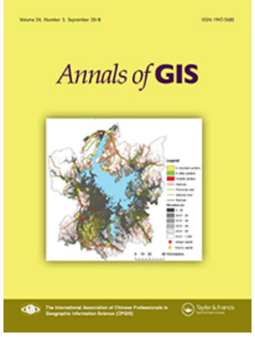

Annals of GIS

ISSN: 1947-5683 (Print) 1947-5691 (Online) Journal homepage: http://www.tandfonline.com/loi/tagi20

The representativeness and spatial bias of volunteered geographic information: a review

Guiming Zhang & A-Xing Zhu

To cite this article: Guiming Zhang & A-Xing Zhu (2018) The representativeness and spatial bias of volunteered geographic information: a review, Annals of GIS, 24:3, 151-162, DOI: 10.1080/19475683.2018.1501607

To link to this article: https://doi.org/10.1080/19475683.2018.1501607

Published online: 26 Jul 2018.

ANNALS OF GIS
2018, VOL. 24, NO. 3, 151–162
https://doi.org/10.1080/19475683.2018.1501607

Taylor & Francis
Taylor & Francis Group

The representativeness and spatial bias of volunteered geographic information: a review

Guiming Zhang[a,b] and A-Xing Zhu[b,c,d,e,f]

[a]Department of Geography & the Environment, University of Denver, Denver, USA; [b]Department of Geography, University of Wisconsin-Madison, Madison, USA; [c]Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China; [d]Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Nanjing, China; [e]State Key Laboratory Cultivation Base of Geographical Environment Evolution, Nanjing, China; [f]State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

ABSTRACT

Many applications of volunteered geographic information (VGI) involve inferring the properties of the underlying population from a sample consisting of VGI observations, i.e. VGI sample. The representativeness of VGI sample is crucial for deciding the fitness for use of VGI in such applications. Due to the volunteers' opportunistic observation efforts, spatial distribution of VGI observations is often biased (i.e. spatial bias). This degrades the representativeness of VGI and impedes the quality of inference made from VGI. Extensive research has been conducted on assessing or assuring VGI quality from the perspective of the fundamental dimensions of spatial data quality. Yet, this perspective alone provides limited insights on the representativeness of VGI. Assessing VGI representativeness and developing novel approaches to accounting for spatial bias in VGI is in need for broadening the spectrum of VGI applications. This article offers a comprehensive survey of the scientific literature from various domains (ecology, statistics, machine learning, etc.) to summarize existing endeavors related to sample representativeness assessment and sample selection bias correction for enlightening the treatment of these issues in VGI applications.

ARTICLE HISTORY
Received 20 November 2017
Accepted 5 June 2018

KEYWORDS
Volunteered geographic information (VGI); representativeness; representative sample; spatial bias; sample selection bias

1. Introduction

Volunteered geographic information (VGI) refers to geographic information created by volunteer citizens (Goodchild 2007). Broadly, VGI includes geographic information generated by volunteer participants in citizen science (Silvertown 2009), social media (Longley and Adnan 2016), etc., as they all share the commonality of voluntary and non-expert geographic information creation.

VGI is now driving many successful applications including global-scale geographic information production (Haklay and Weber 2008), species monitoring (Sullivan et al. 2009), emergency response and disaster monitoring (Zook et al. 2010), digital soil mapping (Rossiter et al. 2015), land cover mapping (Fonte et al. 2015), etc. VGI represents a paradigm shift in the way geographic information is created and shared, and in the content and characteristics of geographic information (Elwood 2008). It is thought of as an innovative phenomenon that will considerably influence geographic information science (GIScience) and geography and its relationship to the society (Goodchild 2007). VGI is also an important source of geospatial big data (Yang et al. 2016; Zhang, Zhu, and Huang 2016) that could significantly contribute to the paradigm shift from traditional geographic research to the emerging 'data-driven geography' (Miller and Goodchild 2014) and more broadly 'data-intensive science' (Kelling et al. 2009).

The representativeness of VGI, referring to the degree to which a sample consisting of VGI observations (i.e. VGI sample) can represent the underlying population, is an important factor for the potential VGI user to decide on the fitness for use of VGI in a particular application. Observations in a VGI dataset is a sample of some underlying population of interest, just like any other spatial data that often result from spatial sampling of the real world and objects in a spatial dataset are a sample drawn from the universe of all such objects (i.e. the population) (De Gruijter et al. 2006; Jensen and Shumway 2010). Many applications of VGI involve inferring properties of the underlying population from a sample consisting of VGI observations (Fink et al. 2010; Zhu et al. 2015). For example, based on the sighting records reported by volunteer birders, ecologists model the relationship between bird species habitat preference and the environmental

CONTACT Guiming Zhang guiming.zhang007@gmail.com 2050 Eeast Iliff Avenue, Denver, Colorado, 80210 USA

conditions and then use such relationship to predict and map the distribution of birds (Fink et al. 2010). In this process, it is desirable that the VGI sighting records, as a sample drawn from the population of the species of interest, can well capture the relationship between species distribution and the environmental conditions. Only under this premise, the species distribution map predicted based upon such a relationship is indicative of the real distribution of the species (Franklin and Miller 2009). Likewise, the opinion of a larger group of people can be inferred from tweets only if the sampled Twitter users form a 'representative' sample of that group of people (Morstatter et al. 2013). In general, any analysis that involves inferring something about the underlying population from a sample requires the sample to be 'representative' (McBratney, Mendonça Santos, and Minasny 2003). Assessing the representativeness (i.e. the degree of being representative) of VGI provides vital information on deciding whether VGI is suitable for such analyses.

Even though information on VGI data quality could provide information for the user to evaluate the general fitness for use of VGI in applications (Zhu 2004), the spatial data quality perspective alone provides little insights on the representativeness of VGI. VGI data quality has drawn extensive research attention (Goodchild and Li 2012; Goodchild 2013). Methods have been developed for VGI quality assessment (Foody et al. 2013; Jackson et al. 2013; Senaratne et al. 2017; Degrossi et al. 2018) and for VGI quality assurance or improvement (Goodchild and Li 2012; Naroditskiy et al. 2012; Ali and Schmid 2014). These methods assess and assure VGI data quality from the perspective of source credibility and the fundamental dimensions of spatial data quality: positional accuracy, attribute accuracy, temporal accuracy, semantic accuracy, logical consistency, completeness and lineage (Goodchild and Li 2012). However, among these dimensions, only completeness, the degree of which 'describes to what extent the entity objects within a dataset represent all instances of the abstract universe' (Brassel et al. 1995), could potentially provide information to assess the representativeness of VGI. Yet, evaluation of the completeness of VGI mostly focuses on the geometric aspects of VGI data (Girres and Touya 2010; Haklay 2010; Zielstra and Zipf 2010; Jackson et al. 2013) and rarely on the thematic attributes, which are also important aspects for assessing the representativeness of VGI (Comber et al. 2013; Fonte et al. 2015).

