A Representativeness Directed Approach to Spatial Bias Mitigation in VGI for Predictive Mapping

By

Guiming Zhang

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Geography)

at the

UNIVERSITY OF WISCONSIN-MADISON

2018

Date of final oral examination: 4/24/2018

The dissertation is approved by the following members of the Final Oral Committee:

A-Xing Zhu (*Advisor*, *Chair*), Professor, Geography

Mark Craven, Professor, Biostatistics & Computer Science

Erika Marín-Spiotta, Associate Professor, Geography

Qunying Huang, Assistant Professor, Geography

Ken Keefover-Ring, Assistant Professor, Botany & Geography

This dissertation is dedicated to

my parents **Shuzhen Tang** (唐树珍) and **Yinfeng Zhang** (张银峰),

and my wife **Xiaoli Zhu** (朱小丽)

Acknowledgements

First and foremost, I want to express my deepest gratitude to my advisor, Professor A-Xing Zhu. I met him in 2010 and have been working with him since then. It has been a great fortune for me to undergo the rigorous trainings to become a scientific researcher under his supervision. He is an incredible advisor and always gives me insightful comments and suggestions on my research. He is more than an academic advisor to me. He has provided me much support and guidance outside research. I am deeply grateful to him for the training and the support I have received. Thank you so much, Professor Zhu!

I also want to thank Professor Mark Craven, Professor Erika Marín-Spiotta, Professor Qunying Huang and Professor Ken Keefover-Ring, for their kindness and willingness to serve in my dissertation committee. They provided many useful inputs on spatial analysis, machine learning, biogeography and statistics, as well as scientific writing to help me complete and improve this dissertation. CS760 Machine Learning, taught by Professor Craven, is among the first courses I took to fulfil the credit requirements for my second M.Sc. in Computer Science at UW-Madison. Some of the initial thoughts on my Ph.D. thesis sprouted from my term project for this course.

I want to express my sincere thanks to Professor Emeritus Jim Burt. It has always been a pleasure to interact with him in our GIS group meetings. His wisdom, insightfulness, and kindness always lights up my inspiration. My sincere thanks also go to Professor Qunying Huang. Her research expertise on geo-computing has opened a new window for my own research. Being the teaching assistant for her first offering of the spatial database course pushed and enlightened me to pursue database courses in computer science, which was the starting point for me to complete my Ph.D. minor and a second M.Sc. in Computer Science. I am also appreciative of the support I received from Professor Robert Roth, Professor Song Gao, GIS Professional M.Sc. Program Director Dr. Ian Muehlenhaus, Graduate Program Director Sharon Kahn, Graduate Program Coordinator Marguerite Roulet, and Department IT Manager Jay Scholz. Best wishes and good luck to Sharon for her new position in the School of Business after serving Geography for almost nine years.

I would also like to extend my appreciation to my friends in the SoLIM research group. They are Dr. Fei Du, Dr. Jing Liu, Dr. Shanxin Guo, Yuying Chen, Han Liu, Aaron Schuck, Jeffrey Hatzel, Starr Moss, and Clara Risk. We shall share all the precious memories of happiness and misery of working and mingling in Rm 440 in Science Hall on the good days and the bad days, accompanied by the eye-catching note sticking on the door-facing wall, “ENGLISH ONLY”. Although Rm 440 is a graduate office without a window through which we could have the luxurious view of Lake Mendota, it is full of shining lights from self-exceling and hardworking fellows, and of course the light tubes on the ceiling.

I also want to thank my close friends and fellow graduate students in Madison. They are Dr. Chaoyi Chang, Dr. Lang Chen, Qianqian Dong, Chaoqun Mei, Dr. Huan Gu, Dr. Lei, Gu, Duanyang Jing, Jack Keel, Dr. Hyun Kim, Xinyi Liu, Chris Scheele, Hui Wang, Gang Yan, Dr. Yang Yang, Liangfei Ye, Zhiwei Ye and Zidong Zhang. They made my life in Madison a joyful journey. Particularly, my sincerest gratitude goes to Captain Jack for his selfless friendship and tremendous efforts of tutoring me and his other tutees on English and the culture. Introduced by Shanxin, I met Jack in the early days after I came to Madison in the Fall of 2013. We have been hanging out for cycling, swimming, driving practice, fixing bicycles and cars, dining on the State Street, drinking beers on the Terrace, tasting ice cream in the Union, gorging premium steaks cooked up by Chef Keel with friends in his condominium, and teaching him the after-dinner poker game Shengji (i.e., Shuangkou) involving up to three decks of cards and five players in my Eagle Heights apartment. Of course, we also do proof reading my writing pieces (including this one), compiling and debugging computer programs, sharing our thoughts on the mechanism of how human brain works, etc. He is truly an Encyclopaedia! Maybe only this legend can properly reflect his tie to the Geography folks. Once upon a time, Jack was the English tutor and friend to all Chinese graduate students and visiting scholars in Science Hall! I will certainly miss him very much after I leave Madison. Jack, please take very good care of yourself and I wish all the best for you!

Finally, I want to thank my parents, Shuzhen Tang and Yinfeng Zhang, for their unconditional support and love, which is beyond any words can describe. I owe a debt of gratitude to my dear wife, Xiaoli Zhu, for her tolerance of my absence of companionship over the course of my Ph.D. study. Thank her for being with me at every difficult moment. I am so fortunate to have her in my life. Half an Earth away, I wish they are in the deepest dreamless slumber.

Table of Contents

Abstract

Information on spatial variation of geographic phenomena is essential to many environmental modeling efforts and geographic decision making. Predictive mapping is a framework for mapping geographic phenomena (e.g., soil, species habitat suitability) whose spatial variation is not directly observable but is highly related to, and thus can be inferred from, their environmental covariates. Establishment of the covariation relationships between the target geographic phenomenon and its environmental covariates is the key for predictive mapping. The relationships are usually derived from *representative* field samples that capture well the spatial variation of the covariates and the geographic phenomenon. Such representative samples are often obtained through well-designed geographic sampling (e.g., random sampling, stratified random sampling, systematic sampling, etc.).

Volunteered geographic information (VGI), referring to geographic information created by volunteer citizens, has the potential of providing field samples at low cost for predictive mapping over large areas. However, there is no sense of geographic sampling design in the provision of samples through VGI. Due to the opportunistic nature of the voluntary observation efforts, VGI observations usually are concentrated more in some geographic areas than others (i.e., *spatial bias*). Due to the spatial bias, field samples consisting of observations from VGI (*VGI-based samples* hereafter) might not be representative. As a result, the accuracy of predictive mapping using VGI-based samples might be unsatisfactory.

This dissertation proposes a representativeness directed approach to spatial bias mitigation in VGI-based samples for predictive mapping. First, *sample representativeness* in this study is defined as the "goodness-of-coverage" of the samples in the covariates space, which in turn is measured by the similarity between the probability density distribution of the samples in the covariates space (i.e., *sample distribution*) and the probability density distribution of all mapping units (raster cells) within the study area (i.e., *population distribution*). Spatial bias is then mitigated by reweighting the samples towards increasing sample representativeness (i.e., decreasing spatial bias). The optimal sample weights that maximize sample representativeness are determined through an optimization procedure based on genetic algorithm. Under the optimal sample weights, samples over-representing the fair-share of their environmental niche are weighted less than samples under-representing their environmental niche. Manifested in the geographic space, spatially clustered samples are weighted less than sparsely distributed samples. Finally, samples weighted with the optimal weights are used to train predictive models for mapping spatial variation of the target geographic phenomenon.

The effectiveness of the proposed representativeness directed spatial bias mitigation approach is thoroughly evaluated through two predictive mapping applications: species habitat suitability mapping and soil mapping. Experiment results show that the accuracy of predictive mapping using samples weighted with the optimal weights is higher than using the original unweighted samples. *Besides*, significance tests suggest that the weight allocation in the optimal weights is statistically meaningful. Accuracy of predictive mapping using samples weighted with the optimal weights is statistically significantly higher than using samples weighted with randomly assigned weights or randomly shuffled optimal weights. *In addition*, a positive relationship between sample representativeness and predictive mapping accuracy was observed, suggesting that the sample representativeness is an effective indictor of predictive mapping accuracy. *In conclusion*, the proposed approach can effectively mitigate spatial bias in VGI-based samples to improve predictive mapping accuracy.

Spatial bias is an issue not only for VGI-based samples. Samples from other sources may also be subject to spatial bias. For example, soil mapping using existing soil samples may suffer unsatisfactory mapping accuracy due to the spatial bias in soil samples obtained from multiple sources. The proposed approach is also applicable for mitigating spatial bias in non-VGI samples for predictive mapping. Besides, spatial bias is one type of sample selection bias. Beyond predictive mapping, the approach is of potential use for sample selection bias correction in many other domains (e.g., machine learning and data mining from biased samples).

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Predictive mapping

Many environmental modeling and geographic decision making efforts require information on spatial variation of geographic phenomena (Goodchild et al. 1993; Franklin 1995; Zhu & Mackay 2001; Zhu et al. 2015b; Zhang et al. 2018). For example, maps containing spatial variation of physical geographic phenomena (e.g., soil, vegetation, temperature, etc.) are indispensable inputs to land surface processes models, such as hydrological models (Zhu & Mackay 2001; Singh & Woolhiser 2002). Biodiversity conservation endeavors call for species distribution or species habitat suitability maps to support decision making in habitat management and restoration, spatial conservation prioritization, and systematic reserve design (Margules & Pressey 2000; Wilson et al. 2005; Elith & Leathwick 2006; Telesco et al. 2007).

Predictive mapping is a framework for mapping spatial variation of geographic phenomena (Zhu et al. 1997; McBratney et al. 2003). Geographic phenomena are influenced by other environmental factors (Zhu et al. 1997; McBratney et al. 2003; Franklin & Miller 2009). As a result, spatial variation of a geographic phenomenon is usually correlated with spatial variation of the influencing environmental factors; There exists covariation between the geographic phenomenon and its environmental factors (covariates). For instance, the formation of soil is influenced by parent material, terrain relief, and vegetation, among others (Dokuchayev 1883; Jenny 1941). Soil distribution thus covaries with the distribution of these environmental factors. *Predictive mapping* maps spatial variation of the target geographic phenomena based on spatial variation of its environmental covariates and their relationships (Franklin 1995; Zhu et al. 1997; Guisan & Zimmerman 2000; Scull et al. 2003; Franklin & Miller 2009), as shown in Equation 1.1.

$$T = f(E)$$

Equation 1.1

where *T* is the target geographic phenomenon, *E* is a set of environmental covariates, and *f* the covariation relationships between T and E. Over the past decades, geographic information acquisition and earth observation techniques have gone through rapid developments. Environmental data sets capturing spatial variation of various environmental factors have been vastly accumulated and are increasingly available (Kerr & Ostrovsky 2003; Hijmans et al. 2005; Gesch et al. 2009). Such data sets provide a large pool of candidate covariates for predictive mapping of geographic phenomena.

The establishment of the *covariation relationships* between the target geographic phenomenon and its covariates thus is the key for predictive mapping and it is often obtained from field samples (Guisan et al. 2002; McBratney et al. 2003; Austin 2007; Franklin & Miller 2009). To achieve high predictive mapping accuracy, it is required that the field samples should capture well the relationship between the spatial variation of the covariates and the spatial variation of the geographic phenomenon over the area to be mapped (*i.e.*, field samples should be *representative*) (Mitchell 1997; Zhu 2000; McBratney et al. 2003; Qi & Zhu 2003; Franklin & Miller 2009).

To ensure that they are representative, field samples are usually 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, etc.) (Brus & de Gruijter 1997; De Gruijter et al. 2006; Gregoire & Valentine 2007; Jensen & Shumway 2010; Wang et al. 2012) and purposive sampling (Yang et al. 2013; Zhang et al. 2016b). Sampling locations are allocated in such a way that the geographic space and/or the covariates space are well covered by the collected field samples (e.g., samples are taken across the complete gradient of the covariates space) (Royle & Nychka 1998; Minasny & McBratney 2006; Gregoire & Valentine 2007; Jensen & Shumway 2010; Wang et al. 2012; Yang et al. 2013; Zhang et al. 2016b).

Obtaining representative samples in geography is *costly*, *labor intensive*, and *time-consuming*. It requires not only careful design in the planning stage but also intensive efforts and labors in the field sampling stage (Gregoire & Valentine 2007; Yang et al. 2013). Logistic constraints such as inaccessibility to designed sampling locations and prohibitive weather conditions further complicate geographic sampling by bringing unexpected adjustments to the sampling plan (e.g., Zhang et al. 2016) and introduce time lags in collecting field samples. Thus, it can be expensive to conduct geographic sampling to collect filed samples, particularly over large areas. In addition, the temporal dynamics of geographic phenomena are of interest in many applications (e.g., environmental monitoring) (e.g., Fink et al. 2010). But it might be impractical to collect field samples through geographic sampling to reflect temporal dynamics because of the time lags in collecting field samples, or because periodically collecting field samples is prohibitively costly. As a result, obtaining field samples through geographic sampling is mostly conducted over small areas for predictive mapping of geographic phenomena that are relatively static (e.g., soil), except for cases where large-scale monitoring networks have been well established (e.g., Hargrove et al. 2003; Brus et al. 2011).

1.2 VGI as a way of obtaining field samples for predictive mapping

Volunteered geographic information (VGI) refers to geographic information created by citizen volunteers (Goodchild 2007a, 2007b). VGI broadly includes geographic information generated by volunteer participants in public participation geographic information system (PPGIS) (Sieber 2006), neogeography (Turner 2006), wikification of GIS (Sui 2008), citizen science (Silvertown 2009; Dickinson et al. 2012), crowdsourcing (Harvey 2013), and social media (Longley & Adnan 2016), as they all share the commonality of voluntary and non-expert geographic information creation. VGI represents a paradigm shift in how geographic information is created and shared, as well as in its content and characteristics (Elwood 2008a, 2008b). VGI is thought of as an innovation that will have profound impacts on geographic information science (GIScience) and more generally on the discipline of geography and its relationship to the general public (Goodchild 2007a).

The remarkable phenomenon of VGI has proliferated in recent years because *technological advancements* have enabled the general public to generate geospatial data (Goodchild 2007a; Elwood 2008a; Graham et al. 2011). With ubiquitous access to the Internet and to positioning technologies such as global positioning system (GPS), average citizens can now easily create and share georeferenced observations of the world through their smartphones, personal computers, and other portable devices (e.g., Haklay & Weber 2008; Sullivan et al. 2009; Gao et al. 2011). Interest in VGI has grown rapidly and it is now driving many successful *ongoing applications*. OpenStreetMap is producing geographic information for every corner of the world that is freely available (Haklay & Weber 2008). eBird is documenting presence and abundance data for hundreds of bird species at continental and global scale, and the collected data have been used in various scientific research (Sullivan et al. 2009). VGI is providing timely information for disaster monitoring and response in emergency management of wildfires and earthquakes (Goodchild & Glennon 2010; Zook et al. 2010). In the world's poor and remote areas, local residents are serving as cost-effective data sources for collecting wildlife distribution data to support conservation programs (Anadón et al. 2009; Zhu et al. 2015a; Zhang et al. 2017b). Soil scientists are promoting the use of VGI in digital soil mapping (Rossiter et al. 2015). There is also growing use of VGI as reference data for land cover map validation in various projects (Fonte et al. 2015). VGI is an important component of the boarder phenomenon of geospatial big data (Xu & Yang 2014; Zhang et al. 2016a, 2017; Yang 2017) that greatly contributes to the paradigm shift from traditional scientific research to the emerging "data-driven geography" (Miller & Goodchild 2014) and, more broadly, "data-intensive science" (Kelling et al. 2009; Hochachka et al. 2012).

VGI has several *advantages* as an alternative mechanism for the acquisition and compilation of geographic information. VGI contains rich local information that spans a wide temporal spectrum because citizens, as local experts and sensors, have long been sensing and accumulating knowledge of their respective areas (Goodchild 2007a). But VGI also has the potential to provide geographic information over large areas, given that billions of networked human sensors are distributed across the globe. In addition, VGI can provide timely updated geographic information that are difficult to obtain through remote sensing techniques but can be easily collected by citizens on the ground (Goodchild 2007a; Fink et al. 2010; Kelling et al. 2013a). Moreover, VGI is much less expensive than traditional scientific data collection protocols (e.g., geographic sampling, biological survey). In many cases citizens contribute geographic information purely voluntarily in the spirit of self-promotion and altruism without any hope of financial reward (Goodchild 2007a, 2007b; Coleman et al. 2009). This low cost is of great practical significance in many real-world applications such as the abovementioned wildlife conservation.

Due to the advantages of VGI, it is possible to obtain timely updated field samples from VGI observations to cover large areas. VGI contains valuable field observations of geographic phenomena, and in some cases, represents the only available data that reflect spatial distribution of geographic phenomena of interest. VGI has the *potential* of providing field samples for predictive mapping of spatial variation of geographic phenomena of interest.

1.3 Issues of using VGI for predictive mapping

Data quality of VGI is the major concern when using VGI for predictive mapping or any other VGI applications. The general public engaged in creating VGI is not composed of well-trained professionals and their voluntary data collection actions are mostly constrained by internal commitment. Thus data collected by volunteers may or may not be accurate (Goodchild 2007a). Three aspects of VGI data quality are particularly relevant to the use of VGI for predictive mapping: *source creditability*, *positional accuracy*, and *spatial bias*.

1.3.1 *Source credibility*

Volunteers need to be trustworthy in reporting the observed geographic phenomenon (e.g., sighting of a bird species) so that VGI can provide ground truth observations that are useful for predictive mapping. Approaches have been developed to assess the source credibility of VGI. (Flanagin & Metzger 2008) proposed a method for assessing VGI source credibility that examines the information environment fostering collective information contribution, explores the environment of information abundance, examines credibility and related notions within this environment, and leverages extant research findings to understand user-generated geographic information. Foody et al. (2013, 2014) derived information on the quality of sources of VGI using latent class analysis. Bimonte et al. (2014) studied the integration of VGI in spatial online analytical processing systems to address precision and credibility problems related to VGI data. Hung et al. (2016) proposed a method using logistic regression to assess the credibility of VGI for time-critical conditions such as disaster response. eBird uses a two-part approach to quality control during data entry (Kelling et al. 2013b): automated filters and a growing network of regional experts. Automated data quality filters flag records for review based on observation date and geographic location. A flagged entry, once confirmed as legitimate by the observer, is then reviewed by a regional expert reviewer again.

With such credibility assessment methods and quality control mechanisms, the confidence that the reported observations in VGI actually reflect ground truth of the geographic phenomenon of interest can be assessed or controlled to an acceptable level.

1.3.2 *Positional accuracy*

Positional accuracy of the VGI observations used for predictive mapping needs to be high so that the locations can be used to accurately obtain the corresponding values of environmental covariates at these locations from environmental databases. Insufficient positional accuracy of field observations leads to mismatch between the value of the observed geographic phenomenon and the values of the covariates, and thus degrades the accuracy of predictive mapping. For instance, low positional accuracy of species records inversely affects the performance of species distribution models (Osborne & Leitão 2009; Moudrý & Šímová 2012).

Positional accuracy of VGI depends on several factors such as the nature of the geographic phenomena under observation and the availability of positioning technology. Stationary geographic features can be accurately located with the aid of high-accuracy positioning techniques. For example, OpenStreetMap data on human tracks (e.g., roads, streets, buildings) and physical geographic features (e.g., rivers, lakes) are of high positional accuracy comparable to authoritative survey products of government mapping agencies (Girres & Touya 2010; Haklay 2010) because these stationary targets were digitized from accurately georeferenced high-resolution remote-sensing imagery. Smart phones equipped with high-accuracy GPS units ensure generated VGI is associated with accurate geographic coordinates. Geospatially enabled and user-friendly and effective geovisualization interfaces also help improve positional accuracy of VGI (Seeger 2008; Newman et al. 2010; Ma et al. 2014; Zhu et al. 2015b). Olteanu-Raimond et al. (2016) found that much VGI data was acquired with a positional accuracy that, while less than that typically acquired by professional mapping agencies, actually exceeded the requirements of the nominal data capture scale used by most agencies.

It is also important to note that the impact of positional accuracy of field observations on predictive mapping depends on the spatial resolution at which predictive mapping is conducted. Predictive mapping at high spatial resolution (e.g., using covariates of 30 m ×

30 m grids) definitely requires field samples of high positional accuracy that is comparable to spatial resolution of the covariates data layers so that values of the covariates at these locations can be accurately extracted from environmental data layers. In contrast, for predictive mapping at coarse spatial resolution (e.g., 1000 m × 1000 m grids), the absolute positional accuracy of field samples does not have to be very high as long as it is high enough relative to the spatial resolution of covariates data used.

In general, with access to high-accuracy positioning technologies (e.g., GPS-equipped smart phones, georeferenced high-resolution satellite imagery), volunteers can georeference VGI observations to a positional accuracy that is often sufficient for predictive mapping. In applications where the positional accuracy of field observations seems to be insufficient, geospatial analysis techniques can be applied to minimize the impact of positional imprecision of VGI (Khalili et al. 2010; Zhu et al. 2015b).

1.3.3 *Spatial bias*

VGI observations are often concentrated more in some geographic areas than others (i.e., spatial bias) because observations made by citizens are opportunistic in nature (Zhu et al. 2015b). Unlike well-designed geographic sampling schemes which allocate sampling locations in a way such that the geographic space and/or the covariates space are well covered by the collected samples, spatial distribution of the observation efforts of volunteers would be considered neither random nor regular in the sense of geographic sampling design. One example to demonstrate this is wildlife sightings elicited from local residents. Local residents are not intentionally tracking wildlife of interest. Instead, they typically spot the wildlife en route to doing something else. The routes on which local citizens spot wildlife would be considered neither random nor regular but 'ad hoc' (Zhu et al. 2015b). As a result, wildlife sightings elicited from local residents are usually concentrated in areas with higher route accessibility. Such roadside bias is a common phenomenon in records of plant and animal distribution (Kadmon et al. 2004).

Spatial bias in VGI has a significant impact on predictive mapping using VGI (Graham et al. 2004; Fink et al. 2010; Leitão et al. 2011; Pardo et al. 2013; Zhu et al. 2015b). Due to spatial bias, field samples consisting of observations from VGI (*VGI-based samples* hereafter) might not be representative. The covariation relationships derived from VGI-based samples thus might not well represent the underlying covariation between the target geographic phenomenon and its environmental covariates. Spatial bias in VGI, if not appropriately accounted for, would adversely affect the accuracy of predictive mapping using VGI-based samples (Thuiller et al. 2004; Graham et al. 2004, 2008; Kadmon et al. 2004; Barry & Elith 2006; Hortal et al. 2008; Ibáñez et al. 2009; Boakes et al. 2010; Fink et al. 2010; Leitão et al. 2011; Pardo et al. 2013; Kramer-Schadt et al. 2013).

This research assumes that source credibility and positional accuracy of VGI are controlled to an acceptable level, for example, by means of quality control mechanisms in the data collection stage (e.g., Khalili et al. 2010; Foody et al. 2013, 2014; Kelling et al. 2013b). This study focuses on mitigating spatial bias in VGI to improve the prediction accuracy of predictive mapping using VGI-based samples.

1.4 Existing work related to spatial bias mitigation

There are few studies focusing on spatial bias mitigation of VGI for predictive mapping. The general problem of sample selection bias is an issue encountered in various fields, and methods developed to correct for sample selection bias can be grouped into: training local predictive models, filtering samples, weighting samples based on cumulative visibility, factoring bias out, modeling sample selection process, and importance weighting (Table 1.1).

Table 1.1. Summary of methods related to spatial bias mitigation.

Domain	*Method*	*Limitations*	*References*
Predictive mapping	Training local predictive models with samples in sub-areas.	Does not account for potential spatial bias in sub-areas.	Fink et al. 2010; Fink et al. 2013.
	Weighting samples based on cumulative visibility at the observation sites.	Applicable only when cumulative visibility is a reasonable approximation of sampling/observation effort.	Zhu et al. 2015b.
	Filtering samples based on the heuristic that removing samples within certain distance of one another would balance the bias.	Reduces sample size. Determination of the distance threshold.	Kramer-Schadt et al. 2013; Boria et al. 2014; Varela et al. 2014.
	Factoring bias out by selecting background samples with the same	Requires sampling/observation effort information to generate background samples.	Dudík et al. 2005; Phillips et al. 2009.

	bias as the presence-only samples.		
Statistics	Modeling the sample selection process.	Needs good understanding and detailed information of the sampling process.	Heckman 1979; Vella 1998; Bethlehem 2010; Bethlehem 2012.
Machine learning	Weighting samples by an importance weighting function in learning classifiers.	Requires sufficiently large sample size to estimate the optimal weighting function. Hard for high dimensional cases.	Shimodaira 2000; Zadrozny et al. 2003; Zadrozny 2004; Sugiyama et al. 2007; Cortes et al. 2008.

1.4.1 Training local predictive models

Fink et al. (2010, 2013) proposed an *AdaSTEM* approach that exploits variation in the density of VGI observations to accommodate spatial bias in broad-scale biological survey data (i.e., eBird data). The continent- or hemisphere-wide study area is partitioned into rectangular spatial units (i.e., sub-areas) of size dependent upon density of VGI observations. Predictive models are trained with only VGI observations in each spatial unit and are later used for prediction in that spatial unit. This approach mitigates spatial bias in the overall data set to a certain degree by training local predictive models in sub-areas, instead of training a global predictive model using data over the whole area. But a sub-area over which a local predictive model is trained still covers a large geographic area (e.g., 3 × 4 latitude by longitude). VGI observations in such a large sub-area potentially have spatial bias as well. This approach does not address the potential spatial bias in VGI observations within each sub-area.

1.4.2 Filtering field samples

Filtering samples in the geographic or environmental space (i.e., remove localities that are within certain distance of one another) is also applied to reduce sample selection bias (Kramer-Schadt et al. 2013; Boria et al. 2014; Varela et al. 2014). This method is based on the heuristic that removing localities (i.e., field samples) that are within certain distance of one another would somehow balance the unequal sampling or observation effort. Yet there is no objective way of determining the distance threshold, which has a profound impact on the filtering process. Moreover, it reduces effective sample size and discards useful information in the removed samples. It thus does not apply to cases where only a paucity of field samples exists.

1.4.3 Weighting samples based on cumulative visibility

If detailed information on sampling or observation effort is available, such information can then be incorporated to correct for spatial bias. Zhu et al. (2015b) proposed to compensate for spatial bias in VGI by weighting VGI observations with weights inversely proportional to the *cumulative visibility* at the observation sites (i.e., the frequency at which a given location can be seen by observers from the routes they take), given that cumulative visibility is a good proxy of the underlying observation effort in VGI genesis. This method is applicable only for cases where cumulative visibility is a reasonable approximation of sampling or observation effort.

1.4.4 Factoring bias out

Spatial bias is a common problem in many biological datasets (e.g., natural history museum animal records) because of unequal sampling efforts in their collection (Franklin & Miller 2009; Phillips et al. 2009; Pardo et al. 2013). Dudík et al. (2005) and Phillips et al. (2009) developed a *FactorBiasOut* method to correct for spatial bias in species presence-only data for species distribution modeling with MAXENT (Phillips et al. 2006). This method first estimates an empirical distribution to approximate the underlying but usually unknown sampling distribution that generated the presence-only data. This approximate sampling distribution is then used to factor out the spatial bias in presence-only data. This is done by feeding MAXENT with background data (i.e., pseudo absences) that have the same spatial bias as the presence data. For instance, occurrence data of a target group of species that are observed by similar methods (if such data are available) are taken as the estimate of the effort information and thus are used as the background data (Dudík et al. 2005; Phillips et al. 2009). The *FactorBiasOut* method works only for species distribution models that require background data. It requires information on sampling or observation effort or its estimate to generate the background data. Yet sampling or observation effort information is generally not available in VGI genesis.

1.4.5 Modeling sample selection process

Nonrandom selection is a source of bias in empirical research, such as surveys with nonresponses and self-selections (Särndal & Lundström 2005; Bethlehem 2010) and species distribution models with presence-only data (Phillips et al. 2009), and a fundamental aspect of many social and economic data collection processes (Heckman 1979;

Winship & Mare 1992; Vella 1998). One approach to correcting for such selection bias is to explicitly model the selection processes (i.e., selection probabilities) using selection rules from domain knowledge or parametric selection models fitted on empirical data. These selection models are then used to account for selection bias in estimation or modeling (Heckman 1979; Vella 1998; Bethlehem 2010, 2012). For instance, survey response propensities and probabilities can be modelled using ancillary variables that might influence one's decision of whether to take the survey (Särndal & Lundström 2010), and then be used to correct for nonresponse bias (Bethlehem 2010, 2012). This approach requires deep understanding of the underlying selection processes in order to come up with reasonable selection models. It is difficult to adopt this approach to correcting for spatial bias in VGI because detailed information on the selection processes underlying VGI genesis is rarely available.

