advantage of our approach was that the ratios between sample weights were bounded to W_{max} (i.e. one sample observation can be treated at most W_{max} times as more important than another), whereas the weight ratios could be unboundedly large using the other two weighting method.

Applicability of the approach

eBird checklist locations representing the observation effort of the birders (regardless of whether Red-tailed hawks were observed at the locations) were treated as a biased sample. The proposed approach was applied to determine the optimal sample weights for these locations that maximize their representativeness. The underlying assumption is

that improved representativeness of the observation effort (i.e. checklist locations) implies better representativeness of the reported species occurrences.

As such, the proposed approach should not be applied directly on a sample consisting of only species occurrence locations of one species, as evaluating the representativeness of the occurrence locations is inappropriate without knowledge of the underlying observation effort (i.e. all locations that have been visited). However, if the occurrences of a group of species that resulted from the same underlying observation campaign are available, these occurrences can be pooled and used as a proxy of the underlying observation effort (Dudik *et al.* 2005, Phillips *et al.* 2009) and the approach can be applied.

Generally, the approach should be applicable for mitigating spatial bias in any field sample, as long as the sample locations represent the sampling effort. For example, the approach is conceivably applicable for habitat suitability mapping using a sample consisting of species occurrences and 'true' absences, as the sample contains all visited locations (sampling effort). The approach should also apply to mitigating spatial bias in soil sample locations for soil mapping. Soil samples can be taken at every location visited given that soil is often considered a continuum. Thus, soil sample locations naturally imply sampling effort.

Conclusions

We propose a representativeness-directed approach to spatial bias mitigation in VGI-based samples for predictive mapping. Spatial bias mitigation is accomplished by weighting sample locations towards increasing sample representativeness. The optimal sample weights that maximize sample representativeness were determined using an optimization procedure. The evaluations of the approach demonstrate that weighting sample locations by optimal weights improves predictive mapping accuracy, sample representativeness is a valid indicator of prediction accuracy, and the optimal weights are informative of the importance of individual sample locations. Overall, the approach can effectively mitigate spatial bias in VGI-based samples to improve predictive mapping accuracy.

For future work, the approach could be extended to support training localized predictive models for mapping geographic phenomena over large areas, where spatial non-stationarity in the covariation relationships needs to be accounted for (Fotheringham *et al.* 2003). From a computational perspective, measures can be taken to improve the computational efficiency of the approach. The genetic algorithm for determining optimal weights is computationally demanding, particularly when the sample size is large. It could be accelerated using parallel computing, harnessing computing powers on multi-core CPUs (central processing units) and GPUs (graphics processing units) (Zhang *et al.* 2016, 2017). Alternatively, other less computationally demanding algorithms (e.g. black-box optimization; Le Digabel 2011) could be adopted to optimize sample weights.

Acknowledgments

Supports to Guiming Zhang through the Faculty Startup Funds from the University of Denver and through the Whitbeck Graduate Dissertator Award from the Department of Geography, University of Wisconsin-Madison are greatly appreciated.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Dr. Guiming Zhang is an assistant professor of GIScience in the Department of Geography and the Environment at the University of Denver. His research is focused on geospatial big data analytics and the related enabling geo-computation technologies. In particular, he is interested in volunteered geographic information (VGI) and its applications in mapping natural resources (e.g., wildlife habitat, soil).

Dr. A-Xing Zhu is a professor of Geography in the Department of Geography at the University of Wisconsin-Madison. He has extensive experience of applying GIS in environmental modeling and resource management (e.g., digital soil mapping, landslide susceptibility mapping, and hydrological modeling).