A VGI record often contains four components: location (where), time (when), observer (who) and attributes or phenomena (what). The 'where' and 'when' components are implicit keys for VGI to reveal the spatiotemporal variations of geographic phenomena (Fink et al. 2010). The representativeness of VGI is mostly discussed on the 'who' and 'what' components: the degree to which the observers can represent a larger group of people (Morstatter et al. 2013) or the degree to which the spatiotemporal variations of the geographic phenomena captured in the VGI observations can represent those in the real world (Fink et al. 2010; Zhu et al. 2015).

On the 'who' component, demographic biases in observers (i.e. not all demographic groups are equally likely to contribute to VGI creation) is the major cause that impedes the representativeness of VGI to represent a larger group of people (Li, Goodchild, and Xu 2013; Hecht and Stephens 2014; Malik et al. 2015). It should be noted that demographic biases sometimes can be attributed to the spatially biased distribution of the background population (Malik et al. 2015). On the 'what' component, VGI records are often more concentrated in some geographic areas than in others (i.e. spatial bias) and such spatial bias reduces the representativeness of VGI to represent the spatial variation of geographic phenomena or the covariation relationship among them (e.g. relationship between a target variable and other environmental variables learned through regression analysis) (Graham et al. 2004; Fink et al. 2010; Leitão, Moreira, and Osborne 2011; Zhu et al. 2015). Volunteers conduct observations wherever they want, and thus, their observation efforts are 'ad-hoc' and opportunistic in nature. This is radically different from scientific geographic sampling (e.g. random, stratified random, or systematic sampling) in which observations of geographic phenomena are taken at carefully chosen sampling sites to ensure the collected samples are representative (De Gruijter et al. 2006; Minasny and McBratney 2006; Gregoire and Valentine 2007; Jensen and Shumway 2010). Due to spatial bias, good representativeness of VGI is not guaranteed (Graham et al. 2004; Zhu et al. 2015).

Thus, spatial bias is an utmost cause that degrades the representativeness of VGI and impedes the quality of inference made from VGI. It is an issue that must be addressed in many geographic analyses using VGI. However, there is limited investigation on this subject in existing GIScience literature. Research on VGI quality currently focuses more on issues at the data collection stage rather than on the impacts of the issues on VGI data analyses. There are few VGI use cases for inferential analysis where population properties are inferred from VGI samples (except Fink et al. 2010; Zhu et al. 2015). This is probably due to the lack of awareness and analytical methods that can accommodate the imperfect representativeness and the spatial bias of VGI.

However, it is the potential of using VGI for inference that marks VGI an invaluable innovation and a prosperous phenomenon. Addressing the issue of representativeness and spatial bias of VGI is worth devoted attention to fulfil the full potential of VGI.

This article provides a comprehensive survey of existing scientific literature related to assessing the representativeness of VGI and correcting for the spatial bias in VGI. The general issue of 'representative sample' and sample selection bias is encountered in other domains such as ecology, statistics and machine learning (Heckman 1979; Cortes et al. 2008; Phillips et al. 2009; Bethlehem 2010). Although methods developed in these domains for representativeness assessment and sample selection bias correction are often designed to deal with non-VGI data, they can shed light upon addressing the issues in VGI data. This survey draws literature from various domains and summarizes existing endeavors related to representativeness assessment and spatial bias correction for VGI.

This article is organized as follows. Section 2 examines the meaning of 'representativeness' in the context of geographic research and demonstrates how spatial bias in VGI may impede the representativeness of VGI samples. Section 3 reviews the general approaches for assessing sample representativeness assessment and related work on VGI representativeness assessment. Section 4 surveys the existing methods for sample selection bias correction that could be adopted to address the spatial bias issue of VGI. Section 5 presents a summary discussion of the current state of research and conclusions.

2. The meaning of representativeness

Population and sample are essential to understand representativeness. A population consists of all units of interest for a study. In many cases, it is the population that we want to learn something about. A complete survey of the entire population is often not feasible or inefficient for various reasons. Thus, a sample consists of a subset of the units in the population that is selected and studied; conclusions about the population is then drawn based on the sample (i.e. inference). To make sound inferences, the selected sample needs to be 'representative' of the population (Jensen and Shumway 2010).

Kruskal and Mosteller (1979a, 1979b) examined in detail the meaning of 'representative sample'. 'Representative sample' could mean: 1) There are no selective forces in sampling which implies random selection of each unit in the population to enter the sample with an equal or unequal but known probability; 2) A miniature of the population which emphasizes that the important characteristics of the population are contained in their proper proportion in the sample; 3) A sample that is composed of typical units with certain known characteristics of the population; 4) A sample with good coverage of the population which requires that the sample contains at least one member from each set of a relevant partition of the population; 5) A sample that is obtained through a specific probabilistic sampling methods; 6) A sample that permits virtuous, or at least satisfactory, estimation of population characteristic where 'virtue' might be described in terms of little or no bias, low sampling error, etc.; 7) A sample that is good enough for a particular purpose. Thus, it might not require a properly executed probabilistic sampling of the population.

The above discussion on 'representative sample' is helpful to understand the meaning of representativeness, although the term 'representativeness' often is not formally defined. Several observations can follow. First, the representativeness of sample should always be discussed regarding specified characteristics of the population. If the population characteristics are known, representativeness can be assessed by quantifying the degree of agreement between the sample distribution and population distribution over the specified characteristics (Kruskal and Mosteller 1979b). Second, representative samples are usually obtained through probabilistic sampling procedures. Last, the representativeness of sample should be assessed regarding the purpose for which the sample is used.

In geographic studies, a population consists of all spatial units (e.g. point, linear or areal units) within a certain study area. A sample consists of measurements taken on a selected subset of spatial units in the study area where each measurement is composed of the values of thematic attributes of interest measured at a selected spatial unit, and the geographic coordinates of the selected spatial unit (Jensen and Shumway 2010). Geographic studies are concerned with the spatial variation of geographic phenomenon and the covariation among geographic phenomena. In this context, the representativeness of geographic sample refers to the degree to which the spatial variation of the attributes captured in the sample (i.e. selected spatial units) represents the underlying spatial variation of the attributes over the whole study area (i.e. all spatial units) or the degree to which the relationship between the attributes captured in the sample represents the underlying relationship that is held over a whole study area. Representative geographic samples are often obtained through well-designed geographic sampling schemes (De Gruijter et al. 2006; Minasny and McBratney 2006; Gregoire and Valentine 2007; Jensen and Shumway 2010; Yang et al. 2013).