1.4.6 Importance weighting

Sample selection bias is also well studied in the machine learning community but under different names such as sample selection bias, covariates shift, data set shift, cost-sensitive learning, and transfer learning (Zadrozny et al. 2003; Zadrozny 2004; Cortes et al. 2008; Pan and Wang 2010; Moreno-Torres et al. 2012). Sample selection bias arises where training and test data are drawn from different distributions. In other words, distribution of the training data (i.e., field samples) in feature space (i.e., covariates space) is different from that of test data (e.g., all spatial units in the study area). The approach to correcting for sample selection bias is importance weighting where, in learning classifiers (e.g., decision trees, support vector machines, logistic regression), training examples are weighted by an importance weighting function in maximizing the log-likelihood function

(Shimodaira 2000) or empirical risk minimization (Sugiyama et al. 2007). *Asymptotically*, the optimal weighting function proves to be the ratio of the probability density function of features on the test data and that on the training data (Zadrozny et al. 2003; Zadrozny 2004; Cortes et al. 2008). The weighting function is estimated based on empirical estimates of the two density functions. This method requires sufficiently large sample size to estimate the optimal weighting function. In addition, density estimation in high dimensional cases is known to be hard (Shimodaira 2000).

1.5 Remaining challenges and research question

Most existing bias mitigation methods rely on information of the underlying sampling or observation process (e.g., selection probabilities, sampling effort) to correct for bias. But generally, such information is not available in VGI genesis because volunteers are not committed to report effort information. The importance weighting method does not apply for correcting for spatial bias in VGI for predictive mapping because predictive mapping using VGI usually involves many environmental covariates (i.e., high dimension) and VGI-based samples of (possibly) small sample size. How to mitigate spatial bias in VGI to improve the accuracy of predictive mapping using VGI-based samples remains a challenge.

The research question of this dissertation is how to mitigate spatial bias in VGI to improve the prediction accuracy of predictive mapping using VGI-based samples.

1.6 Organization of this dissertation

Chapter 1 establishes the research question of this dissertation. The remainder of this dissertation is organized as follows. Chapter 2 develops a representativeness directed approach to spatial bias mitigation in VGI for predictive mapping. The basic idea and implementation of the methodology are discussed in detail. Chapter 3 and Chapter 4 present applications of the representativeness directed spatial bias mitigation approach in two domains: species habitat mapping and predictive soil mapping. The habitat mapping case study presented in Chapter 3 uses species occurrence data contributed by volunteers (i.e., VGI samples). Nonetheless, it is worth noting that the proposed approach is applicable not only for mitigating spatial bias in VGI samples. It is also applicable for spatial bias mitigation in field samples in general (i.e., VGI samples and/or non-VGI samples). The applicability of the approach for spatial bias mitigation in non-VGI field samples was evaluated through a soil mapping case study using existing multi-source soil samples in Chapter 4. Chapter 5 concludes this dissertation and discusses future research directions.

Chapter 2 Methodology

2.1 Basic idea

Spatial bias in VGI adversely affects accuracy of predictive mapping using VGI-based samples because spatial bias impedes the "representativeness" of these samples. The representativeness of a sample set is defined as the degree to which the samples capture the spatial variation of the environmental covariates and the spatial variation of the target geographic phenomenon over the geographic area to be mapped. Thus assessing "representativeness" of VGI-based samples is the basis for mitigating the effects of spatial bias in VGI.

With increasingly available geospatial data sets that are used to characterize environmental covariates (e.g., remote sensing data), it is feasible to assess the "representativeness" of VGI-based samples *w.r.t.* the environmental covariates. Assessing the "representativeness" of VGI-based samples *w.r.t.* the target geographic phenomenon is hard because spatial variation of the geographic phenomenon is unknown (to be predicted). However, given that spatial variation of the target geographic phenomenon and spatial variation of the environmental covariates are correlated (the fundamental assumption of predictive mapping), it is reasonable to expect that the "representativeness" of the samples *w.r.t.* the environmental covariates would be used to approximate the "representativeness" of the samples *w.r.t.* the target geographic phenomenon (Kruskal & Mosteller 1979; Belbin 1993; Hijmans et al. 2000; Minasny & McBratney 2006; Yang et al. 2008, 2013).

The *representativeness* of VGI-based samples is in this research measured as the "goodness-of-coverage" of the samples in the covariates space, which in turn is quantified by the similarity between the probability density distribution of the VGI-based samples in the covariates space (*i.e., sample distribution*) and the probability density distribution of all spatial units in the area (e.g., pixels within a study area) in the covariates space (*i.e., population distribution*) (Figure 2.1). Stronger spatial bias in VGI-based samples would lead to poorer representativeness of the VGI-based samples.

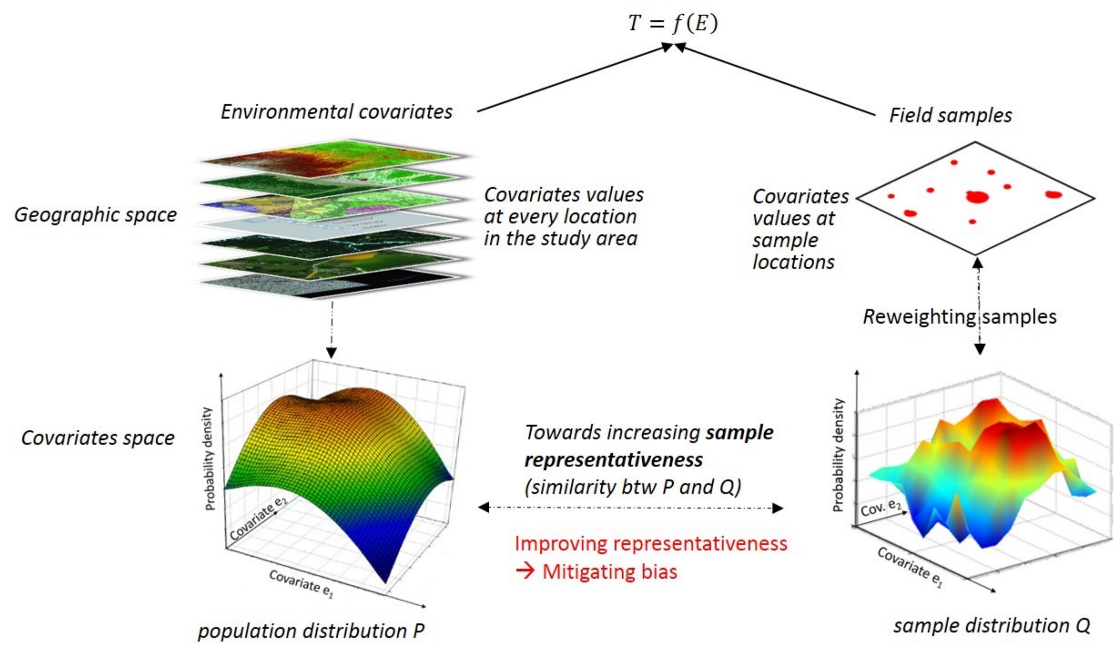

Figure 2.1. Basic idea of representativeness directed spatial bias mitigation.

The effect of spatial bias in VGI is mitigated by improving the representativeness of VGI-based samples. This is achieved by adjusting the sample distribution towards increasing its similarity to the population distribution through reweighting VGI-based samples. VGI observations in an under-represented area would get larger weights and be treated as more important in training predictive models; VGI observations in an over-represented area would get smaller weights and be treated as less important in training predictive models.

Reweighting the samples in this way is expected to reduce bias in the samples and improves representativeness of the samples.

Spatial bias in VGI-based samples cannot always be removed completely. The reweighted VGI-based samples might still not be of perfect representativeness. There might be discrepancies between the adjusted sample distribution (based on the reweighted VGI-based samples) and the population distribution. Sample representativeness thus may also indicate prediction uncertainty.

2.2 Overview of the approach

This dissertation develops a representativeness directed approach to spatial bias mitigation in VGI for predictive mapping (Figure 2.2). The representativeness of VGI-based samples is measured as the similarity between the probability density distribution (in covariates space) of the VGI-based samples (*sample distribution*) and the probability density distribution (in covariates space) of all mapping spatial units within a study area (*population distribution*) (Section 2.3). The approach then mitigates spatial bias in VGI by improving the representativeness of VGI-based samples through reweighting VGI-based samples towards increasing the similarity between the sample distribution and the population distribution. Determination of the optimal weights that maximizes the representativeness of the VGI-based samples is conceived as an optimization problem. A genetic algorithm is adopted to search for the optimal weights.

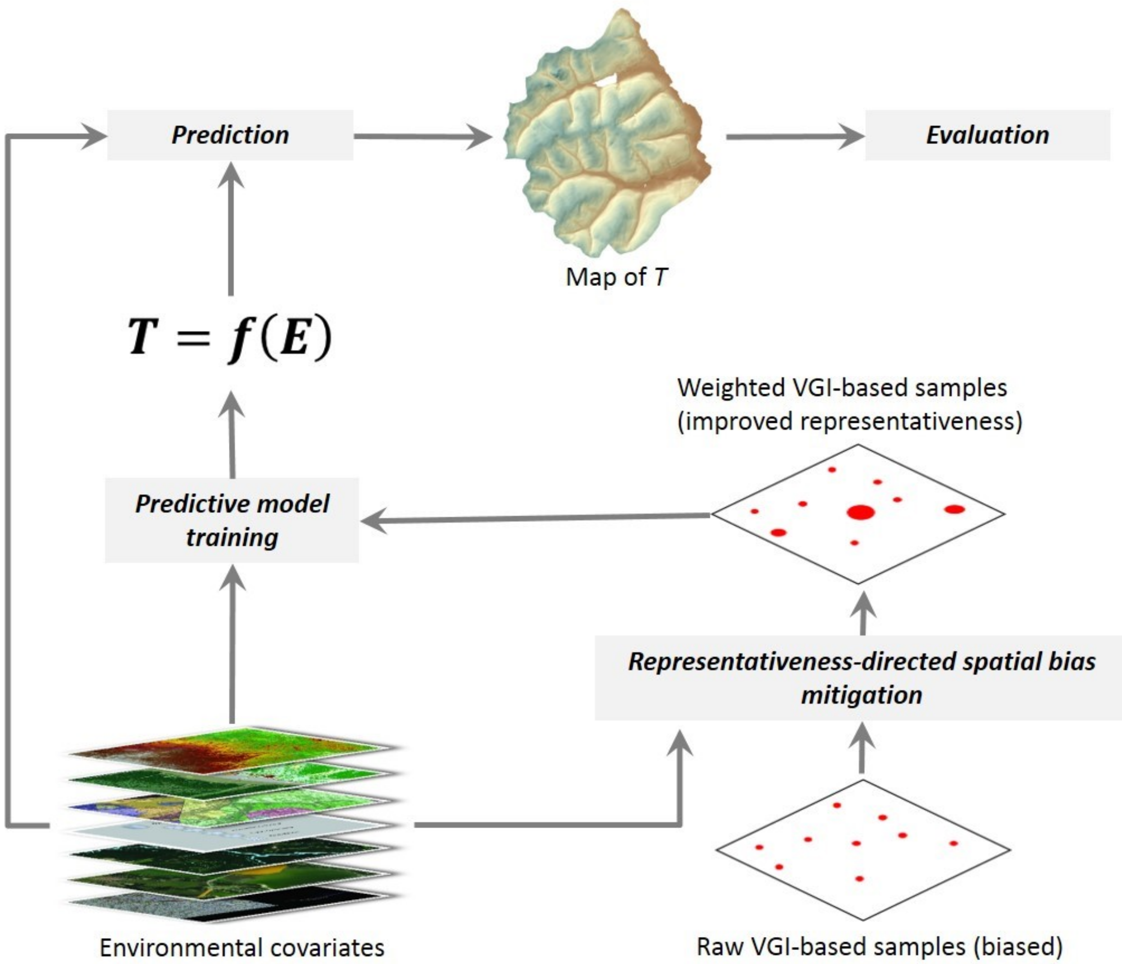

Figure 2.2. Overview of the methodology.

The *weighted* VGI-based samples are then used as input to train predictive models (e.g., statistical, machine learning, data mining) to establish the covariation relationships between the target geographic phenomenon and its environmental covariates. The trained predictive models, which encode the covariation relationships, are then used in combination with environmental covariates data to predict spatial variation of the target geographic phenomenon (Section 2.4).

The effectiveness of the proposed approach is evaluated on two aspects (Section 2.5). Its effectiveness to improve prediction accuracy is evaluated by comparing accuracy of predictive mapping using *weighted* (*representativeness improved*) VGI-based samples to that of using *unweighted* VGI-based samples. The effectiveness of sample representativeness to indicate prediction uncertainty is evaluated by examining the relationship between prediction errors and sample representativeness.

2.3 Representativeness directed spatial bias mitigation for VGI-based samples

2.3.1 Measuring *representativeness* of VGI-based samples

The representativeness of VGI-based samples for predictive mapping is measured by the similarity between the *sample distribution* and the *population distribution* in covariate space.

There exist two difficulties in working with probability density distributions in the covariates space. The first is the *high-dimensionality* of the covariates space. A large number of environmental covariates could be used in predictive mapping due to the fact that the target geographic phenomenon (e.g., soil, habitat suitability) often covaries with many environmental factors and geospatial data that can be used to represent these factors are increasingly available. The curse of dimensionality (Marimont & Shapiro 1979; Raudys & Pikelis 1980; Indyk & Motwani 1998) might arise in working with multivariate probability density distributions in such high-dimensional spaces. The second difficulty is the *multicollinearity* (Farrar & Glauber 1967; Graham 2003) among environmental covariates. Environmental factors are often correlated with each other due to the shared geographic space over which they co-develop (e.g., association between vegetation and precipitation, soil and terrain). The lack of independence (*multicollinearity*) among covariates makes it even harder to work with multivariate portability density distributions.

2.3.1.1 PCA transformation of the covariates space

Principal Component Analysis (PCA) (Jolliffe 2002) is adopted to overcome the above two difficulties. PCA is capable of deriving linearly uncorrelated orthogonal principal components from the (correlated) covariates while preserving the variance of the covariates. Each component is represented by an eigenvector that defines a linear transformation of the original covariates. The proportion of the variance each component represents can be computed from the eigenvalues associated with the components.

The components are *sorted in descending order of the eigenvalues* (the proportion of the variance each variable represents). By selecting the first few components to work with, the dimensionality of the original covariates space is reduced while retaining most variance of the original covariates. The number of components to be selected, L, is determined as follows. First specify a threshold t ($0 < t \leq 1$) indicating the total proportion of the variance to retain (e.g., $t = 0.9$). Then determine the smallest possible L such that the cumulative proportion of the variance of the first L components is no less than t. The original covariates space is transformed to a new covariates space of lower dimensionality that is defined by the L selected components (each component is an axis of the new covariates space).

2.3.1.2 Computing representativeness

There are two steps involved in computing the representativeness of VGI-based samples in the *PCA-transformed* covariates space.

Step 1: Compute similarity regarding each selected component

On *each of the L components*, the probability density distribution of the VGI-based samples (*sample distribution*) and the probability density distribution of all the spatial units to be mapped (*population distribution*) are estimated using kernel density estimation (KDE).

KDE is a nonparametric density estimation method capable of estimating continuous probability density functions (PDF) from discrete samples. It follows Equation 2.1 (Silverman 1986):

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

Equation 2.1

where $f(v)$ is the estimated PDF with respect to variable v, V_i is the value of v at sample location i, n is the total number of samples. K is a kernel density function and here the Gaussian kernel was adopted (Equation 2.2) (Silverman 1986):

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

Equation 2.2

h is a smoothing parameter called bandwidth and it is a crucial parameter for KDE. A bandwidth that is too large would result in a "flat" PDF that fails to reflect the variability in data, and one that is too small would result in a "spiky" PDF that contains too much noise. The "rule-of-thumb" algorithm (Equation 2.3) is often adopted for bandwidth determination (Silverman 1986):

$$h = 1.06 \cdot \sigma_v \cdot n^{-1/5}$$

Equation 2.3

in which σ_v is the standard deviation of values of v at sample locations. This simple algorithm can estimate a robust bandwidth when n (sample size) is large. The computational cost of this algorithm is very low.

Alternatively, the "golden section search optimization procedure" (Brunsdon 1995) can be adopted to find the optimal bandwidth based on maximum likelihood criterion through cross-validation on the sample data. This procedure finds the optimal bandwidth that maximizes the likelihood of observing the sample data by searching within a range of bandwidth values, for example, from $0.01h_0$ to $2h_0$ where h_0 is the "rule-of-thumb" bandwidth. Full details of the "golden section search optimization procedure" can be found in Brunsdon (1995). This procedure can determine a robust bandwidth even when the sample size is relatively small. Its computational cost increases dramatically with increasing sample size (Zhang et al. 2017).

Based on the KDE method, the sample distribution and population distribution *w.r.t.* the l^{th} principal component are estimated using Equation 2.4 and Equation 2.5 respectively:

$$Q^l(v^l) = \sum_{i=1}^{n} w_i \frac{1}{h^{lQ}} K\left(\frac{v^l - V_i^l}{h^{lQ}}\right)$$

Equation 2.4

and

$$P^l(v^l) = \sum_{j=1}^{m} \frac{1}{h^{lP}} K\left(\frac{v^l - V_j^l}{h^{lP}}\right)$$

Equation 2.5

In the above equations, *n* is the number of sample locations in the VGI-based samples, and *m* is the number of locations (pixels) in the study area to be mapped. Q^l and P^l are the estimated sample distribution and population distribution on the l^{th} component, respectively. v^l is some value of the l^{th} component. V_i^l is the value of the l^{th} component at the i^{th} sample location in the VGI-based samples and w_i is a normalized weight (i.e., $\sum_{i=1}^{n} w_i = 1$) of the i^{th} sample location (Equation 2.1; samples can have different weights). V_j^l is the value of the l^{th} component at the j^{th} pixel in the area.

h^{lQ} and h^{lP} are the bandwidths. In this study, h^{lQ} (i.e., bandwidth for estimating sample distribution) is determined using the "golden section search optimization procedure" because this procedure is well suited for finding a robust bandwidth even when *n* (i.e., number of VGI-based samples) is relatively small. h^{lP} (i.e., bandwidth for estimating population distribution) is determined using the "rule-of-thumb" algorithm (Equation 2.3) as it can estimate a robust bandwidth at a very low computational cost when *m* (i.e., number of pixels in the area) is large (Silverman 1986).

The *similarity* between Q^l and P^l, SIM^l, is computed as the overlapping area between the two distributions (Zhu 1999) (Equation 2.6):

$$SIM^l = \frac{2 \times A_{Q^l} \cap A_{P^l}}{A_{Q^l} + A_{P^l}}$$

Equation 2.6

in which A_{Q^l} and A_{P^l} are the areas under the sample distribution curve and the population distribution curves respectively. $A_{Q^l} \cap A_{P^l}$ is the overlapping area under both curves. SIM^l, with a value range of 0 to 1, reflects the "goodness-of-coverage" of the VGI-based samples regarding the l^{th} component. Figure 2.3 shows an example of the sample distribution and the population distribution on one principal component.

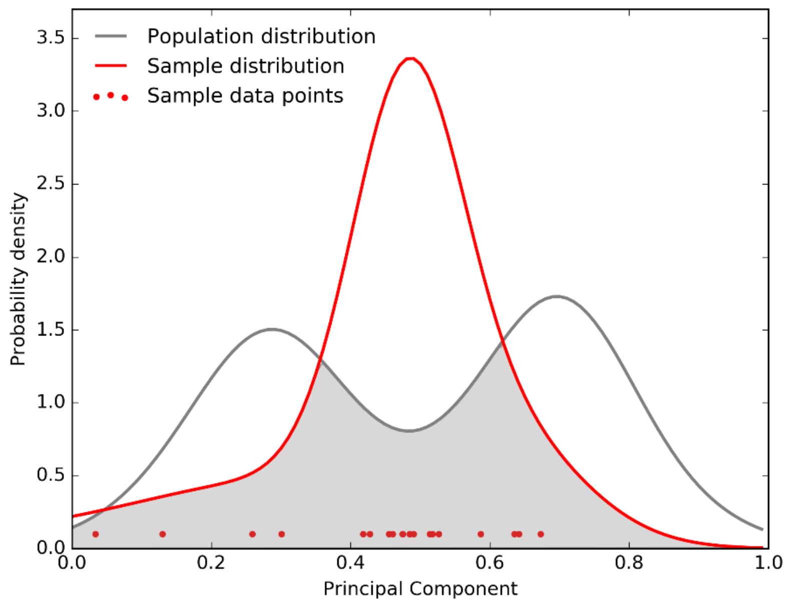

Figure 2.3. An example of population distribution and sample distribution on one principal component. The similarity (i.e., overlapping area) between the two distributions is 0.579.

The similarities between the sample distribution and the population distribution regarding each of the L selected principal components are computed following Equation 2.4 through Equation 2.6.

Step 2: Compute overall similarity regarding all L selected components

The *representativeness* of VGI-based samples is measured as the *overall similarity* between the sample distribution and the population distribution. The overall similarity is a weighted average of the similarities regarding each of the L selected components computed in step 1, with the weight being proportional to the proportion of the variance each component represents (Equation 2.7):

$$R = SIM^{overall} = \sum_{i=1}^{L} \frac{\lambda^i}{\sum_{j=1}^{L} \lambda^j} SIM^i$$

Equation 2.7

in which R is the representativeness of the VGI-based samples. $SIM^{overall}$ is the overall similarity between the sample distribution and the population distribution. SIM^i is the similarity between the two distributions regarding the i^{th} component. λ^i is the eigenvalue of the i^{th} component indicating the amount of variance it represents. The proportionality of the weight to the proportion of the variance each component represents is a desirable property, given that the components representing larger variance are more important for the VGI-based sample to capture variation of the covariates. The value range of R is between 0 and 1, with a larger value indicating higher representativeness of the VGI-based samples.

2.3.2 Representativeness directed spatial bias mitigation

The effect of spatial bias in VGI on predictive mapping using VGI-based samples is mitigated by improving the representativeness of VGI-based samples. This is achieved by adjusting the sample distribution towards increasing its similarity to the population distribution (i.e., improving representativeness) by means of reweighting the VGI-based samples. VGI observations in over-observed (over-represented) area would get smaller weights; observations in under-observed (under-represented) area would get larger weights. The key is to determine the optimal weights that maximize the representativeness of VGI-based samples in a systematic manner.

The determination of the optimal weights for VGI-based samples is conceived as an optimization problem, where the objective is to find a set of optimal weights corresponding to sample locations in the VGI-based samples that maximizes the representativeness of the VGI-based samples.

A Genetic Algorithm (GA) (Davis 1991; Mitchell 1997, 1998) is adopted to search for the optimal weights. Genetic algorithms are motivated by an analogy to biological evolution. Potential solutions to a problem are represented and initialized as a collection of individuals (i.e., population), with each individual in turn composed of genes. A fitness function is evaluated on each individual and a fitness score is computed indicating the quality of an individual (*evaluation*). GAs generate successor individuals by recombining (*crossover*) parts of the best currently known individuals and repeatedly probabilistically muting genes (*mutation*). At each iteration, the current population is updated by probabilistically replacing some fraction of the population (*selection*) by offspring of the most fit current individuals with an intension of improving the average fitness of the updated population. This evolution process continues until a predefined number of generations are gone through, or fitness values of individuals exceeding some fitness threshold (Davis 1991; Mitchell 1997, 1998).

The workflow of a GA used to find a set of optimal weights for VGI-based samples is shown in Figure 2.4. A set of weights corresponding to the sample locations in a VGI-based sample is encoded as an individual that is an array of length *n* (the number of sample locations in VGI-based samples) filled with non-negative floating-point numbers. A floating-point number on the i^{th} position stands for the weight for the i^{th} sample location. The initial weights are assigned random floating-point numbers drawn from a uniform distribution in the interval of [1.0, W_{max}] in this study (W_{max} = 10.0 if not otherwise specified). The fitness of an individual is *evaluated* as the representativeness of the VGI-based samples that is computed as described in Section 2.3.1.

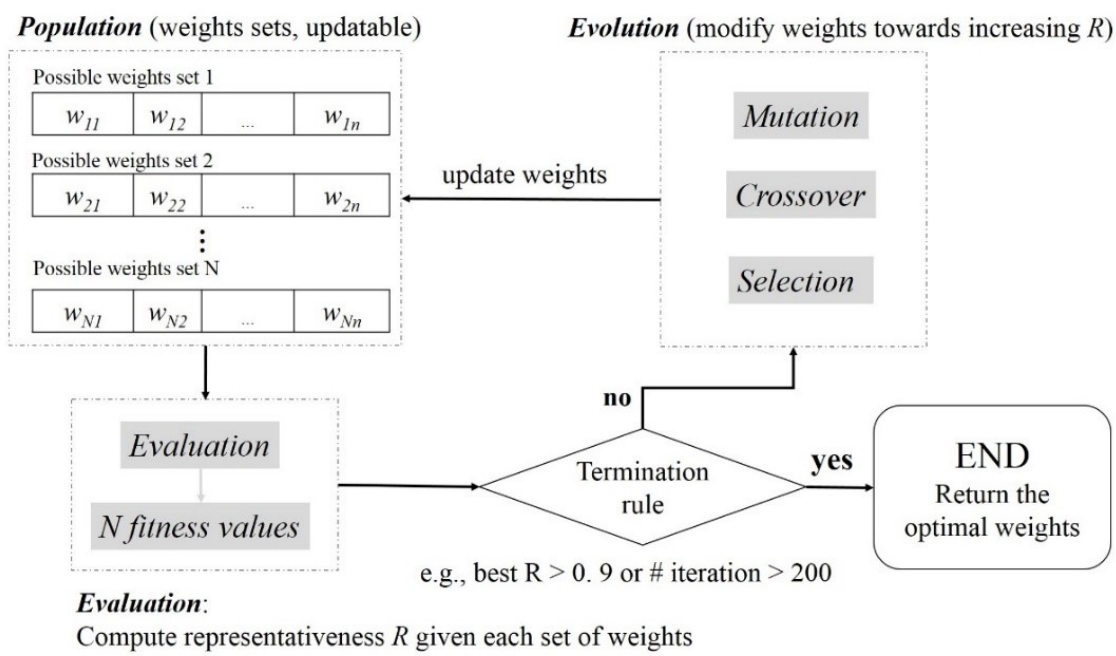

Figure 2.4. Workflow of the genetic algorithm adopted to find the optimal weights for samples.

The *selection* probability of an individual (a possible set of weights) to remain in the population is based on its fitness score (representativeness of VGI-based samples given the set of weights). An individual resulting in a higher fitness score is more likely to remain in the population. The selection probabilities are also used for selecting pairs of individuals to apply the *crossover* operator on each pair to produce new offspring (weights sets). After the offspring are added to the population, a certain percentage (e.g., 5 percent) of the individuals in the population is selected with uniform probability. Each selected individual is *mutated* by adjusting the weight values at randomly selection array positions by a small amount. A Gaussian distribution random number generator is used in this study for weights mutation (mean = 0, standard deviation = 0.5). To this end, the population is *updated*. Another iteration of *evaluation* (i.e., computing representativeness/fitness under updated weights), *selection*, *crossover*, *mutation*, and *update* can be initiated. The GA stops after going through a predefined number of iterations (number of generations = 200 if not otherwise specified), or when the best fitness score of the population exceeds some fitness threshold (e.g., 0.9). The genetic algorithm implemented in the DEAP package (Rainville et al. 2012) is used in this study. Figure 2.5 shows an example of effects of the representativeness directed spatial bias mitigation approach on one principal component.

