Using Volunteered Geographic Information (VGI) in Design-Based Statistical Inference for Area Estimation and Accuracy Assessment of Land Cover

Stephen V. Stehman[a], Cidália C. Fonte[b], Giles M. Foody[c], Linda See[d]

[a] Department of Forest and Natural Resources Management, SUNY College of Environmental Science and Forestry, Syracuse, NY 13210, United States (svstehma@syr.edu)

[b] Departmento de Matemática, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Apartado 3008, EC Santa Cruz, 3001 – 501 Coimbra, Portugal (cfonte@mat.uc.pt)

[c] School of Geography, University of Nottingham, Sir Clive Granger Building, University Park, Nottingham, NG7 2RD, United Kingdom (giles.foody@nottingham.ac.uk)

[d] International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria (see@iiasa.ac.at)

Corresponding Author: Stephen V. Stehman (svstehma@syr.edu)

Abstract

Volunteered Geographic Information (VGI) offers a potentially inexpensive source of reference data for estimating area and assessing map accuracy in the context of remote-sensing based land-cover monitoring. The quality of observations from VGI and the typical lack of an underlying probability sampling design raise concerns regarding use of VGI in widely-applied design-based statistical inference. This article focuses on the fundamental issue of sampling design used to acquire VGI. Design-based inference requires the sample data to be obtained via a probability sampling design. Options for incorporating VGI within design-based inference include: 1) directing volunteers to obtain data for locations selected by a probability sampling design; 2) treating VGI data as a "certainty stratum" and augmenting the VGI with data obtained from a probability sample; and 3) using VGI to create an auxiliary variable that is then used in a model-assisted estimator to reduce the standard error of an estimate produced from a probability sample. The latter two options can be implemented using VGI data that were obtained from a non-probability sampling design, but require additional sample data to be acquired via a probability sampling design. If the only data available are VGI obtained from a non-probability sample, properties of design-based inference that are ensured by probability sampling must be replaced by assumptions that may be difficult to verify. For example, pseudo-estimation weights can be constructed that mimic weights used in stratified sampling estimators. However, accuracy and area estimates produced using these pseudo-weights still require the VGI data to be representative of the full population, a property known as "external validity". Because design-based inference requires a probability sampling design, directing volunteers to locations specified by a probability sampling design is the most straightforward option for use of VGI in design-based inference. Combining VGI from a non-probability sample with data from a probability sample using the certainty stratum approach or the model-assisted approach are viable alternatives that meet the conditions required for design-based inference and use the VGI data to advantage to reduce standard errors.

Key Words: probability sampling; external validity; pseudo-weights; data quality; model-based inference; Volunteered Geographic Information (VGI); crowdsourcing

1. Introduction

Volunteered Geographic Information (VGI) is defined as "tools to create, assemble, and disseminate geographic data provided voluntarily by individuals" (Goodchild 2007). For land-cover studies, VGI may provide the reference condition or the information used to determine the reference condition of a spatial unit. The reference condition, defined as the best available assessment of the ground condition, plays a critical role in accuracy assessment and area estimation (Olofsson et al. 2014). When used in map production, VGI could form all or part of the data used to train the land-cover classification algorithm. The focus of this article is the contribution of VGI to the reference data used for accuracy assessment and area estimation. Accuracy assessment is an essential component of a rigorous mapping-based analysis of remotely sensed data as without it the obtained products are little more than pretty pictures and simply untested hypotheses (McRoberts 2011; Strahler et al. 2006). In addition an accuracy assessment adds value to a study, especially when estimates of class area (e.g. deforestation) are to be obtained (Olofsson et al. 2014). Fonte et al. (2015) examined the use of VGI for land cover validation, including the types of VGI that have been used, the main issues surrounding VGI quality assessment, and examples of VGI projects that have collected data for validation purposes. We build upon this past work to focus on the issue of statistical inference when incorporating VGI in applications of accuracy and area estimation, but our work is also relevant to application of citizen science data in general (Bird et al. 2014).

Map accuracy assessment is a spatially explicit comparison of the map class label to the reference condition on a per spatial unit basis (e.g., pixel, block, or segment). Accuracy assessment typically focuses on producing an error matrix and associated summary measures including overall, user's, and producer's accuracies (see Section 2 for details). Estimates of area of each land-cover class or type of land-cover change based on the reference condition are often produced in conjunction with the accuracy estimates (Olofsson et al. 2013, 2014). Sampling, defined as selecting a subset of the population, is almost always necessary because it is too costly to obtain a census of the reference condition. VGI represents a subset of the population and as such may be viewed as a sample. Whether the VGI data were collected via a probability sampling design is a key consideration when evaluating the utility of VGI for design-based inference. Design-based inference is a standard, widely used approach adopted in environmental science for furthering knowledge and understanding on the basis of a sample of cases rather than a study of the entire population.

We describe options for incorporating VGI into map accuracy assessment and area estimation within the design-based inference framework (Figure 1). We evaluate how the potential cost savings of

VGI can be transformed into more precise estimators (i.e., smaller standard errors, a desirable outcome of an effective sampling strategy) within the scientifically defensible framework provided by design-based inference. If the VGI data are obtained via a probability sampling design, application of design-based inference is straightforward and can be informed by good practice guidelines (Olofsson et al. 2014). Alternatively, if the VGI data are not obtained via a probability sampling protocol, the VGI data can be combined with additional data from a probability sample to produce estimates that satisfy the conditions underlying design-based inference. In such cases the VGI data from a non-probability sample serve as a means to reduce standard errors of estimates rather than as the sole data from which the area and accuracy estimates are produced.

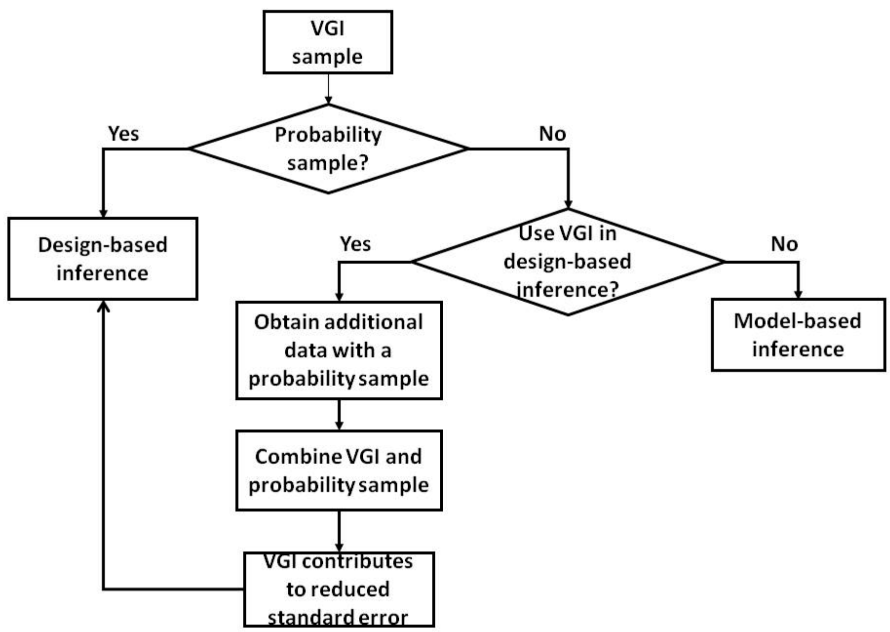

Figure 1. Schema for methodologies using VGI in accuracy assessment and area estimation.

This article has two major objectives. First, it illustrates how statistically rigorous and credible inference may be drawn from studies that use VGI and thereby helps ensure that the vast potential of VGI that has recently arisen is realized fully. This in turn will help remote sensing achieve its full potential as a source of land cover information which is often constrained by lack of ground reference data. Second, the article provides methodological rigor and good practice advice for the use of data acquired via popular sample designs, ranging from judgmental to probability sampling. As such this article articulates methodology for producing credible inference from data sets that often do not conform to the requirements of widely used statistical inferential methods for two common and important application areas of remote sensing, accuracy assessment and area estimation. To do this, we, for the first time, synthesize methods developed in the general sampling literature into a comprehensive treatment of the theory and methods for using VGI in design-based inference. This includes translating methods developed for the use of non-probability samples for accuracy assessment and area estimation applications. As such we will show how VGI may be constructively used to decrease costs and reduce uncertainty (e.g., yield smaller standard errors and hence narrower confidence intervals) while following a methodology that allows for rigorous design-based inference. Throughout this article, guidance for using VGI in design-based inference is framed by examining the direct connection of the inference process to the three component protocols of accuracy assessment, the response design, sampling design, and analysis (Stehman and Czaplewski 1998).

The article is organized as follows. In Section 2, we define inference and describe the conditions needed to satisfy design-based inference. Considerations regarding the use of VGI in design-based inference are then explained in Section 3 in regard to the response design, sampling design and analysis protocols. Section 4 provides the details of two methods for incorporating VGI in estimation of accuracy and area that satisfy conditions of design-based inference, with both methods requiring that an additional probability sample exists or could be acquired if the VGI did not originate from a probability sampling design. Options for analysis when the only data available are VGI from a non-probability sample are discussed in Section 5. Sections 6 and 7 provide discussion and a summary of the article.

2. Inference

Following Baker et al. (2013, p.91), we define statistical inference as "… a set of procedures that produces estimates about the characteristics of a target population and provides some measure of the reliability of those estimates." Statistical inference focuses on the use of sample data to estimate parameters of a target population, where a parameter is defined as a number describing the population (e.g., the population mean and population proportion are two common parameters). Determining the numerical value of a parameter would require a census of the study region, but in practice parameters are estimated from a sample. Statistical inference also includes how bias and variance of these sample-based estimators are defined. Baker et al. (2013, p.91) further specify that "A key feature of statistical inference is that it requires some theoretical basis and explicit set of assumptions for making the estimates and for judging the accuracy of those estimates." Consequently, sampling design and analysis protocols must adhere to certain rules of implementation to ensure that the underlying mathematical basis of the inference framework is satisfied. Failure to adhere to these rules may lead to substantial bias in the estimators of parameters of interest or even nullify the ability to implement design-based inference entirely (see Section 3.3).