However, due the spatial bias in VGI, a geographic sample consisting of VGI observations might not be representative of the underlying spatial variation or of the underlying covariation relationship. Thus, it is always desirable to assess the representativeness of VGI before it is used for examining the spatial variation of geographic phenomenon or for investigating the covariation relationship among geographic phenomena. Spatial bias in VGI has a significant impact on the inference made from sample consisting of VGI observations (Graham et al. 2004; Fink et al. 2010; Leitão, Moreira, and Osborne 2011; Pardo et al. 2013; Zhu et al. 2015). If not appropriately accounted for, spatial bias would adversely affect the validity of inferences made from VGI samples (Graham et al. 2004; Kadmon, Farber, and Danin 2004; Fink et al. 2010; Leitão, Moreira, and Osborne 2011; Kramer-Schadt et al. 2013; Pardo et al. 2013). Assessing the representativeness of VGI, thus, is desirable before using VGI sample for inference. If the representativeness is unsatisfactory, efforts need to be taken to correct for the spatial bias to improve the representativeness of the VGI sample such that VGI can still meet the requirements of the application.

3. Approaches for assessing representativeness

Based on the understanding of sample representativeness, general approaches have been developed for assessing the representativeness of a sample (Table 1). However, few research has been focusing on assessing the representativeness of VGI in particular.

3.1 General approaches

The general approaches for assessing sample representativeness include comparing the sample with the population on target variables, comparing the sample with the population on ancillary comparison variables, comparing the sample against a random sample, and investigating the sampling process to examine whether the selection of sampling units is *random*.

3.1.1 Comparing on target variables

Using this method for assessing the representativeness of sample, the values of the target variables (i.e. population characteristics which the sample is expected to be representative of) need to be known on both the population and the sample. If data on the underlying population is not available, a reference dataset is often used as a surrogate of the population. Román et al. (2009) developed a methodology to evaluate the representativeness of tower albedo measurement (i.e. the degree to which a tower measurement captures the spatial variability at the satellite pixel level) by comparing variogram models estimated from satellite surface albedo retrievals. Yesson et al. (2007) explored the global biodiversity representativeness of biodiversity occurrence data from Global Biodiversity Information Facility (GBIF) by comparing the coverage of GBIF plant occurrence data with data from the International Legume Database and Information Service, which is often considered as of being representative of global plant biodiversity.

The dilemma is that in many cases the values of the target variables are unknown on the population and they are exactly what need to be inferred from the sample given that the sample is representative. Assessing sample representativeness by directly comparing a sample with a population on the target variables is not always practically feasible (Kruskal and Mosteller 1979b). In cases where a reference dataset is used as a surrogate of the underlying population, there is another complication that the representativeness of the reference dataset itself should be assessed first.

3.1.2 Comparing on comparison variables

An alternative way to assess sample representativeness is to compare the sample with the population on comparison variables, rather than on target variables (Kruskal and Mosteller 1979b). Comparison variables are those variables believed to be related to the target variables in certain way, and they should be obtainable for both the sample and the population. Yang et al. (2008) evaluated the representativeness of the AmeriFlux network of eddy covariance towers to represent the environments contained within the coterminous United States by comparing environmental similarity (in terms of climatic, physiographic driver, and vegetation variables) between any ecoregion and the ecoregion containing the most similar network site. Ferrer et al. (2006) examined how avian sampling effort was influenced by environmental conditions, population distribution, and ornithological preferences of birdwatchers. Visit frequency was compared against a set of variables describing environmental condition, population distribution, and species richness through regression analysis. Hijmans et al. (2000) evaluated the representativeness of a Genebank collection of wild potatoes by comparing the distribution of distance to road in the Genebank collection against the distribution of distance to road in the study area.

Note that the selection of comparison variables is essential for this approach of representativeness assessment. The selection depends on the objectives of a specific study and is often carried out by the exercise of subjective judgement in practice. There is always a danger of omitting some important comparison variables: either they are mistakenly excluded or are completely unobserved. One

Table 1. Summary of approaches for assessing sample representativeness.

Domain	Method	Limitations	References
General	Comparing sample and population on target variables	Target variables are unknown on the population (in most cases, they are exactly what need to be inferred from the sample).	Yesson et al. 2007; Román et al. 2009.
	Comparing sample and population on ancillary variables	Selection of ancillary variables is often subjective. There is always a danger of omitting important variables.	Hijmans et al. 2000; Ferrer et al. 2006; Yang et al. 2008.
	Comparing sample against a random sample	Limited by the availability of a representative random sample to compare against.	Reddy and Davalos 2003; Kadmon, Farber, and Danin 2004.
	Examining sampling process to see if selection of sample is random	Requires knowledge and detailed information of the sampling process, which might not be available.	Ponder et al. 2001; Bethlehem 2010; Bethlehem 2012.
VGI	Evaluating representativeness of VGI contributors	Provides no information regarding the representativeness of VGI w.r.t. thematic attributes.	Brown, Kelly, and Whitall 2014; Hecht and Stephens 2014; Malik et al. 2015.
	Evaluating the completeness of VGI	Mostly focuses on the geometric aspects of VGI data. Provides little information regarding the representativeness of VGI w.r.t. thematic attributes.	Girres and Touya 2010; Haklay 2010; Zielstra and Zipf 2010; Jackson et al. 2013.
	Comparing VGI to a reference dataset	Limited by the availability of a up-to-date reference dataset to compare against.	Snäll et al. 2011; Brown, Kelly, and Whitall 2014.

can never robustly identify such danger nor quantify its probability. Avoidance of such danger is one motivation for using probabilistic sampling procedures to obtain representative sample (Kruskal and Mosteller 1979b).

3.1.3 *Comparing against a random sample*

A random sample obtained through probabilistic sampling procedure is often treated as of being representative. The representativeness of other sample can thus be assessed by comparing the sample to a random sample (on either target variables or comparison variables). Kadmon, Farber, and Danin (2004) examined the roadside bias in the distribution of woody plants species. The spatial distribution of plant observations was compared to a hypothetical spatially random distribution of plant observations. They also investigated climatic bias of the road network by comparing the climatic characteristics of locations randomly distributed along roadsides against locations randomly spread over the study area. Reddy and Davalos (2003) designed and applied statistical tests to assess sampling bias in a dataset comprising occurrence localities of passerine birds. The statistical tests were used to test whether the distribution of these localities was significantly different from random distributions.