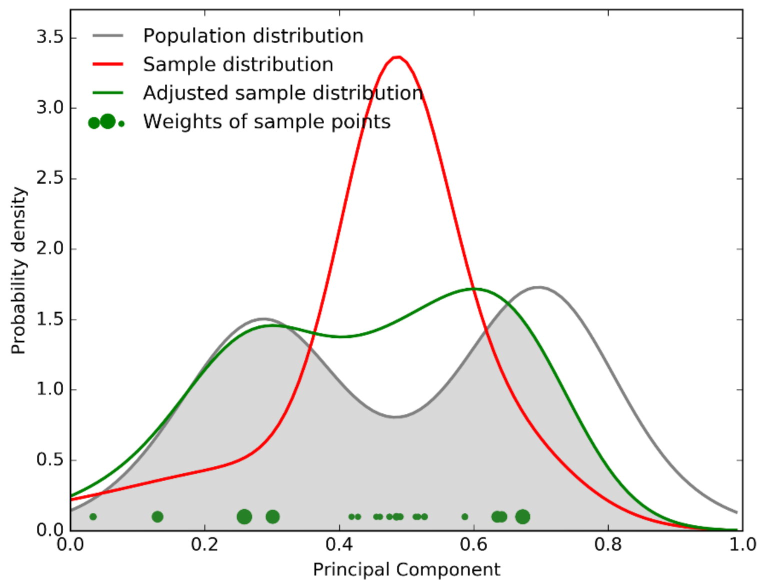

Figure 2.5. An example illustrating the representativeness directed spatial bias mitigation. After reweighting, the similarity between the sample distribution and the population distribution increased from 0.579 to 0.835 (optimal sample weights returned after 25 generations).

It is worth emphasizing the rationale of setting the weight range as [1.0, W_{max}] instead of [0.0, 1.0]. With a weight range of [1.0, W_{max}], it is ensured that every sample would have a weight of at least 1.0 and at most W_{max}. This has several implications. To begin with, every sample contributes in training predictive models, i.e., no samples are excluded. This is desirable in cases where we want to make full use of the samples. More importantly, it also implies that a sample can be treated at most W_{max} times as important than another sample (i.e., the ratio of the relative importance between two samples is bounded). On the other hand, if a weight range of [0.0, 1.0] is adopted, samples of weight 0's would be excluded, and the ratio of the relative importance of samples can be infinitely large (i.e., weight ratio is unbounded). Besides, compared to [0.0, 1.0], the weight range of [1.0, W_{max}] is a wider interval which allows more flexibility for the genetic algorithm to explore the optimal weights. One may argue that weights in the range of [1.0, W_{max}] could be normalized to be in the range of [0.0, 1.0] while maintaining bounded weight ratios, for example through linear stretching. Yet for the same reason as aforementioned (i.e., to allow more flexibility for the genetic algorithm to explore the optimal weights), weights are not normalized to [0.0, 1.0], although they can be normalized where necessary in training predictive models using weighted samples.

It is also worth noting that heuristics based on domain knowledge can be flexibly incorporated in GA. For example, instead of assigning random initial weights, larger weights can be set to array positions that correspond to the sample locations that are deemed more "*important*" (e.g., sample locations in under-observed geographic areas). Similarly, in the mutation step, the probability of mutating weights on the positions corresponding to the important VGI observations can be set lower than others. Such heuristics incorporate prior domain knowledge into the GA. They help accelerate the convergence of the population to optimal individuals (i.e., optimal weights that result in high representativeness of the VGI-based samples). It is most beneficial for problems with large sample sizes, where using GA to determine optimal weights is computationally intensive.

2.4 Predictive mapping using weighted VGI-based samples

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

The trained or learned predictive models, which encode the covariation relationships, are then used in combination with the environmental covariates data to predict spatial variation of the target geographic phenomenon. For every location in the study area (e.g., a pixel), values of environmental covariates at that location are extracted from the GIS covariates data layers. The value of the target geographic phenomenon (e.g., soil class) at that location is then predicted based the covariates values at that location and the covariation relationships encoded in predictive models.

2.5 Evaluation on effectiveness of the proposed approach

The representativeness directed spatial bias mitigation approach is designed for improving accuracy of predictive mapping using VGI-based samples. Evaluation on effectiveness of the approach is conducted on two aspects: its effectiveness to improve prediction accuracy and the effectiveness of sample representativeness to indicate prediction uncertainty.

The effectiveness of the proposed approach to improve prediction accuracy is evaluated as follows. Two predictive maps of the target geographic phenomenon are generated. One is generated by predictive mapping using *unweighted* VGI-based samples. The other is generated by predictive mapping using *weighted* VGI-based samples (the weights are determined as described in Section 2.3.2). The accuracy of the two maps is validated against independent validation samples. If the map based on weighted samples has higher accuracy than the map based on unweighted samples, the proposed approach is proved to be effective in improving prediction accuracy.

The representativeness of the VGI-based samples is an indicator of the overall prediction certainty. The higher representativeness of VGI-based samples is, the higher the prediction accuracy, and thus the lower the prediction error. Based on this observation, the effectiveness of sample representativeness to indicate prediction uncertainty is evaluated by comparing the overall prediction error against the quantified sample representativeness (e.g., through a scatter plot). If there is a negative relationship between prediction error and sample representativeness, this indicates that sample representativeness is an effective indicator to prediction uncertainty.

Chapter 3 Habitat Mapping Application

3.1 Introduction

3.1.1 Background on habitat mapping

Habitat mapping predicts the spatial variation of species habitat suitability and thus is also referred to as habitat suitability mapping, environmental niche modeling, and species distribution modeling (Franklin & Miller 2009). Habitat suitability maps can be used to support a wide range of applications such as ecological monitoring, biodiversity assessment, biological reserve design, habitat restoration, invasive species management, etc. (Ferrier et al. 2002; Thuiller et al. 2005; Telesco et al. 2007; Thorn et al. 2009; Lindenmayer & Likens 2010).

Habitat suitability mapping requires two inputs: environmental data characterizing the spatial variation of environmental conditions (i.e., maps of environmental variables), and species data indicating species habitat use at sampled sites (e.g., occurrence, abundance, absence). The relationships between species habitat suitability and environmental conditions are derived from species data and environmental covariates. Habitat suitability mapping is then achieved by projecting the relationships from environmental space to geographic space to map spatial variation of species habitat suitability (Guisan & Zimmerman 2000; Hirzel & Lay 2008; Warren 2012). With the rapid development of geospatial technologies such as global positioning system (GPS), geographic information systems (GIS), and remote sensing, environmental data are now increasingly available (van Zyl 2001; Kerr & Ostrovsky 2003; Gillespie et al. 2008; Viña et al. 2008).

Obtaining species data, particularly wildlife data, often requires much more effort than environmental data. Generally, wildlife are collected using techniques such as transects and distance sampling (Anderson et al. 1979; Buckland et al. 2001), radio telemetry (Campbell & Sussman 1994), infrared trapping cameras (Trolle & Kéry 2003; Burton et al. 2012), and GPS collars (Hemson et al. 2005). Wildlife data collected through these techniques are of high quality as the data collection efforts follow certain sampling designs. Yet, these techniques are often prohibitively expensive for habitat mapping projects with limited budgetary support and projects over large mapping areas. The high cost renders them impractical and unsustainable for wildlife habitat mapping projects over large areas and projects conducted in poor and remote regions of the world where most of the world's biodiversity hot spots occur (Myers et al. 2000; Danielsen et al. 2003).

Geospatial technologies have enabled the general public to contribute species data (Goodchild 2007a; Sullivan et al. 2009). With ubiquitous access to the Internet and positioning technologies such as GPS, average citizens can now easily contribute georeferenced species observations through their smartphones. Volunteers are contributing a vast amount of species observations around the world on a daily basis (Sullivan et al. 2009). For instance, amateur birders around the world are reporting occurrences of thousands of bird species to the eBird citizen science project on daily basis (Sullivan et al. 2009). In the world's poor and remote areas, local residents are serving as cost-effective data sources for collecting wildlife data to support conservation programs (Anadón et al. 2009; Zhu et al. 2015b). Such VGI can provide species data for mapping species habitat suitability over large areas at low cost.

Species observations from VGI are often concentrated more in some geographic areas than others (i.e., spatial bias) because observations made by citizens are opportunistic in nature (Zhu et al. 2015b). Volunteers do not follow well-designed geographic sampling schemes to conduct observation. The spatial distribution of their observation efforts would be considered neither random nor regular in the sense of geographic sampling design. As a result, species observations are usually spatially biased towards areas with denser population or higher route accessibility (Kadmon et al. 2004). Spatial bias in species observations, if not appropriately accounted for, would adversely affect the accuracy of habitat mapping (Thuiller et al. 2004; Graham et al. 2004, 2008; Kadmon et al. 2004; Barry & Elith 2006; Hortal et al. 2008; Ibáñez et al. 2009; Boakes et al. 2010; Fink et al. 2010; Leitão et al. 2011; Pardo et al. 2013; Kramer-Schadt et al. 2013).

3.1.2 Uniqueness of habitat mapping

The uniqueness of species habitat mapping (i.e., differences from other predictive mapping such as soil mapping) lies in the samples used for mapping. Unlike many other geographic phenomena or features that distribute continuously over space, species occurrences are discrete on the landscape. For example, soils are everywhere on the landscape whilst species only occur at certain locations. This difference has important implications on the way of obtaining samples for mapping. For soil mapping, soil samples can be taken at any location the sampler visits in the study area. For species habitat mapping on the contrary there is no guarantee that species occurrence can be observed at every site the observer visits.

One may argue that the observer could record either presence or absence of the species at visited site, in which case a response (presence/absence) can be recorded at any location the observer visits. However, recorded absences are dubious due to various reasons (Hirzel et al. 2002; Gu & Swihart 2004; Li & Hilbert 2008). For instance, species absence recorded during one visit does not imply the species avoid the habitat at that location. Considering species mobility (e.g., wildlife), it could be that the species prefer the habitat and presented at that location, but the presence was not observed. It could also be that the environmental conditions at that location are suitable for the species, but they do not present there because of accessibility constraints (i.e., geographic or ecological barriers). As a result, species absence data are often regarded unreliable and only species occurrence data are used for species habitat suitability mapping (Hirzel et al., 2002; Phillips et al., 2006).

Species occurrence data result from observers' observation effort (e.g., volunteers, citizen scientists). It is reasonable to expect that, if the underlying observation effort is spatially biased, spatial bias would also exist in species occurrence data. In other words, better representativeness of the underlying observation effort can imply better representativeness of the recorded species occurrences.

Thus, the proposed representativeness direct approach should be applied on "samples" representing observation effort of the observers (instead of species occurrence locations). For example, *locations at which the observers carried out observations* (regardless of the specific species observed) can be treated as the *biased samples* on which the approach should be applied. The optimal sample weights determined from the approach for these observation locations can then be used to weight species occurrence locations in training predictive models to mitigate spatial bias for improving habitat suitability mapping accuracy.

This chapter examines the applicability of the representativeness directed approach for mitigating spatial bias in species occurrence data from VGI to improve the accuracy of habitat suitability mapping.

3.2 Materials and methods

3.2.1 Study area and data

3.2.1.1 Study area and species

The study area for this case study is the state of Wisconsin located in the north-central United States (in the Midwest and Great Lakes regions) (Figure 3.1). The 169,640 km^2 study area can be divided into five regions with distinct geographic features. The Lake Superior Lowland in the north occupies a belt-shape area along Lake Superior. Just to the south, the Northern Highland is covered by massive mixed hardwood and coniferous forests and hosts thousands of glacial lakes. The Central Plain in the middle of the study area provides rich farmland. The Western Upland occupies the western part. The Eastern Ridges and Lowlands in the southeast are home to several large cities. Most of the study area has warm-summer humid continental climate except for the southern and southwestern portions with hot-summer humid continental climate. This area receives a large amount of regular snowfall averaging around 100 cm in the south with up to 410 cm in the north annually (Martin 1965).

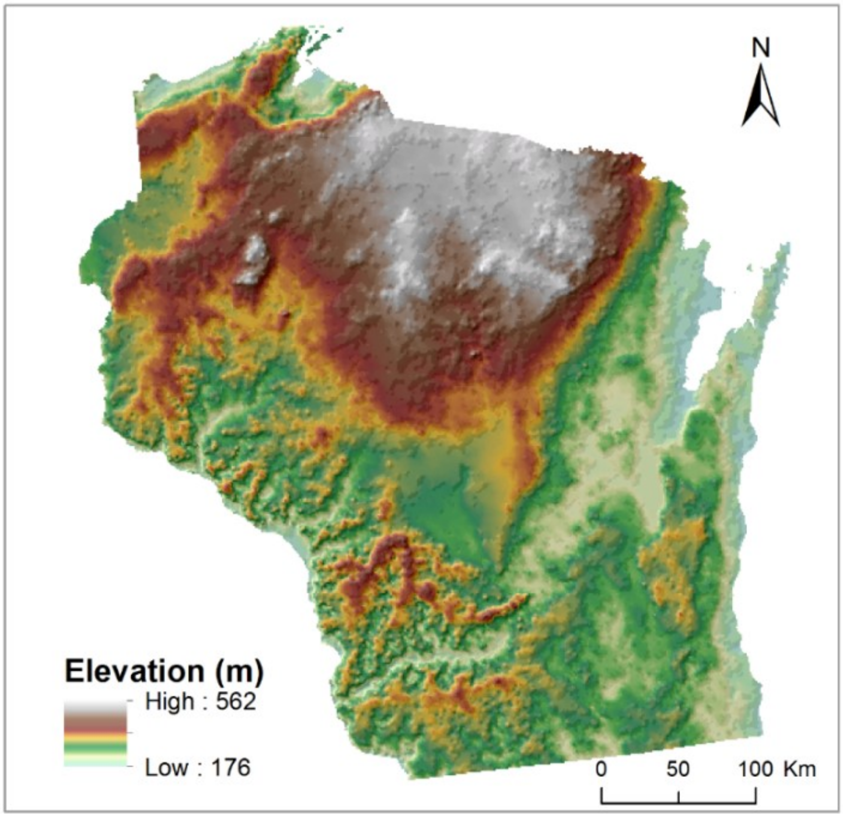

Figure 3.1. Hillshade map of Wisconsin study area.

Habitat suitability of the red-tailed hawk (*Buteo jamaicensis*) was mapped in the study area. The red-tailed hawk is a large bird species that is easy to identify, typically weighing from 690 to 1,600 g and measuring 45–65 cm in length, with a wingspan from 110–145 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 (Wikipedia 2017).

3.2.1.2 Biased VGI samples

VGI data from the eBird citizen science project (Munson et al. 2012; Wood et al. 2011) were used for habitat suitability mapping. *eBird checklist locations* indicating bird watchers' observation effort in the study area were treated as biased VGI samples in the representativeness directed approach.

eBird checklist data are freely available at http://www.ebird.org. Each record in the checklist data is a sampling event. A sampling event contains the geographic locations (latitude and longitude) at which birders carried out bird watching, information describing the sampling event (date, time, duration, locality type, distance traveled, protocol type, observer, trip comments, etc.), and fields indicating the observed count, presence or absence of bird species. The *protocol type* field indicates the type of survey associated with this sampling event. There are four main protocol types: Traveling Count where bird observations are made over a known period of time while traveling a known distance; Stationary Count where the observer should remain in an area approximately 30 m in diameter while recording birds; Area Count where observations are made while thoroughly searching a given location or area; and Casual Observation where observations are incidental sightings that involve no time or distance/area components. The *locality type* field is a code used to define the type of location used. Birders can plot specific locations on a map (P), choose existing locations from a map (H), or choose to submit data for a town (T), postal code (PC), county or state (S). The geographic location (latitude/longitude) of the checklists and information regarding occurrences of the ted-tailed hawk are used in data analysis.

Checklist locations in the study area satisfying the following requirements were selected. *First*, the observation date should fall in June 2012. This condition was imposed because in this case study validation data used for evaluating prediction accuracy were collected only in the breeding season, June (the validation data are from the North America Breeding

Bird Survey, BBS) (see Section 3.2.3.1 for details). Th imposed time frame is to match the time frame of the validation data. *Second*, protocol type was limited to Stationary Count, Casual Observation, or Traveling Count with a distance traveled less than 500 m. This restriction intends to reduce positional uncertainty in the checklist locations. *Third*, locality type must be P (plot locations on a map) or H (choose existing locations on a map). This restriction is also for reducing positional uncertainty in the checklist locations. *Fourth*, the key words "BBS" and "Breeding Bird Survey" and their variants were not mentioned in the trip comment of a checklist. This requirement was imposed because some BBS surveyors did submit their survey data to eBird database; Removing eBird checklists contributed by BBS surveyors is necessary to keep the validation data independent from data used for training predictive models.

Selection according to the above requirements resulted in 1415 checklists. Some of the selected checklists are geographically redundant as they have the same geographic coordinates. After removing such geographically redundant checklists, 655 checklists with unique geographic locations were left (Figure 3.2). The 655 checklist locations indicating bird watchers' observation effort on the landscape are spatially biased. They tend to be clustered in areas with denser population (e.g., the vicinity of large cities) and better accessibility (e.g., along roads).

It is worth pointing out beforehand that in this study the 655 checklist locations were treated as biased "samples" in the proposed representativeness directed approach. The approach was applied to determine the optimal sample weights for these checklist locations (see Section 3.2.4 for more details). It should also be pointed out that these checklist locations themselves were not used to train predictive models. Species occurrence locations were used as training samples to train predictive models (Section 3.2.1.3). The optimal weights associated with checklist locations at which the species occurrences were reported (species occurrence locations; Section 3.1.2.3) were used to weight the species occurrence locations in training predictive models to mitigate spatial bias (see Section 3.2.4 for details).

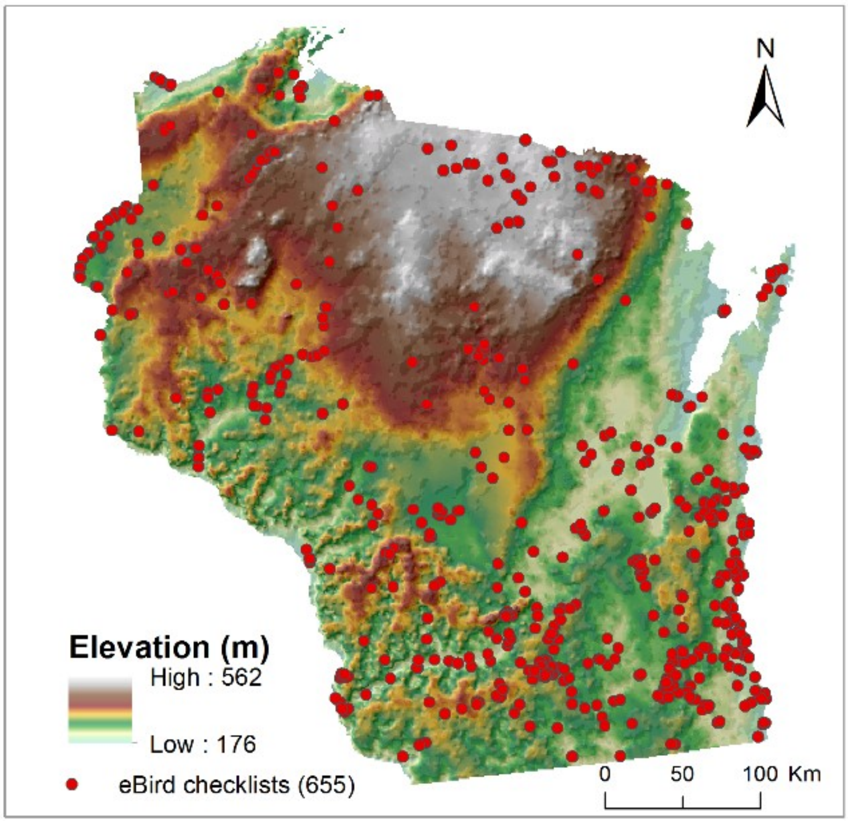

Figure 3.2. Selected eBird checklist locations in June 2012.

3.2.1.3 Training samples

Samples used to train predictive models for habitat suitability mapping include *occurrence locations of the red-tailed hawk* extracted from eBird checklists and *background locations* (i.e., pseudo absences) randomly selected from the study area.

Red-tailed hawk occurrence locations: Among the 655 checklist locations, occurrences of the red-tailed hawk were reported at 75 locations (Figure 3.3). These red-tailed hawk occurrence locations, along with background locations (see next paragraph), were used as training samples to train predictive models for habitat mapping. These occurrence locations were reviewed and approved by regional experts (Sullivan et al. 2009). The occurrences of the red-tailed hawk also seem to concentrate near large cities. However, this may simply be an artifact of the spatially biased observation effort of the bird watchers and it may not necessarily reflect habitat preferences of the red-tailed hawk in the study area.

Background locations: When modeling from species presence-only data, predictive habitat mapping models require both species occurrence locations and background locations (pseudo absences) as training samples (Franklin & Miller 2009). A general method for generating background locations is to randomly select locations from the study area (Franklin & Miller 2009). In this study, 1000 locations were chosen uniformly at random from the study area (i.e., every location/pixel in the area has an equal probability of being selected) and were used as background locations. Environmental conditions at these 1000 background locations were used to represent the environmental conditions in the whole area (about 224,000 pixels). These *background locations* (negative samples) together with the *red-tailed hawk occurrence locations* (positive samples) were used to train predictive models for habitat suitability mapping (see Section 3.2.2 for details).

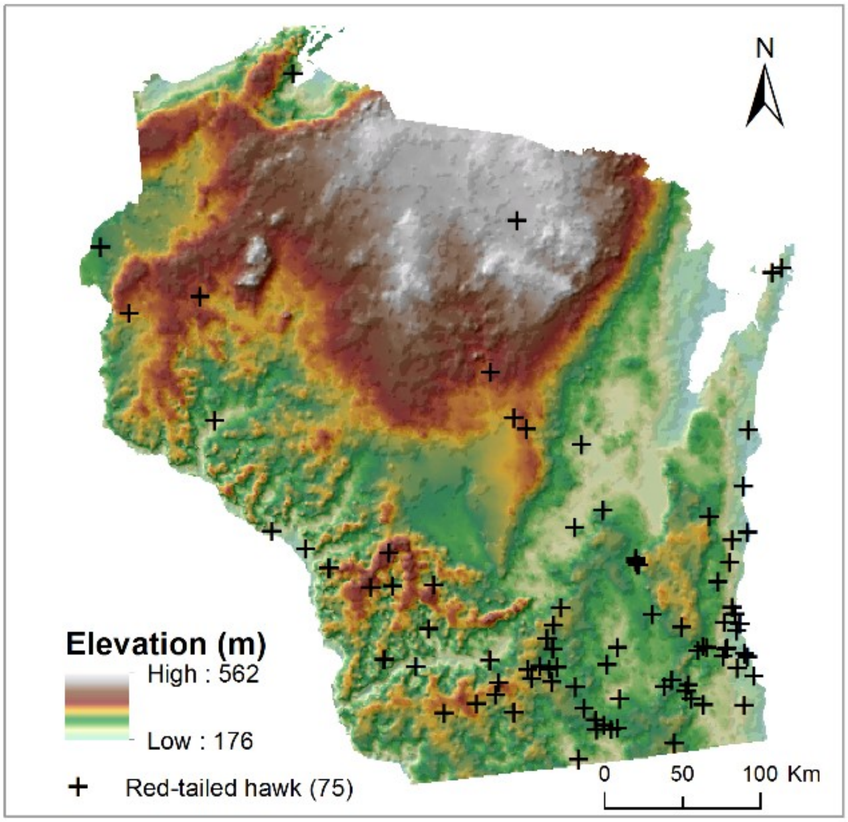

Figure 3.3. Occurrence locations of the red-tailed hawk in June 2012.

3.2.1.4 Environmental covariates

eBird recommends a suite of climatic and landscape variables including statistics about habitat configuration, fine-grained climate measurements, etc. This set of environmental covariates are believed to be a priori most important for most of the species that should be used as a starting point for analyses and species distribution modeling (Munson et al. 2012). Seventy-one environmental variables were selected for predictive habitat suitability mapping in this case study. Amongst them, *housing density*, *housing percent vacant* and *population density* represent population related variables. *Elevation* represents terrain conditions. *Average, minimum, maximum temperature* and *total precipitation* for June represent climatic conditions. *Edge density*, *largest patch index* and *patch density* derived from the 2006 national land cover map represent landscape level indices and statistics. *Edge density*, *largest patch index, patch density* and *percent of surrounding landscape of the same land cover class* computed for each of the fifteen land cover classes represent habitat class specific landscape indices and statistics.

eBird provides a *stratified random design* (SRD) dataset. SRD contains environmental covariates data for random locations within the contiguous US. Locations were chosen using a stratified random design. SRD contains approximately 933,700 random locations roughly based on a 3-km grid of the 48 states in the contiguous US. Covariates values at the random locations were extracted from GIS data layers in the covariates database (Munson et al. 2012). Landscape indices and statistics at a location were computed within a moving circular neighborhood of radius 750 m based on the 2006 national land cover map (Munson et al. 2012).

Based on the covariates values at the SRD random locations, sixty-seven environmental covariates layers (except for the four climatic variables) for the study area were obtained through spatial interpolation using the inverse distance weighted interpolation method (Isaaks & Srivastava 1989). These environmental covariates layers were resampled to 1 km spatial resolution after interpolation. Covariates layers representing the four climatic variables (*average, minimum, maximum temperature, total precipitation*) at 1 km spatial resolution were obtained from the WorldClim database (http://www.worldclim.org/) (Hijmans et al. 2005). These four layers were obtained from WorldClim instead of interpolating from the SRD locations because SRD provides only categorized interval codes rather than the original numerical measurements of these climatic variables at the SRD random locations, which renders spatial interpolation inappropriate.

The covariates layers were standardized by subtracting the mean of a covariate layer from the original covariate layer and then dividing by the standard deviation of the original covariate layer. PCA was then applied to the standardized covariates layers to derive linearly independent principal components. The first 11 principal components retaining 80.1% of the total variance of the original 71 covariates layers (Table 3.1) were selected as new environmental covariates and were used throughout this case study (Figure 3.4).

Table 3.1. Selected principal components for habitat suitability mapping in the study area.

Principal component	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Variance explained (%)	23.1	16.4	8.5	6.4	5.3	4.5	3.9	3.6	3.0	2.8	2.6
Cumulative (%)	23.1	39.5	48.0	54.3	59.6	64.2	68.0	71.7	74.6	77.4	80.1

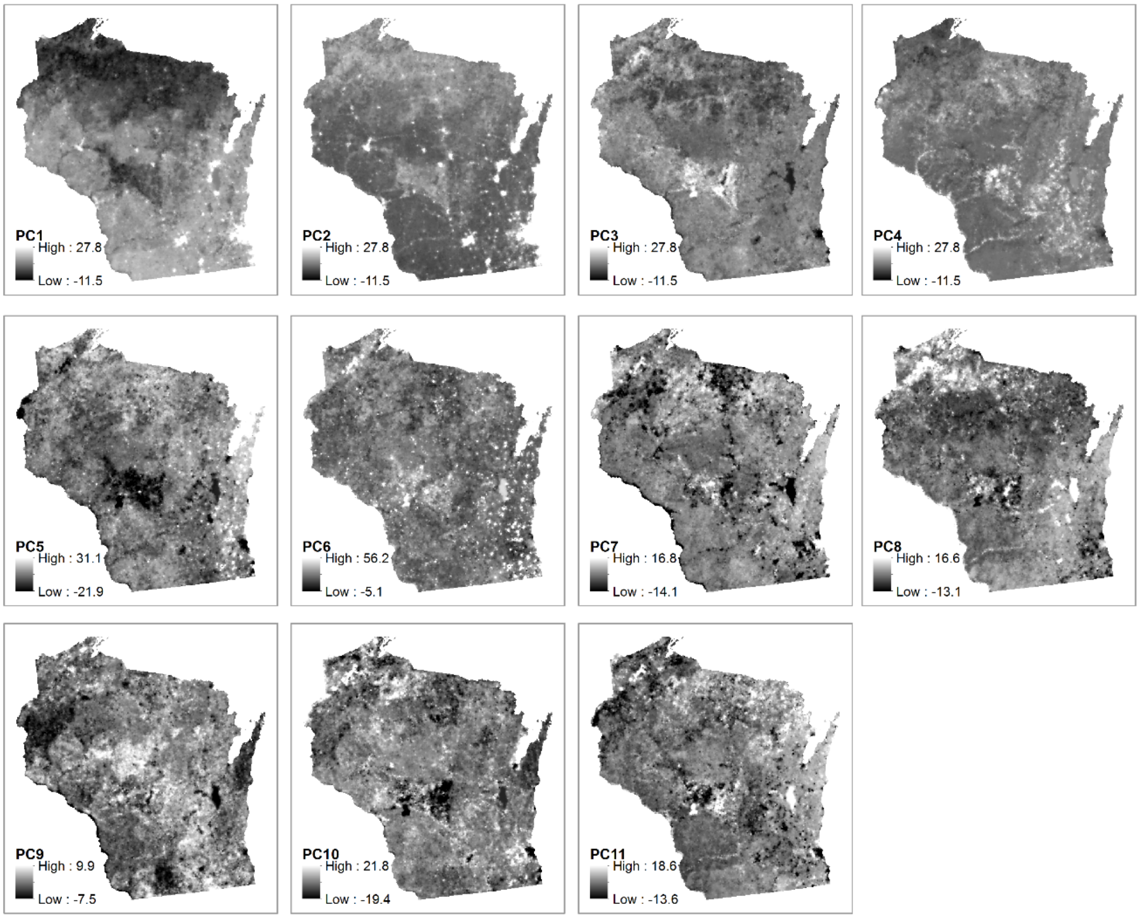

Figure 3.4. Principal components used for predictive habitat mapping in Wisconsin study area.