Two general types of inference are design-based inference and model-based inference (De Gruijter and Ter Braak 1990; Särndal et al. 1992; Gregoire 1998; Stehman 2000; McRoberts 2010, 2011). In design-based inference, bias and variance of an estimator are determined by the randomization distribution of the estimator which is represented by the set of all possible samples that could be selected from the population using the chosen sampling design. This randomization distribution is completely dependent on the sampling design hence the origin of the name "design-based" inference. The inclusion probabilities of the sampling design are the critical link to the randomization distribution that underlies design-based inference (Särndal et al. 1992, section 2.4). The practical considerations for using VGI in design-based inference are explained in detail in Section 4.

A probability sampling design must satisfy two criteria related to the inclusion probabilities determined by the sample selection protocol. The inclusion probability of a particular element of the population (e.g., a pixel) is defined as the probability of that element being included in the sample. An inclusion probability is defined in the context of all possible samples that could be selected for a given sampling design. For example, if the design is simple random sampling of n elements selected from the N elements of the population, the inclusion probability of each element u of the population is $\pi_u = n/N$. That is, in the context of all possible simple random samples of size n from this population, element u has the probability of n/N of being included in the sample selected. The two requirements of a probability sampling design are that π_u must be known for each element of the sample and $\pi_u > 0$ for each element of the population (Särndal et al. 1992; Stehman 2000). Probability sampling requires a randomization mechanism to be present in the selection protocol. Convenience, judgment, haphazard, and purposive selection of sample elements are examples of protocols that do not satisfy the criteria defining a probability sampling design (Cochran 1977, Sec. 1.6). Use of such samples for inference carries considerable risk due to lack of representation of the population.

An alternative to design-based inference is model-based inference (Valliant et al. 2000). As the name implies, model-based inference requires specification of a statistical model and inference is dependent on the validity of the model. Consequently, verifying model assumptions is a critical and often challenging feature of model-based inference. Model-based inference does not require a probability sampling design, although implementation of a probability sampling design is often recommended to ensure objectivity in sample selection because of the randomization (Valliant et al. 2000, p.20). Applications of model-based inference are briefly discussed in Section 5.3.

3. Component Protocols of Accuracy Assessment and Area Estimation

We describe the role of each of the three components of the methodology (response design, sampling design, and analysis) in determining how VGI can be incorporated in rigorous design-based inference. The response design is the protocol for determining the reference condition (i.e., the best available assessment of the ground condition). The response design includes all steps leading to assignment of the reference condition label of a point or spatial unit (e.g., a land-cover class or change versus no change label). The sampling design is the protocol for selecting the sample units at which the response design will be applied. Lastly, the analysis consists of defining parameters to describe properties of the population (e.g., overall accuracy, proportion of area of each class) and the formulas required to estimate these population parameters from the sample data. To justify the requirements of each step to achieve the final accuracy or area estimates, our description starts with the analysis (Section 3.1) focusing on how the VGI data would be used, followed by the steps of the response design (Section 3.2) and the sampling design (Section 3.3).

3.1 Analysis: Accuracy and Area Estimation Based on Totals

The details of the analysis protocol that specify how the estimates of accuracy and area are produced yield insights into how VGI should be evaluated for use in design-based inference. The analysis focuses on summarizing information contained in an error matrix. We define the population to be a collection of *N* equal-area units partitioning the region of interest. The population error matrix resulting from a census can be constructed in terms of area as illustrated by the numerical example in Table 1 for a simple two-class legend, “crop” and “not crop” for a population (target region) of 1000 km^2. The error matrix expressed in terms of area (Table 1) could easily be converted to proportion of area by dividing each cell of the error matrix by 1000 km^2. However, it is useful to focus on the error matrix expressed in terms of area because we can formulate the population parameters of interest for accuracy and area as totals or ratios of totals of areas. For example, overall accuracy is the total area of agreement obtained from the sum of the area of the diagonal cells (930 km^2) divided by the total area of the target region (1000 km^2) to yield overall accuracy of 0.93 or 93%. User's accuracy for the crop class is the total area where both the map and reference condition are crop (840 km^2) divided by the total area mapped as crop (890 km^2) to yield the parameter 0.94 or 94%. Producer's accuracy for the crop class is the total area where both the map and reference condition are crop (840 km^2) divided by the total area of reference condition of crop (860 km^2) to yield the parameter 0.98 or 98%. Lastly, the area of reference condition of the crop class is also simply a total, in this case the sum of the two cells in the "crop" column of reference condition (840+20 = 860 km^2).

Table 1. Population error matrix expressed in terms of area (km^2) for a hypothetical target region of 1000 km^2. Overall accuracy is 93% (930/1000).

	Reference Condition			
Map	Crop	Not Crop	Total	User's
Crop	840	50	890	0.94
Not Crop	20	90	110	0.82
Total	860	140	1000	
Producer's	0.98	0.64		

Given that the parameters of interest for accuracy and area can be expressed in terms of totals, the analysis focuses on estimating these totals. Basic sampling theory provides an unbiased estimator of a population total in the form of the Horvitz-Thompson estimator (Horvitz and Thompson 1952). The population total of the variable y_u is defined as

$$Y = \sum_P y_u \quad [1]$$

where the summation is over all *N* elements of the population, *P*. For example, if y_u is the area of crop (as determined from the reference condition) for element *u*, then *Y* is the total area of crop. The population total *Y* can be estimated from a sample using the Horvitz-Thompson estimator

$$\hat{Y} = \sum_s \frac{y_u}{\pi_u} \quad [2]$$

where the summation is over all elements of the sample *s*.

The Horvitz-Thompson estimator is an unbiased estimator of a population total for any sampling design as long as the inclusion probabilities of the sample elements are known for that design. A useful re-expression of the Horvitz-Thompson estimator highlighting the sample estimation weights is

$$\hat{Y} = \sum_s w_u y_u \quad [3]$$

where $w_u = 1/\pi_u$ is the estimation weight for element *u* of the sample. Because $w_u \geq 1$, the y_u value for each sampled element is multiplied by an "expansion factor" w_u to estimate a total. In effect each sample element must account for itself along with some additional elements of the population that were not selected into the sample. For example, for simple random sampling $w_u = N/n$ so y_u for each sampled element is "expanded" by the multiplier w_u to account for N/n elements of the population. The critical importance of known inclusion probabilities for rigorous design-based inference is evident via the role of the weights $w_u = 1/\pi_u$ in the estimator $\hat{Y}$ (equations 2 and 3).

Parameters such as user's accuracy and producer's accuracy are ratios of totals and consequently can be estimated by the corresponding ratio of estimated totals (Särndal et al. 1992, section 5.3). For example, if we define *Y* as the total area of the population for which both the map and reference condition are crop and *X* as the total area mapped as crop, the ratio of population totals *Y*/*X* would be the population parameter for user's accuracy of crop. User's accuracy could then be estimated from the sample data using a ratio of Horvitz-Thompson estimators, $\hat{Y}/\hat{X}$, where both $\hat{Y}$ and $\hat{X}$ are estimated totals based on equation (2), considering, respectively, y_u=area of pixel *u* with both map and reference condition of crop and x_u=area of pixel u mapped as crop. In the case of a pixel-based assessment and assuming all pixels are equal area, user's accuracy of crop estimated using a ratio of Horvitz-Thompson estimators would simply require defining y_u=1 if pixel u has both map and reference labels of crop (y_u=0 otherwise) and defining x_u=1 if pixel u has map label of crop (x_u=0 otherwise). In this formulation of user's accuracy, the ratio Y/X is the proportion of pixels mapped as the target class that have the reference label of that class.

Formulas for the variance and estimated variance of the Horvitz-Thompson estimator are provided by Särndal et al. (1992, section 2.8). The square root of the estimated variance (standard error) would be used to construct a confidence interval for the parameter of interest so issues of inference obviously extend to variance and confidence interval estimation. Although we do not delve into the details of the formulas for variance estimators, we emphasize that known inclusion probabilities are an essential feature of variance estimation. Consequently, the requirement of implementing probability sampling to ensure known inclusion probabilities for estimating a total applies as well to estimating the variance of an accuracy or area estimator.

The conditions required for VGI to be used in design-based inference are apparent from the analysis protocol. The accuracy and area parameters of interest can be expressed as population totals or ratios of population totals and these totals can be estimated using the Horvitz-Thompson estimator. From the Horvitz-Thompson estimator formula (equations 2 and 3) we observe that the key features of VGI relevant to estimating a total are quality of the observation y_u and knowledge of the inclusion probability π_u. In other words, the questions pertinent to evaluating the utility of VGI for design-based inference are: 1) What is the quality of y_u (an issue to address in the response design) and 2) Is π_u known (an issue to address in the sampling design)? The following two subsections address issues of VGI related to the response and sampling designs.

3.2 Response Design

The response design is the protocol for determining the reference condition of an element of the population. In the case of a land-cover legend based on a conventional hard classification, the response design results in a reference land-cover label assigned to each pixel (i.e., if the legend consists of *C* classes, one and only one of these class labels is assigned to the pixel). The reference class labels can be translated to a quantity by the simple process of defining $y_u = 1$ if pixel *u* has reference class *c* and $y_u = 0$ otherwise. Thus for example if class *c* is forest, all pixels with reference class forest would be assigned $y_u = 1$ and all non-forest pixels would have $y_u = 0$. Evaluating and assuring the quality of VGI is critical because high quality reference data are absolutely essential to accuracy and area estimation. If the reference labels are not accurate, these errors can have a substantial impact on accuracy and area estimates (Foody 2009, 2010). Very accurate reference data obtained within a timeframe corresponding to the date of remote sensing image acquisition are a necessity for every application of accuracy assessment and area estimation from remote sensing. VGI has considerable potential as a source of reference data, notably in facilitating the collection of a large set of observations over broad geographical regions. However, the use of volunteers rather than experts in assigning the reference class labels may exacerbate concerns regarding label accuracy, although amateurs can sometimes be as accurate as experts in labeling (See et al. 2013). Further, VGI tends to be collected continuously rather than within a narrow time frame which can limit its value, especially for studies of land-cover change.