3.1.4 *Examining the sampling process*

Based on the assumption that a sample obtained through probabilistic sampling procedures should be representative, the representativeness of a sample can be assessed by investigating the sampling process to examine whether the selection of sampling units is random. Ponder et al. (2001) evaluated the representativeness of specimen-based museum collection biodiversity data. One shortcoming of such data is the geographic gaps caused by the opportunistic collecting effort. The biodiversity occurrence data for the target taxon was compared to the occurrence records of organisms that are possibly captured by the same collectors or using the same methods as the target taxon. The sampling effort was then assessed by examining spatial statistics describing the spatial distribution of sampling points. It is implicitly assumed that the sampling effort that could result in a representative sample should be of sampling density that is equivalent to spatially uniform random sampling.

Non-response is a major cause of biased sample in survey studies. Bethlehem (2012) used response probabilities to assess sample representativeness in surveys. Response probabilities are modelled based on ancillary variables that are thought to be influential on the subject's decision process of whether to respond to the survey. For example, logistic regression is adopted to model response probabilities using certain specified socio-economic factors as influential variables, given that the values of the ancillary variables are available for both the respondents and non-respondents. The coefficient of variation of the response probabilities is then used as an indicator for the representativeness of the survey response (Bethlehem 2010; Bethlehem 2012).

This approach to assessing sample representativeness by examining the underlying sampling process requires knowledge and detailed information of the sampling process. Yet in many cases such information might not be available.

3.2 *Approaches to assess the representativeness of VGI*

There exist studies that evaluate the representativeness of the contributors of VGI to represent a larger group of people, with an emphasis on VGI generated on social media platforms. Research has been conducted to assess the completeness of VGI, an aspect of spatial data quality that is related to the representativeness of VGI. In many studies, the representativeness of VGI is assessed by comparing the VGI dataset against a reference dataset that is treated as of being representative of the underlying population.

3.2.1 Evaluating the representativeness of VGI contributors

There are studies evaluating the representativeness of the contributors of VGI to represent a larger group of people. Brown, Kelly, and Whitall (2014) compared the group responses about identifying national forest values and use preferences from a group of volunteers and those from a sample of randomly selected households. They found responses from the two groups are quite different. Thus, it is suggested to include scientific sampling in the VGI methods of collecting responses. Hecht and Stephens (2014) evaluated the potential urban bias of VGI generated on social media platforms (e.g. Twitter, Flickr, and Foursquare). Spatial distributions of county-level summary statistics on tweets count, check-ins count, and other attributes were compared against the distribution of the urban/rural population ratio by computing correlation measures. They found that VGI generated on social media platforms were biased towards urban areas. Malik et al. (2015) evaluated the representativeness of Twitter geotag users to represent the general population within the United States. They linked the counts of unique geotag users to census population counts at the block group level and found that the distribution of users over the population are nonrandom. They then used a simultaneous autoregressive model to investigate explanatory factors that could result in this nonrandom distribution. They found out that the nonrandom distribution can be explained by socioeconomic and geographic factors including income, urban or rural, distance to coast, and ethnic compositions.

Assessing the representativeness of VGI contributors (i.e. the 'who' component of VGI) to represent a larger group of people does not provide information for assessing the representativeness of VGI to represent spatial variation of the observed thematic attributes (i.e. the 'what' component of VGI). Yet in many cases it is the thematic attributes (e.g. species occurrence, land cover) that are of most interest in VGI applications.

3.2.2 Evaluating the completeness of VGI

Studies have been conducted to evaluate the completeness of VGI datasets, which could potentially provide information for assessing the representativeness of VGI. Completeness indicates 'whether the entity objects within a database represent all the entity instances in the real world' (Brassel et al. 1995). The comparison between VGI and the 'real world' is operationalized by comparing the VGI dataset (e.g. OpenStreetMap) to a reference dataset that is thought to be a good surrogate of the real world (e.g. an authoritative dataset produced by professional mapping agencies) (Girres and Touya 2010; Haklay 2010; Zielstra and Zipf 2010; Jackson et al. 2013). In the evaluation, measures of completeness at certain spatial unit level are computed based on the overlapping and difference between geometric features in the VGI dataset and those in the reference dataset. The distribution of completeness is then used to identify areas that are under-represented in the VGI dataset. In general, studies have found that there is a strong heterogeneity in VGI datasets in terms of completeness (e.g. the coverage of OpenStreetMap data in rural areas is much smaller than in urban areas).

Evaluation of the completeness of VGI was conducted extensively on the OpenStreetMap dataset, and mostly focuses on the geometric aspects of VGI data. Evaluating the completeness of VGI thus provides little information for assessing the representativeness of VGI in terms of the thematic attributes that reveal the spatial distribution of geographic phenomena.

3.2.3 Comparing VGI to a reference dataset

The representativeness of VGI is often assessed by comparing the VGI dataset against a reference dataset that is treated as of being representative of the underlying population. Brown, Kelly, and Whitall (2014) compared the group responses about identifying national forest values and use preferences from a group of volunteers and those from a random sample of households. Studies on evaluating the completeness of VGI also adopted this idea (e.g. comparing OpenStreetMap data with authoritative data). Snäll et al. (2011) provided another example by comparing annual dynamics of bird species modelled based on VGI data against those modelled based on a reference dataset that was collected using rigorous scientific protocols. The method performs regression analysis on the annual estimates of an abundance index of each species computed based on the VGI dataset and the abundance index computed based on the reference dataset. They found that at population level there exists positive relationship in inter-annual variation between the two datasets, but at species level it did not display consistent correlations.

Because data on the underlying population is rarely available, a reference dataset is used as a surrogate of the underlying population for assessing representativeness of a VGI dataset. Up-to-date authoritative map data exist in many countries, but limited budgets simply prohibit mapping agencies from keeping all themes of data up-to-date in all areas. In fact, they may be considered 'financially responsible' to only update data in urban areas and/or areas of intensive or extensive new development where the likelihood of change is much higher. In many cases VGI is the only data available reflecting the spatial distribution of the interested geographic phenomenon over large areas (e.g. eBird for global bird species monitoring); an authoritative reference dataset does not even exist because gathering and maintaining such a dataset is prohibitively expensive and unpractical.