3.2.2 Habitat mapping methods

Logistic regression (LR) was adopted to map habitat suitability for the red-tailed hawk in the study area.

Background on LR: LR is a general method for modeling nonlinear relationships between a dependent variable that ranges from 0 to 1 (habitat suitability) and a set of independent variables (principal components).

Original version of LR (unweighted samples): A LR model takes the form of Equation 3.1:

$$\hat{S}_j = \frac{1}{1 + e^{-\left(\beta^0 + \sum_{l=1}^{L} \beta^l \times V_j^l\right)}}$$

Equation 3.1

where $\hat{S}_j$ is the predicted habitat suitability at location $\boldsymbol{j}$, V_j^l is the value of the l^{th} principal component at location $\boldsymbol{j}$, β^0 is the intercept, and β^l is the coefficient for the l^{th} principal component.

The intercept β^0 and coefficients β^l's are determined by fitting the model on training samples. Here training samples include species occurrence locations (positive samples; value of the dependent variable is 1) and background locations (negative samples; value of the dependent variable is 0). The values of β^0 and β^l's are determined by minimizing the following cost function (S_i = 1 at occurrence locations, S_i = 0 at background locations) (Equation 3.2).

$$\beta^0, \beta^1, \ldots, \beta^L = argmin_{\beta^0, \beta^1, \ldots, \beta^L} \left\{ \frac{1}{2} \sum_{j=1}^{L} {\beta^j}^2 + \sum_{i=1}^{n} log \left[1 + e^{-S_i \times \left(\beta^0 + \sum_{l=1}^{L} \beta^l \times V_i^l\right)} \right] \right\}$$

Equation 3.2

The procedures implemented in the *Scikit-learn* package (Pedregosa et al. 2012) were adopted to fit a LR model from the training samples. The model was then applied to every location (pixel) in the study area to predict a habitat suitability map.

Revised version of LR (weighted samples): In the original LR method, species occurrence locations are not weighted. In the revised version of LR, species occurrence locations are weighted by sample weights. The weighted occurrence locations (positive samples) along with the background locations (negative samples) were then used as training samples to train a LR model. The weights are used to weight individual log likelihood terms corresponding to positive samples/species occurrence locations (w_i = 1 for all negative samples/background locations) in determining the LR model parameters. Species occurrence locations with larger weights are treated more important in this model fitting process (Equation 3.3).

$$\beta^0, \beta^1, \dots, \beta^L = argmin_{\beta^0,\beta^1,\dots,\beta^L}\left\{\frac{1}{2}\sum_{j=1}^{L}{\beta^j}^2 + \sum_{i=1}^{n} w_i \times log\left[1 + e^{-S_i \times (\beta^0 + \sum_{l=1}^{L}\beta^l \times V_i^l)}\right]\right\}$$

Equation 3.3

The procedures implemented in the Scikit-learn package (Pedregosa et al. 2012), capable of accounting for sample weights, were used to train a LR model. The model was then applied to every location (pixel) in the study area to predict a habitat suitability map.

3.2.3 Evaluation

3.2.3.1 Validation data

Validation data used to evaluate the accuracy of the predicted habitat suitability maps include *red-tailed hawk presence locations* obtained from the North American Breeding Bird Survey (BBS) (Pardieck et al. 2016) and *background locations* randomly chosen from the study area.

Red-tailed hawk presence locations from BBS: The Breeding Bird Survey (BBS) project monitors the status and trends of bird populations in North America since 1966. It is a joint project of the United States Geological Survey (USGS) and the Canadian Wildlife Service (Wikipedia 2016). BBS routes are distributed following a stratified random design to ensure roughly uniform spatial coverage and to sample habitats that are representative of the entire region (Butcher et al. 1986; Sauer et al. 2013; Pardieck et al. 2016). The BBS routes are 24.5 miles long and there are 50 stops at every 0.5 mile (800 m) along the route. The surveys take place during the peak of the nesting season, June (or May in countries with warmer temperatures). At each stop the observer stands near his or her car and records on prepared forms the total number of each bird species heard and seen within a radius of 0.25 mile (400 m). BBS survey data in 2012 were downloaded from this website: https://www.pwrc.usgs.gov/bbs/rawdata. Figure 3.5 shows the active BBS routes in Wisconsin and the first stop on each route. The routes uniformly spread across the study area. The results of bird survey along these routes are thus regarded as representative of the true distribution of the bird species.

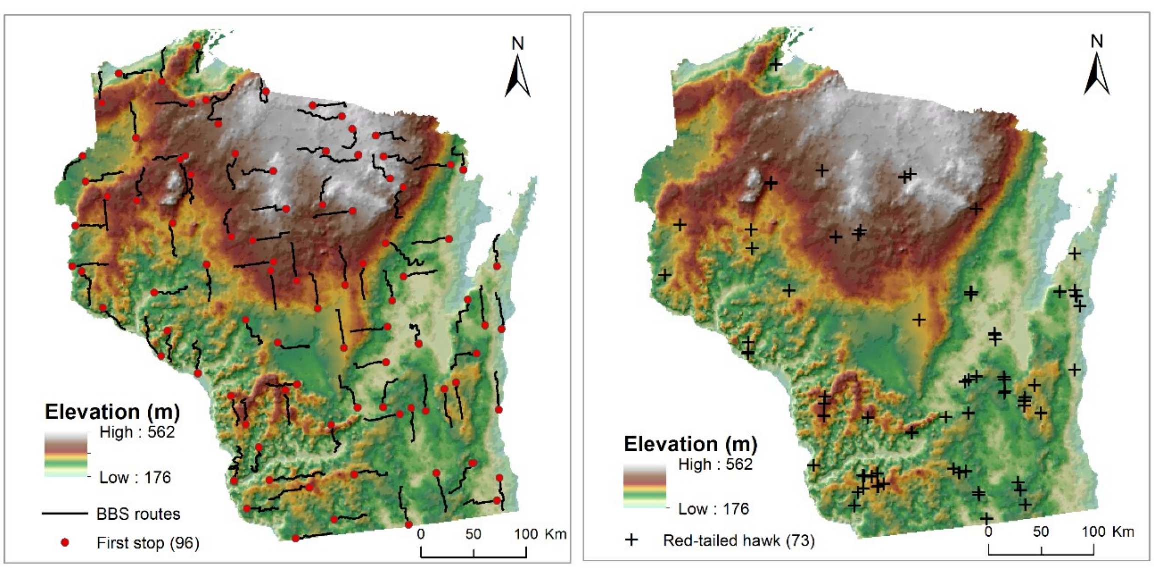

Figure 3.5. Active BBS routes in Wisconsin (left) and occurrences of the red-tailed hawk observed on these routes in June 2012 (right).

BBS provides only the geographic coordinates of the first stop on each route, but not the geographic coordinates of the other 49 stops on each route. Based on the route paths (https://www.mbr-pwrc.usgs.gov/bbs/geographic_information/GIS_shapefiles_2012.html) and geographic coordinates of the first stops, the other 49 stops along each route can be roughly located according to the fact that two consecutive stops are about 0.5 mile apart. 73 stops at which the occurrences of the red-tailed hawk were observed in BBS route survey conducted in June 2012 were selected (Figure 3.5).

Background locations: 1000 background locations were chosen uniformly at random from the study area (i.e., every location/pixel in the area has an equal probability of being selected). It is worth noting that this set of background locations were only used in validation; They are different from the set of background locations used in training predictive models, although both sets of background locations were chosen uniformly at random from the study area.

3.2.3.2 Evaluation metric

The area under the ROC (receiver operating characteristic) curve (AUC) was adopted as an accuracy measure of the predicted suitability map. AUC can be computed for a suitability map given the validation data including presence locations of the red-tailed hawk from BBS (positive) and the background locations (negative) chosen uniformly at random from the study area (Phillips & Dudík 2008).

A ROC curve is constructed with the following steps (Fielding & Bell 1997). *Step 1*: determine a series of habitat suitability thresholds, for example, from 0.05 to 1.0 with an increment of 0.1. *Step 2*: for each of the suitability thresholds, do the following. *First*, αcut the suitability map using the suitability threshold: Assign value 1 (positive) to pixels with suitability value greater than or equal to the threshold; Assign value 0 (negative) to pixels with suitability values less than the threshold. *Second*, compute true positive rate (TPR) and false positive rate (FPR) of the binary map using validation data. TPR is the ratio of the number of species presence locations that are predicted positive on the binary map to the total number of species presence locations in the validation data. FPR is the ratio of the number of background locations that are predicted positive on the binary map to the total number of background locations in the validation data. *Step 3*: ROC curve is obtained by plotting TPR values on the *y* axis against their corresponding FPR values on the *x* axis for all suitability thresholds.

The AUC is the area under the ROC curve. It is interpreted as the probability that the predicted suitability at a randomly chosen species presence location will be higher than that at a randomly chosen background location (Phillips et al. 2006). The AUC ranges from 0.5 to 1.0. A value of 0.5 indicates that the prediction is no better than random predictions. A value of 1.0 indicates perfect model performance, although with species presence-only validation data (i.e., no true absence data) the maximum achievable AUC is less than 1.0 (Wiley et al. 2003). AUC provides a single accuracy measure that is independent of any choice of suitability threshold.

3.2.4 Experiment design

The 655 eBird checklist locations indicate observation effort of the birders in the area. It is assumed that representativeness of the observation effort implies representativeness of the occurrence locations of bird species reported by the birders. Thus, in this study the checklist locations were treated as biased VGI "samples" for which the proposed representativeness directed approach was applied to determine the optimal weights. In applying the approach, sample distribution in the covariate space was computed from covariate values at the checklist locations. Population distribution was computed from covariate values at all pixels in the study area. Representativeness is the similarity between these two distributions (Section 2.3). Default parameter settings for the approach were: Population size for the genetic algorithm = 500; Number of generations for the genetic algorithm = 500; Upper limit of sample weight W_{max} = 10.0.

The optimal weights for checklist locations determined from the proposed approach were then used to mitigate spatial bias in occurrence locations of the red-tailed hawk. Specifically, the optimal weights associated with checklist locations at which the species was observed (i.e., the 75 species occurrence locations) were extracted and used to weight the occurrence locations in training predictive models (Section 3.2.2) for habitat suitability mapping.

3.2.4.1 Effectiveness of the approach

Two habitat suitability maps were produced. One was predicted using the LR model trained using unweighted occurrence locations of the species. The other was predicted using the LR model training using species occurrence locations weighted by the optimal weights. Accuracies of the predicted suitability maps were evaluated by computing the AUCs of the suitability maps based on the validation data. AUCs of the suitability maps predicted from the unweighted species occurrence locations and from the weighted occurrence locations were compared.

Two tests were performed to examine the statistical significance of the effects of weighting the species occurrence locations by the optimal weights. *First*, prediction accuracy achieved under the optimal weights was compared to prediction accuracy achieved under randomly assigned weights. For this purpose, 100 sets of random weights uniformly distributed over the range [1, W_{max}] were generated. The occurrence locations weighted by each set of the random weights were used in training LR models for predicting habitat suitability maps. AUCs of the predicted suitability maps were computed based on the validation data. One sample *t-test* was then applied to test if the AUC achieved under the optimal weights is significantly higher than the mean AUC achieved under random weights.

Second, prediction accuracy achieved under the optimal weights was compared to prediction accuracy achieved under shuffled optimal weights. 100 sets of weights were generated by randomly shuffling the optimal weights (i.e., changing the order of the weights). Species occurrence locations weighted by each set of the random weights were used in training LR models for predicting habitat suitability maps. AUCs of the predicted suitability maps were computed based on the validation data. Again, one sample *t-test* was then applied to test if the AUC achieved under the optimal weights is significantly higher than the mean AUC achieved under randomly shuffled optimal weights.

3.2.4.2 Representativeness vs. prediction accuracy

The relationship between prediction accuracy and sample representativeness over the generations of the genetic algorithm was also examined (recall that the checklist locations were treated as biased "samples" in this study). Weights for the checklist locations (and thus representativeness of the checklist locations) evolve gradually over the generations of the genetic algorithm. At each generation of the genetic algorithm, the weights corresponding to the best representativeness (fitness score) among the individuals in the current population were recorded; Weights for checklist locations at which the red-tailed hawk was observed were extracted and used to weight the species occurrences locations in training a LR model for mapping species habitat suitability. Accuracy of the predicted suitability map was evaluated by computing the AUC using the validation data. A scatter plot was produced by plotting AUC of the predicted suitability map against the representativeness over the generations of the genetic algorithm.

3.2.4.3 Impact of W_{max}

To examine impact of the upper weight limit W_{max} on effectiveness of the representativeness directed approach, the approach was run under various settings of W_{max} (the value range of weights is [1, W_{max}]). Specifically, the approach was applied to determine optimal weights for the checklist locations under the settings of W_{max} = 5, 10, 20, 50, and 100, respectively. The optimal weights obtained under various W_{max} were used to weight the species occurrence locations in training LR models for habitat suitability mapping. AUCs of the predicted suitability maps were computed using the validation data and were compared.

3.3 Results

3.3.1 Effectiveness of the approach

Figure 3.6 (left) shows the optimal weights for the 655 checklist locations determined from the approach. Spatially clustered checklist locations tend to receive smaller weights than sparsely distributed locations. Weighted by the optimal weights, the overall representativeness of the checklist locations increases from 0.855 to 0.935. Figure 3.6 (right) shows the weights corresponding to the red-tailed hawk occurrence locations (i.e., weights associated with the checklist locations at which the species was observed). Densely distributed species occurrence locations (e.g., occurrences in the Milwaukee areas) tend to get smaller weights than sparsely distributed occurrence locations (e.g., occurrences in northern areas).

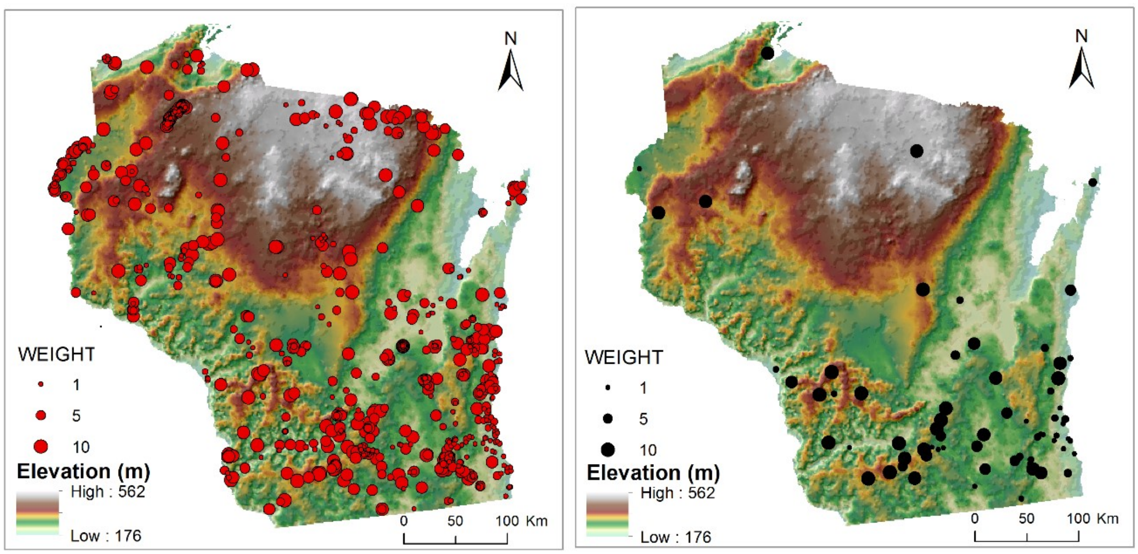

Figure 3.6. Optimal weights returned by the approach for the eBird checklist locations (left) and the weights associated with red-tailed hawk occurrence locations (right).

The habitat suitability maps predicted using LR models trained using unweighted species occurrence locations and using occurrence locations weighted by the optimal weights are shown in Figure 3.7. Using unweighted species occurrence locations, areas predicted to be of higher habitat suitability (e.g., suitability above 0.5) are limited to densely populated urban and suburban areas surrounding big cities such as the Milwaukee, Madison, and Green Bay areas. This spatial pattern of red-tailed hawk habitat suitability most likely reflects an artifact in the training data rather than the ecological reality of the species. There tends to be a larger number of birders carrying out observations in densely populated areas. As a result, observed occurrences of the species are more frequent in these areas than other less observed areas (i.e., biased). When such biased species occurrence locations are used as positive samples in training a predictive model, the model overfits to the majority of the samples that are distributed in those densely populated areas. The predicted habitat suitability map manifests this artifact. In contrast, weighting the species occurrence locations by the optimal weights in training the predictive model can counteract this artifact to better reveal the underlying ecological reality of the species. Using the weighted species occurrence locations, areas predicted to be of higher habitat suitability (e.g., suitability above 0.5) are of a much wider geographic range.

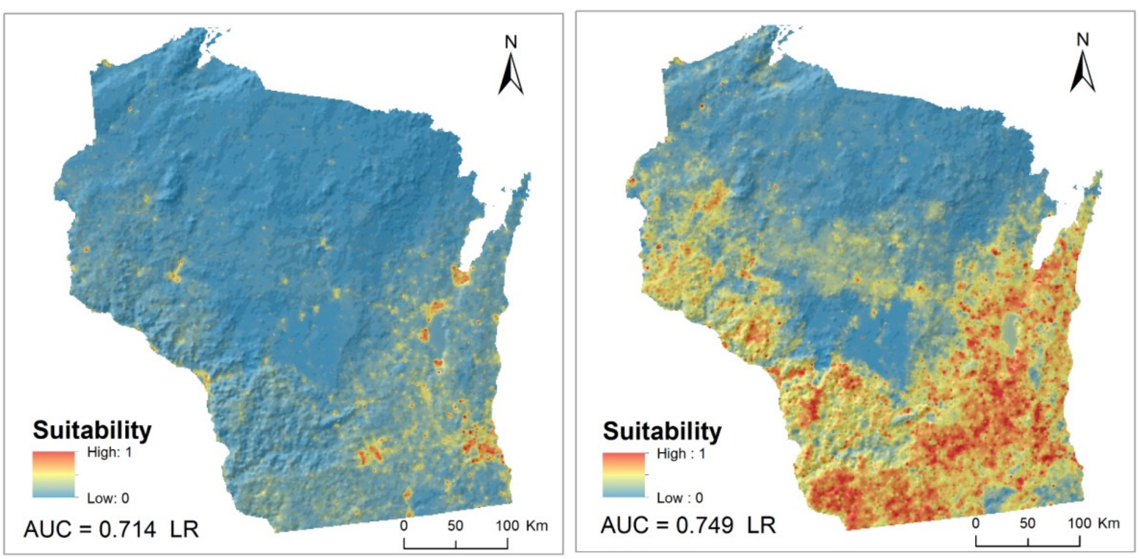

Figure 3.7. Habitat suitability maps predicted using unweighted species occurrence locations (left) and using occurrence locations weighted by the optimal weights (right).

Evaluation results based on the validation data (Table 3.2) reveal that weighting species occurrence locations by the optimal weights improves accuracy of the predicted suitability map. AUC computed on the validation data increases from 0.714 to 0.749 when training the predictive model using occurrence locations weighted by the optimal weights. As revealed by the ROC curves (Figure 3.8), weighting the species occurrence locations reduces overfitting the predictive model to the biased training samples (AUC computed based on the training samples decreases from 0.829 to 0.789) and thus effectively increases validation accuracy.

Table 3.2. Accuracies (AUCs) of habitat suitability maps predicted using unweighted or weighted species occurrence locations.

	Unweighted occurrence locations	*Weighted occurrence locations*
AUC	0.714	0.749

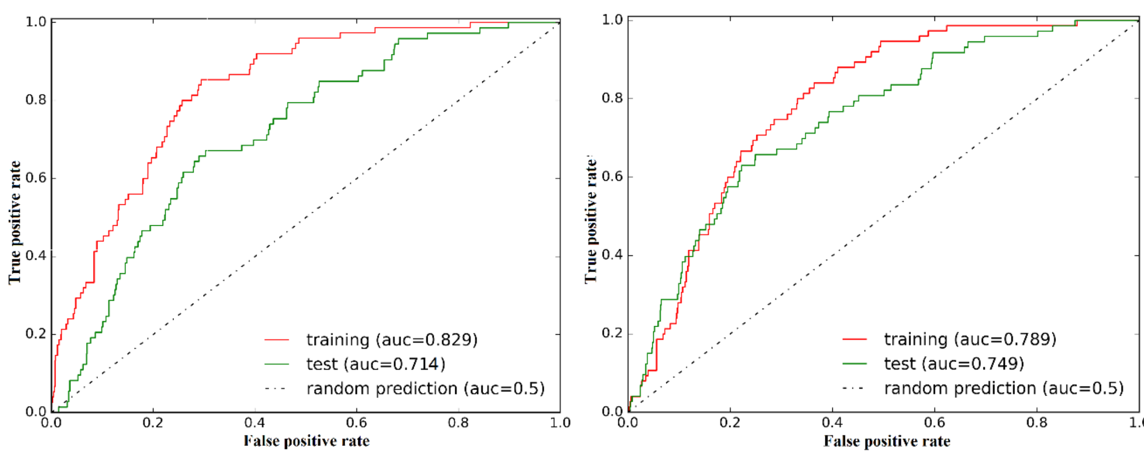

Figure 3.8. ROC curves of the red-tailed hawk habitat suitability maps predicted using unweighted species occurrence locations (left) and using occurrence locations weighted by the optimal weights (right).

Statistical significance tests (Table 3.3) show that the prediction accuracy achieved under the optimal sample weights (AUC = 0.749) is significantly higher than prediction accuracy achieved under randomly assigned weights (mean AUC = 0.717, Std = 0.005). It is also significantly higher than prediction accuracy achieved under shuffled optimal weights (mean AUC = 0.716, Std = 0.007). This suggests that the weight configuration as reflected in the optimal weights returned by the representativeness directly approach is meaningful. The effect of the optimal weights in improving prediction accuracy is better than what would be expected purely by chance.

Table 3.3. Accuracy of suitability maps (AUC) predicted from species occurrence locations weighted by the optimal weights, random weights, and shuffled optimal weights.

Optimal weights AUC	*Random weights*				*Shuffled optimal weights*			
	Mean AUC	*Std AUC*	*t*	*p*	*Mean AUC*	*Std AUC*	*t*	*p*
0.749	0.717	0.004	-74.531	0.000	0.716	0.007	-49.082	0.000

Notes: The mean and standard deviation of AUC were computed based on 100 values. t and p are the one sample *t-test* statistics and p value for the null hypothesis that the AUC achieved under the optimal weights is higher than the mean AUC achieved under random weights or shuffled optimal weights.

3.3.2 Representativeness vs. prediction accuracy

The scatter plot in Figure 3.9 shows the relationship between representativeness of the checklist locations and accuracy of the suitability map (AUC computed using the validation data) predicted using the weighted species occurrence locations over generations of the genetic algorithm (500 generations). A clear positive relationship between representativeness and prediction accuracy was observed (*Pearson*'s r = 0.961). This suggests that representativeness of the checklist locations can effectively indicate accuracy of the suitability map predicted from species occurrence locations.

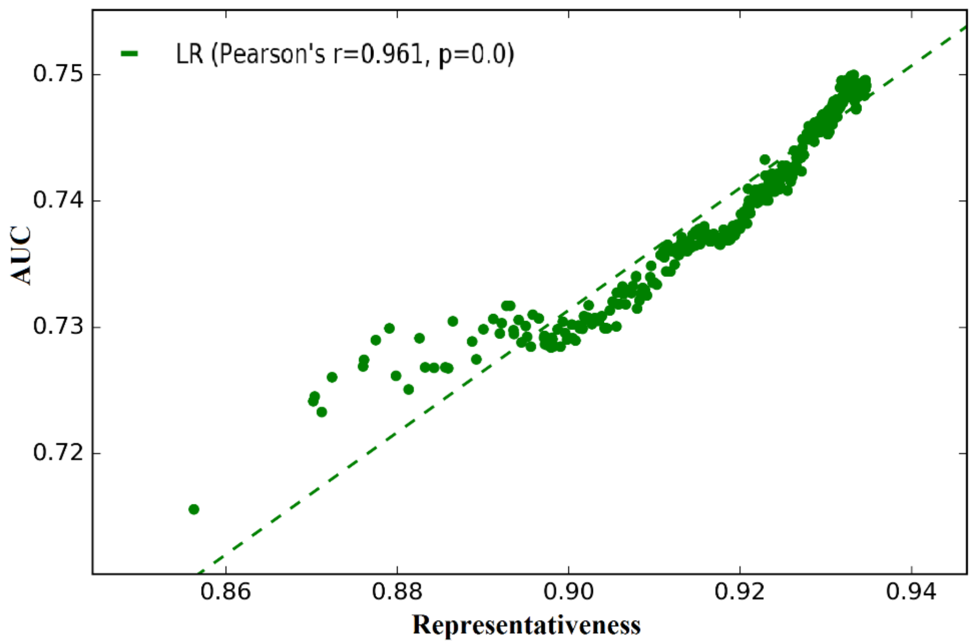

Figure 3.9. Relationship between representativeness of the checklist locations and prediction accuracy of the suitability map (AUC) over generations of the genetic algorithm.

3.3.3 Impact of W_{max}

Figure 3.10 shows the evolution of representativeness of the checklist locations and accuracy (AUC) of suitability maps predicted using weighted species occurrence locations over the generations of the genetic algorithm. As shown in Table 3.4, under all W_{max} settings (W_{max} = 5, 10, 20, 50, 100), accuracy of the habitat suitability maps predicted using species occurrence locations weighted with the optimal weights (AUC = 0.744, 0.749, 0.745, 0.743, 0.724 respectively) is higher than using unweighted species occurrence locations (AUC = 0.714). The AUCs achieved under $W_{max} \leq 50$ (AUC generally above 0.740) are comparable and is generally higher than AUC achieved under W_{max} = 100. The AUC reaches the highest when W_{max} = 10.

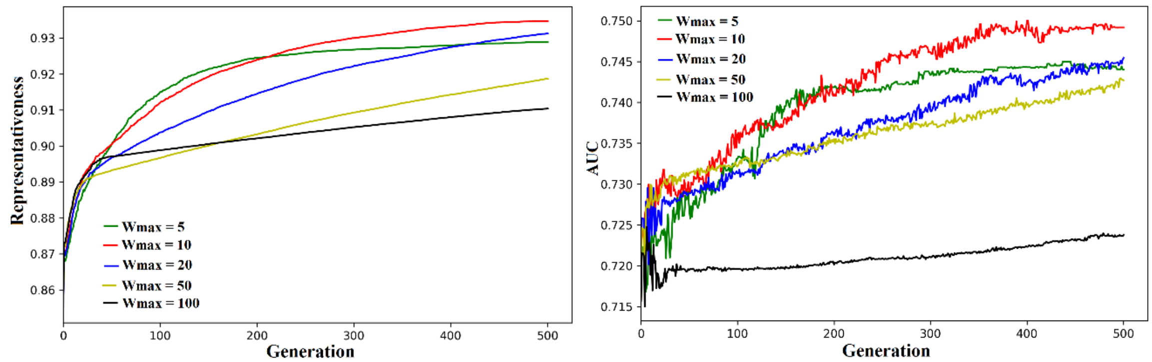

Figure 3.10. Evolution of representativeness of the checklist locations (left) and accuracy (AUC) of suitability maps predicted using weighted species occurrence locations (right) over the generations of the genetic algorithm.

Table 3.4. Accuracy of suitability maps (AUC) predicted from species occurrence locations weighted by optimal weights determined from the proposed approach under different W_{max} settings.

Wmax	5	10	20	50	100
AUC	0.744	0.749	0.745	0.743	0.724

Strong positive correlations were observed among the optimal weights obtained under various W_{max} (*Spearman's r* > 0.49) (Table 3.5). Notably, the correlation among the optimal weights obtained under W_{max} = 5, 10, or 20 are very strong (*Spearman's r* > 0.7). It suggests that the general pattern of optimal weight allocation (i.e., relative importance of the samples) was fairly consistent across various W_{max} settings, especially when W_{max} = 5, 10, or 20.

