Remote Sensing of Environment 198 (2017) 150–159

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

# Effects of temporally external auxiliary data on model-based inference

Zhengyang Hou [a,b,*], Qing Xu [b,c], Ronald E. McRoberts [d], Jonathan A. Greenberg [b], Jinxiu Liu [e], Janne Heiskanen [e], Sari Pitkänen [a], Petteri Packalen [a]

[a] University of Eastern Finland, Faculty of Science and Forestry, School of Forest Sciences, P.O. Box 111, FI-80101 Joensuu, Finland
[b] University of Nevada, Reno, Natural Resources & Environmental Science, Reno, NV 89667, United States
[c] University of Illinois at Urbana-Champaign, Department of Geography and Geographic Information Science, IL 61801, United States
[d] Northern Research Station, U.S. Forest Service, Saint Paul, MN, United States
[e] University of Helsinki, Department of Geosciences and Geography, P.O. Box 64, FI-00014 Helsinki, Finland.

ARTICLE INFO

*Article history:*
Received 16 February 2017
Received in revised form 21 May 2017
Accepted 7 June 2017
Available online xxxx

*Keywords:*
Natural resource inventory
Model-based inference
Sampling
Uncertainty
Bootstrapping
Covariance matrix estimator

ABSTRACT

One of the benefits of model-based inference relative to design-based inference is that probability samples are not required which means that models can be constructed using data external to the area of interest. Although "external" usually means spatially or geographically external, it could also be used in the temporal sense that the model is constructed using data whose dates are temporally external to the dates of the data to which the model is applied. This study focuses on assessing the effects of such temporally external application data on model-based inference using remotely sensed auxiliary information. The study area was in Burkina Faso, and the variable of interest was firewood volume ($m^3$/ha). A sample of 160 field plots was selected from the population and measured, and auxiliary datasets from Landsat 8 were acquired. Models were fit using weighted least squares; the population mean, $\mu$, was estimated; and the variance of the population mean, $Var(\hat{\mu})$, was estimated using both an analytical variance estimator, $\overline{\hat{V}}(\hat{\mu})_{an}$, and an empirical bootstrap estimator, $V(\hat{\mu})_{boot}$. The estimates, $\hat{\mu}$ and $\widehat{Var}(\hat{\mu})$, were compared for models constructed using calibration and application data of the same date and models constructed using calibration and application data whose dates differed. The primary results were two-fold. First, for cases for which the dates of the model calibration and application data were the same, $\hat{\mu}$, $\overline{\hat{V}}(\hat{\mu})_{an}$, $V(\hat{\mu})_{boot}$ and $\widehat{Bias}(\hat{\mu})$ were similar across datasets. These results suggest that the particular date of the dataset from which the calibration and application data are obtained may be mostly arbitrary assuming the relation between the dependent and independent variables does not change over time. Second, for a model for which the calibration and application data were obtained from temporally different datasets, $\overline{\hat{V}}(\hat{\mu})_{an}$, $V(\hat{\mu})_{boot}$, and $\widehat{Bias}(\hat{\mu})$ were all greater than when the calibration and application data were not temporally different. Further, the criterion for screening candidate models must be based on estimation of $\hat{\mu}$ and $\widehat{Var}(\hat{\mu})$ rather than the model prediction accuracy or goodness of fit. The adverse effects of differing dates for the calibration and application data were exacerbated as the difference in dates increased. Finally, because the temporal differences also affected the analytical variance calculation, the bootstrapping procedure is recommended.



## 1. Introduction

When multiple sets of remotely sensed auxiliary data are available for modeling a forest attribute, which set to select? The term *model* here refers to the combination of the mathematical expression representing the relationship between a dependent and independent variables, the selected independent variables, and the parameter estimates. Because different datasets often lead to different models, which model to choose? Should the criterion be goodness of fit as expressed by $R^2$, the prediction accuracy as expressed by RMSE, or confidence interval widths for the population parameter estimates? How robust is the estimation when the dates for values of the independent variables used for applying a model differ from the dates of the data used to construct the model? These sampling questions can be properly evaluated in the context of model-based inference.

In forestry, a spatial area of interest can be tessellated by smaller units of a given size that then serve as population units. The population parameters of interest are usually the population total ($\tau$) or mean ($\mu$) of the forest attribute, and development of estimators for these parameters is a common topic of sampling theory. The desired properties of

* Corresponding author at: University of Eastern Finland, Faculty of Science and Forestry, School of Forest Sciences, P.O. Box 111, FI-80101 Joensuu, Finland.
*E-mail address:* zhou@unr.edu (Z. Hou).

http://dx.doi.org/10.1016/j.rse.2017.06.013
0034-4257/

an estimator are with respect to the sampling distribution of estimates, and are primarily unbiasedness and small variance (Hou et al., 2015). Forest inventories rely heavily on sampling theory and have been established and conducted in many developed countries. Common practice is to conduct these inventories using field surveys based on specific probabilistic sampling designs that support estimators with sufficient precision. This design-based inference relies on sample sizes that are sufficiently large that the central limit theorem can be invoked to ensure the desired properties. Design-based inference is free from assumptions regarding the structure and distribution of the population, because it is based on the distribution of all possible estimates permissible under the strict terms of the sampling design (Cochran, 1977). However, a large sample consisting entirely of field measurements following a probabilistic design is apt to become unaffordable and less cost-efficient, whereas a reduced sample size risks failure to satisfy precision criteria. This dilemma could introduce a serious economic burden on developing countries and would hinder them from implementing repetitive inventories to update the status of forest resources regularly as is required today (MEDD, 2012).