4. Methods for correcting for spatial bias in VGI

There are few studies focusing on correcting for spatial bias in VGI in particular. The more general problem of sample selection bias, however, is an issue encountered in various domains, and methods have been developed to address the issue. Although these methods are not necessarily design to address specifically the spatial bias issue in VGI data, they might be adopted to correct for spatial bias in VGI. These methods include training local models, filtering samples, weighting samples by effort information, factoring bias out, modeling sample selection process, and importance weighting (Table 2).

4.1 *Training local instead of global models*

Spatial bias is a common phenomenon in broad-scale biological survey data (e.g. data from the eBird citizen science project; VGI data). Fink et al. (2010) proposed an AdaSTEM (adaptive spatiotemporal exploratory models) approach that can exploit variation in the density of observations to accommodate the spatial bias. The continent- or hemisphere-wide study area is partitioned into small rectangular spatial units (i.e. sub-areas) of size dependent upon the density of observations. Predictive models are trained with only observations in each small spatial unit and are later used for prediction in that spatial unit. By training local predictive models using data in sub-areas, instead of training a global predictive model using data over the whole study area, this approach mitigates the overall spatial bias in the dataset to certain extent.

It should be noted that a sub-area over which a local predictive model is trained might still covers a large geographic area (e.g. 3 × 4 latitude by longitude). Potentially, observations in such a large sub-area might still have spatial bias that is not accounted for in training the local model.

4.2 *Filtering samples*

Filtering species occurrence data in the geographic or attribute 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 sample localities that are within certain distance (in either geographic or attribute space) of one another would somehow balance the unequal sampling or observation effort.

It is challenging, however, to objectively determine the distance threshold, which has a profound impact on the filtering process. Moreover, filtering samples reduces effective sample size and discards useful information in the removed samples. It is thus not applicable to cases where only a paucity of samples exists.

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. (2015) proposed an approach for predictive mapping using VGI (e.g. mapping wildlife habitat suitability based on wildlife sighting records elicited from the local residents). When extracting the suitability-environment relationships from VGI records, they developed a method to correct for spatial bias in VGI by inversely weighting VGI observations with weights proportional to the cumulative visibility at the observation sites. Here 'cumulative visibility' is the frequency of a given location being seen by observers from the routes taken by the local residents, which is calculated using viewshed analysis based on the routes and a digital elevation model of the study area. It is used as a proxy of the underlying observation effort of the local residents in observing the wildlife.

Obviously, this method is applicable only for cases where cumulative visibility is a reasonable approximation of the underlying sampling or observation effort. More generally, if sampling or observation effort information can be quantified, it could be used to weigh VGI observations in this way to correct for spatial bias in VGI.

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 and Miller 2009; Pardo et al. 2013; Phillips et al. 2009). Dudík, Schapire, and Phillips (2005) and Phillips

Table 2. Summary of methods for sample selection bias correction.

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.
	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. 2015.

	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 bias as the presence-only samples.	Requires sampling/observation effort information to generate background samples.	Dudík, Schapire, and Phillips 2005; Phillips et al. 2009.

Statistics	Modelling the sample selection process.	Needs good understanding and detailed information of the sampling process.	Heckman 1979;
			Bethlehem 2010;
			Bethlehem 2012.

Machine learning	Weighting samples by an importance weighting function to compute the loss in learning classifiers.	Requires sufficiently large sample size to estimate the optimal weighting function.
		Hard for high dimensional cases.
			Shimodaira 2000;
			Zadrozny 2004;
			Cortes et al. 2008.

et al. (2009) developed a FactorBiasOut method to correct for spatial bias in species presence-only data for species distribution modeling with MAXENT (Phillips, Anderson, and Schapire 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 samples (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 samples (Dudík, Schapire, and Phillips 2005; Phillips et al. 2009).

The FactorBiasOut method works only for species distribution models that require background samples. It requires information on sampling effort or its estimate to generate the background samples properly. However, sampling effort information underlying VGI genesis might not be available as the volunteers who submit observations are not committed to report detailed effort information.

4.5 Modeling sample selection process

Nonrandom selection is a source of bias in empirical research, such as surveys with nonresponses and self-selections (Bethlehem 2010; Särndal and Lundström 2005) and species distribution models with presence-only data (Phillips et al. 2009), and a fundamental aspect of many social and economic data collection processes (Winship and Mare 1992; Heckman 1979). 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 modelling (Bethlehem 2012; Bethlehem 2010; Heckman 1979). 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 and Lundström 2010), and then be used to correct for nonresponse bias (Bethlehem 2012; Bethlehem 2010).

This approach requires deep understanding of the underlying selection processes to come up with appropriate selection models. It might be 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.

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, and transfer learning (Cortes et al. 2008; Zadrozny 2004; Pan and Wang 2010). Sample selection bias arises where the underlying distributions from which the training and test data are drawn from are different. In other words, the distribution of the training data in feature space is different from the distribution of the test data. 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 to compute the loss (Shimodaira 2000). Asymptotically, the optimal weighting function proves to be the ratio of the probability density function of features on the test data and the density function on the training data (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). VGI applications might involve many variables (i.e. high dimension) and VGI samples of possibly small sample size. It is, thus, challenging to apply the importance weighting method to correcting for spatial bias in VGI.

Conclusions

5.1 Summary

Making inferences from samples is still the effective way to learn about the underlying population of interest wherever a complete enumeration of the population is impractical (Jensen and Shumway 2010). A sample consists of individuals or units in the population that are obtained by either formal sampling (e.g. geographic sampling) or casual observation (e.g. VGI). Sample representativeness remains a valid concern wherever samples are relied on to make inferences. Compared to samples collected through well designed geographic sampling schemes, samples consisting of VGI observations (VGI samples) are more prone to representativeness problem due to the spatial bias in VGI. Assessing the representativeness of VGI and accounting for the spatial bias in VGI samples are crucial for any VGI applications that involve sample-to-population inferences.

Current research on VGI data quality focuses more on the data collection stage than on the data analysis stage (e.g. making inferences). Consequently, extensive studies have been conducted on assessing or assuring VGI data quality from the perspectives of source credibility and the fundamental dimensions of spatial data quality (i.e. positional, attribute, temporal, semantic accuracy, logical consistency, completeness, and lineage) (Senaratne et al. 2017). Yet there are relatively fewer VGI data analyses examining the impacts of VGI data quality issues on inferences made from VGI samples, not to mention efforts on assessing the representativeness of VGI and on tackling the spatial bias in VGI (Fink et al. 2010; Zhu et al. 2015). This review makes a novel contribution by comprehensively surveying scientific literature to summarize existing endeavors related to sample representativeness assessment and sample selection bias correction, in general, and existing efforts on VGI sample representativeness assessment and spatial bias correction, in particular.