Table 3.5. *Spearman*'s rank correlation coefficient between the optimal weights for the checklist locations obtained under various W_{max} settings.

W_{max}	*5*	*10*	*20*	*50*	*100*
5	1.000	**0.778**	**0.742**	0.561	0.493
10		1.000	0.720	0.647	0.513
20			1.000	0.620	0.601
50				1.000	0.594
100					1.000

3.4 Discussion

3.4.1 Effectiveness of the approach

In this study, the checklist locations indicating observation effort were treated as biased VGI "samples" and the proposed representativeness directed approach was applied to determine the optimal weights that maximize representativeness of the checklist locations. Checklist locations over-representing their fair-share environmental niche are weighted less than checklist locations under-representing their fair-share environmental niche. Under the assumption that representativeness of the observation effort implies representativeness of reported species occurrence locations, the optimal weights for the checklist locations were then used to mitigate spatial bias in species occurrence locations for predictive habitat suitability mapping. This was done through weighting species occurrence locations by the optimal weights in training predictive models.

Experiment results show that the prediction accuracy of predictive models trained using species occurrence locations weighted by the optimal weights is higher than predictive models trained using unweighted species occurrence locations. Moreover, significance tests reveal that the effect of the optimal weights in improving prediction accuracy is statistically significantly better than what would be expected purely by chance. In addition, a strong positive relationship between sample representativeness and prediction accuracy was observed, suggesting that the representativeness is an effective indicator of prediction accuracy.

3.4.2 Parameter settings

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

The value range of the weights [1.0, W_{max}] is a key parameter for the approach. The physical meaning of W_{max} is that a species occurrence location with weight W_{max} will be treated W_{max} times as important as a species occurrence location with weight 1.0 in training predictive models. Experiment results in this case study show that the optimal weights obtained under various W_{max} settings are positively correlated, indicating that the weight allocations (i.e., order of sample importance) reflected in different sets of optimal weights are fairly consistent. For experiments in this study weighting species occurrence locations with the optimal weights obtained under W_{max} = 10.0 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, atop of taking its physical meaning into consideration.

3.4.3 Applicability of the approach

As was made clear in the Introduction (Section 3.1.2), it is worth pointing out again that the checklist locations representing observation effort of the birders were treated as biased samples in this case study. The proposed representativeness direct approach was applied to determine the optimal weights that maximize the representativeness of the checklist locations (observation effort). The underlying assumption is that improved representativeness of observation effort can imply better representativeness of the reported species occurrence locations.

The proposed approach should not be applied directly on a sample consisting of species occurrence locations. Evaluating the representativeness of species occurrence locations is inappropriate without knowledge of the underlying observation effort. When applied to species habitat suitability mapping based on species occurrence data, the approach requires data representing the underlying observation effort (i.e., checklist locations in this case study). It is thus inapplicable for cases where occurrence locations of only one species are available (the underlying observation effort is unknown). However, if occurrence locations of a group of species resulted from the same underlying observation campaign are available, occurrence locations of the group of species can be pooled together and used as a surrogate of the underlying observation effort (Dudík et al. 2005; Phillips et al. 2009).

The approach is conceivably applicable for species habitat suitability mapping applications where both species occurrence and absence data are available. A sample consisting of species occurrence locations and absence locations reflects the underlying observation effort. Thus, the approach can be directly applied to determine the optimal weights that maximize the representativeness of such a sample for habitat mapping. Yet in practice the availability and reliability of species absence data are always challenging issues for species habitat suitability mapping (Franklin & Miller 2009).

3.5 Chapter conclusions

This chapter examines the effectiveness of the representativeness directed spatial bias mitigation approach in species habitat suitability mapping applications. The approach was thoroughly evaluated through a case study of mapping red-tailed hawk habitat suitability in the Wisconsin study area, United States. Using VGI data from the eBird citizen science project, experiments were conducted to examine the effectiveness of the proposed approach in mitigating spatial bias in species occurrence data to improve accuracy of habitat suitability mapping.

Experiment results show that the representativeness directed approach can effectively mitigate spatial bias in species occurrence data to improve habitat suitability mapping accuracy through weighting species occurrence locations in training predictive models by the optimal weights determined through the approach. Besides, the effect of the optimal weights in improving prediction accuracy is statistically significantly better than the effect of randomly assigned weights. Furthermore, a strong positive relationship between sample representativeness and prediction accuracy was observed, which indicates that the representativeness is an effective indicator of prediction accuracy.

The representativeness directed spatial bias mitigation approach effectively mitigates the adverse effects of spatial bias in VGI samples and improves accuracy of predictive habitat suitability mapping. This approach is useful for species habitat suitability mapping using species occurrence records contributed by volunteers participating in citizen science projects where sampling or observation effort of the volunteers is likely to suffer from spatial bias.

Chapter 4 Soil Mapping Application

4.1 Introduction

The proposed approach is applicable not only for mitigating spatial bias in field samples contributed by volunteers (i.e., VGI samples). It should also apply for spatial bias mitigation in field samples in general (i.e., VGI samples and/or non-VGI samples). The applicability of the approach for spatial bias mitigation in non-VGI field samples was evaluated through a soil mapping case study using existing multi-source soil samples in this chapter.

Soil is an important natural resource on earth. Soil maps provide inventories of this precious resource. Soil information is also crucial ingredient for environmental modeling. For example, soil maps are key inputs to land surface processes models such as hydrological models (Zhu & Mackay 2001; Singh & Woolhiser 2002). Predictive mapping is often adopted to predict soil property and soil class maps based on soil samples and environmental covariates data (Zhu et al. 1997; McBratney et al. 2003).

Soil samples might suffer from spatial bias due to various reasons. To begin with, even soil samples collected through designed geographic sampling campaigns might be subject to spatial bias and would not be perfectly representative. Constrained by field conditions and logistic support, planned sampling sites may not always be accessible, e.g., sampling sites are physically unreachable, or they reside on private land. Such unexpected conditions bring adjustments on the original sampling plan and soil samples are taken at spare locations chosen at the sampler's discretion (e.g., Zhang et al. 2016). Such deviations from the original sampling plan may introduce spatial bias into the samples.

In addition, existing soil samples from multiple sources may suffer from spatial bias. Collecting new soil samples is always expensive and time consuming. It is thus common to use existing soil samples available in the study area for predictive mapping. Legacy soil samples resulted from past soil surveys are often used for soil mapping (Vaysse & Lagacherie 2015). However, in soil surveys, sampling methods adopted by surveyors are generally empirical and lack statistical criteria, which may introduce biases in the sampled areas (Carré et al. 2007). Existing soil samples may also be collected by groups of researchers conducting related investigations in the study area. Yet such soil samples are often for different purposes and may cover different parts of the area at varying density. Moreover, citizen scientist volunteers can contribute soil samples (Rossiter et al. 2015). But, due to the opportunistic nature of the voluntary sampling effort, samples contributed by volunteers are also often spatially biased, for example, towards areas of better accessibility (Kadmon et al. 2004). As a result, pooling all existing soil samples together for predictive mapping makes better use of the available data but may result in a soil sample set that is spatially biased (Liu 2017).

This chapter examines the applicability of the proposed representativeness directed bias mitigation approach for mitigating spatial bias in soil samples to improve the accuracy of predictive soil mapping.

4.2 Materials and methods

4.2.1 Study area and data

4.2.1.1 Study area

The 60 km^2 Heshan study area is located at Heshan farm (116°12′E, 48°57′N) in Heilongjiang province, northeastern China (Figure 4.1). It has a maximum terrain relief of 87 m (276 to 363 m) and is generally flat with slope gradient less than 4°. The soils in this area are mostly formed on deposits of silt loam loess except the valley where the underlying parent material is fluvial deposits. The farm has been cultivated for over 40 years to grow soybeans and wheat. There is a thick *A*-horizon (top-layer of soil) with high organic matter content. The land use and soil management have been uniform throughout the area and no organic fertilizer has been applied to these soils to maintain agricultural productivity because of the naturally high organic matter content (Zhu et al. 2010).

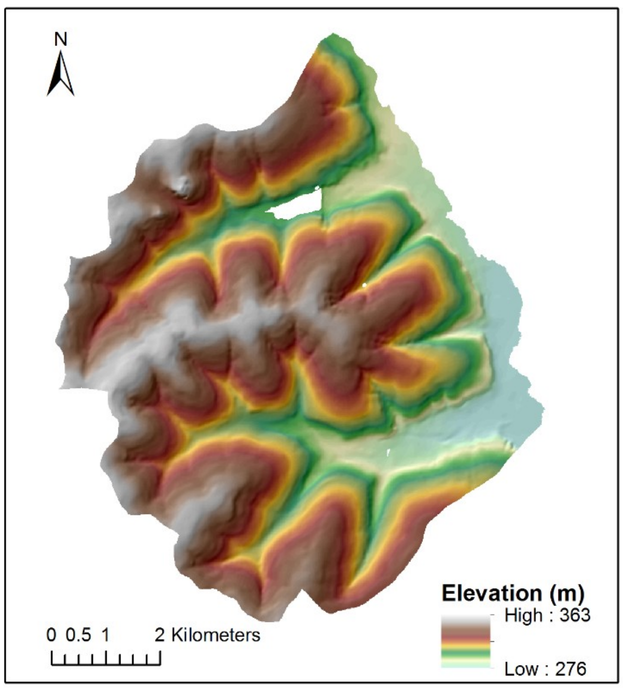

Figure 4.1. Hillshade map of Heshan study area.

4.2.1.2 Soil samples

There are 59 existing soil samples in the study area that were obtained through three sampling strategies (Figure 4.2) (Zhu et al. 2010; Yang et al. 2013; Zeng et al. 2016). 29 samples through integrative hierarchical stepwise sampling, 10 samples through subjective sampling, and 20 samples through transect sampling. The soil organic matter (SOM) content (%) in *A*-horizon soil was measured for each of the soil samples. The *A*-horizon SOM content values at the 59 sample locations have a mean of 4.454 and a standard deviation of 1.638.

The sampling campaigns were designed for different purposes (Zhu et al. 2010; Yang et al. 2013). The 59 soil samples pooled from these sampling campaigns thus are subjected to spatial bias. As can be seen from the spatial distribution of these soil samples, there are areas of clusters of samples where soil samples are more concentrated than other areas. The proposed approach was applied on these 59 samples to mitigate spatial bias and to improve SOM content mapping accuracy using these samples.

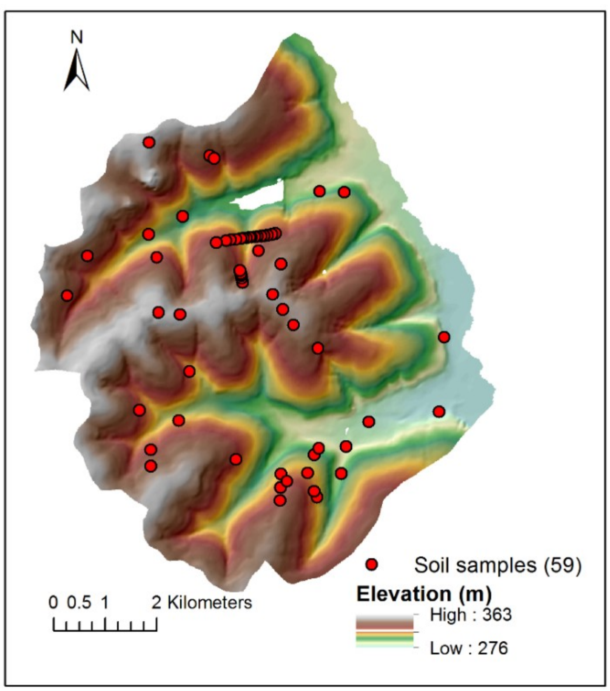

Figure 4.2. Soil samples in Heshan study area.

4.2.1.3 Environmental covariates

Environmental covariates were selected based on soil forming factors (Dokuchayev 1883; Jenny 1941). There are five categories of soil forming factors: climate, organisms, terrain, parent materials, and time. For this small study area, macro-climatic conditions and parent materials are fairly uniform and thus were not considered. Micro-climate conditions can be reflected by terrain variables. Thus, six topographic variables and one vegetation variable were selected as environmental covariates that indicate the spatial variation of the *A*-horizon SOM content in this area: elevation, slope gradient, contour curvature, profile curvature, relative slope position, topographic wetness index (TWI), and normalize difference vegetation index (NDVI).

A digital elevation model (DEM) of the study area was created from the 1:10,000 topographic map of the area. Elevation, slope gradient, contour curvature, profile curvature, relative slope position (Qin et al. 2009), and TWI (Qin et al. 2007) were then derived from this DEM. NDVI was derived from a Landsat ETM+ image of the area obtained on September 25, 2000 (Zhu et al. 2015a). Map of the elevation covariate is shown in Figure 4.1 and maps of other covariates are shown in Figure 4.3.

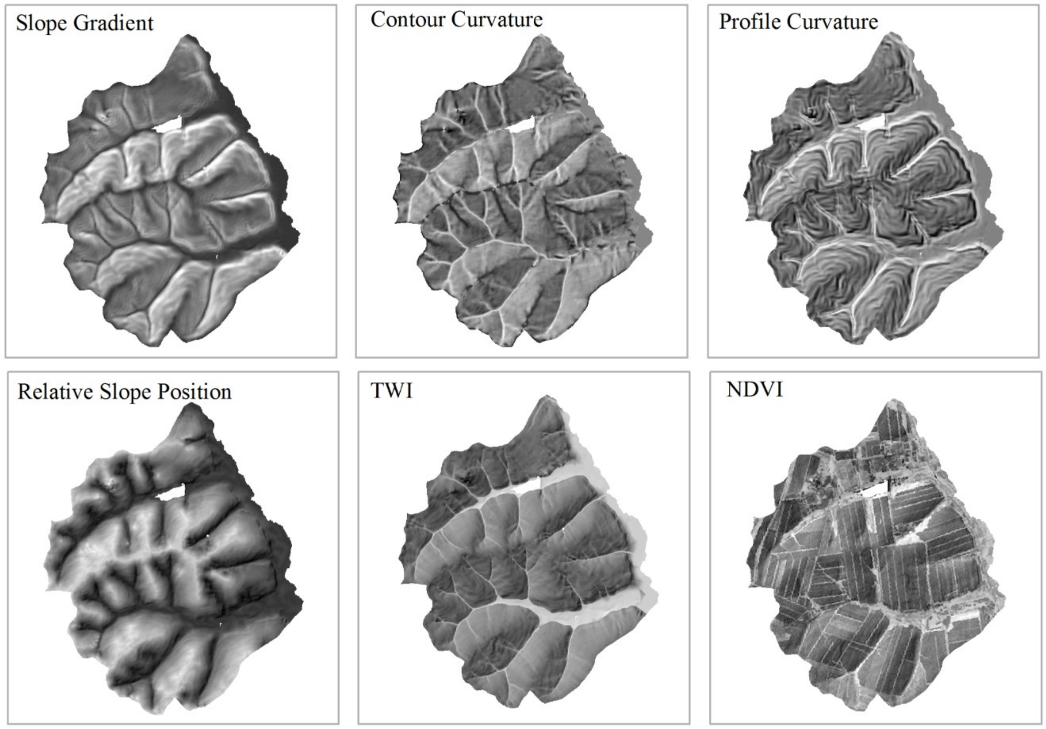

Figure 3.3. Environmental covariates for Heshan study area (brighter colors indicate higher values).

Outliers in the covariates data layers were removed and the covariates layers were linearly stretched to either range 0 to 100 (elevation, slope gradient, relative slope position, TWI, NDVI) or range -50 to 50 (contour curvature, profile curvature) (Yang et al. 2013). PCA (principal component analysis) was then applied to the covariates layers to derive linearly independent principal components (PCs). The first 3 PCs retaining 91.7% of the total variance (66.6%, 17.7%, and 7.4% respectively) were selected as new environmental covariates and were used throughout this case study (Figure 4.4).

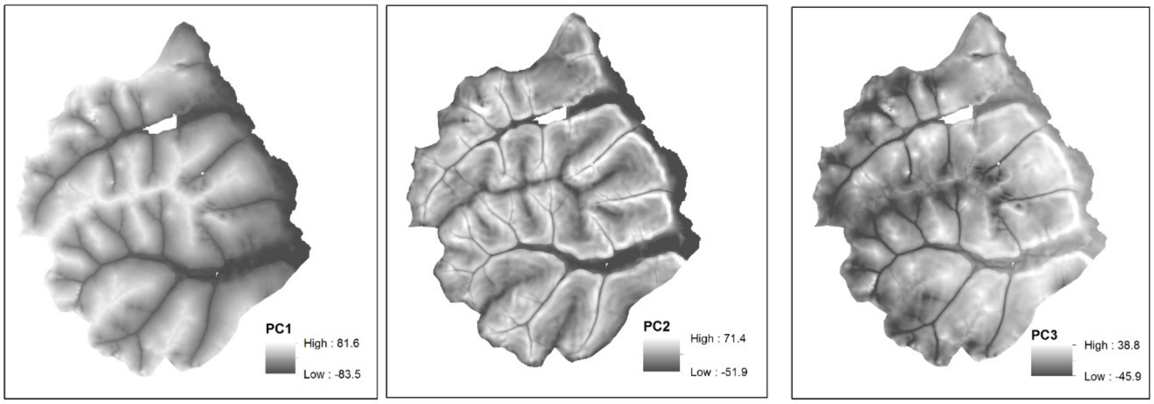

Figure 3.4. Principal components used for predictive soil mapping in Heshan study area.

4.2.2 Soil mapping methods

Two methods were adopted for mapping the *A*-horizon SOM content in the Heshan study area. One is the individual predictive soil mapping (iPSM) method developed specifically for mapping soil properties (Zhu et al. 2015a). The other is the multiple linear regression (MLR) method, a general approach for modeling multivariate linear relationships. Soil mapping using these two methods allows examination of the effectiveness of the proposed representativeness direct approach on domain-specific predictive soil mapping methods as well as general predictive mapping methods.

4.2.2.1 Individual predictive soil mapping (iPSM)

Background on iPSM: iPSM is specially design for predictive soil mapping (Zhu et al. 2015a). It uses the soil-environment relationship at each individual soil sample location to predict soil properties at un-sampled locations. With the assumption that the more similar environment conditions between two locations the more similar soil property values, iPSM predicts soil properties of un-sampled locations based on the environmental similarity between un-sampled locations and sample locations. iPSM imposes no requirements on sample size and does not require the set of sample locations being representative. It is reported to perform well even when existing soil samples are limited.

Original version of iPSM (unweighted samples): An overview of the iPSM method is provided as follows. Interested readers may refer to (Zhu et al. 2015a) for full details of the method. The iPSM method includes two main steps. The first step is to calculate environmental similarity. Environmental similarity between an un-sampled location ***j*** and sample location ***i*** is first evaluated at the individual environmental variable (i.e., principal component) level, and then similarities based on all environmental variables are integrated to represent the overall similarity between un-sampled location ***j*** and sample location ***i***.

The environmental similarity between un-sampled location ***j*** and sample location ***i*** w.r.t. the l^{th} principal component, $S_{i,j}^{l}$, is calculate using Equation 4.1:

$$S_{i,j}^{l} = \exp\left[-\frac{\left(V_i^l - V_j^l\right)^2}{2 \times \left(\frac{SD^l}{SD_j^l} \times SD^l\right)^2}\right]$$

Equation 4.1

in which, V_i^l and V_j^l are the value of the l^{th} principal component at sample location ***i*** and un-sampled location ***j*** respectively. SD^l is the standard deviation of the l^{th} principal component. SD_j^l is the standard deviation of the l^{th} principal component from V_j^l (instead of the mean of l^{th} principal component) (Equation 4.2):

$$SD_j^l = \sqrt{\frac{\sum_{p=1}^{m}\left(V_p^l - V_j^l\right)^2}{m}}$$

Equation 4.2

where V_p^l is the value of the l^{th} principal component at sample location *p*. *m* is the total number of un-sampled locations (pixels) in the study area.

The overall environmental similarity between un-sampled location ***j*** and sample location ***i*** considering all *L* selected principal components, $S_{i,j}$, is then determined following a limiting factor approach based on the assumption that the least favorite environmental condition determines soil formation and thus soil property at a given location. A minimum operator was applied to take the minimum of the environmental similarities to individual principal components (i.e., $S_{i,j}^1, S_{i,j}^2, \dots, S_{i,j}^L$) as the overall environmental similarity $S_{i,j}$ (Zhu & Band 1994) (Equation 4.3):

$$S_{i,j} = \min\left(S_{i,j}^1, S_{i,j}^2, \dots, S_{i,j}^L\right)$$

Equation 4.3

The environmental similarity between un-sampled location ***j*** and each of the *n* sample locations can be computed following Equation 4.1 through Equation 4.3.

The second step of iPSM is to compute soil property value at un-sampled location ***j*** based on its environmental similarities to the *n* sample locations. A weighted average approach is adopted for this purpose (Equation 4.4):

$$\hat{T}_j = \frac{\sum_{i=1}^{n} S_{i,j} \times T_i}{\sum_{i=1}^{n} S_{i,j}}$$

Equation 4.4

where $\hat{T}_j$ is the predicted value of the soil property (i.e., A-horizon SOM content) at un-sampled location $\boldsymbol{j}$. T_i is the observed value of the soil property at sample location $\boldsymbol{i}$. Predicting SOM content at every location (pixel) in the study area using unweighted soil samples following Equation 4.4 results in a SOM content map.

Revised version of iPSM (weighted samples): In the original iPSM method, soil samples are not weighted by sample weights. In the revised version of iPSM, however, soil samples are weighted with the optimal sample weights determined through the representativeness directed spatial bias mitigation approach (Section 2.3.2), following Equation 4.5.

$$\hat{T}_j = \frac{\sum_{i=1}^{n} w_i \times S_{i,j} \times T_i}{\sum_{i=1}^{n} w_i \times S_{i,j}}$$

Equation 4.5

in which w_i is the weight of the soil sample at location $\boldsymbol{i}$ determined from the proposed method (other notations are the same as in Equation 4.4). Everything else being equal, soil property values at sample locations associated with larger weights have larger contributions to the estimated soil property value at an unvisited location. Predicting SOM content at every location (pixel) in the study area using soil samples weighted by the weights determined from the proposed method following Equation 4.5 results in a SOM content map.

4.2.2.2 Multiple linear regression (MLR)

Background on MLR: Multiple linear regression (MLR) is a general method for modeling multivariate linear relationships between a dependent variable and independent variables.

Unlike iPSM, MLR has stricter requirements on sample size and the representativeness of the sample set for building a statistically robust MLR model.

Original version of MLR (unweighted samples): A MLR model takes the form of Equation 4.6:

$$\hat{T}_j = \beta^0 + \sum\nolimits_{l=1}^{L} \beta^l \times V_j^l$$

Equation 4.6

where $\hat{T}_j$ is the predicted value of the soil property (i.e., *A*-horizon SOM content) at un-sampled location $\boldsymbol{j}$, V_j^l is the value of the l^{th} principal component at location $\boldsymbol{j}$, β^0 is the intercept, and β^l is the coefficient for the l^{th} principal component.

The intercept β^0 and coefficients β^l's are determined by fitting the model on training data (i.e., soil property values and values of the principal component at the *n* sample locations) based on the ordinary least squares (OLS) criterion, i.e., finding the values of β^0 and β^l's that minimize the sum of squared residuals between predicted soil property values and observe soil property values at the *n* sample locations (Equation 4.7).

$$\beta^0, \beta^1, \dots, \beta^L = argmin_{\beta^0, \beta^1, \dots, \beta^L} \left\{ \sum\nolimits_{i=1}^{n} \left[T_i - \left(\beta^0 + \sum\nolimits_{l=1}^{L} \beta^l \times V_i^l \right) \right]^2 \right\}$$

Equation 4.7

The OLS procedures implemented in the Scikit-learn package (Pedregosa et al. 2012) were adopted to train MLR models. A MLR model was trained based on unweighted soil samples. The model was then applied to every location (pixel) in the study area to produce a SOM content map.

Revised version of MLR (weighted samples): In the original MLR method, soil samples are not weighted by sample weights. In the revised version of MLR, soil samples are weighted with the optimal sample weights determined from the proposed method. Specifically, sample weights are used to weight individual squared residuals in determining the MLR model parameters using OLS. Samples with larger weights are treated more important in this model fitting process (Equation 4.8).

$$\beta^0, \beta^1, \dots, \beta^L = argmin_{\beta^0, \beta^1, \dots, \beta^L} \left\{ \sum_{i=1}^{n} w_i \left[T_i - \left(\beta^0 + \sum_{l=1}^{L} \beta^l \times V_i^l \right) \right]^2 \right\}$$

Equation 4.8

The OLS procedures implemented in the Scikit-learn package (Pedregosa et al. 2012), capable of accounting for sample weights, were used to train MLR models. A MLR model was trained based on soil samples weighted by the weights determined from the proposed method. The model was then applied to every location (pixel) in the study area to predict a SOM content map.

4.2.3 Evaluation

4.2.3.1 Validation data

44 samples were collected through systematic sampling on a 1100 m × 740 m grid in the study area (Figure 4.5) (Zhu et al. 2010). The SOM content (%) in *A*-horizon soil was measured for each of the soil samples. The SOM content values at the 44 sample locations

have a mean of 4.348 and a standard deviation of 0.982. These 44 systematic soil samples are representative of the study area and thus were used as validation samples to evaluate the accuracy of the predicted SOM content maps.

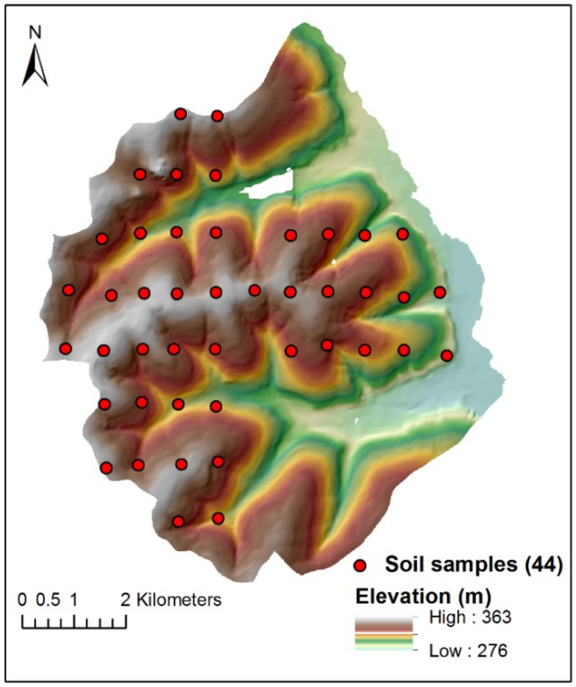

Figure 4.5. Validation soil samples in Heshan study area.

4.2.3.2 Evaluation metrics

Root mean square error (RMSE) and mean absolute error (MAE) were adopted as evaluation metrics to measure the accuracy of predicted SOM content maps. RMSE and MAE are computed based on the predicted and observed SOM content values at the 44 validation soil sample locations following Equation 4.9 and Equation 4.10:

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(T_i - \hat{T}_i\right)^2}$$

Equation 4.9

and

$$MAE = \frac{1}{k}\sum_{i=1}^{k}\left|T_i - \hat{T}_i\right|$$

Equation 4.10

where k is the number of validation soil samples (k = 44), and $\hat{T}_i$ and T_i are the predicted SOM content and observed SOM content at validation sample location ***i***, respectively.

Both RMSE and MAE express average model prediction error in units of the predicted variable of interest (i.e., SOM content). They are indifferent to the direction of errors and are negatively-oriented scores, which means lower values indicate higher prediction accuracy. They differ in that the RMSE gives a relatively high weight to large prediction errors since the errors are squared before they are averaged. This means the RMSE is more indicative of large prediction errors, which makes the RMSE more useful for measuring the accuracy of the predicted SOM content maps where large prediction errors are undesirable (Chai & Draxler 2014).

4.2.4 Experiment design

Experiments were designed to investigate three aspects of the representativeness directed spatial bias mitigation approach: its effectiveness in improving the accuracy of SOM content mapping, its sensitivity to parameter settings, and the impact of sample size.

4.2.4.1 Effectiveness of the approach

4.2.4.1.1 Effectiveness under different sample availability conditions

Effectiveness of the proposed approach was examined under two scenarios: the limited samples scenario and the all available samples scenario. Evaluation of the approach under these two scenarios allows investigating the effectiveness of the approach under different sample availability conditions.