Applications in which VGI has been collected for land cover and land use studies are becoming increasingly common. Fonte et al. (2015) reviewed several applications including:

1) Geo-Wiki project, which uses the crowd for interpretation of very high resolution satellite imagery (Fritz et al. 2012);

2) VIEW-IT, which is a validation system for MODIS land cover (Clark and Aide 2011); and

3) geo-tagged photographs for land cover validation from different applications such as the Degree Confluence Project, Geograph, Panoramio and Flickr (Antoniou et al. 2016; Fonte et al. 2015; Iwao et al. 2006).

Another source of VGI for land-cover studies is the LACO-Wiki system, an online land cover validation tool intended as a repository of openly available validation data crowdsourced from different users (See et al. 2017). More recently, land cover and land use have been crowdsourced in the field through the FotoQuest Austria app, which sends users to specific locations and loosely follows the LUCAS protocol for data collection (Laso Bayas et al. 2017). Hou et al. (2015) describe geo-tagged web texts as an alternative to photographs as yet another source of VGI useful for land-cover studies.

The quality of the VGI data collected for land cover and land use studies has received recent attention. A substantial body of literature focuses on the positional quality and completeness of OpenStreetMap (OSM), the most commonly cited VGI project (e.g., Ciepłuch et al. 2010; Girres and Touya 2010; Haklay 2010). Other elements of quality include thematic accuracy (which is relevant to land cover and land use), temporal quality, logical consistency, and usability, all of which are set out in ISO 19157 (Fonte et al. 2017a). In addition, Antoniou and Skopeliti (2015) outline quality indicators that are tailored to VGI such as data indicators, demographic and other socio-economic indicators, and indicators about the volunteers. Due to the specificities of VGI when compared to traditional geographic information and the diversity of uses of these data, additional methodologies are starting to be developed that aim to integrate several quality measures and indicators into quality assessment workflows, enabling quality data to be combined to produce more reliable quality information (e.g., Bishr and Mantelas 2008; Jokar Arsanjani and Bakillah 2015; Meek et al. 2016).

Although concern with reference data error may be heightened when VGI is used, there are methods such as latent class analysis, which can be used to characterize volunteers in terms of their quality in labeling classes and could therefore be used to filter or weight the data when used subsequently in applications (Foody et al. 2013, 2015). These issues of data quality associated with the response design are critical to the overall process of accuracy and area estimation. In reality, reference data quality issues are equally impactful whether the source of the reference classification is VGI or expert interpretation (See et al. 2013).

3.3 Sampling Design

The sampling design is the protocol used to select the subset of locations (e.g., pixels) at which the reference condition is determined. As noted earlier, the inclusion probability of pixel *u* is denoted as π_u, and the two criteria defining a probability sampling design are: 1) π_u is known for all pixels in the sample and 2) $\pi_u > 0$ for all pixels in the population. Because probability sampling is a requirement of rigorous design-based inference, the sample selection protocol must ensure that these two conditions of π_u are satisfied. Moreover, randomization of the sample selection is required of all probability sampling designs as it is this randomization that creates the probabilistic foundation for design-based inference. The sampling design is linked to the analysis via the inclusion probabilities that are incorporated in the Horvitz-Thompson estimator (equations 2 and 3).

Because design-based inference requires known inclusion probabilities, it is critical to establish whether a probability sampling design was the basis for collecting VGI data. The distinction between active and passive VGI is relevant in this regard. Active VGI refers to directing volunteers to specific sample locations (e.g., See et al. 2016) and therefore allows for implementing a probability sampling design for collecting VGI. Conversely, passive VGI refers to allowing volunteers to choose where they will collect data and typically leads to purposive or convenience sampling with attendant concern regarding lack of representation of the full population. The protocols that determine where VGI data are collected span a continuum ranging from rigorous probability sampling to selection by judgment or convenience without an underlying random mechanism.

The Degree Confluence Project (Iwao et al. 2006) is an example in which VGI data are collected via a probability sampling protocol. These data are obtained at locations defined by the intersection of lines of latitude and longitude and therefore originate from a design akin to systematic sampling (due to the Earth's shape the distances between sample points vary with latitude so the inclusion probabilities would not all be equal but would still be known).   A second example of VGI based on a probability sampling design is the FotoQuest Austria app which uses the Land Use/Cover Area frame Survey (LUCAS) sample (which is based on a systematic sample of points spaced 2 km apart in the four cardinal directions across the European Union) followed by a stratified sample (Martino et al. 2009). That is, land cover and land use were crowdsourced via the FotoQuest Go mobile app in which volunteers were sent to specific locations that formed part of the LUCAS systematic sample for Austria, and the LUCAS sample was then augmented with additional sample units (Laso Bayas et al. 2016).

Several VGI applications include sample data originating from both probability sampling designs and volunteer chosen locations. The Geo-Wiki project is used to collect land cover and land use data via different campaigns (See et al. 2015). These campaigns have all had different purposes and hence were driven by different sampling designs. For example, the first campaign to validate a map of land availability for biofuels was driven by a stratified random sample with equal sample size in both the land available stratum and the land unavailable stratum. To this an additional sample from cropland areas was added although the data were not used to undertake an accuracy assessment as such but to modify the statistics on how much land is available (Fritz et al. 2013). Other studies have made use of Geo-Wiki data from previous campaigns for validation that were not obtained using a probability sampling approach for the specific product to be validated (see, for example, Schepaschenko et al. (2015) and Tsendbazar et al. (2015) for review of reference datasets including those from Geo-Wiki). The VIEW-IT application (Clarke and Aide 2011) either directs users to specific locations selected based on a probability sampling design or users can provide information about the land cover at any location, which means these latter sample locations would not be part of a probability sampling design. The LACO-Wiki system (See et al. 2017) has built-in probability sampling schemes although users can upload their own sample locations that do not necessarily conform to a probability sampling design.

Photograph repositories such as Panoramio, Flickr, and Instagram are examples of passive VGI and therefore do not conform to any probability sampling design. For example, photographs made available by citizens may be positioned at any location chosen by the volunteer (such as the photographs available in Flickr or Instagram), or collected at predefined locations. Similarly, the data available in collaborative projects such as OSM are created at locations of interest to the citizen volunteers, and consequently these data have no underlying probability sampling design. The amount and quality of the OSM data are known to be correlated with demographic or socio-economic factors (e.g., Mullen et al. 2014; Elwood et al. 2013) and this offers some possibility for adjusting estimates to account for misrepresentation of the population (see Section 5.1).

The Geograph project asks users to take photographs in every square kilometer of the United Kingdom and classify them (now also extended to other locations in the world). Since 2005, 83.4% of the 1 km^2 squares in Great Britain and Ireland have photographs (http://www.geograph.org.uk/, accessed 29 October 2017) and nearly 5.5 million images are available within this time period. Volunteers may choose locations within each square kilometer at which photographs are taken. Therefore, if each photograph is viewed as representing a point location or, for example, the 30 m x 30 m pixel surrounding the photograph's location, the data would not meet the criteria defining a probability sampling design due to the lack of randomization in the selection protocol. Directing the volunteers to cover the 1 km^2 squares provides a better degree of spatial representation of the VGI than might otherwise occur if volunteers are allowed to choose locations completely on their own. Specifically, the 1 km^2 squares effectively serve as spatial (geographic) strata, and with over 83% of these strata visited, the Geograph project data achieve the desirable design criterion of being spatially well distributed (Stehman 1999, Figure 3). The Geograph project data collection protocol illustrates the fact that within the class of non-probability sample designs, features can be built into the protocol to enhance representation of the VGI data.

4. Methods to Use VGI in Design-based Inference

In this section, we address how to incorporate VGI into design-based inference focusing on sampling design and estimation considerations (Figure 2). The label quality issues of VGI remain a concern but are not addressed in this section. The most straightforward approach to ensure the utility of VGI for design-based inference is to direct volunteers to collect data at locations specified by a probability sampling design (which is possible with "active VGI"). Several examples of VGI collections based on a probability sampling design were documented in Section 3.3. Specifying sample locations selected via probability sampling has the potential drawback that volunteer participation may be reduced if volunteers are unable to choose locations of personal interest. Consequently, additional effort may be necessary to obtain y_u at those locations neglected by volunteers.

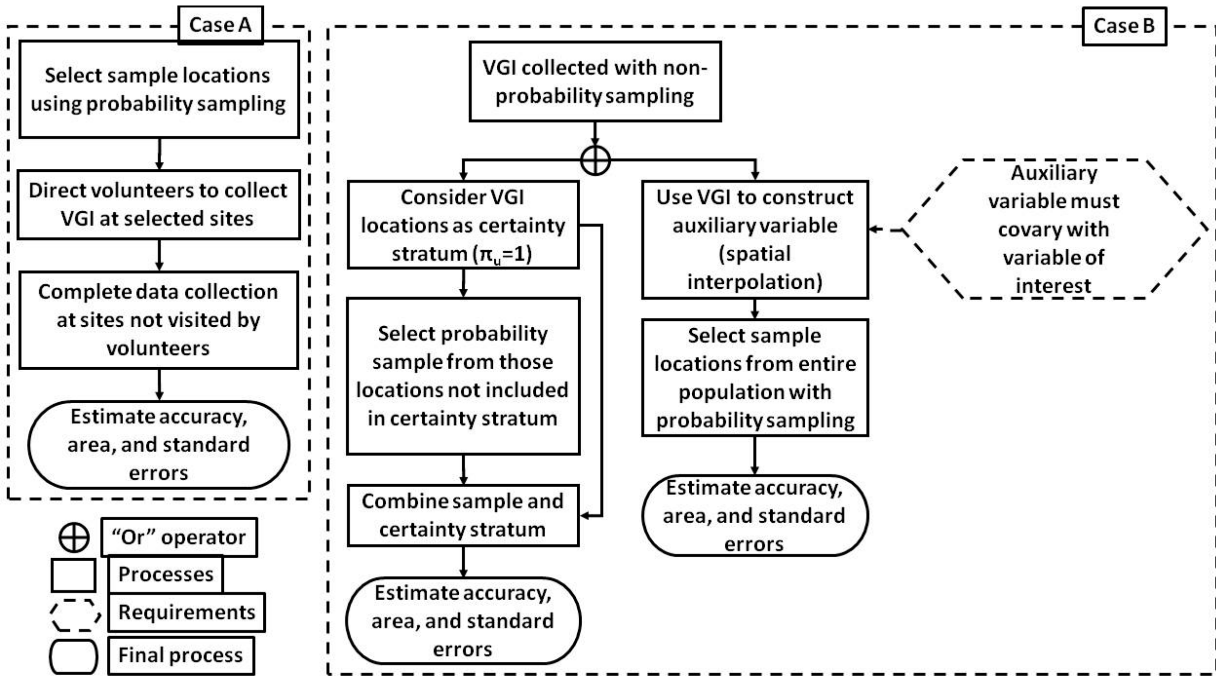

Figure 2. Schema for using VGI in design-based inference.