Assessing sample representativeness takes two general approaches. The first approach is examining the sample selection process to see whether the sample is obtained through probabilistic sampling procedures. This approach is rarely applicable to VGI because VGI observations are most often not gathered through probabilistic sampling. The second approach is comparing the sample against another sample that is deemed representative of the population or directly against the population on a target variable or ancillary variables. VGI representativeness assessment can be easily operationalized provided that a reference representative sample is available or an authoritative reference dataset serving as a surrogate of the population exists. Yet, in many cases such reference datasets do not exist due to the high cost of data collection, maintenance and update, which imposes a practical challenge for assessing VGI representativeness.

Studies have been conducted to assess the representativeness of VGI contributors on social media platforms to represent a larger group of people and to evaluate the completeness of VGI datasets (e.g. OpenStreetMap). However, assessing the representativeness of VGI contributors focuses exclusively on the 'who' component of VGI. Evaluation of the completeness of VGI mostly considers only the geometric aspects of VGI data. These studies, thus, provide little insights for assessing the representativeness of VGI in terms of the thematic attributes, which are of core interest to many geographic studies concerning the spatial variation of geographic phenomenon or the covariation among geographic phenomena. In addition, the representativeness of VGI was often assessed indirectly by qualitatively examining the biases of VGI. There lacks a clear definition for the representativeness of VGI. Quantitative measures of the representativeness of VGI are missing.

Existing methods could potentially be adopted to correct for or mitigate the spatial bias in VGI, although each method has its own data requirements to which VGI application may not meet (Section 4). For example, many bias correction methods (i.e. modeling sample selection process, weighting samples by effort, factoring bias out) rely on detailed information of the sampling or observation process (e.g. selection probabilities, sampling or observation effort) to correct for bias. Such information, however, might not be available in VGI genesis as volunteers are not committed to report effort information. The fitness of use of the bias correction methods need to be evaluated on a case by case basis.

5.2 Outlook

Assessing representativeness of VGI samples and mitigating spatial bias in VGI samples is in need for many VGI applications. More efforts are called for to develop methodologies for assessing and measuring the representativeness of VGI samples to make reliable inferences from VGI. Representativeness measures can be helpful in many respects. As a starting point, for instance, they can be used for quantitatively assessing VGI representativeness, which is desirable for evaluating the fitness for use of VGI in applications. On the other hand, such representativeness measures can be used for developing new approaches to mitigating the spatial bias in VGI. For example, bias correction by filtering samples is based on a simple distance heuristic that is not directly related to sample representativeness. Yet, given representativeness measures, more informative heuristics could be designed to direct the filtering process towards increasing the representativeness to correct for the spatial bias in VGI samples.

Acknowledgments

Support to Guiming Zhang through the Whitbeck Graduate Dissertator Award from the Department of Geography, University of Wisconsin-Madison is appreciated. Supports to A-Xing Zhu from the National Natural Science Foundation of China [grant number 41431177], National Basic Research Program of China [grant number 2015CB954102], Natural Science Research Program of Jiangsu [grant number 14KJA170001], PAPD, National Key Technology Innovation Project for Water Pollution Control and Remediation [grant number 2013ZX07103006], and Outstanding Innovation Team in Colleges and Universities in Jiangsu Province are acknowledged. Supports to A-Xing Zhu through the Vilas Associate Award, the Hammel Faculty Fellow Award, the Manasse Chair Professorship from the University of Wisconsin-Madison, and the "One-Thousand Talents" Program of China are greatly appreciated.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [41431177].

ORCID

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

References

Ali, A. L., and F. Schmid. 2014. "Data Quality Assurance for Volunteered Geographic Information." *Proceedings of the Eighth International Conference on Geographic Information Science*. Austria: Vienna. 126–141.

Bethlehem, J. 2010. "Selection Bias in Web Surveys." *International Statistical Review* 78: 161–188. doi:10.1111/(ISSN)1751-5823.

Bethlehem, J. 2012. "Using Response Probabilities for Assessing Representativity." *Statistics Netherlands, International Statistical Review* 80: 382-399.

Boria, R. A., L. E. Olson, S. M. Goodman, and R. P. Anderson. 2014. "Spatial Filtering to Reduce Sampling Bias Can Improve the Performance of Ecological Niche Models." *Ecological Modelling* 275: 73–77. doi:10.1016/j.ecolmodel.2013.12.012.

Brassel, K., F. Bucher, E. Stephan, and A. Vckovski. 1995. "Completeness." In *Elements of Spatial Data Quality*, eds S. C. Guptill and J. L. Morrison, 81–108. Oxford: Elsevier.

Brown, G., M. Kelly, and D. Whitall. 2014. "Which "Public"? Sampling Effects in Public Participation GIS (PPGIS) and Volunteered Geographic Information (VGI) Systems for Public Lands Management." *Journal of Environmental Planning and Management* 57: 190–214. doi:10.1080/09640568.2012.741045.

Comber, A., L. See, S. Fritz, M. Van Der Velde, C. Perger, and G. Foody. 2013. "Using Control Data to Determine the Reliability of Volunteered Geographic Information about Land Cover." *International Journal of Applied Earth Observation and Geoinformation* 23: 37–48. doi:10.1016/j.jag.2012.11.002.

Cortes, C., M. Mohr, M. Riley, and A. Rostamizadeh. 2008. "Sample Selection Bias Correction Theory." *International Conference on Algorithmic Learning Theory*, 38-53. Berlin, Heidelberg: Springer.

De Gruijter, J., D. J. Brus, M. F. Bierkens, and M. Knotters. 2006. *Sampling for Natural Resource Monitoring*. Berlin/Heidelberg: Springer Science & Business Media.

Degrossi, L. C., J. Porto de Albuquerque, R. dos Santos Rocha, and A. Zipf. 2018. "A Taxonomy of Quality Assessment Methods for Volunteered and Crowdsourced Geographic Information." *Transactions in GIS* 22: 542–560. doi:10.1111/tgis.2018.22.issue-2.

Dudík, M., R. E. Schapire and S. J. Phillips. 2005. "Correcting Sample Selection Bias in Maximum Entropy Density Estimation." *Proceedings of the 18th International Conference on Neural Information Processing Systems*, 323–330. Cambridge, MA: MIT Press.

Elwood, S. 2008. "Volunteered Geographic Information: Key Questions, Concepts and Methods to Guide Emerging Research and Practice." *GeoJournal* 72: 133–135. doi:10.1007/s10708-008-9187-z.

Ferrer, X., L. M. Carrascal, Ó. Gordo, and J. Pino. 2006. "Bias in Avian Sampling Effort Due to Human Preferences: An Analysis with Catalonian Birds (1900–2002)." *Ardeola* 53: 213–227.