For the *limited samples scenario*, 10 soil samples on a transect line (Figure 4.6) were selected as training samples to map SOM content. This sample set is of limited representativeness for the study area because it has a small sample size and very limited coverage in the geographic space and environmental covariates space (the 10 samples on the hill-slope transect were not representative of the floodplain areas). It was also used in Zhu et al. (2015b) to demonstrate the strength of the iPSM method facing limited samples.

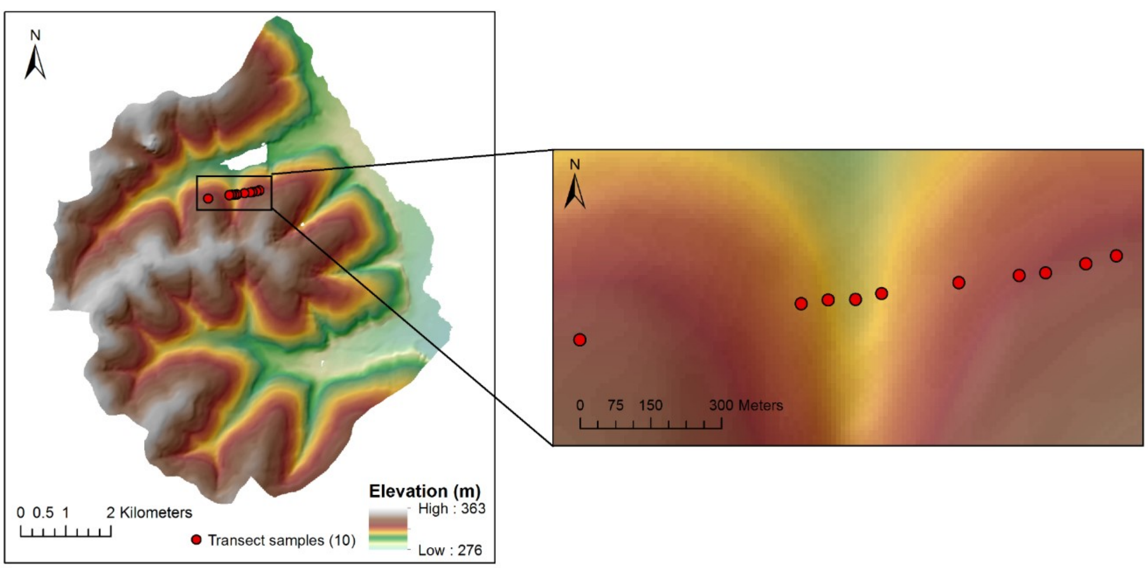

Figure 4.6. Soil samples on a transect line (10 samples) in Heshan study area.

For the *all available samples scenario*, the 59 soil samples (Figure 4.2) were used as training samples to map SOM content. The 59 samples spread across the study area including the floodplains and could represent the area much better than the 10 samples.

For both scenarios, prediction accuracies (i.e., RMSE, MAE) of the SOM content maps were evaluated based on the 44 validation samples. Accuracies of SOM content maps predicted using unweighted training samples were compared to those predicted using training samples weighted by the optimal weights.

4.2.4.1.2 Statistical significance tests

Two tests were performed to examine the statistical significance of the effects of the optimal sample weights determined from the proposed approach. *First*, prediction accuracy achieved under the optimal sample weights was compared to prediction accuracy achieved under randomly assigned weights. For this purpose, 100 sets of random sample weights uniformly distributed over the range [1, W_{max}] were generated. Training samples weighted by each set of the random weights were used to map SOM content and the prediction accuracy was computed. *One sample t-test* was then applied to test if the accuracy achieved under the optimal weights is significantly higher than under random sample weights.

Second, prediction accuracy achieved under the optimal weights was compared to prediction accuracy achieved under shuffled optimal weights. 100 sets of weights were generated by randomly shuffling the optimal weights (i.e., changing the order of the weights). Training samples weighted by each set of the shuffled weights were used to map SOM content and prediction accuracy was computed. *One sample t-test* was then applied to test if the accuracy achieved under the optimal weights is significantly higher than under randomly shuffled optimal weights.

4.2.4.1.3 Representativeness vs. prediction accuracy

The relationship between prediction accuracy and sample representativeness over the generations of the genetic algorithm was also examined to assess the effectiveness of sample representativeness to indicate prediction uncertainty. The optimal sample weights are final results returned by the genetic algorithm after reaching a prescribed number of generations. Sample weights (and thus sample representativeness) evolve gradually over the generations of the genetic algorithm. The best sample weights at each generation were used to weight the training samples to compute sample representativeness and to map SOM content. By examining the relationship between prediction accuracy and sample representativeness, the effectiveness of sample representativeness to indicate prediction uncertainty can then be evaluated.

4.2.4.2 Sensitivity to parameter settings

Several parameters including the population size (i.e., number of weights sets) in the genetic algorithm and the maximum sample weight (i.e.., W_{max}) may affect performance of the proposed approach. To examine the impacts of these parameters, the approach was run on the 59 training samples with varying population sizes (population size = 50, 100, 200, 300, 500) and varying maximum sample weight W_{max} (W_{max} = 5, 10, 20, 50, 100). The accuracies of the SOM content maps predicted under different parameters settings were evaluated based the 44 validation samples and compared.

4.2.4.3 Impact of sample size

To investigate the impact of sample size on the performance of the approach, the approach was applied on training samples of varying sample sizes (sample size = 10, 20, 30, 40, 50). Samples were subjectively selected from the 59 soil samples in a way such that the sample sets maintain certain characteristics of spatial bias. For each of the sample sizes, one sample set was subjectively selected from the 59 samples (Figure 4.7). The procedures for selecting these subjective sample sets are as follows. A sample location on the flood plain in the south part of the study area was chosen as the seed sample (highlighted on the first map in Figure 4.7). Then samples were drawn randomly from the 59 training samples (without replacement) at selection probabilities being inversely related to their distances to the seed sample (a sample closer to the seed sample have a higher probability of being selected) (Equation 4.11):

$$p_i \propto \frac{1}{d_i^q}$$

Equation 4.11

in which p_i is the selection probability of sample at location i, d_i is the distance between the sample at location i to the seed sample, and q is a distance decay factor determining how quickly selection probability (p_i) decreases with increasing distance (d_i). After experimenting with a few q values, q was set to 4.0 in this study as it generally results in selected samples of desired spatial distribution patterns. For each sample size, multiple sets of samples were generated following the above procedures. Spatial distribution patterns of the sample sets were then visually examined. One sample set was subjectively chosen following the principle that the samples should have a relatively wide spatial coverage while concentrating in some areas more than others. For example, the selected sample set of size 10 spreads across most part of the study area but most of the samples are on the floodplain and foot-slopes. Similarly, the sample set of size 20 has a wide spread but most samples are clustered on the foot-slopes and hill-slopes in the south part of the study area.

The proposed approach was applied to find the optimal sample weights for each subjective sample set. Accuracies of SOM content maps predicted using unweighted subjective samples were compared to those predicted using subjective samples weighted by the optimal weights.

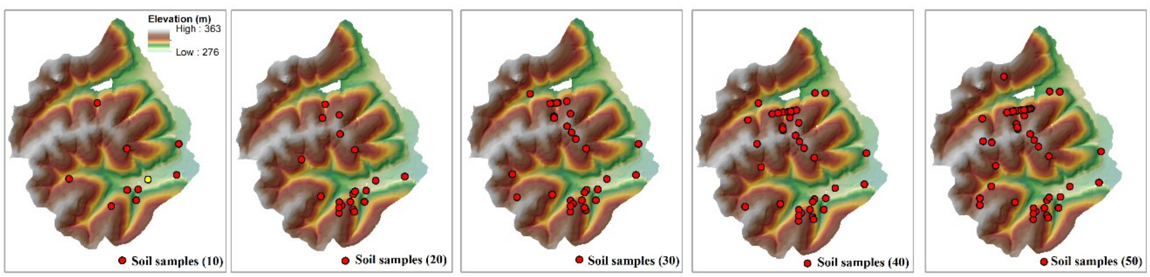

Figure 4.7. Sample sets of varying sizes subjectively selected from the 59 soil samples in Heshan study area.

4.3 Results

4.3.1 Effectiveness of the approach

4.3.1.1 Limited samples scenario

For the limited samples scenario, the optimal weights returned by the genetic algorithm are mapped in Figure 4.8. Samples close to each other tend to get smaller weights than samples that are relatively distant from others. With the optimal weights, the overall representativeness of the samples increases from 0.883 to 0.909. The SOM content maps predicted using iPSM based on unweighted samples and using samples weighted by the optimal weights are shown in Figure 4.9. SOM content maps predicted using the MLR method were not shown because the MLR model built from this limited sample set is not meaningful, e.g., predicting negative SOM content values (Zhu et al. 2015a).

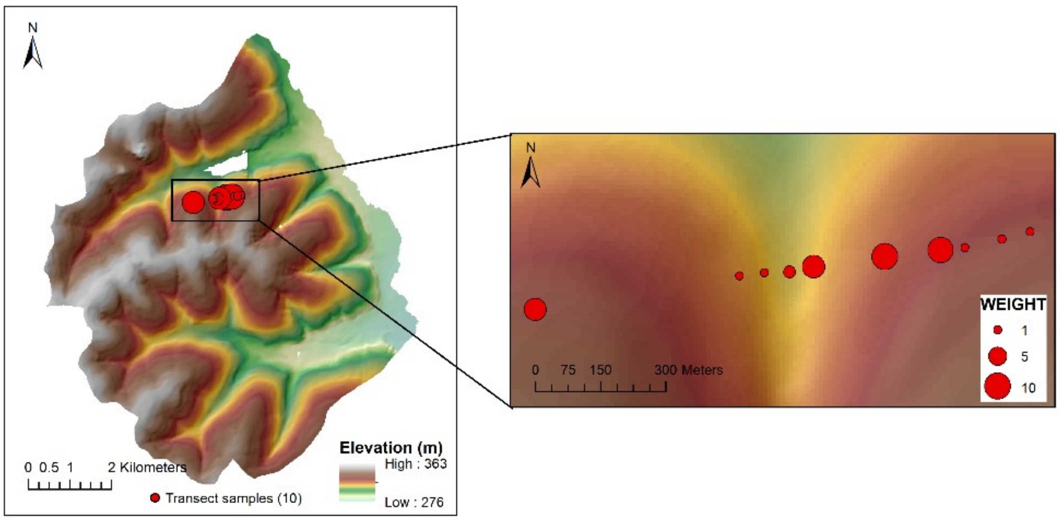

Figure 4.8. Optimal weights of the 10 transect samples returned by the genetic algorithm.

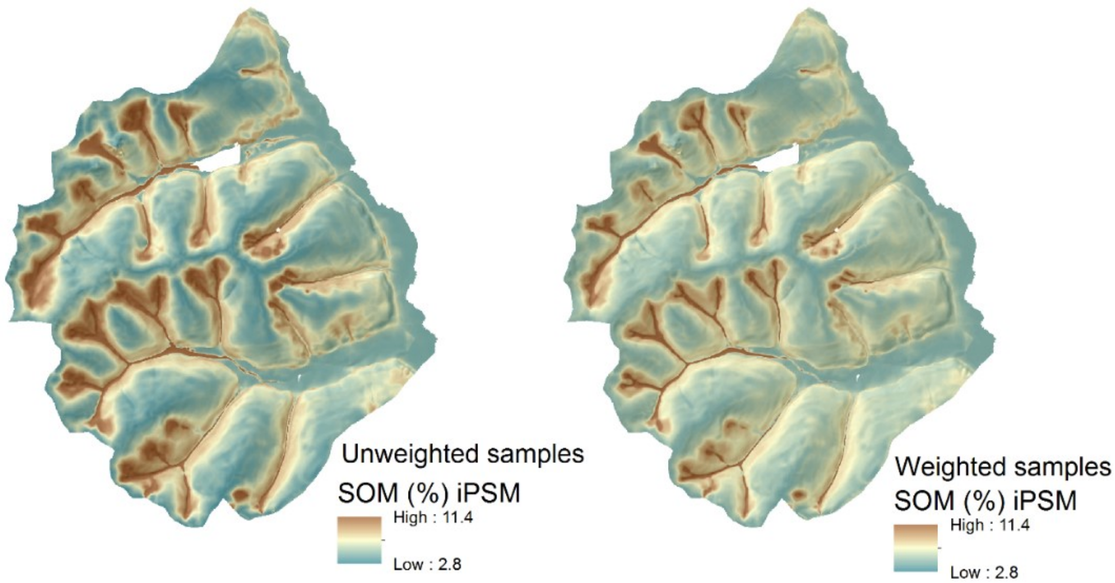

Figure 4.9. SOM content maps predicted using the 10 transect samples.

The general spatial pattern of the two SOM content maps are similar. Areas predicted with high SOM content are lower-to-toe slopes and areas predicted with low SOM content are upper-to-middle slopes. This matches our understanding of how the terrain influence SOM content in the study area. On lower-to-toe slopes, gentle depositional processes tend to be the dominant processes that usually lead to higher *A*-horizon SOM content. On upper-to-middle slopes, erosive processes tend to be the dominant processes and that reduce the A-horizon SOM content. The floodplain areas were predicted to be of low SOM content, which contradicts our understanding that these areas in fact have high SOM content. This is because the 10 samples on the hill-slope transect were not representative of the floodplain area. The range of SOM content predicted using weighted samples is narrower than that using unweighted samples.

Validation of SOM maps predicted using the 10 transect samples based on the 44 validation samples (Table 4.1) reveals that accuracy of the SOM map predicted using iPSM based on weighted samples is higher than that predicted using unweighted samples: RMSE is 21.2% lower and MAE is 16.4% lower when predicting SOM content using weighted samples. Accuracies of SOM content maps predicted using the MLR method were very low (i.e., very large RMSE and MAE) because the MLR model built from this limited sample set is not meaningful, e.g., predicting negative SOM content values. Nevertheless, there was still a 2% accuracy improvement by weighting the samples.

Table 4.1. Accuracy of SOM maps predicted using unweighted samples and samples weighted with the optimal weights (limited samples scenario).

Method		*Unweighted samples*	*Weighted samples*	*Accuracy Improvement*
iPSM	RMSE	1.450	1.143	21.2%
	MAE	1.019	0.852	16.4%
MLR	RMSE	8.55	8.38	2.0%
	MAE	7.34	7.19	2.0%

4.3.1.2 All available samples scenario

For the all available samples scenario, Figure 4.10 shows the evolution of sample weights and sample representativeness over the generations of the genetic algorithm. Figure 4.11 shows how the optimal sample weights improves sample representativeness over each of the three principal components. The optimal weights of the 59 samples returned by the genetic algorithm (after 200 generations) are mapped in Figure 4.12. Spatially clustered samples tend to get smaller weights than sparsely distributed samples. With the optimal weights, the overall representativeness of the samples increases from 0.906 to 0.964.

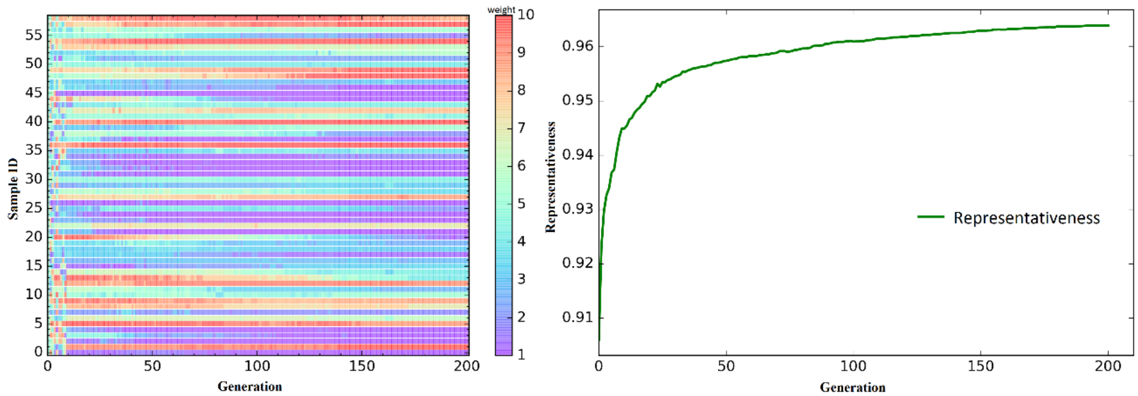

Figure 4.10. Evolution of sample weights (left) and samples representativeness (right) over the generations of the genetic algorithm.

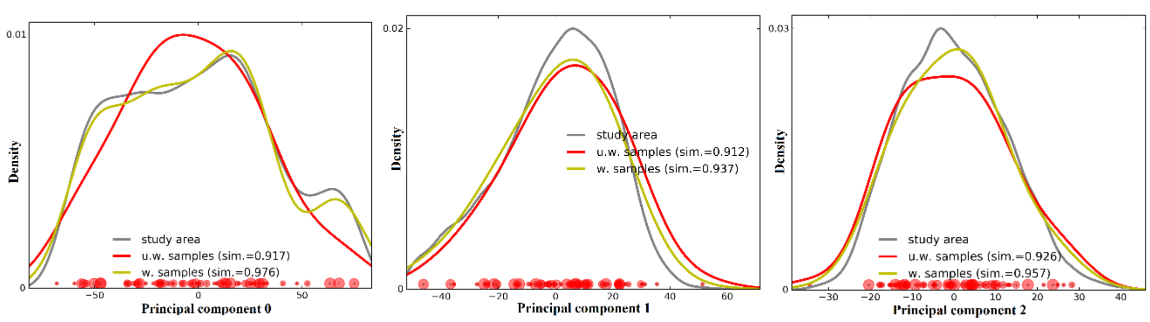

Figure 4.11. Probability density distributions estimated based on the 59 samples over the three principal components.

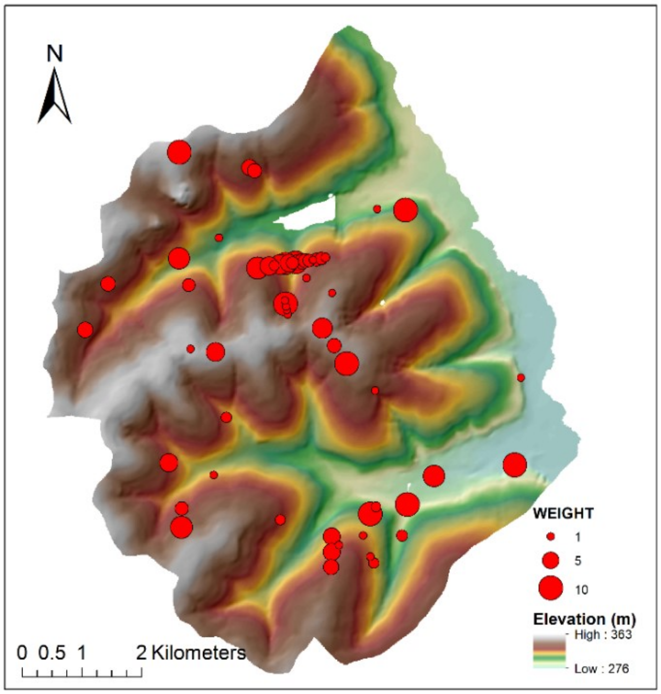

Figure 4.12. Optimal weights of the 59 samples returned by the genetic algorithm.

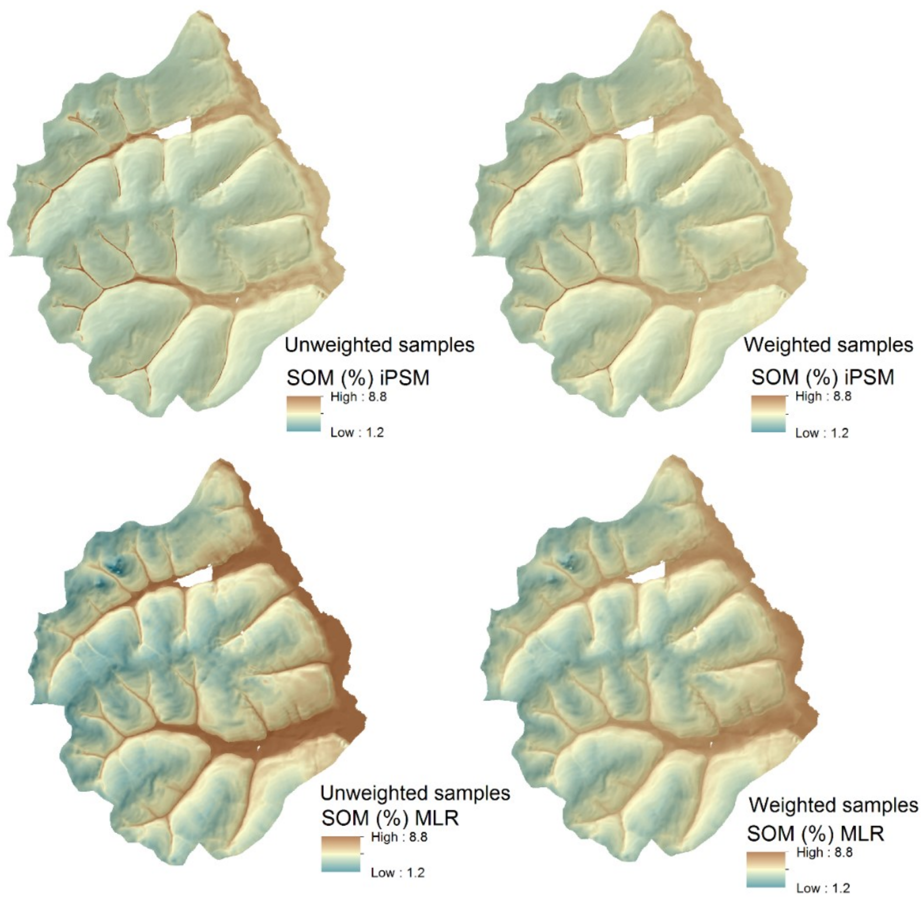

Figure 4.13. SOM content maps predicted using the 59 samples.

The SOM content maps predicted using unweighted samples and using samples weighted by the optimal weights are shown in Figure 4.13. The 59 samples spread across the study area (including the floodplains) and could represent the area much better than the 10 samples. The general spatial pattern of the SOM content maps predicted using the 59 samples are similar. Lower-to-toe slopes and floodplain areas were predicted to have high SOM content and upper-to-middle slopes were predicted to have low SOM content. This agrees with our understanding of the study area. SOM content maps predicted with the iPSM method tend to have fewer variations and smaller numerical ranges than those predicted with the MLR method.

Validation of the predicted SOM maps (Table 4.2) reveals that accuracies of the SOM maps predicted using weighted samples are higher than those predicted using unweighted samples for both iPSM and MLR. iPSM achieved an accuracy improvement of 5.1% in terms of RMSE and 1.7% in terms of MAE. MLR achieved an accuracy improvement of 13.9% in terms of RMSE and 11.3% in terms of MAE. SOM content maps predicted with iPSM were of higher accuracy than MLR. Yet MLR achieved larger accuracy improvements than iPSM. iPSM did not achieve as much accuracy improvements because the accuracies of SOM maps predicted with iPSM using unweighted samples were very high already (RMSE=1.035, MAE = 0.682).

Table 4.2. Accuracy of SOM maps predicted using unweighted samples and samples weighted with the optimal weights (all available samples scenario).

Method		*Unweighted samples*	*Weighted samples*	*Accuracy Improvement*
iPSM	RMSE	1.035	0.982	5.1%
	MAE	0.682	0.670	1.7%
MLR	RMSE	1.198	1.032	13.9%
	MAE	0.799	0.709	11.3%

4.3.1.3 Statistical significance tests

Results of statistical significance tests on the performance of the proposed approach (i.e., prediction accuracy improvement with optimal weights) are shown in Table 4.3. For the two scenarios, accuracies of the SOM content maps predicted under the optimal sample weights returned by the proposed approach are significantly higher than either the accuracies achieved under the randomly assigned sample weights or under the shuffled optimal sample weights. This observation is consistent for iPSM and MLR in terms of both RMSE and MAE measures. It indicates that the prediction accuracy improvements achieved by the proposed representativeness directed approach are statistically significant.

Table 4.3. Statistical significance tests on the accuracy of SOM maps predicted using samples weighted with the optimal weights.

Scenario			*Optimal weights*	*Random weights*				*Shuffled optimal weights*			
				Mean	*Std*	*t*	*p*	*Mean*	*Std*	*t*	*p*
Limited samples (10)	iPSM	RMSE	1.143	1.439	0.172	17.047	0.000	1.460	0.323	9.729	0.000
		MAE	0.852	1.019	0.098	16.856	0.000	1.036	0.189	9.676	0.000
All available samples (59)	iPSM	RMSE	0.982	1.045	0.031	20.002	0.000	1.051	0.058	11.808	0.000
		MAE	0.670	0.692	0.021	10.512	0.000	0.701	0.031	9.655	0.000
	MLR	RMSE	1.032	1.216	0.081	22.609	0.000	1.243	0.116	18.085	0.000
		MAE	0.709	0.820	0.048	22.980	0.000	0.839	0.077	16.740	0.000

Notes: The Mean and Std (standard deviation) values were computed based on 100 simulations. t and p are the one sample t-test statistics and p value for the null hypothesis that the RMSE or MAE achieved under the optimal weights is lower than the mean RMSE or MAE achieved under random weights or shuffled optimal weights.

4.3.1.4 Representativeness vs. prediction accuracy

Figure 4.14 shows how accuracies of the SOM content maps predicted using the 59 samples (weighted by the best sample weights at each generation) changed over the generations of the genetic algorithm. RMSE and MAE fluctuate dramatically at the early generations of the genetic algorithm but tend to decrease and stabilize towards the later generations of the genetic algorithm. Accuracies of the predicted SOM content maps generally improved over the generations of the genetic algorithm.

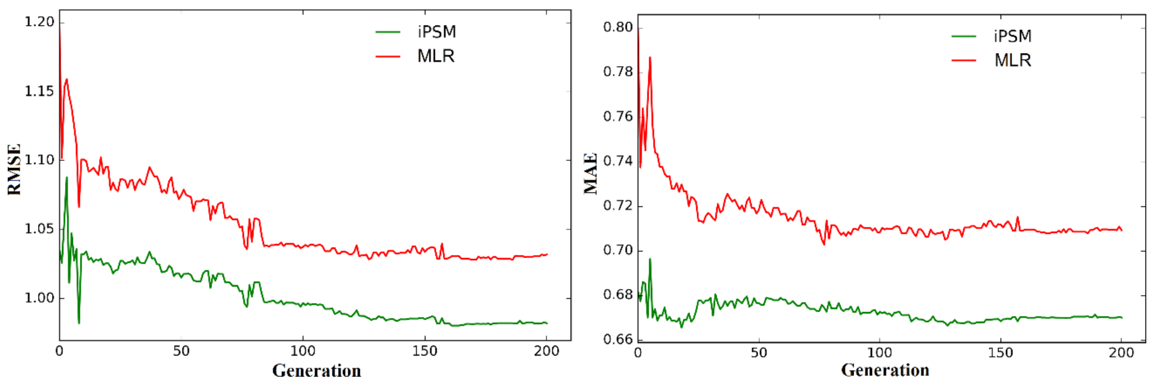

Figure 4.14. Evolution of the accuracies of SOM content maps predicted using the 59 samples over the generations of the genetic algorithm.

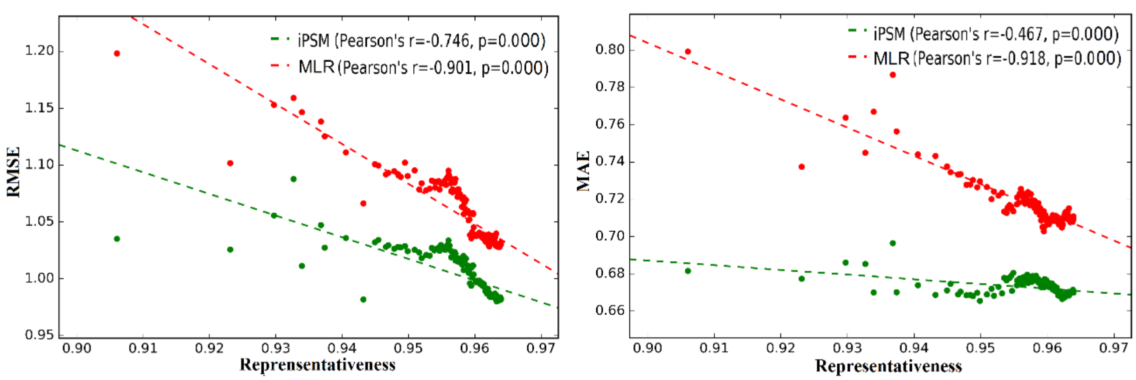

Figure 4.15. Relationship between sample representativeness and prediction accuracy over the generations of the genetic algorithm.