If a large quantity of VGI obtained from a non-probability sampling design exists, the VGI data may be augmented with data from a probability sampling design (Figure 2). Two options are described in the following subsections. In the first option, the VGI data are treated as a "certainty stratum" and combined with data from a probability sample selected from the locations not already included in the VGI data. In the second option, the probability sample is selected from the full population and the VGI data are used to construct an auxiliary variable that is then incorporated in a model-assisted estimator to reduce the standard errors of the estimates based on the data from the probability sample.

4.1 VGI Incorporated as a Certainty Stratum

VGI data can be combined with data obtained from a probability sample by treating each VGI sample unit (e.g., a pixel) as belonging to a "certainty stratum" in which the inclusion probability is $\pi_u=1$ (Overton et al. 1993). By assigning $\pi_u=1$ to each VGI sample unit, we acknowledge that these sample units were not selected via a randomized selection protocol, and instead we view these units as having been purposely selected to be included with certainty in the sample. From the remaining units of the population not included in the VGI certainty stratum, a probability sampling design is implemented and these newly selected sample units are combined with the VGI data to produce the accuracy and area estimates. In this approach the VGI data are used directly in the estimation of accuracy and area, so the quality of the VGI data is a critical concern.

All sample units selected via the probability sampling design will have a known inclusion probability and the data from these sample units can be combined with the VGI data using the Horvitz-Thompson estimator. Specifically, suppose there are N_1 elements for which we have no VGI and N_2 elements for which VGI provides y_u ($N=N_1+N_2$). Further, let G denote the subset for which VGI is available (the "G" is from the middle letter of VGI) and $\tilde{G}$ denote the subset of the population for which VGI is not available. The population total Y can then be partitioned into summations over the two subpopulations $\tilde{G}$ and G,

$$Y = \sum_{\tilde{G}} y_u + \sum_G y_u = Y_{\tilde{G}} + Y_G \qquad [4]$$

Because Y_G (total of y_u for the VGI data) is known, it is only necessary to estimate $Y_{\tilde{G}}$ from the sample. Therefore, an estimator of Y can be expressed as

$$\hat{Y} = \sum_s y_u/\pi_u + \sum_G y_u = \hat{Y}_{\tilde{G}} + Y_G \qquad [5]$$

where the first summation is over the elements selected in the sample from the N_1 elements of the population $\tilde{G}$ for which VGI is not available. The variance of $\hat{Y}$ is $V\left(\hat{Y}\right) = V(\hat{Y}_{\tilde{G}})$ because the total of the VGI data is a known quantity with no uncertainty attributable to sampling. That is, the only uncertainty attributable to sampling arises from estimating the total $Y_{\tilde{G}}$ for the non-VGI portion of the population, $\tilde{G}$.

The benefit of the VGI data when incorporated as a certainty stratum is to reduce the standard errors of the accuracy and area estimators and accordingly to decrease the width of confidence intervals for the parameters of interest. To illustrate the potential reduction in standard error, we focus on the objective of estimating area based on the reference condition obtained for each sample unit. The benefit of the VGI data can then be quantified by comparing the variance of the estimator of total area without using VGI data to the variance of the estimator using the certainty stratum approach (equation 5). Several conditions are imposed to simplify the variance comparison: 1) the sample of non-VGI units is selected by simple random sampling; 2) the VGI data have the same variability as the non-VGI data (i.e., the variance of y_u for the VGI subpopulation G is the same as the variance of y_u for the non-VGI subpopulation $\tilde{G}$); and 3) the sample size n is the same regardless of whether VGI is present (i.e., the VGI data are viewed as obtained at no cost so n is the same with or without VGI). If no VGI data are available and a simple random sample is selected from the full population of N elements (i.e., N_2=0 because no VGI data exist), the variance of the estimated total is

$$V(\hat{Y}) = N^2\left(1-\frac{n}{N}\right)V_y/n \qquad [6]$$

The variance of $\hat{Y}$ when VGI is available for N_2 elements of the subpopulation G is derived as follows. A simple random sample of n elements is selected from the N_1 non-VGI units. The variance of the estimated total combining the VGI data with the non-VGI sample (equation 5) depends only on the variance of the total estimated from the non-VGI sample units,

$$V(\hat{Y}_{\tilde{G}}) = N_1^2\left(1-\frac{n}{N_1}\right)V_y/n \qquad [7]$$

To quantify the reduction in variance achieved by the VGI data, we examine the ratio of the two variances,

$$R = \frac{V(\hat{Y}_{\tilde{G}})}{V(\hat{Y})} = \frac{N_1^2\left(1-\frac{n}{N_1}\right)}{N^2\left(1-\frac{n}{N}\right)} \qquad [8]$$

The V_y/n term common to both equations (6) and (7) cancels in the ratio R by virtue of the assumption that the variability of y_u is the same in the VGI and non-VGI subpopulations (if V_y is different in the two subpopulations, R will be impacted by the ratio of the variances of the two subpopulations, G and $\tilde{G}$).

Under the assumption of equal variance for the two subpopulations, the benefit of VGI to reduce variance depends on the proportion of the population that is covered by the VGI data, which is defined as $k=N_2/N$. If we define $f=n/N$ to be the proportion of the total population selected for the probability sample, then R can be re-written as

$$R = (1-k)(1-f-k)/(1-f). \qquad [9]$$

If no VGI data exist, then k=0 and R=1 as expected because there would be no reduction in variance from VGI. Conversely, if k=1, then R=0 as expected because the VGI would constitute a census and the population total Y would be known yielding a variance of 0. As the quantity of VGI gets larger (i.e., $k=N_2/N$ increases), R decreases indicating a greater benefit accruing to the availability of the VGI data. Numerical values of $\sqrt{R}$ (ratio of standard errors) for several combinations of k and f are presented in Table 2. For a fixed value of $f=n/N$, $\sqrt{R}$ decreases approximately linearly with increasing k. For a fixed value of k, the decrease in $\sqrt{R}$ is much less prominent as f increases except for the case with f=0.25 and k=0.75 which represents a census so $V(\hat{Y}_{\tilde{G}}) = 0$. To simplify the problem still further, assume that the spatial unit of the assessment is a pixel and that N is so large that $f = n/N = 0$. Then setting $f = 0$ in equation (9), we obtain $R = (1 - k)^2$ which leads directly to

$$\sqrt{R} = 1 - k \qquad [10]$$

Thus for very large populations the reduction in standard error achieved by VGI will be directly related to k, the proportion of the population for which VGI is available – the greater the quantity of VGI available (i.e., larger k) the greater the reduction in standard error.

Table 2. Reduction in standard error achieved by using VGI in the certainty stratum approach. Values shown in the table are $\sqrt{R}$ where R is the ratio of the variance of the estimated total with VGI data incorporated in a certainty stratum divided by the variance of the estimated total in the absence of VGI (see equations 8 and 9). Ratios are provided for different combinations of $k=N_2/N$ (the proportion of the region of interest covered by VGI) and $f=n/N$ (proportion of the study region covered by the simple random sample).

	$f = n/N$				
k	0.00	0.01	0.05	0.10	0.25
0.01	0.99	0.99	0.99	0.99	0.99
0.05	0.95	0.95	0.95	0.95	0.94
0.10	0.90	0.90	0.90	0.89	0.88
0.25	0.75	0.75	0.74	0.74	0.71
0.50	0.50	0.50	0.49	0.47	0.41
0.75	0.25	0.25	0.23	0.20	0.00
0.90	0.10	0.10	0.07	0.00	0.00

Equation (9) and the results of Table 2 can be used to examine the benefit of VGI arising from photographs contributed by volunteers (Antoniou et al. 2016), a common source of VGI for land-cover studies. Suppose we assume a photograph to be representative of a 30 m x 30 m pixel and consider a region of interest that covers 8 million km^2 (roughly the size of the conterminous United States, excluding Alaska and Hawaii). This region would have approximately N = 9 billion pixels. To achieve a 5% reduction in the standard error of the estimated area of a targeted class (i.e., $\sqrt{R}$ changes from 1 to 0.95) the certainty stratum approach would require $k=N_2/N=0.05$ which translates to needing N_2 = 450 million photographs. As a second example, suppose the target region of interest covers 100,000 km^2 (area slightly larger than Portugal). This population would have N = 100 million pixels (30 m x 30 m) so for VGI data to contribute a 5% reduction in standard error we would need N_2 = 5 million photographs. Typically the VGI photographs will have to be processed to obtain the land-cover information of interest (e.g., a land-cover class). Consequently, the large number of photographs needed in these examples to achieve only a 5% reduction in standard error would require substantial computer processing capability and possibly automated methods to identify the land-cover class from the photographs. Accordingly, the response design effort to process such large numbers of photographs may make this use of VGI cost prohibitive in some applications.