References

- Anadón, J.D., *et al.*, 2009. Evaluation of local ecological knowledge as a method for collecting extensive data on animal abundance. *Conservation Biology*, 23, 617–625. doi:10.1111/j.1523-1739.2008.01145.x
- Boria, R.A., *et al.*, 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275, 73–77. doi:10.1016/j.ecolmodel.2013.12.012
- Brunsdon, C., 1995. Estimating probability surfaces for geographical point data: an adaptive kernel algorithm. *Computer Geoscience*, 21, 877–894. doi:10.1016/0098-3004(95)00020-9
- Robbins, C.S., Bystrak, D. & Geissler, P.H. (1986). *The breeding bird survey: its first fifteen years, 1965-1979* (No. FWS-PUB-157), Patuxent Wildlife Research Center. Laurel, MD.
- Coleman, D.J., Georgiadou, Y., and Labonte, J., 2009. Volunteered Geographic Information: the nature and motivation of producers. *International Journal of Spatial Data Infrastructure Research*, 4, 332–358.
- Cortes, C., *et al.*, 2008. Sample selection bias correction theory. *International Conference of Algorithmic Learn. Theory 2008 Oct 13*. Berlin, Heidelberg: Springer, 38–53.
- Davis, L. (1991). *Handbook of genetic algorithms*.
- Dudik, M., Schapire, R.E., and Phillips, S.J., 2005. Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems*, 18 (17), 323–330.
- Elith, J., *et al.*, 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography (Cop.)*, 29, 129–151. doi:10.1111/j.2006.0906-7590.04596.x
- Fielding, A.H. and Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24, 38–49. doi:10.1017/S0376892997000088
- Fink, D., *et al.*, 2010. Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, 20, 2131–2147. doi:10.1890/09-1340.1
- Fotheringham, A.S., Brunsdon, C., and Charlton, M., 2003. *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester, UK: John Wiley & Sons, Limited.
- Franklin, J. and Miller, J.A., 2009. *Mapping species distributions: spatial inference and prediction*. Cambridge: Cambridge University Press.
- Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, 211–221. doi:10.1007/s10708-007-9111-y
- Goodchild, M.F., Parks, B.O., and Steyaert, L.T. (1993). *Environmental modeling with GIS*.
- Graham, E.A., Henderson, S., and Schloss, A., 2011. Using mobile phones to engage citizen scientists in research. *Eos, Trans. Am. Geophys. Union* 92, 313–315.
- Gregoire, T.G. and Valentine, H.T., 2007. *Sampling strategies for natural resources and the environment*. Boca Raton, FL: CRC Press.

- Guisan, A. and Zimmerman, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135, 147–186. doi:10.1016/S0304-3800(00)00354-9
- Heckman, J.J., 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–161.
- Hijmans, R.J., et al., 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. *Conservation Biology*, 14, 1755–1765. doi:10.1111/cbi.2000.14.issue-6
- Jensen, R.R. and Shumway, J.M., 2010. Sampling our world. In: B. Gomez and J. III, J.P., eds.. *Res. Methods Geogr. A Crit. Introd.*. Malden, MA: John Wiley & Sons, 77–90.
- Jolliffe, I.T., 2002. Principal component analysis and factor analysis. In: *Principal Component Analysis* (pp. 150–166). New York, NY: Springer.
- Kadmon, R., Farber, O., and Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, 14, 401–413. doi:10.1890/02-5364
- Kelling, S., et al., 2013. Estimating species distributions-across space, through time, and with features of the environment. In: M. Atkinson, et al., eds.. *DATA Bonanza Improv. Knowl. Discov. Sci. Eng. Bus.* Hoboken, NJ: John Wiley & Sons, Inc., 441–458.
- Le Digabel, S., 2011. Algorithm 909: NOMAD: nonlinear optimization with the MADS algorithm. *ACM Transction Mathematical Software*, 37, 44. doi:10.1145/1916461.1916468
- Leitão, P.J., Moreira, F., and 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, 439–454. doi:10.1080/13658816.2010.531020
- Margules, C.R. and Pressey, R.L., 2000. Systematic conservation planning. *Nature*, 405, 243–253. doi:10.1038/35012251
- McBratney, A., Mendonça Santos, M., and Minasny, B., 2003. On digital soil mapping. *Geoderma*, 117, 3–52. doi:10.1016/S0016-7061(03)00223-4
- Minasny, B. and McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32, 1378–1388. doi:10.1016/j.cageo.2005.12.009
- Mitchell, M., 1998. *An introduction to genetic algorithms*. Cambridge, MA: MIT press.
- Munson, A.M., et al., 2012. The ebird reference dataset, version 4.0. *Cornell Lab Ornithol. Natl. Audubon Soc.*, Ithaca, NY. 1–11.
- Pardieck, K.L., Jr., et al., 2016. *North American Breeding Bird Survey Dataset 1966-2015, version 2015.1*. U.S. Geological Survey, Patuxent Wildlife Research Center. Laurel, MD.
- Pardo, I., et al., 2013. A novel method to handle the effect of uneven sampling effort in biodiversity databases. *PLoS One*, 8, e52786. doi:10.1371/journal.pone.0052786
- Pedregosa, F., et al., 2012. Scikit-learn: machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Phillips, S.J., et al., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19, 181–197. doi:10.1890/07-2153.1
- Phillips, S.J., Anderson, R.P., and Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231–259. doi:10.1016/j.ecolmodel.2005.03.026
- Phillips, S.J. and Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography (Cop.)*, 31, 161–175. doi:10.1111/j.0906-7590.2008.5203.x
- Preston, C.R., 2000. *Red-tailed hawk*. 1st. Mechanicsburg, PA: Stackpole Books.
- Qi, F. and Zhu, A.-X., 2003. Knowledge discovery from soil maps using inductive learning. *International Journal of Geographical Information Science*, 17, 771–795. doi:10.1080/13658810310001596049
- Sauer, J.R., et al., 2013. The North American Breeding Bird Survey 1966-2011: summary Analysis and Species Accounts. *North American Fauna*, 79, 1–32. doi:10.3996/nafa.79.0001
- Scott, D.W., 2015. *Multivariate density estimation: theory, practice, and visualization*. 2nd. Hoboken, New Jersey: John Wiley & Sons.
- Scull, P., et al., 2003. Predictive soil mapping: a review. *Progress in Physical Geography: Earth and Environment*, 27, 171–197. doi:10.1191/0309133303pp366ra

- Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90, 227–244. doi:10.1016/S0378-3758(00)00115-4
- Silverman, B.W., 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, UK.
- Sullivan, B.L., et al., 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142, 2282–2292. doi:10.1016/j.biocon.2009.05.006
- Varela, S., et al., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography (Cop.)*, 37, 1084–1091.
- Vella, F., 1998. Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources*, 33, 127–169. doi:10.2307/146317
- Wilson, K.A., et al., 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, 122, 99–112. doi:10.1016/j.biocon.2004.07.004
- Yang, F., et al., 2008. Assessing the representativeness of the AmeriFlux network using MODIS and GOES data. *Journal of Geophysical Research: Biogeosciences*, 113, 1–11. doi:10.1029/2007JG000627
- Yang, L., et al., 2013. An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping. *International Journal of Geographical Information Science*, 27, 1–23. doi:10.1080/13658816.2012.658053
- Zhang, G., et al., 2016. Enabling point pattern analysis on spatial big data using cloud computing: optimizing and accelerating Ripley's K function. *International Journal of Geographical Information Science*, 30, 2230–2252. doi:10.1080/13658816.2016.1170836
- Zhang, G., et al., 2018a. Validity of historical volunteered geographic information: evaluating citizen data for mapping historical geographic phenomena. *Transactions in GIS*, 22, 149–164. doi:10.1111/tgis.2018.22.issue-1
- Zhang, G., et al., 2018b. A heuristic-based approach to mitigating positional errors in patrol data for species distribution modeling. *Transactions in GIS*, 22, 202–216. doi:10.1111/tgis.2018.22.issue-1
- Zhang, G., et al., 2018c. Modelling species habitat suitability from presence-only data using kernel density estimation. *Ecological Indicators*, 93, 387–396. doi:10.1016/j.ecolind.2018.04.002
- Zhang, G. and Zhu, A.-X., 2018. The representativeness and spatial bias of volunteered geographic information: a review. *Annals of GIS*, 24, 151–162. doi:10.1080/19475683.2018.1501607
- Zhang, G., Zhu, A.-X., and Huang, Q., 2017. A GPU-accelerated adaptive kernel density estimation approach for efficient point pattern analysis on spatial big data. *International Journal of Geographical Information Science*, 31, 2068–2097. doi:10.1080/13658816.2017.1324975
- Zhu, A.-X., et al., 1997. Derivation of soil properties using a soil land inference model (SoLIM). *Soil Science Society of America Journal*, 61, 523–533. doi:10.2136/sssaj1997.03615995006100020022x
- Zhu, A.-X., 1999. A personal construct-based knowledge acquisition process for natural resource mapping. *International Journal of Geographical Information Science*, 13, 119–141. doi:10.1080/136588199241382
- Zhu, A.-X., et al., 2015. A citizen data-based approach to predictive mapping of spatial variation of natural phenomena. *International Journal of Geographical Information Science*, 29, 1864–1886. doi:10.1080/13658816.2015.1058387
- Zhu, A.X. and Mackay, D.S., 2001. Effects of spatial detail of soil information on watershed modeling. *Journal of Hydrology*, 248, 54–77. doi:10.1016/S0022-1694(01)00390-0