Scatter plots of accuracy measures against sample representativeness over the generations of the genetic algorithm are shown in Figure 4.15. There were strong negative linear relationships between the accuracy measures (i.e., RMSE and MAE) and sample representativeness, as revealed by the high negative *Pearson's* correlation coefficients between RMSE and sample representativeness, and between MAE and sample representativeness. It suggests that sample representativeness can effectively indicate prediction uncertainty. Samples of higher the representativeness would result in predicted SOM content maps of lower prediction uncertainty (i.e., lower RMSE and MAE; higher accuracy).

4.3.2 Sensitivity to parameter settings

4.3.2.1 Population size

The parameter population size, i.e., number of weight sets or "individuals" in the genetic algorithm, had an impact on the convergence behavior of the genetic algorithm as it affects how quickly the genetic algorithm converges to the optimal representativeness. It is worth noting that population size is different from sample size (i.e., number of gens in an individual; problem size), although a larger population size is often needed for optimization problems involving a larger number of samples. As revealed in Figure 4.16, the genetic algorithm converged to optimal sample weights corresponding to higher sample representativeness at a faster speed (generation-wise) under larger population sizes. The differences were small when population size was greater than or equal to 200. But it should be noted that running the genetic algorithm with larger populations size requires longer computing time because it takes longer to evaluate a larger number of individuals at each generation. Population size was set to 200 as default when running the genetic algorithm, as it appears to be a good balance between optimal representativeness and required computing time.

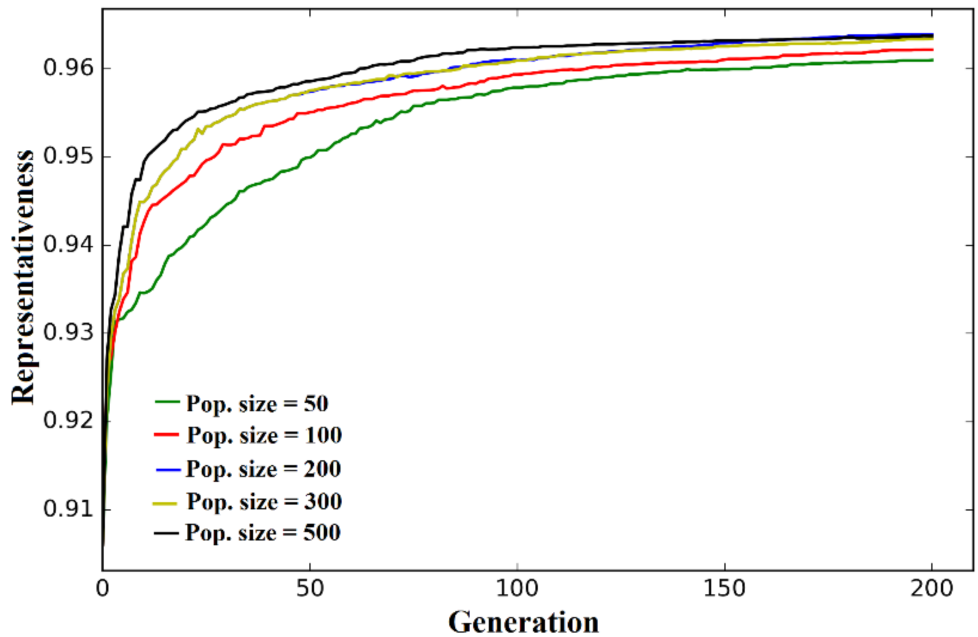

Figure 4.16. Impact of population size on the representativeness of the 59 samples over the generations of the genetic algorithm.

Population size also had an impact on accuracy of the SOM content maps predicted using samples weighted by the best sample weights over the generations of the genetic algorithm (Figure 4.17). The general trend was the larger population sizes (e.g., 200, 300, 500) resulted in SOM content maps of higher prediction accuracy (i.e., lower RMSE) than smaller population sizes (e.g., 50 ,100). When population size was set to 200 or 300, the proposed approach achieved good performance for both iPSM and MLR.

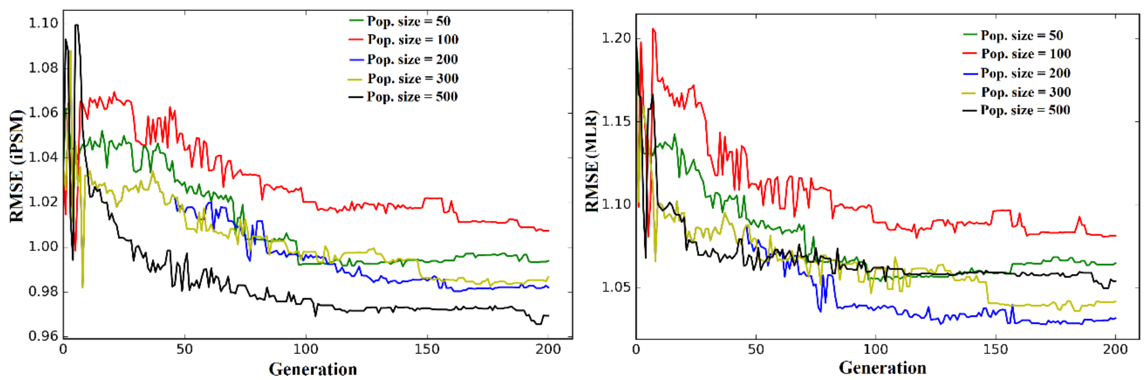

Figure 4.17. Impact of population size on the accuracy of SOM content maps predicted using the 59 samples.

The optimal sample weights for the 59 samples (returned after 200 generations) obtained under different population sizes were shown in Figure 4.18. The general pattern of optimal weight assignments across the samples under various population sizes was consistent, although the specific weight values may differ. For example, under different population sizes, sample 5, 11, and 23 were all assigned a large weight close to10, and sample 13, 14, and 15 were all assigned a small weight close to1.0. This consistent pattern of sample weights allocation was confirmed by the high correlation coefficients (*Spearman's r* > 0.5) between the optimal sample weights obtained under various population sizes (Table 4.4).

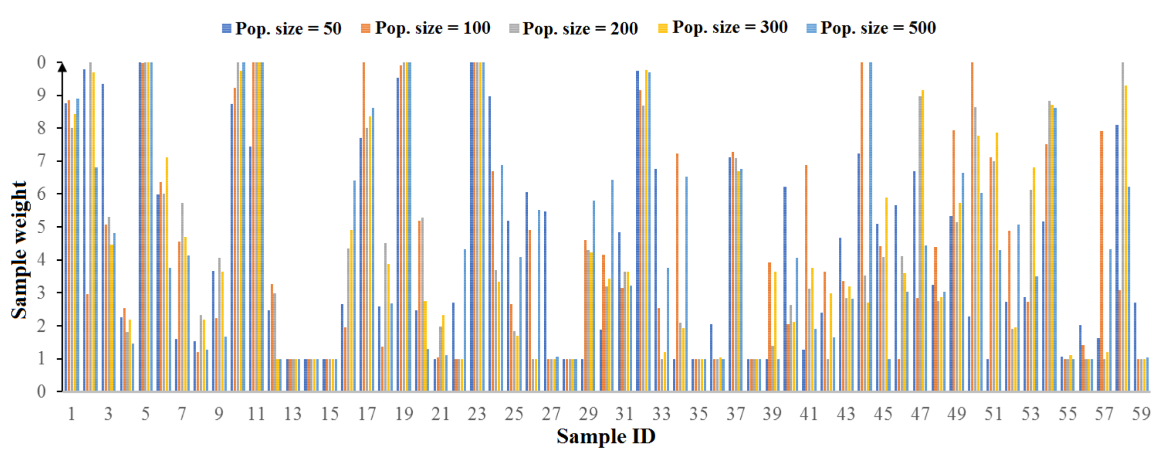

Figure 4.18. Optimal sample weights for the 59 samples obtained under different populations sizes.

Table 4.4. Spearman's correlation coefficients between optimal weights for the 59 samples obtained under different population sizes in the genetic algorithm.

Pop. size	*50*	*100*	*200*	*300*	*500*
50	1.000	0.502	0.625	0.597	0.694
100		1.000	0.676	0.690	0.779
200			1.000	0.947	0.727
300				1.000	0.711
500					1.000

4.3.2.2 Sample weight range

The upper limit of sample weight range W_{max} set had an impact on the returned optimal weights and thus the corresponding sample representativeness. As shown in Figure 4.19, setting W_{max} to 10 led to optimal weights corresponding to the highest sample representativeness, although the optimal representativeness values were close when setting W_{max} to 10 or 20. W_{max} was set to 10 as default when running the genetic algorithm.

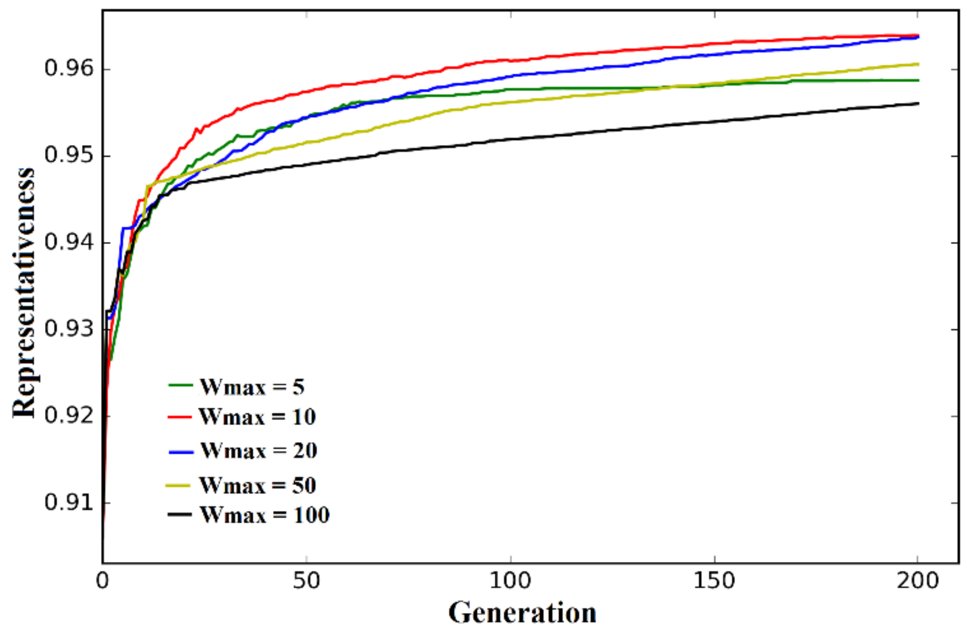

Figure 4.19. Impact of sample weight range on the representativeness of the 59 samples over the generations of the generations of the genetic algorithm.

Sample weight range also had an impact on the accuracy of the SOM content maps predicted using samples weighted by the best sample weights over the generations of the genetic algorithm (Figure 4.20). When W_{max} was set to 10 or 20, the proposed approach achieved good performance for both iPSM and MLR.

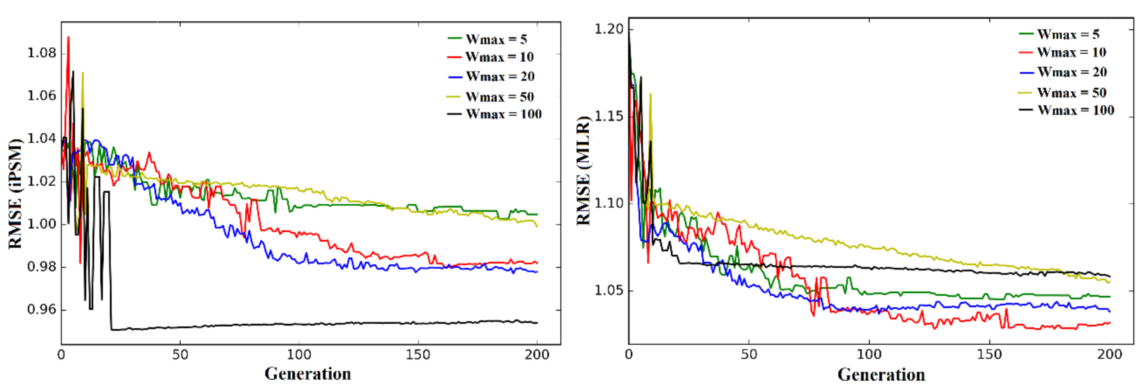

Figure 4.20. Impact of sample weight range on the accuracy of SOM content maps predicted using the 59 samples.

The optimal sample weights for the 59 samples obtained under weight ranges were shown in Figure 4.21. The general pattern of optimal weight assignments (relative importance) across the samples under various weight ranges was consistent. For example, under different weight ranges, sample 5, 19, and 23 were all assigned a large weight close to W_{max}, and sample 13, 35, and 36 were all assigned a small weight close to1.0. The correlation coefficients between the optimal sample weights obtained under various weight ranges were high (*Spearman's r* > 0.5) (Table 4.5), which also confirms this consistent pattern of sample weights allocation.

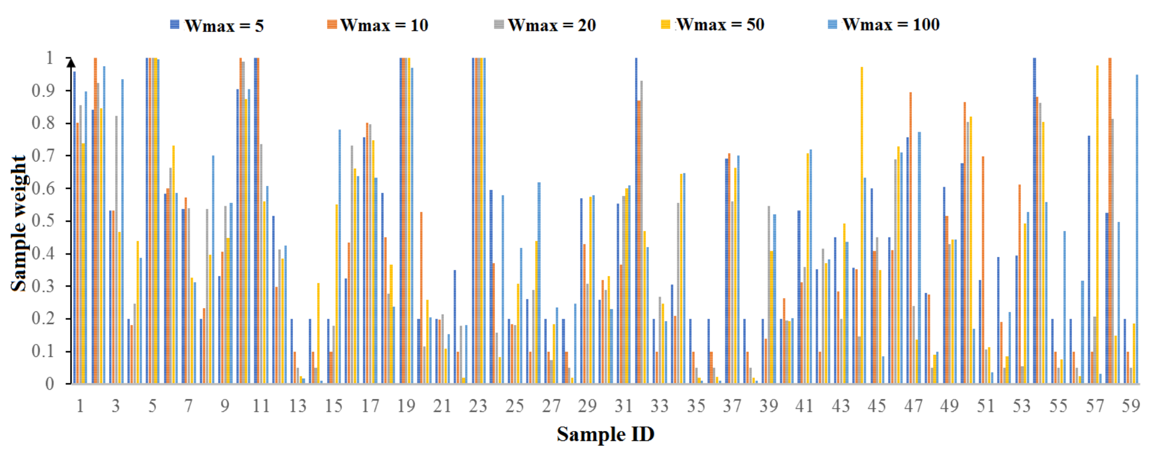

Figure 4.21. Optimal sample weights for the 59 samples obtained under different weight ranges (weights were standardized to [0, 1] for better visualization).

Table 4.5. Spearman's rank correlation coefficients between optimal weights for the 59 samples obtained under different sample weight ranges in the genetic algorithm.

W_{max}	*5*	*10*	*20*	*50*	*100*
5	1.000	0.801	0.734	0.691	0.495
10		1.000	0.752	0.596	0.513
20			1.000	0.765	0.633
50				1.000	0.653
100					1.000

4.3.3 Impact of sample size

Both representativeness of the unweighted samples and representativeness of the samples weighted with the optimal weights generally increases with increasing sample size (Table 4.6). The accuracies of SOM content maps predicted using the unweighted samples and samples weighted with the optimal weights were shown in Table 4.7. Prediction accuracy using unweighted samples and prediction accuracy using weighted samples increases with increasing sample size. Prediction accuracy using weighted samples is higher than prediction accuracy using unweighted samples across various sample sizes, which means weighting the samples with the optimal sample weights effectively improves prediction accuracy. The proposed approach achieved accuracy improvements between about 10 to 43% on smaller sample sizes (10, 20, and 30). The accuracy improvements were below 10% on larger sample sizes (40 and 50). The magnitude of prediction accuracy improvement achieved by the approach on the subjective samples decreases with increasing sample size.

Table 4.6. Representativeness of the subjective samples sets computed based on unweighted samples and samples weighted with the optimal weights.

Sample size	*Unweighted samples*	*Weighted samples*
10	0.696	0.864
20	0.884	0.908
30	0.878	0.919
40	0.899	0.958
50	0.903	0.960

Table 4.7. Accuracies of SOM content maps predicted using unweighted subjective samples and samples weighted with the optimal weights.

	Sample size	*RMSE*			*MAE*		
		Unweighted samples	*Weighted samples*	*Accuracy improvement*	*Unweighted samples*	*Weighted samples*	*Accuracy improvement*
iPSM	10	1.603	1.357	15.4%	1.031	0.891	13.6%
	20	1.461	1.113	23.8%	0.852	0.706	17.1%
	30	1.287	1.048	18.5%	0.803	0.727	9.4%
	40	1.151	1.039	9.7%	0.773	0.719	6.9%
	50	1.002	0.971	3.1%	0.675	0.655	2.8%
MLR	10	2.103	1.334	36.5%	1.730	0.990	42.8%
	20	1.465	1.075	26.7%	1.101	0.736	33.1%
	30	1.236	1.021	17.4%	0.861	0.676	21.4%

40	1.144	1.086	5.0%	0.833	0.781	6.3%
50	1.109	1.033	6.9%	0.750	0.705	5.9%

4.3.4 Spatial pattern of optimal sample weights

Figure 4.22 shows the optimal weights for the subjective samples returned by the genetic algorithm. Samples in areas that are (relatively) over-represented by the sample set tend to be allocated smaller weights than samples in areas that are (relatively) under-represented. For example, for the sample set of size 10, samples on the floodplain (majority of the sample set) have smaller weights compared to samples on hill-slopes. This spatial pattern of optimal sample weights is consistent with the spatial pattern of optimal sample weights in the limited samples scenario (Section 4.3.1.1) and the all available samples scenario (Section 4.3.1.2).

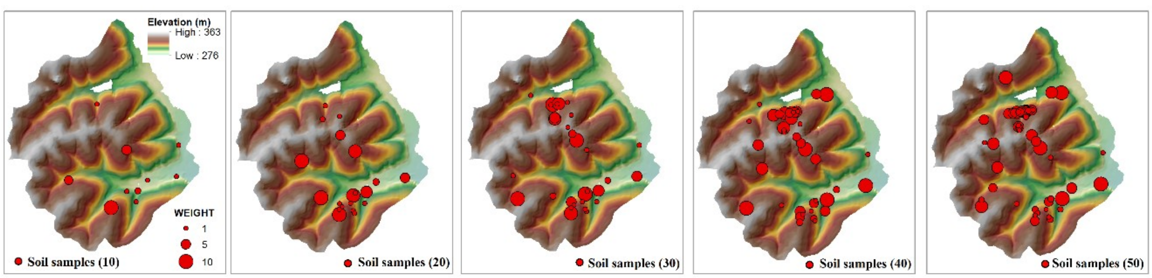

Figure 4.22. Optimal weights of the subjective samples returned by the genetic algorithm.

4.4 Discussion

4.4.1 Effectiveness of the approach

A biased sample distribution in covariates space is often a consequence of biased distribution of the samples in geographic space (spatial bias). Using representativeness of the sample set as a heuristic and criterion, the proposed approach mitigates spatial bias by weighting the samples such that the weighted sample distribution in covariate space better approximates the population distribution. This weighting scheme down-weights samples that over-represent the covariates space and up-weights samples that under-represent the covariates space. It is reflected in the geographic space that spatially clustered samples are assigned smaller weights and sparsely distributed samples are assigned larger weights.

Evaluation of the approach through predictive mapping experiments reveals that weighting samples with the optimal weights that maximize sample representativeness effectively improves prediction accuracy in both a limited samples scenario and an all available samples scenario. Noticeably, the approach brings a larger magnitude of prediction accuracy improvement for the limited samples scenario where sample representativeness is poor than the all available samples scenario where sample representativeness is already sufficient. This is expected given that there is more space for improvement on limited samples than samples whose representativeness is already good enough to achieve high prediction accuracy.

Effects of the optimal sample weights were compared to effects of randomly assigned sample weights to test statistical significance of the proposed approach. As the results have shown, the prediction accuracy achieved with the optimal weights is significantly higher than what is achievable with random weights. This indicates that the pattern of weight allocation as reflected in the optimal weights is meaningful, which in turn suggests that the representativeness heuristic is indeed effective in mitigating spatial bias and improving prediction accuracy.

Sample representativeness, which in this study is quantified as similarity between the sample distribution and the population distribution in covariates space, can effectively indicate prediction uncertainty. A negative relationship between sample representativeness and prediction error was observed in the evaluation experiments. Over the generation of the genetic algorithm, as sample representativeness evolves towards higher values, prediction error decreases. Additionally, it was observed that sample representativeness increases with increasing sample size whilst prediction error decreases.

Overall, evaluations of the proposed representativeness directed approach have demonstrated that the approach can effectively mitigate spatial bias in samples to improve predictive mapping accuracy. The quantified sample representativeness is an effective indicator of prediction uncertainty.

4.4.2 Parameter settings

Performance of the proposed approach might be sensitive to its parameters (e.g., population size and number of generations in the genetic algorithm, sample weight range). In general, *population size* in a genetic algorithm influences the convergence rate of the objective function (i.e., sample representativeness) and the quality of the solution (i.e., optimal sample weights). As a basic requirement, a large enough population size (relative to the problem size) is needed to find a good enough solution because large population size allows the initialization, selection, crossover and mutation operations to evaluate a larger variety of candidate solutions. Beyond certain population size threshold, the genetic algorithm is likely to converge to similar solutions. Larger population size also implies longer computing time. For experiments in this case study, it was found that the optimal sample weights obtained under various population sizes are positively correlated, suggesting that the weight allocation in different sets of optimal weights follows similar patterns. A population size of 200 was used in the experiments in this case study because it is large enough relative to the problem size (59 sample weights to determine) and is a good balance between prediction performance and computing time. For cases of larger problem size, a larger population size should be adopted.

The parameter *number of generations* has a similar impact. Going through a larger number of generations of the genetic algorithm tends to return a better solution (given a large enough population size) but also takes longer to compute. In general, number of generations should be determined along with other parameters such as population size and problem size. As a “rule-of-thumb”, the number of generations can be set to the generation at which the objective function no longer improves much. In the experiments, the genetic algorithm was terminated after going through 200 generations before returning the optimal weights. This number was determined based on the observation that the evolution of sample representativeness stalls and reaches a plateau in fewer than 200 generations. However, for cases of larger problem size, a larger number of generations along with a larger population size may be necessary. Alternatively, the genetic algorithm can also be terminated when the objective function (i.e., representativeness) exceeds a threshold (e.g., 0.90).

The value range of the sample weights [1.0, W_{max}] is a key parameter for the genetic algorithm. Here W_{max} is the maximum possible weight of a sample. The physical meaning of this parameter is that a sample with weight W_{max} will be treated W_{max} times as important as a sample with weight 1.0 in estimating the sample distribution and training predictive models. For experiments in this case study, it was found that the optimal sample weights obtained under various W_{max} settings are positively correlated, indicating that the order of sample importance in different sets of optimal weights follows similar patterns. W_{max} was subjectively set to 10.0 for experiments in this study (a sample can be at most 10 times as important as another sample). Weighting samples with the optimal weights obtained under this setting achieved satisfactory prediction accuracy improvement. In cases where data availability allows, W_{max} may be determined through data-driven procedures such as cross-validation, atop of taking its physical meaning into consideration.

4.4.3 Impact of sample size

Representativeness of the samples generally increases with increasing sample size, so does the accuracy of predictive mapping using either unweighted samples or samples weighted by the optimal weights.

The magnitude of prediction accuracy improvement achieved by weighting samples with the optimal weights decreases with increasing sample size. Recall that the subjective samples were determined in a way such that they maintain spatial bias of certain characteristics. At smaller sample sizes, spatial bias in the samples is severer and sample representativeness is poorer. Yet there is more space for the representativeness-directed approach to improve sample representativeness to achieve larger amount of prediction accuracy improvement.

4.5 Chapter conclusions

This chapter applies the representativeness directed spatial bias mitigation approach to a soil property mapping application. The approach was thoroughly evaluated through a case study of mapping *A*-horizon SOM content in Heshan study area. Using the 103 existing soil samples available in the study area (44 systematic samples as validation samples and the other 59 as training samples) and environmental covariates data, a series of experiments were designed to examine many aspects of the proposed approach regarding its effectiveness to improve prediction accuracy, sensitivity to parameter settings, and response to sample size.

Experiment results show that the representativeness directed approach can effectively mitigate spatial bias in samples to improve predictive mapping accuracy in both limited samples and all available samples scenarios. The effect of the optimal sample weights determined through the approach is statistically significant compared to randomly assigned weights. The weight allocation pattern as reflected in the optimal weights is meaningful in that it puts lower weights on samples that over-represent the covariates space and higher weights on samples that under-represent the covariates space. *Moreover*, the quantified sample representativeness is also an effective indicator of prediction uncertainty.

The sensitivity of the approach to parameter settings (e.g., population size and number of generations in the genetic algorithm, sample weight range) was investigated and suggestion on parameter setting was provided based on empirical evidence. Under different population size and number of generation settings the genetic algorithm converges to similar weight allocation patterns. A larger populations size and a larger number of generation usually lead to better solutions, but these parameters should be determined considering problem size and the balance between solution quality and computing time. The maximum possible weight of a sample W_{max} is a key parameter (a sample with weight W_{max} will be treated W_{max} times as important as a sample with weight 1.0). In cases where data availability allows, W_{max} may be determined through data-driven procedures such as cross-validation, atop of taking its physical meaning into consideration. For this study the best settings for these parameters are: Population size = 200, number of generations = 200, and W_{max} = 10.0.

The representativeness directed spatial bias mitigation approach effectively mitigates the adverse effects of spatial bias in samples and improves accuracy of predictive soil property mapping. This approach is useful for soil mapping using existing soil samples (potentially from multiple sources) or soil samples contributed by volunteers participating in citizen science projects, where in both cases, the soil samples are likely to suffer from spatial bias.

Chapter 5 Conclusions

5.1 Conclusions

This dissertation proposes a representativeness directed approach to spatial bias mitigation in VGI samples for predictive mapping. The *representativeness* of a set of samples is defined as the "goodness-of-coverage" of the samples in the covariates space, which in turn is quantified as the similarity between the probability density distribution of the samples in the covariates space (*sample distribution*) and the probability density distribution of all mapping units (raster cells) within the study area (*population distribution*). Sample representativeness is then used as a heuristic to mitigate the spatial bias in samples. Spatial bias mitigation is accomplished by reweighting samples towards increasing sample representativeness. Determination of the sample weights is conceived as an optimization problem. The optimal sample weights maximizing the sample representativeness are determined through an optimization procedure based on the genetic algorithm.

The effect of the sample representativeness directed reweighting is that samples which are over-representing their fair-share of environmental niche are weighted less than samples which are under-representing their fair-share of environmental niche. As manifested in the geographic space, densely distributed samples over similar geographic locations tend to get smaller weights than sparsely distributed sample locations. The reweighted samples along with the environmental covariates data are then used to train predictive models for mapping spatial variation of the target geographic phenomena.

The effectiveness of the proposed representativeness directed spatial bias mitigation approach is demonstrated through two predictive mapping applications: species habitat suitability mapping and soil property mapping. Experiment results in both applications show that the accuracy of predictive mapping using samples weighted with the optimal weights is higher than using the original unweighted samples. *Besides*, tests suggest that sample weight allocation in the optimal weights was meaningful. Accuracy of predictive mapping using samples weighted with the optimal weights is statistically significantly higher than using samples weighted with randomly assigned weights or randomly shuffled optimal weights. *In addition*, a positive relationship between sample representativeness and predictive mapping accuracy was observed, suggesting that the so-defined and quantified sample representativeness is an effective indictor of predictive mapping accuracy. *In conclusion*, the proposed approach can effectively mitigate spatial bias in VGI-based samples to improve predictive mapping accuracy.

Spatial bias is an issue not only for VGI-based samples. Samples from other sources may also be subject to spatial bias. For example, soil mapping using existing soil samples may suffer unsatisfactory mapping accuracy due to the spatial bias in soil samples obtained from multiple sources. The proposed approach is also applicable for mitigating spatial bias in non-VGI samples for predictive mapping. Besides, spatial bias is one type of sample selection bias. Beyond predictive mapping, the approach is of potential use for sample selection bias correction in many other domains (e.g., machine learning and data mining from biased samples).