The certainty stratum approach may have greater utility when the VGI data are in the form of fully mapped areas classified to a land-cover or change type (i.e., in contrast to individual, unlabeled photographs as in the previous paragraph). For example, Fonte et al. (2017b) described an application in which OSM provided land-cover information for two study areas of 100 km^2 in London and Paris. OSM coverage was 88% for the London region and 97% for the Paris region. Because of the substantial portion of area covered by OSM (k=0.88 for London and k=0.97 for Paris) a large reduction in standard error of accuracy and area estimates would be expected by using these OSM data in the certainty stratum approach. For example, if k=0.88 and f=0.1 (the London example), we obtain R=0.00266 ($\sqrt{\mathrm{R}}$=0.05) indicating that the standard error of the certainty stratum estimator would be 5% of the standard error of the estimated area when not using the VGI from OSM. Obviously the areas of the regions of interest for the OSM examples in this paragraph are much smaller than for the examples in the previous paragraph and k would surely be smaller if OSM were to be used for national estimates.

4.2 Use of VGI in a Model-Assisted Estimator

Brus and de Gruijter (2003) developed an approach to use data from a non-probability sampling design to produce estimates within the design-based inference framework. In this approach, a spatial interpolation method is applied to the non-probability sample of VGI data to construct an auxiliary variable for all N elements of the population. The auxiliary variable is then used in a model-assisted estimator to achieve a reduction in standard error. Model-assisted estimators represent a broad class of estimators in which one or more auxiliary variables are incorporated in the estimator. Common examples of model-assisted estimators include difference, ratio, and regression estimators as well as post-stratified estimators (Särndal et al. 1992; Gallego 2004; Stehman 2009; McRoberts 2011; Sannier et al. 2014). The auxiliary variables are expected to covary with the target variable of interest and the information in the auxiliary variables, when incorporated in the model-assisted estimator, thus serves to reduce standard errors (Särndal et al. 1992, Chapter 6).

The Brus and de Gruijter (2003) approach could be applied to VGI as follows. Consider the objective of estimating the proportion of area of a class (e.g., area of forest) based on the reference condition. Suppose the spatial unit of the analysis is a pixel and the VGI data consist of N_2 pixels labeled as forest or non-forest. The Brus and de Gruijter (2003) approach uses these VGI data to construct an auxiliary variable x_u for all N pixels in the population. For example, for a binary classification of forest / non-forest, the auxiliary variable would be defined as x_u=1 if the class is forest and x_u=0 if the class is non-forest. The auxiliary variable x_u is known for the N_2 pixels comprising the VGI, and the Brus and de Gruijter (2003) approach would then implement a spatial interpolation method such as indicator kriging (e.g., Isaaks and Srivastava 1989) to predict values of x_u for the N-N_2 pixels not included in the VGI subset of the population. The binary forest / non-forest classification of the region predicted from the VGI data could be used in the same manner as auxiliary data from any forest / non-forest map. For example, to estimate the proportion of area of forest based on the reference condition (y_u), a probability sample from all N pixels would be selected for which the reference class of each sampled pixel would be obtained. If the reference observation is also a binary forest / non-forest classification (i.e., y_u=1 if the reference condition is forest, y_u=0 otherwise), an error matrix could be estimated from the sample based on the reference class data and the map classification of forest or non-forest created from the VGI data. The error matrix information could then be combined with the VGI generated forest / non-forest map information to produce a post-stratified estimator of the proportion of area (Card 1982; Stehman 2013). The expectation is that the auxiliary variable created from the VGI would yield a reduction in standard error of the post-stratified estimator relative to an estimator that did not incorporate the VGI. That is, the map generated via spatial interpolation of the VGI data would be used in the same way that a forest / non-forest map derived from remotely sensed data would be used in a post-stratified estimator.

The Brus and de Gruijter (2003) method requires a probability sample to provide the reference data (y_u) for the accuracy and area estimates. This probability sample must be selected from the full population of N units, including those units for which VGI is available. In contrast, the certainty stratum use of VGI (section 4.1) does not require a sample from the subpopulation G that has VGI. The Brus and de Gruijter (2003) approach does not use the VGI data as the observed response (i.e., the reference data value, y_u) so the quality of the class labels associated with the VGI data will not impact the estimates in terms of potential bias attributable to labeling error of the VGI. However, better quality (i.e., more accurate) VGI data would likely yield a greater reduction in standard error in the same manner that a more accurate map yields a greater reduction in standard error when the map data are used in a post-stratified estimator (Stehman 2013). In the context of land-cover accuracy and area estimation applications, remote sensing information is almost always available to produce a map that would provide auxiliary information that could be used in a model-assisted estimator. Spatial interpolation of VGI using the methods described by Brus and de Gruijter (2003) provides another option for producing a map of auxiliary information, and incorporating remote sensing imagery in linear spatial models (Diggle et al. 1998) might further enhance the precision benefit of the Brus and de Gruijter (2003) approach.

To summarize, the model-assisted estimator based on spatially interpolated data does not rely on the VGI data to provide the y_u values that are the basis of the parameter estimates thus decreasing the concern with bias attributable to inaccurately labeled VGI data. Instead, the approach employs the VGI to create an auxiliary variable x_u that is then used in a model-assisted estimator to reduce the standard errors of the accuracy and area estimates. The magnitude of the reduction in standard error would depend on the quality of the VGI. While this approach would have great utility if no other auxiliary information were available, we typically have access to remotely sensed data that could be used to produce a classification that would serve the same purpose as a map derived from spatially interpolating VGI data. Consequently, for land-cover studies the primary benefit obtained by spatial interpolation of VGI may occur in circumstances where a map produced from remotely sensed data is not available.

5. Use of VGI from Non-Probability Samples

If the VGI data are the only source of reference data (i.e., there is no probability sample and unable to acquire one), it will be challenging to use these VGI data in the manner of design-based inference (Figure 3). One option for using VGI in this context is to replace the estimation weights $w_u=1/\pi_u$ (equation 3) by pseudo weights that depend on assuming the sample can be treated as though it had been obtained via a probability sampling design. For example, suppose the reference data for accuracy assessment and area estimation are land-cover interpretations extracted from a non-probability sample of photographs. If the inclusion probabilities (π_u) of the spatial units represented by these photographs are unknown, one approach to estimate totals is to assume that the VGI locations represent a stratified random sample (see Section 5.1 for details). Using this approach it is possible to construct pseudo-weights such that estimated totals will match known parameters of the population. Although this weighted estimation approach can adjust a VGI sample to achieve estimates that correspond to the correct proportional representation of the population, the question of "external validity" of the VGI data must be addressed. External validity is defined and applied in Section 5.2.

Model-based inference is a second option for using VGI data that were not obtained from a probability sampling design. The application of model-based inference to accuracy and area estimation is discussed in Section 5.3.

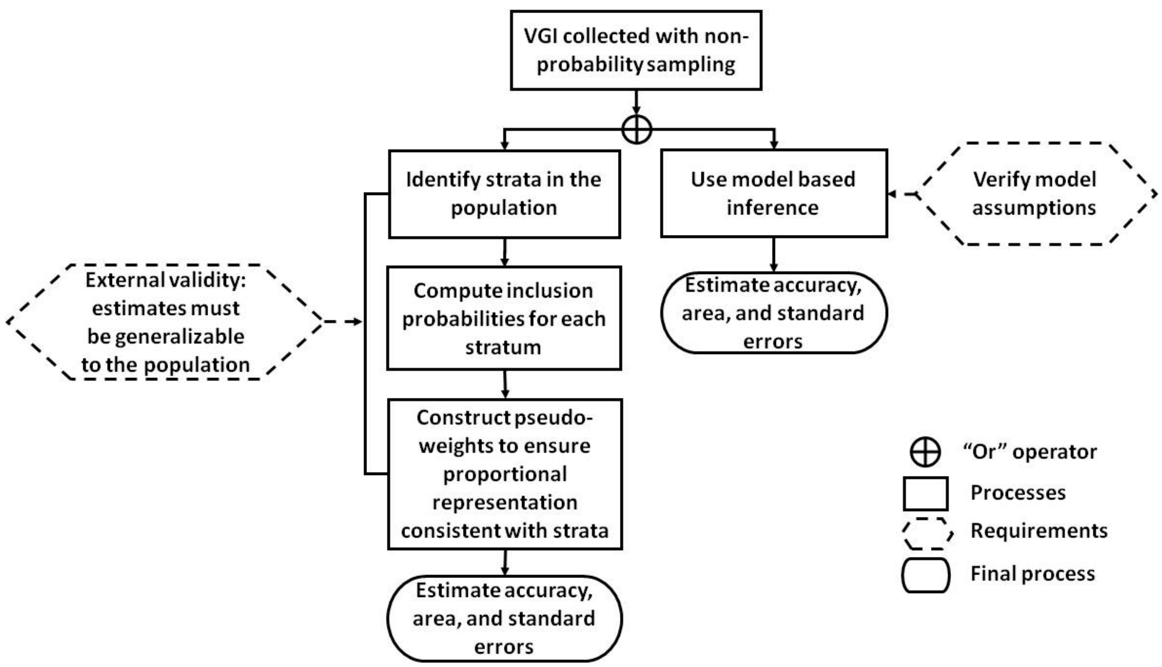

Figure 3. Schema for using VGI collected via a non-probability sampling design.

5.1 Estimation Based on Pseudo-Weights

If the only reference data available for accuracy and area estimation are VGI that did not originate from a probability sampling design, an obvious initial step in the analysis is to examine the proportional distribution of the VGI sample relative to known characteristics of the population. For example, using a land-cover map of the study region, we could compare the proportion of the VGI data found within each land-cover class to the proportion of each class in the entire population. For the hypothetical numerical example of Table 3, the VGI sample shows preferential selection from the developed and crop classes at the expense of representation of the "other" and natural vegetation classes reflecting the relative ease of access to the classes associated with the transport network. Representativeness of the VGI data could also be assessed by examining the distribution of distances to the nearest road or distances to the nearest population center. For example, we could compare the mean distance to the nearest road for the VGI locations to the mean distance for all *N* pixels in the population. If the mean for the VGI locations was less than the mean for the population, this discrepancy would indicate preferential selection of VGI closer to a road. A relevant question is then whether this preferential selection could introduce bias because map accuracy may differ depending on proximity to a road.