Fink, D., W. M. Hochachka, B. Zuckerberg, D. W. Winkler, B. Shaby, M. A. Munson, G. Hooker, M. Riedewald, D. Sheldon, and S. Kelling. 2010. "Spatiotemporal Exploratory Models for Broad-Scale Survey Data." *Ecological Applications* 20: 2131–2147. doi:10.1890/09-1340.1.

Fonte, C. C., L. Bastin, L. See, G. Foody, and F. Lupia. 2015. "Usability of VGI for Validation of Land Cover Maps." *International Journal of Geographical Information Science* 29: 1269–1291. doi:10.1080/13658816.2015.1018266.

Foody, G. M., L. See, S. Fritz, M. Van Der Velde, C. Perger, C. Schill, and D. S. Boyd. 2013. "Assessing the Accuracy of Volunteered Geographic Information Arising from Multiple Contributors to an Internet Based Collaborative Project." *Transactions in GIS* 17: 847–860. doi:10.1111/tgis.2013.17.issue-6.

Franklin, J., and J. A. Miller. 2009. *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge: Cambridge University Press.

Girres, J. F., and G. Touya. 2010. "Quality Assessment of the French OpenStreetMap Dataset." *Transactions in GIS* 14: 435–459. doi:10.1111/j.1467-9671.2010.01203.x.

Goodchild, M. F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *Geojournal* 69: 211–221. doi:10.1007/s10708-007-9111-y.

Goodchild, M. F. 2013. "The Quality of Big (Geo) Data." *Dialogues in Human Geography* 3: 280–284. doi:10.1177/2043820613513392.

Goodchild, M. F., and L. Li. 2012. "Assuring the Quality of Volunteered Geographic Information." *Spatial Statistics* 1: 110–120. doi:10.1016/j.spasta.2012.03.002.

Graham, C. H., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. "New Developments in Museum-Based Informatics and Applications in Biodiversity Analysis." *Trends in Ecology & Evolution* 19: 497–503. doi:10.1016/j.tree.2004.07.006.

Gregoire, T. G., and H. T. Valentine. 2007. *Sampling Strategies for Natural Resources and the Environment*. Boca Raton: CRC Press.

Haklay, M. 2010. "How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning. B, Planning & Design* 37: 682–703. doi:10.1068/b35097.

Haklay, M., and P. Weber. 2008. "OpenStreetMap: User-Generated Street Maps." *Pervasive Computing, IEEE* 7: 12–18. doi:10.1109/MPRV.2008.80.

Hecht, B., and M. Stephens. 2014. "A Tale of Cities: Urban Biases in Volunteered Geographic Information." Proceedings of the 8th International AAAI Conference on Weblogs & Social Media, Ann Arbor, MI, June 1-4.

Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica: Journal of the Econometric Society* 153–161. doi:10.2307/1912352.

Hijmans, R. J., K. A. Garrett, Z. Huaman, D. P. Zhang, M. Schreuder, and M. Bonierbale. 2000. "Assessing the Geographic Representativeness of Genebank Collections: The Case of Bolivian Wild Potatoes." *Conservation Biology* 14: 1755–1765. doi:10.1046/j.1523-1739.2000.98543.x.

Jackson, S. P., W. Mullen, P. Agouris, A. Crooks, A. Croitoru, and A. Stefanidis. 2013. "Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information." *ISPRS International Journal of Geo-Information* 2: 507–530. doi:10.3390/ijgi2020507.

Jensen, R. R., and J. M. Shumway. 2010. "Sampling Our World." In *Research Methods in Geography: A Critical Introduction*, eds B. Gomez and J. P. Jones III, 77–90. Hoboken, New Jersey: John Wiley & Sons.

Kadmon, R., O. Farber, and A. Danin. 2004. "Effect of Roadside Bias on the Accuracy of Predictive Maps Produced by Bioclimatic Models." *Ecological Applications* 14: 401–413. doi:10.1890/02-5364.

Kelling, S., W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker. 2009. "Data-Intensive Science: A New Paradigm for Biodiversity Studies." *Bioscience* 59: 613–620. doi:10.1525/bio.2009.59.7.12.

Kramer-Schadt, S., J. Niedballa, J. D. Pilgrim, B. Schröder, J. Lindenborn, V. Reinfelder, M. Stillfried, et al. 2013. "The Importance of Correcting for Sampling Bias in MaxEnt Species Distribution Models." *Diversity and Distributions* 19: 1366–1379. doi:10.1111/ddi.12096.

Kruskal, W., and F. Mosteller. 1979a. "Representative Sampling, II: Scientific Literature, Excluding Statistics." *International Statistical Review* 47: 111–127. doi:10.2307/1402564.

Kruskal, W., and F. Mosteller. 1979b. "Representative Sampling, III: The Current Statistical Literature." *International Statistical Review* 47: 245–265. doi:10.2307/1402647.

Leitão, P. J., F. Moreira, and P. E. Osborne. 2011. "Effects of Geographical Data Sampling Bias on Habitat Models of Species Distributions: A Case Study with Steppe Birds in Southern Portugal." *International Journal of Geographical Information Science* 25: 439–454. doi:10.1080/13658816.2010.531020.

Li, L., M. F. Goodchild, and B. Xu. 2013. "Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr." *Cartography and Geographic Information Science* 40: 61–77. doi:10.1080/15230406.2013.777139.

Longley, P. A., and M. Adnan. 2016. "Geo-Temporal Twitter Demographics." *International Journal of Geographical Information Science* 30: 369–389. doi:10.1080/13658816.2015.1089441.

Malik, M. M., H. Lamba, C. Nakos, and J. Pfeffer. 2015 "Population Bias in Geotagged Tweets." Proceedings of the 9th International AAAI Conference on Weblogs & Social Media, Oxford, England, May 26–29.

McBratney, A., M. Mendonça Santos, and B. Minasny. 2003. "On Digital Soil Mapping." *Geoderma* 117: 3–52. doi:10.1016/S0016-7061(03)00223-4.

Miller, H. J., and M. F. Goodchild. 2014. "Data-Driven Geography." *GeoJournal* 80: 449–461. doi:10.1007/s10708-014-9602-6.

Minasny, B., and A. B. McBratney. 2006. "A Conditioned Latin Hypercube Method for Sampling in the Presence of Ancillary Information." *Computers and Geosciences* 32: 1378–1388. doi:10.1016/j.cageo.2005.12.009.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2013. "Is the Sample Good Enough?" Comparing Data from Twitter's Streaming API with Twitter's Firehose." In Proceedings of the 7th International AAAI Conference on Weblogs & Social Media, Cambridge, MA, July 8–11.