The proposed approach has certain *advantages* compared to existing bias mitigation methods (as reviewed in Section 1.4). *First*, this approach requires only field samples and environmental covariates to mitigate spatial bias. It needs no information about the underlying sampling process. Such information is often lacking in practice, especially in

VGI genesis. This makes it more practically applicable than the method that mitigates sample selection bias by modelling the sampling process, and the method that compensates spatial bias by weighting samples with terrain cumulative visibility. *Second*, it can cope with the high dimensionality and multicollinearity issues that pose challenges to the importance weighting method. *Third*, besides its applicability to spatial bias mitigation in samples used to train a global predictive model, it can also be applied to mitigate spatial bias in samples used to train local predictive models. *Fourth*, compared to the method that mitigates spatial bias by filtering samples in the geographic space or covariate space, the proposed approach does not exclude useful information in the available samples, which is desirable for cases where samples are of limited availability or samples are a precious resource. *Lastly*, the factoring bias method is applicable only to species habitat mapping (or species distribution modelling) whereas the proposed approach is generally applicable to predictive mapping applications.

5.2 Future research

Current implementation of the proposed approach does not accommodate *categorical covariates*. Categorical covariates are important in many predictive mapping applications (e.g., parent material type for digital soil mapping). The principal component analysis adopted for reducing dimensionality and multicollinearity of the covariates is not very good at dealing with categorical variables. There are several options worth future exploration. *First*, categorical variables can be binary encoded and then input to PCA. *Second*, categorical variables can be used to stratify the mapping area and the proposed approach is then applied for predictive mapping in each strata using only numerical covariates. *Third*, autoencoder can be used as an alternative to PCA for deriving new features from the covariates to achieve dimensionality reduction. Autoencoder is an artificial neural network used for unsupervised learning of an efficient representation (encoding) for a set of data (covariates) (Holden et al. 2006). It better deals with categorical covariates data.

The proposed approach could be extended to *support local modeling methods* (e.g., regression using samples in a local geographic area), in addition to global modeling methods used in this dissertation (e.g., regression using samples in the whole study area). The localized version of the approach would be very useful for mapping geographic phenomena over large geographic areas where the covariation relationship between the target geographic phenomenon and its covariates is spatially varying (non-stationary) (Fotheringham et al. 2003). The localized version of the approach resembles the geographically weighted regression (Fotheringham et al. 2003) in that only samples within a local area around the prediction location are used to train the predictive model. But it differs from GWR in that the sample weights are determined through the representativeness directed procedures, not based on distances from the samples to the prediction location.

In this study, the approach was only used for regression methods that predict continuous target variables (e.g., SOM content, habitat suitability). It would be interesting to examine the applicability of the approach for predictive mapping applications involving classification, such as soil type prediction and mapping.

The proposed approach is computationally intensive and advanced high-performance computing technologies may be adopted to speed it up. The most computationally demanding part of the proposed approach is the optimization process for finding the optimal sample weights, especially when the sample size is large (e.g., on spatial big data).

For example, when running the approach on the 655 eBird checklist locations, it took about 2 days to return the optimal sample weights (population size = 500, number of generations = 500 for the genetic algorithm). Currently a genetic algorithm is used to find the optimal weights. When sample size is large, a large population (number of individuals) and a large number of generations for the genetic algorithm is desired in order to find good solutions (sample weights configurations). In such cases, the genetic algorithm demands more computing resource to return the solution within a reasonably short period of time. As a possible alternative, the more computationally efficient *black-box optimization* (*i.e.*, *derivative-free optimization*) tool (Rios & Sahinidis 2013) such as NOMAD (Le Digabel 2011) could be used to find the optimal weights.

The key step in the optimization (either a genetic algorithm or a black-box optimization) is evaluating the objective function (i.e., sample representativeness). Computing sample representativeness is also very computationally demanding when the sample size is large. It involves using KDE (kernel density estimation) to estimate the population distribution and the sample distribution on each covariate and then compute the similarity between the two distributions. In this process, estimating the sample distribution is expensive as a cross-validation procedure is adopted to find the optimal bandwidth for KDE. This procedure can be accelerated by conducting it in parallel using multiple CPU (central processing unit) cores or using the massively parallel computing recourses on GPUs (graphics processing units) (Zhang et al. 2016a, 2017).

With the above methodological extensions implemented and the computational issues resolved, the proposed representativeness directed spatial bias mitigation approach could be used to support predictive mapping application (i.e., regression and classification) over very large areas, using VGI samples, non-VGI samples, or a mixture of samples obtained from multiple sources.

Bibliography

Anadón, J.D., Giménez, A., Ballestar, R. & Pérez, I. (2009). Evaluation of local ecological knowledge as a method for collecting extensive data on animal abundance. *Conserv. Biol.*, 23, 617–625.

Anderson, D.R., Laake, J.L., Crain, B.R. & Burnham, K.P. (1979). Guidelines for line transect sampling of biological populations. *J. Wildl. Manage.*, 43, 70–78.

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecol. Modell.*, 200, 1–19.

Barry, S. & Elith, J. (2006). Error and uncertainty in habitat models. *J. Appl. Ecol.*, 43, 413–423.

Belbin, L. (1993). Environmental representativeness: Regional partitioning and reserve selection. *Biol. Conserv.*, 66, 223–230.

Bethlehem, J. (2010). Selection bias in web surveys. *Int. Stat. Rev.*, 78, 161–188.

Bethlehem, J. (2012). Using response probabilities for assessing representativity. *Stat. Netherlands, Discuss. Pap.*

Bimonte, S., Boucelma, O., Machabert, O. & Sellami, S. (2014). A new Spatial OLAP approach for the analysis of Volunteered Geographic Information. *Comput. Environ. Urban Syst.*, 48, 111–123.

Boakes, E.H., McGowan, P.J.K., Fuller, R. a., Ding, C., Clark, N.E., O’Connor, K., Mace, G.M., O'Connor, K. & Mace, G.M. (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol.*, 8, e1000385.

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. *Ecol. Modell.*, 275, 73–77.

Brunsdon, C. (1995). Estimating probability surfaces for geographical point data: An adaptive kernel algorithm. *Comput. Geosci.*, 21, 877–894.

Brus, D.J. & de Gruijter, J.J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80, 1–44.

Brus, D.J., Kempen, B. & Heuvelink, G.B.M. (2011). Sampling for validation of digital soil maps. *Eur. J. Soil Sci.*, 62, 394–407.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L. (2001). *Introduction to distance sampling estimating abundance of biological populations.* Oxford University Press, Oxford, United Kingdom.

Burton, A.C., Sam, M.K., Balangtaa, C. & Brashares, J.S. (2012). Hierarchical multi-species modeling of carnivore responses to hunting, habitat and prey in a West African protected area. *PLoS One*, 7, e38007.

Butcher, G.S., Robbins, C.S., Bystrak, D. & Geissler, P.H. (1986). *The breeding bird survey: its first fifteen years, 1965-1979. Condor*. DTIC Document.

Campbell, A.F. & Sussman, R.W. (1994). The value of radio tracking in the study of neotropical rain forest monkeys. *Am. J. Primatol.*, 32, 291–301.

Carré, F., McBratney, A.B. & Minasny, B. (2007). Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma*, 141, 1–14.

Chai, T. & Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.*, 7, 1247–1250.

Coleman, D.J., Georgiadou, Y. & Labonte, J. (2009). Volunteered Geographic Information: the nature and motivation of produsers. *Int. J. Spat. Data Infrastructures Res.*, 4, 332–358.

Cortes, C., Mohri, M., Riley, M. & Rostamizadeh, A. (2008). Sample selection bias correction theory. In: *Int. Conf. Algorithmic Learn. Theory 2008 Oct 13 (pp. 38-53). Springer, Berlin, Heidelberg.* Springer, pp. 38–53.

Danielsen, F., Mendoza, M.M., Alviola, P., Balete, D.S., Enghoff, M., Poulsen, M.K. & Jensen, A.E. (2003). Biodiversity monitoring in developing countries: what are we trying to achieve? *Oryx*, 37, 407–409.

Davis, L. (1991). Handbook of genetic algorithms.

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T. & Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Front. Ecol. Environ.*, 10, 291–297.

Le Digabel, S. (2011). Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. *ACM Trans. Math. Softw.*, 37, 44.

Dokuchayev, V. V. (1883). *Russkiy Chernozem (The Russian Chernozem)*. St. Petersburg.

Dudík, M., Phillips, S.J., Schapire, R.E., Dudik, M., Schapire, R.E., Phillips, S.J., Dudık, M., Schapire, R.E. & Phillips, S.J. (2005). Correcting sample selection bias in maximum entropy density estimation. *Adv. neural Inf. Process. Syst. 18*, 17, 323–330.

Elith, J. & Leathwick, J. (2006). Conservation prioritisation using species distribution modelling. In: *Spat. Conserv. prioritization Quant. methods Comput. tools* (eds. Moilanen, A., Wilson, K.A. & Possingham, H.P.). Oxford University Press, pp. 1–31.

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

Elwood, S. (2008b). Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, 72, 173–183.

Farrar, D.E. & Glauber, R.R. (1967). Multicollinearity in regression analysis: the problem revisited. *Rev. Econ. Stat.*, 92–107.

Ferrier, S., Watson, G., Pearce, J. & Drielsma, M. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodivers. Conserv.*, 11, 2275–2307.

Fielding, A.H. & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.*, 24, 38–49.

Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B., Munson, M.A., Hooker, G., Riedewald, M., Sheldon, D. & Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecol. Appl.*, 20, 2131–2147.

Fink, D., Theodoros Damoulas, Dave, J., Damoulas, T. & Dave, J. (2013). Adaptive Spatio-Temporal Exploratory Models: Hemisphere-wide species distributions from massively crowdsourced eBird data. In: *Twenty-Seventh AAAI Conf. Artif. Intell.* pp. 1284–1290.

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

Fonte, C.C., Bastin, L., See, L., Foody, G. & Lupia, F. (2015). Usability of VGI for validation of land cover maps. *Int. J. Geogr. Inf. Sci.*, 29, 1269–1291.

Foody, G.M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C. & Boyd, D.S. (2013). Assessing the accuracy of volunteered geographic information arising from multiple contributors to an Internet based collaborative project. *Trans. GIS*, 17, 847–860.

Foody, G.M., See, L., Fritz, S., van der Velde, M., Perger, C., Schill, C., Boyd, D.S. & Comber, A. (2014). Accurate attribute mapping from volunteered geographic information: Issues of volunteer quantity and quality. *Cartogr. J.*, 1743277413Y.0000000070.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, Limited.

Franklin, J. (1995). Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Prog. Phys. Geogr.*, 19, 474–499.

Franklin, J. & Miller, J.A. (2009). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge.

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

Gesch, D., Evans, G., Mauck, J., Hutchinson, J. & Carswell Jr, W.J. (2009). The national map: Elevation. *US Geol. Surv. fact sheet*, 3053.

Gillespie, T.W., Foody, G.M., Rocchini, D., Giorgi, a. P. & Saatchi, S. (2008). Measuring and modelling biodiversity from space. *Prog. Phys. Geogr.*, 32, 203–221.

Girres, J.-F. & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Trans. GIS*, 14, 435–459.

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

Goodchild, M.F. (2007b). Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *Int. J. Spat. Data Infrastructures Res.*, 2, 24–32.

Goodchild, M.F. & Glennon, J.A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *Int. J. Digit. Earth*, 3, 231–241.

Goodchild, M.F., Parks, B.O. & Steyaert, L.T. (1993). Environmental modeling with GIS.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Peterson, A.T., Loiselle, B.A. & Group, T.N.P.S.D.W. (2008). The influence of spatial errors in species occurrence data used in distribution models. *J. Appl. Ecol.*, 45, 239–247.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evol.*, 19, 497–503.

Graham, E.A., Henderson, S. & Schloss, A. (2011). Using mobile phones to engage citizen scientists in research. *Eos Trans. AGU*, 92.

Graham, M.H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, 84, 2809–2815.

Gregoire, T.G. & Valentine, H.T. (2007). *Sampling strategies for natural resources and the environment*. CRC Press.

De Gruijter, J., Brus, D.J., Bierkens, M.F.P. & Knotters, M. (2006). *Sampling for natural resource monitoring*. Springer Science & Business Media.

Gu, W.D. & Swihart, R.K. (2004). Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biol. Conserv.*, 116, 195–203.

Guisan, A., Edwards, T.C., Jr & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Modell.*, 157, 89–100.

Guisan, A. & Zimmerman, N.E. (2000). Predictive habitat distribution models in ecology. *Ecol. Modell.*, 135, 147–186.

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

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

Hargrove, W.W., Hoffman, F.M. & Law, B.E. (2003). New analysis reveals representativeness of the AmeriFlux network. *Trans. Am. Geophys. Union*, 84, 529.

Harvey, F. (2013). To volunteer or to contribute locational information? Towards truth in labeling for crowdsourced geographic information. In: *Crowdsourcing Geogr. Knowl.* Springer Netherlands, pp. 31–42.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econom. J. Econom. Soc.*, 153–161.

Hemson, G., Johnson, P., South, A., Kenward, R., Ripley, R., Macdonald, D. & Mcdonald, D. (2005). Are kernels the mustard? Data from global positioning system (GPS) collars suggests problems for kernel home-range analyses with least-squares cross-validation. *J. Anim. Ecol.*, 74, 455–463.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.*, 25, 1965–1978.

Hijmans, R.J., Garrett, K.A., Huamán, Z., Zhang, D.P., Schreuder, M. & Bonierbale, M. (2000). Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. *Conserv. Biol.*, 14, 1755–1765.

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, 83, 2027–2036.

Hirzel, A.H. & Lay, G. Le. (2008). Habitat suitability modelling and niche theory. *J. Appl. Ecol.*, 45, 1372–1381.

Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W.-K. & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends Ecol. Evol.*, 27, 130–137.

Holden, a J., Robbins, D.J., Stewart, W.J., Smith, D.R., Schultz, S., Wegener, M., Linden, S., Hormann, C., Enkrich, C., Wegener, M., Soukoulis, C.M., Linden, S., Schurig, D., Smith, D.R., Enkrich, C., Wegener, M., Soukoulis, C.M., Linden, S., Taylor, a J., Highstrete, C., Lee, M., Averitt, R.D., Schultz, S., Markos, P., Soukoulis, C.M., Mcpeake, D., Ramakrishna, S. a, Pendry, J.B., Shalaev, V.M., Maksimchuk, M., Umstadter, D., Chen, W., Shen, Y.R. & Moloney, J. V. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science (80-.).*, 313, 504–508.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, 117, 847–858.

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. *Appl. Geogr.*, 68, 37–47.

Ibáñez, Inés, Silander, J.A., Wilson, A.M., LaFleur, N., Tanaka, N. & Tsuyama, I. (2009). Multivariate forecasts of potential distributions of invasive plant species. *Ecol. Appl.*, 19, 359–375.

Indyk, P. & Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In: *Proc. thirtieth Annu. ACM Symp. Theory Comput.* ACM, pp. 604–613.

Isaaks, E.H. & Srivastava, R.M. (1989). An introduction to applied geostatistics.

Jenny, H. (1941). *Factors of soil formation: a system of quantitative pedology*. 1st edn. Dover Publication, New York.

Jensen, R.R. & Shumway, J.M. (2010). Sampling our world. In: *Res. Methods Geogr. A Crit. Introd.* (eds. Gomez, B. & Jones III, J.P.). John Wiley & Sons, pp. 77–90.

Jolliffe, I.T. (2002). Principal component analysis and factor analysis. *Princ. Compon. Anal.*, 150–166.

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

Kelling, S., Fink, D., Hochachka, W., Rosenberg, K., Cook, R., Damoulas, T., Silva, C. &

Michener, W. (2013a). Estimating species distributions-across space, through time, and with features of the environment. In: *DATA Bonanza Improv. Knowl. Discov. Sci. Eng. Bus.* (ed. M. Atkinson, R. Baxter, M. Galea, M. Parsons, P. Brezany, O. Corcho, J. van H. and D.S.). John Wiley & Sons, Inc., pp. 441–458.

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

Kelling, S., Lagoze, C., Wong, W.-K., Yu, J., Damoulas, T., Gerbracht, J., Fink, D. & Gomes, C. (2013b). eBird: A Human/Computer Learning Network to Improve Biodiversity Conservation and Research. *AI Mag.*, 34.

Kerr, J.T. & Ostrovsky, M. (2003). From space to species: ecological applications for remote sensing. *Trends Ecol. Evol.*, 18, 299–305.

Khalili, N., Wood, J. & Dykes, J. (2010). Analysing uncertainty in home location information in a large volunteered geographic information database. In: *Proc. GIS Res. UK 18th Annu. Conf. Univ. Coll. London, London, UK*. Citeseer, pp. 14–16.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H. & Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Divers. Distrib.*, 19, 1366–1379.

Kruskal, W. & Mosteller, F. (1979). Representative Sampling, III: The Current Statistical Literature. *Int. Stat. Rev.*, 47, 245–265.

Leitão, P.J., Moreira, F. & Osborne, P.E. (2011). Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. *Int. J. Geogr. Inf. Sci.*, 25, 439–454.

Li, J. & Hilbert, D.W. (2008). LIVES: a new habitat modelling technique for predicting the distribution of species' occurrences using presence-only data based on limiting factor theory. *Biodivers. Conserv.*, 17, 3079–3095.

Lindenmayer, D.B. & Likens, G.E. (2010). The science and application of ecological monitoring. *Biol. Conserv.*, 143, 1317–1328.

Liu, J. (2017). *Integration of samples from multiple sources for predictive mapping over large areas*.

Longley, P.A. & Adnan, M. (2016). Geo-temporal Twitter demographics. *Int. J. Geogr. Inf. Sci.*, 30, 369–389.

Ma, C., Huang, Z.-P., Zhao, X.-F., Zhang, L.-X., Sun, W.-M., Matthew B., S., Wang, X.-W., Cui, L.-W. & Xiao, W. (2014). Distribution and conservation status of Rhinopithecus strykeri in China. *Primates*, 55, 377–382.

Margules, C.R. & Pressey, R.L. (2000). Systematic conservation planning. *Nature*, 405, 243–253.

Marimont, R.B. & Shapiro, M.B. (1979). Nearest neighbour searches and the curse of dimensionality. *IMA J. Appl. Math.*, 24, 59–70.

Martin, L. (1965). *The physical geography of Wisconsin*. 3rd edn. The University of Wisconsin Press, Madison.

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

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

Minasny, B. & McBratney, A.B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Comput. Geosci.*, 32, 1378–1388.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

Mitchell, T.M. (1997). *Machine learning*. McGraw Hill, Burr Ridge, IL.

Moreno-Torres, J.G., Raeder, T., Alaiz-Rodr??guez, R., Chawla, N. V. & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognit.*, 45, 521–530.

Moudrý, V. & Šímová, P. (2012). Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *Int. J. Geogr. Inf. Sci.*, 26, 2083–2095.

Munson, A.M., Webb, K., Sheldon, D., Fink, D., Hochachka, W.M., Iliff, M., Riedewald, M., Sorokina, D., Sullivan, B., Wood, C., Munson, M.A., Webb, K., Sheldon, D., Fink, D., Hochachka, W.M., Iliff, M., Riedewald, M., Sorokina, D., Sullivan, B., Wood, C. & Kelling, S. (2012). The ebird reference dataset, version 4.0. *Cornell Lab Ornithol. Natl. Audubon Soc. Ithaca, NY*, 1–11.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B. & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403, 853–858.

Newman, G., Zimmerman, D., Crall, A., Laituri, M., Graham, J. & Stapel, L. (2010). User-friendly web mapping: lessons from a citizen science website. *Int. J. Geogr. Inf. Sci.*, 24, 1851–1869.

Olteanu-Raimond, A.-M., Hart, G., Foody, G.M., Touya, G., Kellenberger, T. & Demetriou, D. (2016). The Scale of VGI in Map Production: A Perspective on European National Mapping Agencies. *Trans. GIS*, 0, n/a-n/a.

Osborne, P.E. & Leitão, P.J. (2009). Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Divers. Distrib.*, 15, 671–681.

Pan and Wang. (2010). A survey on transfer learning. *Knowl. Data Eng. IEEE Trans.*, 22, 1345–1359.

Pardieck, K.L., Jr., D.J.Z., Hudson, M.-A.R. & Campbell, K. (2016). *North American Breeding Bird Survey Dataset 1966 - 2015, version 2015.1.*

Pardo, I., Pata, M.P., Gómez, D. & García, M.B. (2013). A novel method to handle the effect of uneven sampling effort in biodiversity databases. *PLoS One*, 8, e52786.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12, 2825–2830.

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

Phillips, S.J. & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography (Cop.).*, 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. *Ecol. Appl.*, 19, 181–197.

Qi, F. & Zhu, A.-X. (2003). Knowledge discovery from soil maps using inductive learning. *Int. J. Geogr. Inf. Sci.*, 17, 771–795.

Qin, C., Zhu, A.X., Pei, T., Li, B., Zhou, C. & Yang, L. (2007). An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. *Int. J. Geogr. Inf. Sci.*, 21, 443–458.

Qin, C.Z., Zhu, A.X., Shi, X., Li, B.L., Pei, T. & Zhou, C.H. (2009). Quantification of spatial gradation of slope positions. *Geomorphology*, 110, 152–161.

Rainville, D., Fortin, F.-A., Gardner, M.-A., Parizeau, M. & Gagné, C. (2012). DEAP: A python framework for evolutionary algorithms. In: *Proc. 14th Annu. Conf. companion Genet. Evol. Comput.* ACM, pp. 85–92.

Raudys, S. & Pikelis, V. (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 242–252.

Rios, L.M. & Sahinidis, N. V. (2013). Derivative-free optimization: A review of algorithms and comparison of software implementations. *J. Glob. Optim.*, 56, 1247–1293.

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

Royle, J.A. & Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Comput. Geosci.*, 24, 479–488.

Särndal, C.-E. & Lundström, S. (2005). *Estimation in surveys with nonresponse*. John Wiley & Sons.

Särndal, C.-E. & Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Surv. Methodol.*, 36, 131–144.

Sauer, J.R., Hines, J.E., Fallon, J.E., Link, W. a., Fallon, J.E., Pardieck, K.L. & Ziolkowski, D.J. (2013). The North American Breeding Bird Survey 1966-2011: Summary Analysis and Species Accounts. *North Am. Fauna*, 79, 1–32.

Scull, P., Franklin, J., Chadwick, O. a. & McArthur, D. (2003). Predictive soil mapping: a review. *Prog. Phys. Geogr.*, 27, 171–197.

Seeger, C.J. (2008). The role of facilitated volunteered geographic information in the landscape planning and site design process. *GeoJournal*, 72, 199–213.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference*, 90, 227–244.

Sieber, R. (2006). Information Public Geographic Participation: A Literature Review and Framework Systems. *Ann. Assoc. Am. Geogr.*, 96, 491–507.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, UK.

Silvertown, J. (2009). A new dawn for citizen science. *Trends Ecol. Evol.*, 24, 467–471.

Singh, V.P. & Woolhiser, D.A. (2002). Mathematical modeling of watershed hydrology. *J. Hydrol. Eng.*, 7, 270–292.

Sugiyama, M., Krauledat, M. & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8, 985–1005.

Sui, D.Z. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Comput. Environ. Urban Syst.*, 32, 1–5.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.*, 142, 2282–2292.

Telesco, R., Van Manen, F.T., Clark, J.D. & Cartwright, M.E. (2007). Identifying Sites for Elk Restoration in Arkansas. *J. Wildl. Manage.*, 71, 1393–1403.

Thorn, J.S., Nijman, V., Smith, D. & Nekaris, K. a. I. (2009). Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: Nycticebus). *Divers. Distrib.*, 15, 289–298.

Thuiller, W., Brotons, L., Araújo, M.B. & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography (Cop.).*, 2, 165–172.

Thuiller, W., Richardson, D.M., Pyšek, P., Midgley, G.U.Y.F., Hughes, G.O. & Rouget, M. (2005). Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Glob. Chang. Biol.*, 11, 2234–2250.

Trolle, M. & Kéry, M. (2003). Estimation of ocelot density in the Pantanal using capture-recapture analysis of camera-trapping data. *J. Mammal.*, 84, 607–614.

Turner, A. (2006). *Introduction to Neogeography*. First. O'Reilly.

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 (Cop.).*, 37, 1084–1091.

Vaysse, K. & Lagacherie, P. (2015). Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). *Geoderma Reg.*, 4, 20–30.

Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *J. Hum. Resour.*, 33, 127–169.

Viña, A., Bearer, S., Zhang, H., Ouyang, Z. & Liu, J. (2008). Evaluating MODIS data for mapping wildlife habitat distribution. *Remote Sens. Environ.*, 112, 2160–2169.

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

Warren, D.L. (2012). In defense of "niche modeling." *Trends Ecol. Evol.*, 27, 497–500.

Wikipedia. (2016). Breeding Bird Survey. *Wikipedia*.

Wikipedia. (2017). Red-tailed hawk. *Wikipedia*.

Wiley, E.O., McNyset, K., Peterson, T., Robins, R. & Stewart, A. (2003). Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography*, 16, 120–127.

Wilson, K. a., Westphal, M.I., Possingham, H.P. & Elith, J. (2005). Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biol. Conserv.*, 122, 99–112.

Winship, C. & Mare, R.D. (1992). Models for Sample Selection Bias. *Annu. Rev. Sociol.*, 18, 327–350.

Xu, C. & Yang, C. (2014). Introduction to big geospatial data research. *Ann. GIS*, 20, 227–232.

Yang, C.P. (2017). Geospatial cloud computing and big data. *Comput. Environ. Urban Syst.*, 61, 119.

Yang, F., Zhu, a. X., Ichii, K., White, M. a., Hashimoto, H. & Nemani, R.R. (2008). Assessing the representativeness of the AmeriFlux network using MODIS and GOES data. *J. Geophys. Res. Biogeosciences*, 113, 1–11.

Yang, L., Zhu, A., Qi, F., Qin, C., Li, B. & Pei, T. (2013). An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping. *Int. J. Geogr. Inf. Sci.*, 27, 1–23.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In: *Twenty-first Int. Conf. Mach. Learn. - ICML '04*. ACM Press, New York, New York, USA, p. 114.

Zadrozny, B., Langford, J. & Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. *Data Mining, 2003. ICDM 2003. Third IEEE Int. Conf.*

Zeng, C., Yang, L., Zhu, A.-X., Rossiter, D.G., Liu, J., Liu, J., Qin, C. & Wang, D. (2016). Mapping soil organic matter concentration at different scales using a mixed geographically weighted regression method. *Geoderma*, 281, 69–82.

Zhang, G., Huang, Q., Zhu, A.-X. & Keel, J. (2016a). Enabling point pattern analysis on spatial big data using cloud computing: Optimizing and accelerating Ripley's K function. *Int. J. Geogr. Inf. Sci.*, 30, 2230–2252.

Zhang, G., Zhu, A.-X. & Huang, Q. (2017). A GPU-accelerated adaptive kernel density estimation approach for efficient point pattern analysis on spatial big data. *Int. J. Geogr. Inf. Sci.*, 31, 2068–2097.

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

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

Zhu, A.-X. (1999). A personal construct-based knowledge acquisition process for natural resource mapping. *Int. J. Geogr. Inf. Sci.*, 13, 119–141.

Zhu, A.-X. (2000). Mapping soil landscape as spatial continua: the neural network approach. *Water Resour. Res.*, 36, 663–677.

Zhu, A.-X., Band, L., Vertessy, R. & Dutton, B. (1997). Derivation of soil properties using a soil land inference model (SoLIM). *Soil Sci. Soc. Am. J.*, 61, 523–533.

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

Zhu, A.-X., Yang, L., Li, B., Qin, C., Pei, T. & Liu, B. (2010). Construction of membership functions for predictive soil mapping under fuzzy logic. *Geoderma*, 155, 164–174.

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

Zhu, A.X. & Band, L.E. (1994). A knowledge-based approach to data integration for soil mapping. *Can. J. Remote Sens.*, 20, 208–218.

Zhu, A.X. & Mackay, D.S. (2001). Effects of spatial detail of soil information on watershed modeling. *J. Hydrol.*, 248, 54–77.

Zook, M., Graham, M., Shelton, T. & Gorman, S. (2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Med. Heal. Policy*, 2, 6–32.

van Zyl, J.J. (2001). The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography. *Acta Astronaut.*, 48, 559–565.