Table 3. Hypothetical data illustrating evaluation of the proportional representation of VGI. The distribution of the percent area of the map classes is compared between the VGI sample (n=100) and the population (i.e., entire region) known from a land-cover map of the study region.

	Area (%)	
Map Class	VGI	Population
Developed	25	10
Crop	35	20
Natural vegetation	30	50
Other	10	20

In general, we could attempt to adjust estimates to account for recognized non-proportionality of the VGI data relative to known population characteristics (Dever et al. 2008). For the example data of Table 3, the difference between the distribution of the VGI and population data suggests that weighting the data to adjust for this discrepancy would be a good idea when producing estimates. One approach would be to construct weights such that the estimates based on the weighted analysis of the VGI data correspond to known population quantities. A simple way to achieve this is to treat the non-probability sample as having arisen from a stratified design (e.g., Loosveldt and Sonck 2008). Inclusion probabilities for each stratum are then defined as $\pi_u = n_h/N_h$ where n_h is the observed sample size (from the VGI sample) in stratum *h* and N_h is the population size in stratum *h*. The estimation weight for pixel *u* is then $w_u = 1/\pi_u$, and these weights could be used in the Horvitz-Thompson estimator. These stratified estimation pseudo-weights for the hypothetical data of Table 3 are presented in Table 4. Referring to weights constructed in this manner as "pseudo-weights" highlights the fact that they are not derived from inclusion probabilities generated by a probability sampling protocol.

Table 4. Pseudo-weights for VGI sample units based on distributions by class shown in Table 3 (n_h and N_h represent the number of pixels for each class in the VGI sample and in the population).

	n_h	N_h	
Class	VGI	Map	$w_u = N_h/n_h$
Developed	25	1000	40
Cultivated	35	2000	57
Natural veg	30	5000	167
Other	10	2000	200
Total	100	10000	

To illustrate how the stratified estimation approach using pseudo-weights is implemented, consider estimating the proportion of area mapped as the developed class. From Table 3, we know this proportion is 0.10 because we have the map for the entire population. How well does the VGI sample estimate this parameter? We observe that 25 out of 100 VGI pixels are mapped as developed so the estimated proportion of mapped developed is then 0.25 from the VGI data, greater than the known parameter of 0.10 for the population. To produce the estimator using the stratified pseudo-weights of Table 4 we define y_u=1 if the sample pixel has the map label of developed and y_u=0 otherwise. Then for the developed class stratum, y_u=1 for all 25 sample pixels and each of these pixels has a weight of w_u=40, so the estimated total contributed from this stratum is 40 x 25 = 1,000 pixels (using equation 3). For the other three strata, y_u=0 for all sample pixels so these strata contribute no additional pixels to the estimated number of mapped developed pixels. Dividing the estimated total number of map pixels labeled as developed (1,000) by the number of pixels in the population (N=10,000) yields an estimated proportion of 0.10 which matches the population proportion of mapped developed area from Table 3. Thus the sample estimate using the pseudo-weights matches this known population proportion.

In general, the pseudo-weights can be constructed so that the sample estimates will equal known population values. In the example of Table 4, the pseudo-weights reproduce the known values N_h=population size of each stratum, a property known as "proportional representation." These same estimation pseudo-weights are then applied to estimate the target population parameters and the assumption is that estimation weights that effectively adjust the VGI sample data to match known population parameters will also work well when estimating the target parameters for which we do not have full population information. Other more complex methods for creating estimation weights include raking, general calibration estimators (Deville and Särndal 1992), and propensity scores (Valliant and Dever 2011). Models can be used to produce the pseudo-weights used in lieu of weights that are the inverse of the inclusion probabilities of a probability sampling design, but Valliant (2013, p.108) points out that this approach has not yielded promising results because the models are weak and the requirements excessive for covariates to be used in the models.

5.2 External validity

Pseudo-estimation weights can be used to produce estimates that capture the proportional distribution of known population characteristics (i.e., covariates). However, another important aspect of representativeness of non-probability sample data is external validity, defined as the parameter estimates being "generalizable outside the sample, say to the population of interest" (Dever and Valliant 2014). For the pseudo-weight estimation approach described in the previous section, establishing external validity would require that accuracy for the subset of the population represented by the VGI locations be equivalent to accuracy of the full region. Proportional representation of the estimates (Table 4) produced from non-probability sample data is one aspect of external validity, but proportional representation is not sufficient to establish external validity (Dever and Valliant 2014).

External validity may also require establishing that the population represented by the VGI is the same as the population of the full study region. Two examples are provided to illustrate this practical issue. In both examples, the objective is to estimate the accuracy of a map. For the first example, suppose that volunteers avoid locations of complex land cover and provide reference data exclusively for locations that are surrounded by homogeneous land cover. Antoniou et al. (2016) suggest such a strategy may be beneficial when using photographs to avoid difficulties of determining the ground condition. Because homogeneous regions are typically more likely to be classified correctly, the accuracy estimates produced from such data would be expected to have higher accuracy than is true of the study region as a whole. Consequently external validity of these data would be suspect because the estimates based on the non-probability sample would not be generalizable to the target population. As a second example, suppose because of convenient access the VGI data have been collected primarily at locations near roads. Evaluating external validity would then require determining whether accuracy near roads was equivalent to accuracy distant from roads.

Verifying external validity of VGI may be extremely challenging and in some cases impossible (Dever and Valliant 2014). Verification requires comparing characteristics of the VGI data with characteristics of the full study region. Consider the example of VGI data concentrated along roads. To establish that accuracy does not vary with distance from a road, we could collect additional reference data distant from roads based on a probability sampling design, and compare the accuracy estimates from this sample to accuracy estimates for sample data constrained to locations near roads. But the additional effort to obtain the sample data distant from roads would negate much of the value of VGI for reducing the cost of accuracy assessment. That is, to definitively establish the equivalence of accuracy near roads to accuracy distant from roads, we may need a large probability sample, and the primary value of VGI is to reduce the cost and effort of collecting sample data.

Alternatively, it may be possible to cite previous studies to establish external validity. For example, if previous research has demonstrated that distance from a road is not strongly related to accuracy, we would have some assurance of external validity to support use of VGI data collected preferentially near roads. In general, to more fully exploit the potential benefit of VGI, it may be necessary to document typical features of VGI that would commonly need to be addressed to establish external validity and then conduct the necessary studies to inform the decision of whether external validity is tenable. Distance from road, characteristics of volunteers, and complexity of landscape are just a few examples of features that might be explored to determine whether characteristics of populations (e.g., accuracy) differ by these features. If in general there are no such differences, external validity of non-probability sample data is supported to some degree. Developing a cohesive strategy to design and conduct such studies for a broadly applicable assessment of external validity of VGI would likely require a major research initiative.

5.3 VGI and Model-Based Inference

Model-based inference is not predicated on probability sampling so it is a potentially attractive option for using VGI data that did not originate from a probability sampling design. Model-based inference requires specification of a model that relates y_u to a set of covariates (predictors) available for the full population (Valliant et al. 2000). Developing appropriate models and evaluating the underlying assumptions may be difficult and time-consuming (Baker et al. 2013) with the difficulties exacerbated by the fact that in most surveys, numerous estimates are produced from a single sample. In the case of VGI, estimates of accuracy and area for several land-cover or land-cover change types will typically be of interest, and each of these estimates may be desired for several subregions within the target region of interest. A model will need to be developed and assumptions evaluated for all estimates as a model that works well for some estimates may not work well for others. An additional challenge to the model-based approach is that non-probability samples may have an inherent selection bias, so a substantial risk exists that the distribution of important covariates in the sample will differ from the distribution of these covariates in the target population (Baker et al. 2013). Methods to account for preferential sampling (e.g., Diggle et al. 2010) in a model-based framework may be considered in such cases of non-probability sampling.

Numerous model-based methods can be applied to non-probability samples and evaluating the utility of model-based methods is case specific because it is difficult to ascribe general properties to these methods (Baker et al. 2013). An advantage of probability sampling and design-based inference is that a standard general approach is used to produce the complete array of estimates (see Section 2.1). Yet another challenge of model-based inference and non-probability sampling is how to define and quantify uncertainty. A widely accepted measure of precision does not exist for estimates from non-probability samples (Baker et al. 2013, p.97), whereas the standard error (or appropriately scaled version of standard error) is generally accepted for quantifying precision of estimates in design-based inference. Clearly, some of the cost savings achieved by non-probability sampling is lost due to the more complex analyses needed to develop models and test their assumptions (Baker et al. 2013). Because model-based inference encompasses an array of methods, establishing transparency of the methodology is also more demanding because it is necessary to describe the specific model-based approach used and the possible limitations of inference uniquely associated with that approach (Baker et al. 2013, p.100).

6. Discussion

The increasing availability of large quantities of data obtained via non-probability sampling has garnered interest of survey methodologists in a variety of subject areas, so it is relevant to examine issues addressed in the broader survey sampling literature that go beyond just use of VGI in the remote sensing context. For example, internet surveys comprised of volunteer opt-in panels that use social media to extract information result in large quantities of data that are obtained quickly and conveniently but via a selection protocol that has no underlying probability sampling design. Review articles by Baker et al. (2013) and Elliott and Valliant (2017) provide an excellent general overview of methods and issues affecting inference when using data from such non-probability samples. In the broad context of survey sampling, the conventional practice of relying on design-based inference has been questioned because of the tremendous increase in non-response rates. Even if a probability sampling design is implemented, severe non-response will make the application of design-based inference questionable (Baker et al. 2013). Fortunately, in land-cover studies non-response is generally not a major problem. The availability of remote sensing platforms usually allows us to obtain the necessary observations that might otherwise be very difficult if a ground visit were required. Non-response rates are typically very small in accuracy assessment and area estimation applications so the dilemma of severe non-response that impacts current survey practice in other fields of application is typically not a problem in land-cover studies.

