Remote Sensing of Environment 212 (2018) 47–59

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

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
[b] Department of Mathematics, University of Coimbra, P-3001 501 Coimbra, Portugal/Institute for Systems Engineering and Computers at Coimbra, Portugal
[c] School of Geography, University of Nottingham, Sir Clive Granger Building, University Park, Nottingham, NG7 2RD, United Kingdom
[d] International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria

ARTICLE INFO

Keywords:
Probability sampling
External validity
Pseudo-weights
Data quality
Model-based inference
Volunteered geographic information (VGI)
Crowdsourcing

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.

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

* Corresponding author.
E-mail addresses: svstehma@syr.edu (S.V. Stehman), cfonte@mat.uc.pt (C.C. Fonte), giles.foody@nottingham.ac.uk (G.M. Foody), see@iiasa.ac.at (L. See).

https://doi.org/10.1016/j.rse.2018.04.014
Received 29 October 2017; Received in revised form 18 February 2018; Accepted 8 April 2018

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 (Fig. 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.

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

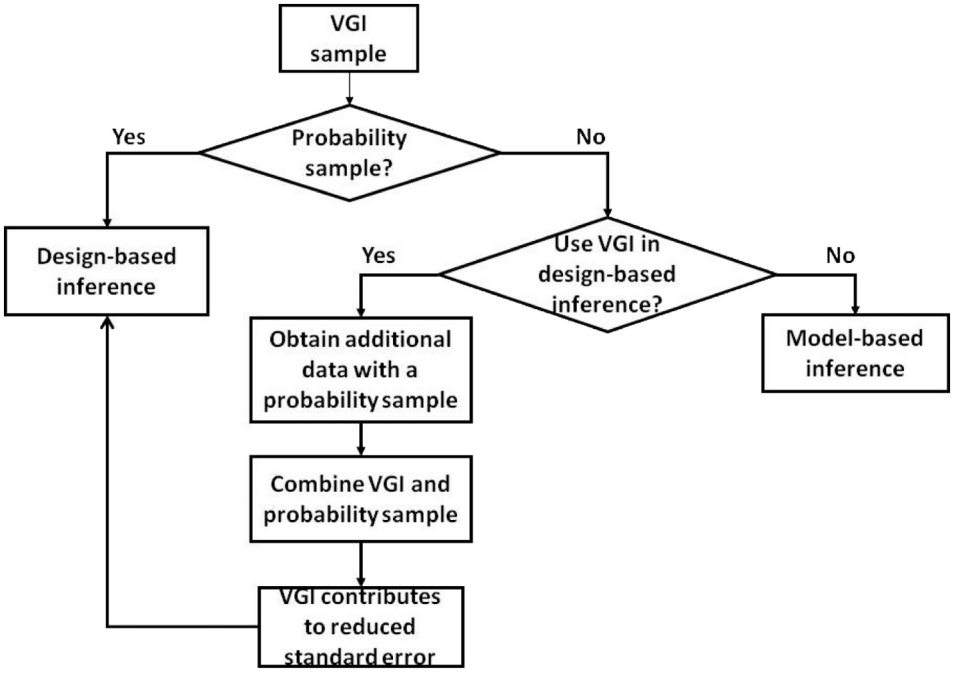

Fig. 1. Schema for methodologies using VGI in accuracy assessment and area estimation.