Naroditskiy, V., I. Rahwan, M. Cebrian, and N. R. Jennings. 2012. "Verification in Referral-Based Crowdsourcing." *PloS One* 7: 45924. doi:10.1371/journal.pone.0045924.

Panand Wang. 2010. "A Survey on Transfer Learning." *Knowledge and Data Engineering, IEEE Transactions on 22* 1345–1359. doi:10.1109/TKDE.2009.191.

Pardo, I., M. P. Pata, D. Gómez, and M. B. García. 2013. "A Novel Method to Handle the Effect of Uneven Sampling Effort in Biodiversity Databases." *PloS One* 8: 52786. doi:10.1371/journal.pone.0052786.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. "Sample Selection Bias and Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data." *Ecological Applications* 19: 181–197. doi:10.1890/07-2153.1.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. "Maximum Entropy Modeling of Species Geographic Distributions." *Ecological Modelling* 190: 231–259. doi:10.1016/j.ecolmodel.2005.03.026.

Ponder, W. F., G. A. Carter, P. Flemons, and R. R. Chapman. 2001. "Evaluation of Museum Collection Data for Use in Biodiversity Assessment." *Conservation Biology* 15: 648–657. doi:10.1046/j.1523-1739.2001.015003648.x.

Reddy, S., and L. M. Davalos. 2003. “Geographical Sampling Bias and Its Implications for Conservation Priorities in Africa.” *Journal of Biogeography* 30: 1719–1727. doi:10.1046/j.1365-2699.2003.00946.x.

Román, M. O., C. B. Schaaf, C. E. Woodcock, A. H. Strahler, X. Yang, R. H. Braswell, P. S. Curtis, et al. 2009. “The MODIS (Collection V005) BRDF/albedo Product: Assessment of Spatial Representativeness over Forested Landscapes.” *Remote Sensing of Environment* 113: 2476–2498. doi:10.1016/j.rse.2009.07.009.

Rossiter, D. G., J. Liu, S. Carlisle, and A.-X. Zhu. 2015. “Can Citizen Science Assist Digital Soil Mapping?” *Geoderma* 259-260: 71–80. doi:10.1016/j.geoderma.2015.05.006.

Särndal, C. E., and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. Hoboken, New Jersey: John Wiley & Sons.

Särndal, C. E., and S. Lundström. 2010. “Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias.” *Survey Methodology* 36: 131–144.

Senaratne, H., A. Mobasheri, A. A L, C. Capineri, and M. Haklay. 2017. “A Review of Volunteered Geographic Information Quality Assessment Methods.” *International Journal of Geographical Information Science* 31: 139–167. doi:10.1080/13658816.2016.1189556.

Shimodaira, H. 2000. “Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function.” *Journal of Statistical Planning and Inference* 90: 227–244. doi:10.1016/S0378-3758(00)00115-4.

Silvertown, J. 2009. “A New Dawn for Citizen Science.” *Trends in Ecology & Evolution* 24: 467–471. doi:10.1016/j.tree.2009.03.017.

Snäll, T., O. Kindvall, J. Nilsson, and T. Pärt. 2011. “Evaluating Citizen-Based Presence Data for Bird Monitoring.” *Biological Conservation* 144: 804–810. doi:10.1016/j.biocon.2010.11.010.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. “eBird: A Citizen-Based Bird Observation Network in the Biological Sciences.” *Biological Conservation* 142: 2282–2292. doi:10.1016/j.biocon.2009.05.006.

Varela, S., R. P. Anderson, R. García-Valdés, and F. Fernández-González. 2014. “Environmental Filters Reduce the Effects of Sampling Bias and Improve Predictions of Ecological Niche Models.” *Ecography* 37: 1084–1091.

Winship, C., and R. D. Mare. 1992. “Models for Sample Selection Bias.” *Annual Review of Sociology* 18: 327–350. doi:10.1146/annurev.so.18.080192.001551.

Yang, L., A. Zhu, F. Qi, C. Qin, B. Li, and T. Pei. 2013. “An Integrative Hierarchical Stepwise Sampling Strategy for Spatial Sampling and its Application in Digital Soil Mapping.” *International Journal of Geographical Information Science* 27: 1–23

Yang, A., H. Fan, N. Jing, Y. Sun, and A. Zipf. 2016. “Temporal Analysis on Contribution Inequality in OpenStreetMap: A Comparative Study for Four Countries.” *ISPRS International Journal of Geo-Information* 5: 5. doi:10.3390/ijgi5010005.

Yang, F., A. Zhu, K. Ichii, M. A. White, H. Hashimoto, and R. R. Nemani. 2008. “Assessing the Representativeness of the AmeriFlux Network Using MODIS and GOES Data.” *Journal of Geophysical Research: Biogeosciences* 113: 1–11. doi:10.1029/2007JG000627.

Yesson, C., P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, et al. 2007. “How Global Is the Global Biodiversity Information Facility?” *PLoS ONE* 2: 1124. doi:10.1371/journal.pone.0001124.

Zadrozny, B. 2004. “Learning and Evaluating Classifiers under Sample Selection Bias.” *Twenty-first international conference on Machine learning - ICML ’04*. New York: ACM Press. 114.

Zhang, G., A. X. Zhu, and Q. Huang. 2016. “Enabling Point Pattern Analysis on Spatial Big Data Using Cloud Computing: Optimizing and Accelerating Ripley’s K Function.” *International Journal of Geographical Information Science* 30: 2230–2252. doi:10.1080/13658816.2016.1170836.

Zhu, A. X. 2004. “Research Issues on Uncertainty in Geographic Data and GIS-Based Analysis.” In *A Research Agenda for Geographic Information Science*, eds R. B. McMaster and E. L. Usery, 197–223. Boca Raton: CRC Press.

Zhu, A. X., G. Zhang, W. Wang, W. Xiao, Z. P. Huang, G. S. Dunzhu, G. Ren, et al. 2015. “A Citizen Data-Based Approach to Predictive Mapping of Spatial Variation of Natural Phenomena.” *International Journal of Geographical Information Science* 29: 1864–1886. doi:10.1080/13658816.2015.1058387.

Zielstra, D. and A. Zipf. 2010. “A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany.” Proceedings of 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, May 10–14.

Zook, M., M. Graham, T. Shelton, and S. Gorman. 2010. “Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake.” *World Medical & Health Policy* 2: 6–32. doi:10.2202/1948-4682.1069.