Ensuring accurate observations (y_u) is perhaps the most challenging aspect of using VGI because it depends on the volunteers to provide good quality data. Accurate interpretation of reference labels for land cover or land-cover change is challenging even for trained experts so label quality of VGI data needs to be scrutinized closely. A great deal of effort has been invested in improving and evaluating the quality of VGI used in land-cover studies, including the assessment of traditional quality measures such as positional, thematic or temporal accuracy (Fonte et al. 2017a), the development of new quality indicators that are applicable specifically to VGI (Meek et al. 2014; Antoniou and Skopeliti 2015; Senaratne et al. 2017), and even combinations of indicators (Bishr and Mantelas 2008; Jokar Arsanjani et al. 2015). The investment in these methods will not only yield better quality VGI data but may also contribute to improved data quality and assessment procedures applicable to reference data obtained by experts.

Baker et al. (2013) make the helpful distinction between "describers" whose purpose is to describe the population and "modelers" whose purpose is to characterize relationships between variables. Accuracy assessment and area estimation applications typically fall within the "describer" class because of the strong focus on descriptive parameters such as user's and producer's accuracies of the different classes and the area or proportion of area of the land-cover or land-cover change classes. Describers generally rely on probability sampling because of the importance of representing the target population. Elliott and Valliant (2017, p.262) provide a strong statement in support of probability sampling for descriptive objectives: "… when critical estimates of descriptive quantities such as means, quantiles or cell probabilities are required, nonprobability designs should be avoided or utilized only when it is reasonably certain that there are available covariates in both datasets related to the nonprobability selection mechanism that can be used to appropriately incorporate information from the nonprobability sample. If a sufficiently large probability sample is available for estimating descriptive statistics, methods to incorporate nonprobability data are likely not warranted."

Although design-based inference requires a probability sampling design, it is not reasonable to assert a recommendation that probability sampling must always be used. Other considerations such as cost and “fit for purpose “may be relevant, the latter including dimensions such as “accuracy, timeliness, and accessibility” (Baker et al. 2013, p. 98). A quote from Kish (1965, pp. 28-29) extracted by Baker et al. (2013, p.92) has direct bearing on this issue: “No clear rule exists for deciding exactly when probability sampling is necessary, and what price should be paid for it … Probability sampling for randomization is not a dogma, but a strategy, especially for large numbers.” Probability sampling offers the strong advantage that it provides the basis for rigorous design-based inference, but there may be exceptional cases in which fit for purpose criteria will be such that VGI from a non-probability sample will suffice. While an unmistakable conclusion from our assessment of VGI for use in design-based inference is that probability sampling should be used, we recognize that occasionally circumstances may exist where not following this recommendation is justifiable.

VGI has great potential value within remote sensing beyond its use to produce accuracy and area estimates within design-based inference. For example, VGI can greatly augment traditional sources of training data used in the classification algorithms of land cover and land use maps. The exact design of the training stage of a supervised classification should, however, be highly classifier-specific as classifiers vary greatly in how they use the training set. While conventional statistical classifiers may benefit from the use of a probability sample in the acquisition of training statistics to obtain a representative and unbiased description of each class, other classifiers, such as machine learning classifiers, may require only very small and distinctly non-random sample. Thus, for example, an effective approach to training data acquisition for a classification by a support vector machine may be to direct citizens to a small number of highly atypical training sites (Pal and Foody 2012). Classifiers also vary in their sensitivity to mis-labeling of training cases (Foody et al. 2016) which may be relevant if VGI is to be used.

Land cover data from several Geo-Wiki campaigns are now available in the openly accessible repository Pangaea and these data could be used as training data (Fritz et al. 2017; Laso Bayas et al.

2017). VGI is also useful in the development of hybrid land-cover maps, where methods such as geographically weighted regression can use VGI to determine the most appropriate land cover class at a given location among several existing products. Such an approach has been demonstrated in the development of global land cover and forest masks (Schepaschenko et al. 2015; See et al. 2015). Finally, VGI can provide a preliminary check on the accuracy of a land-cover product and guide the collection of additional training data in areas where there is visual evidence of confusion between land-cover classes.

7. Summary

The increasing availability and quantity of VGI has generated great interest in how these data might be used in applications requiring land-cover data, specifically area estimation and map accuracy assessment. Scientifically credible use of VGI raises many of the same issues related to inference that McRoberts (2011) discussed pertaining to use of land-cover maps, stating that "…rules must be rigorously followed to produce valid scientific inferences." The requirements for using VGI in rigorous design-based inference are identifiable from the analysis protocol (Sec. 3.1) used to produce the area and map accuracy estimates. Specifically, the estimates are derived from totals, and the Horvitz-Thompson estimator provides an unbiased estimator of a population total if the response design generates accurate observation of the attribute or measurement of interest (y_u) and the sampling design is such that the inclusion probabilities (π_u) are known. If y_u is accurate and π_u is known then we can produce unbiased estimators of the totals that form the basis for accuracy and area estimates. We reviewed recent literature describing methods for obtaining VGI and assessing its quality (Sec. 3.2), and we anticipate that ongoing research will improve reference data quality whether collected by volunteers within a VGI framework or by expert interpreters.

The primary focus of this article has been on the sampling design issues related to using VGI in design-based inference, with attention addressing three primary cases: 1) VGI data are from a probability sampling design; 2) VGI data from a non-probability sampling design are combined with data from a probability sampling design; and 3) the only data available are VGI data from a non-probability sampling design. The most direct approach to ensure that design-based inference can be invoked is to specify that the VGI data will be collected at locations (sample units) selected by a probability sampling design ("active VGI"). Implementing a probability sampling design ensures that the inclusion probabilities (π_u) for the sampled units are known and thus the corresponding estimation weights (w_u=1/π_u) required for the analysis are known. The more common situation is that the VGI data do not originate from a probability sampling design. Implementing design-based inference in this situation requires combining the VGI data with data obtained from a probability sampling design, and the benefit of the VGI data is to reduce the standard errors of the accuracy or area estimates. Two approaches for combining VGI with a probability sample are to treat the VGI as a certainty stratum (i.e., set π_u=1 for each unit from the VGI sample) or to use the VGI to create an auxiliary variable for the population and incorporate this variable in a model-assisted estimator. The certainty stratum approach is the more promising of these two options particularly if a large proportion of the population is covered by VGI. For land-cover studies the model-assisted estimator use of VGI likely will also incorporate maps produced from remote sensing imagery.

If VGI data collected from a non-probability sampling design are the only data available, rigorous design-based inference is not available. Estimates of accuracy and area can be produced using the same estimator formulas of design-based inference by defining pseudo-estimation weights based on treating the VGI as if a stratified random sample had been implemented. Estimates produced in this fashion mimic the proportional representation of the feature of the population used to create the pseudo-weights. However, in contrast to the case where the weights are the inverse of known inclusion probabilities from a probability sampling design, the estimates based on pseudo-weights require the additional step of verifying that the condition of external validity is satisfied. External validity requires that the population for which the VGI data are representative must have the same characteristics (e.g., model relationships) as the full population that is the target of inference. Establishing external validity is often impractical so the pseudo-weight approach to using VGI from a non-probability sample will have limited utility. Model-based inference is perhaps the more promising avenue for using VGI from non-probability samples. Explication of model-based methods and specific example applications of accuracy and area estimation (McRoberts 2006; Magnussen 2015) are needed to make model-based inference more accessible to practitioners.

Invoking design-based inference as the scientific basis to support the validity of inference for estimating area and map accuracy from sample data imposes the requirement that the sampling and estimation protocols implemented must satisfy certain conditions. As is apparent from the methods and discussion presented in this article, the requirement of a probability sampling design places fairly strong restrictions on how VGI can be used in design-based inference. The methods presented in this article for incorporating VGI in design-based inference expand the potential utility of this growing body of data for applications of accuracy assessment and area estimation.

Acknowledgments

This research was supported in part by the Portuguese Foundation for Science and Technology (FCT) under project grant UID/MULTI/00308/2013 (CF); EU-funded FP7 project CrowdLand No. 617754 and the Horizon2020 LandSense project No. 689812 (LS); Cooperative Agreement G12AC20221provided by the United States Geological Survey and NASA Carbon Monitoring System program grant NNX13AP48G (SS). We thank the reviewers for their constructive comments that led to improvements in the manuscript.

References

Antoniou, V., Morley, J., and Haklay, M. (2010). Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica*, *64*, 99–110.

Antoniou, V., Fonte, C. C., See, L., Estima, J., Arsanjani, J. J., Lupia, F., Minghini, M., Foody, G., and Fritz, S. (2016). Investigating the feasibility of geo-tagged photographs as sources of land cover input data. *ISPRS International Journal of Geo-Information*, 5, 64.

Antoniou, V., and Skopeliti, A. (2015). Measures and indicators of VGI quality: An overview. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Presented at the ISPRS Geospatial Week 2015, ISPRS Annals, La Grande Motte, France, pp. 345–351.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013). Summary report of the AAPOR Task Force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-105.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F., Pecl, G. T., Barrett, N., and Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, 173, 144-154.

Bishr, M., and Mantelas, L. (2008). A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal*, 72, 229–237.

Brus, D. J., and de Gruijter, J. J. (2003). A method to combine non-probability sample data with probability sample data in estimating spatial means of environmental variables. *Environmental Monitoring and Assessment*, 83, 303-317.

Card, D. H. (1982). Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering & Remote Sensing*, 48, 431-439.

Ciepłuch, B., Jacob, R., Mooney, P., and Winstanley, A. (2010). Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, 20-23 July, 2010, 337.

Clark, M.L. and Aide, T.M. (2011). Virtual Interpretation of Earth Web-Interface Tool (VIEW-IT) for collecting land-use/land-cover reference data. *Remote Sensing,* 3, 601–620.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. (New York: John Wiley & Sons).

De Gruijter, J. J., and Ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology*, 22, 407-415.

Dever, J. A., Rafferty, A., and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods,* 2, 47-62.

Dever, J. A., and Valliant, R. (2014). Estimation with non-probability surveys and the question of external validity. *Proceedings of Statistics Canada Symposium 2014*, 8 pp.

Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Diggle, P., Menezes, R., and Su, T. (2010). Geostatistical inference under preferential sampling. *Applied Statistics*, 59, 191-232.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Applied Statistics*, 47, 299-350.

Elliott, M. R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 249-264.

Elwood, S., Goodchild, M.F., and Sui, D. (2013). Prospects for VGI research and the emerging fourth paradigm. *Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice*, Springer, Netherlands, pp. 361-375.

Fonte, C. C., Bastin, L., See, L., Foody, G., and Lupia, F. (2015). Usability of VGI for validation of land cover maps. *International Journal of Geographical Information Science*, 29, 1269-1291.

Fonte, C.C., Antoniou, V., Bastin, L., Estima, J., Arsanjani, J.J., Laso-Bayas, J.-C., See, L., and Vatseva, R. (2017a). Assessing VGI data quality, in: Foody, G.M., See, L., Fritz, S., Fonte, C.C., Mooney, P., Olteanu-Raimond, A.-M., Antoniou, V. (Eds.), Mapping and the Citizen Sensor. Ubiquity Press, London, UK, pp. 137-164.

Fonte, C. C., Patriarca, J. A., Minghini, M., Antoniou, V., See, L., and Brovelli, M. A. (2017b). Using OpenStreetMap to create land use and land cover maps: Development of an application. In: *Volunteered Geographic Information and the Future of Geospatial Data* (Ed.: Campelo, C. E. C., Bertolotto, M., and Corcoran, P.). IGI Global. DOI: 10.4018/978-1-5225-2446-5.ch007

Foody, G. M. (2009). The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing*, 30, 3275-3281.

Foody, G.M. (2010). Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment*, 114, 2271-2285.

Foody, G.M., Pal, M., Rocchini, D., Garzon-Lopez, C.X. and Bastin, L. (2016). The sensitivity of mapping methods to reference data quality: Training supervised image classifications with imperfect reference data. *ISPRS International Journal of Geo-Information*, *5*(11), p.199.

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

Foody, G.M., See, L., Fritz, S., Velde, M. van der, Perger, C., Schill, C., Boyd, D.S., and Comber, A. (2015). Accurate attribute mapping from Volunteered Geographic Information: Issues of volunteer quantity and quality. *The Cartographic Journal*, 52, 336–344.

Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., van der Velde, M., Kraxner, F., and Obersteiner, M. (2012). Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling & Software*, 31, 110–123.

Fritz, S., See, L., van der Velde, M., Nalepa, R.A., Perger, C., Schill, C., McCallum, I., Schepaschenko, D., Kraxner, F., Cai, X., Zhang, X., Ortner, S., Hazarika, R., Cipriani, A., Di Bella, C., Rabia, A.H., Garcia, A., Vakolyuk, M., Singha, K., Beget, M.E., Erasmi, S., Albrecht, F., Shaw, B., and Obersteiner, M. (2013). Downgrading recent estimates of land available for biofuel production. *Environmental Science and Technology*, 47, 1688–1694.

Fritz, S., See, L., Perger, C., McCallum, I., Schill, C., Schepaschenko, D., Duerauer, M., Karner, M., Dresel, C., Laso Bayas, J.C., Lesiv, M., Moorthy, I., Salk, C., Danylo, O., Sturn, T., Albrecht, F., You, L., Kraxner, F. and Obersteiner, M. (2017). A global dataset of crowdsourced land cover and land use reference data. *Scientific Data*, 4:170075. doi: 10.1038/sdata.2017.75.

Gallego, F. J. (2004). Remote sensing and land cover area estimation. *International Journal of Remote Sensing*, 25, 3019-3047.

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

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

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

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

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Hou, D., Chen, J., Wu, H., Li, S., Chen, F., and Zhang, W. (2015). Active collection of land cover sample data from geo-tagged web texts. *Remote Sensing*, 7, 5805-5827.

Isaaks, E. H., and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, New York.

Iwao, K., Nishida, K., Kinoshita, T., and Yamagata, Y. (2006). Validating land cover maps with Degree Confluence Project information. *Geophysical Research Letters*, 33, L23404, doi:10.1029/2006GL027768.

Jokar Arsanjani, J. and Bakillah, M. (2015). Understanding the potential relationship between the socio-economic variables and contributions to OpenStreetMap. *International Journal of Digital Earth*, 8, 861–876.

Kish, L. (1965). *Survey Sampling*. Wiley, New York.

Laso Bayas, J.C., Lesiv, M., Waldner, F., Schucknecht, A., Duerauer, M., See, L. et al. (2017). A global reference database of crowdsourced cropland data collected using the Geo-Wiki platform. *Scientific Data*. 4: 170136. doi:10.1038/sdata.2017.136.

Laso Bayas, J.-C., See, L., Fritz, S., Sturn, T., Perger, C., Duerauer, M., Karner, M., Moorthy, I., Schepaschenko, D., Domian, D., and McCallum, I. (2017). Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology. *Remote Sensing* 8(11), 905. doi:10.3390/rs8110905.

Loosveldt, G., and Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods,* 2, 93-105.

Magnussen, S. (2015). Arguments for a model-dependent inference? *Forestry*, 88, 317-325.

Martino, L., Palmieri, A., and Gallego, J. (2009). Use of auxiliary information in the sampling strategy of a European area frame agro-environmental survey. Available from: http://ec.europa.eu/eurostat/documents/205002/769457/LUCAS2009_S2-Sampling_20090000.pdf. Eurostat.

McRoberts, R. E. (2006). A model-based approach to estimating forest area. *Remote Sensing of Environment*, 103, 56-66.

McRoberts, R. E. (2010). Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. *Remote Sensing of Environment*, 114, 1017-1025.

McRoberts, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sensing of Environment*, 115, 715–724.

Meek, S., Jackson, M.J., and Leibovici, D.G. (2014). A flexible framework for assessing the quality of crowdsourced data. In: Huerta, J., Schade, S., and Granell, C. (Eds.), *Connecting a Digital Europe through Location and Place: Proceedings of the AGILE'2014 International Conference on Geographic Information Science*. Presented at the AGILE'2014 International Conference on Geographic Information Science, AGILE Digital Editions, Castellón, Spain.

Meek, S., Jackson, M., and Leibovici, D.G. (2016). A BPMN solution for chaining OGC services to quality assure location-based crowdsourced data. *Computers & Geosciences*, 87, 76–83.

Mullen, W. F., Jackson, S. P., Croitoru, A., Crooks, A., Stefanidis, A., and Agouris, P. (2014). Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features. *GeoJournal*, 80, 587-605.

Olofsson, P., Foody, G. M., Stehman, S. V., and Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129, 122-131.

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., and Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42-57.

Overton, J. M., Young, T. C., and Overton, W. S. (1993). Using 'found' data to augment a probability sample: Procedure and case study. *Environmental Monitoring and Assessment*, 26, 65-83.

Pal, M. and Foody, G.M. (2012). Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5, 1344-1355.

Sannier, C., McRoberts, R. E., Fichet, L.-V., & Makaga, E. M. K. (2014). Using the regression estimator with Landsat data to estimate proportion forest cover and net proportion deforestation in Gabon. *Remote Sensing of Environment*, 151, 138-148.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer-Verlag, New York.

Schepaschenko, D., See, L., Lesiv, M., McCallum, I., Fritz, S., Salk, C., Moltchanova, E., Perger, C., Shchepashchenko, M., Shvidenko, A., Kovalevskyi, S., Gilitukha, D., Albrecht, F., Kraxner, F., Bun, A., Maksyutov, S., Sokolov, A., Dürauer, M., Obersteiner, M., Karminov, V., and Ontikov, P. (2015). Development of a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics. *Remote Sensing of Environment*, 162, 208–220.

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F., and Obersteiner, M. (2013). Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLoS ONE* 8, e69958. doi:10.1371/journal.pone.0069958

See, L., Fritz, S., Perger, C., Schill, C., McCallum, I., Schepaschenko, D., Duerauer, M., Sturn, T., Karner, M., Kraxner, F., and Obersteiner, M. (2015). Harnessing the power of volunteers, the internet and Google Earth to collect and validate global spatial information using Geo-Wiki. *Technological Forecasting and Social Change* 98, 324–335.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., and Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, *5*(5), 55. doi: 10.3390/ijgi5050055

See, L., Laso Bayas, J.C., Schepaschenko, D., Perger, C., Dresel, C., Maus, V., Salk, C., Weichselbaum, J. Lesiv, M., McCallum, I., Moorthy, I. and Fritz, S. (2017). LACO-Wiki: A new online land cover validation tool demonstrated using GlobeLand30 for Kenya. *Remote Sensing* 9(7), 754; doi:10.3390/rs9070754.

Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C., and and Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31, 139-167.

Stehman, S. V. (1999). Basic probability sampling designs for thematic map accuracy assessment. *International Journal of Remote Sensing*, 20, 2423-2441.

Stehman, S. V. (2000). Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sensing of Environment*, 72, 35-45.

Stehman, S. V. (2009). Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. *Remote Sensing of Environment*, 113, 2455-2462.

Stehman, S. V. (2013). Estimating area from an accuracy assessment error matrix. *Remote Sensing of Environment*, 132, 202-211.

Stehman, S. V., and Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment*, 64, 331-344.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., Mayaux, P., Morisette, J. T., Stehman, S. V., and Woodcock, C. E. (2006). Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps, EUR 22156 EN – DG, Office for Official Publications of the European Communities, Luxembourg, 48 pp.

Tsendbazar, N.E., de Bruin, S., and Herold, M. (2015). Assessing global land cover reference datasets for different user communities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103, 93–114.

Valliant, R. (2013). Comment on "Summary report of the AAPOR Task Force on non-probability sampling." *Journal of Survey Statistics and Methodology*, 1, 105-111.

Valliant, R., and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105-137.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons, Inc., New York.

LIST OF FIGURE CAPTIONS

Figure 1.

Figure 2.

Figure 3.