Inventories enhanced by remotely sensed data have become increasingly popular. In particular, remotely sensed auxiliary data that are correlated with forest attributes facilitate use of model-based inference. Distinct from design-based inference, model-based inference relies on a model as the basis for constructing inferences in the form of confidence intervals for the population parameters (Cassel et al., 1977). The finite population is regarded as a realization of a random process called a superpopulation, and every finite population is seen as a sample of the infinite superpopulation (Särndal, 1978). This superpopulation is characterized as infinite because it refers to the distribution of all possible values for each unit in the finite population (McRoberts, 2010). The superpopulation is defined by the superpopulation model wherein the remotely sensed auxiliary information enters as independent variables. The model here refers to the estimated superpopulation model of a given form according to prior knowledge, and model parameters are estimated using sample data collected from the finite population. The true superpopulation model parameters are the only fixed values, whereas the estimated model parameters are random variables contingent on the collected sample. Properties of a model-based population parameter estimator are deduced conditionally with respect to the observed sample and the stipulated model, not the sampling design.

The importance of properly dealing with uncertainty was explicitly advocated by the Intergovernmental Panel on Climate Change in its good practice guidance for greenhouse gas reporting (IPCC, 2006). In some instances, however, it remains unclear how uncertainty assessments should be conducted. For example, McRoberts (2011) demonstrated that maps, accuracy assessments, and models do not directly produce inferences. A major concern with model-based inference is the potential for serious bias in the population parameter estimator if the presumed or stipulated model is misspecified (Gregoire, 1998). Many studies have focused on selection of estimators and sampling designs that are robust with regard to model misspecification, particularly for linear models (e.g. Breidenbach et al., 2014; Chambers and Clark, 2012; McRoberts, 2010; Saarela et al., 2015; Valliant et al., 2000). Dating back to early days, various methods such as designing the sample to be balanced in terms of independent variables (Royall and Herson, 1973a, 1973b) have been proposed to enhance inferential robustness despite model misspecification. Nowadays, requirements for statistically rigorous uncertainty estimates are steadily increasing (Gregoire et al., 2016), particularly when there are multiple sources of uncertainty.

One of the benefits of model-based inference relative to design-based inference is that probability samples are not required. Thus, models can be constructed using data external to the area of interest (McRoberts et al., 2014). Although the term *external* usually means spatially or geographically external, it can also be used in the temporal sense that the model is constructed using data whose dates are temporally external to the dates of the application data. McRoberts et al. (2016a) considered a similar problem for design-based inference, but to our knowledge the problem of temporal externality has not been considered for model-based inference. Standard regression theory assumes that model predictor variables are observed without error (Gregoire and Valentine, 2008), and that if measurement errors and model-related errors cannot be ignored, uncertainty estimates such as variance will be underestimated (Särndal et al., 1992). Therefore, the effects of temporal differences between model calibration and application data can be pernicious.

Consequently, the aims of the study were twofold: (1) to evaluate and compare how $\hat{\mu}$ and $\widehat{Var}(\hat{\mu})$ vary when temporally different remotely sensed data are used for constructing and applying a model; and (2) to summarize rules of thumb that help to reduce the uncertainty caused by an arbitrary selection of dates of auxiliary data. The forest attribute of interest is firewood volume ($m^3$/ha), the main commercial product obtained from the forests in the Burkina Faso study area.

## 2. Materials

### 2.1. Study area

The study area is situated in the rural commune of Kou in southeastern Burkina Faso (11°45′N, 1°57′W) (Fig. 1, left). The topography is a plain with low relief and mean elevation of 350 m above sea level. The soil has a sandy clay texture and mainly consists of plinthosols with a subsurface accumulation of plinthite with small nutrient content (Jonsson et al., 1999). The mean annual precipitation is 790 mm/year, and the mean annual temperature 28 °C. The climate is semi-arid and bimodal (Peel et al., 2007). The monsoonal rainy season lasts seven months from April to October and accounts for 80% of the annual precipitation received, whereas the dry spell season covers the rest of a year (Nicholson, 2009).

Controlled fires are common in the study area (Fig. 2) and are intended to promote a fresh growth of grass for the grazing herds, to make the wildlife more visible for the tourists, and to prevent more destructive wildfires later (Gessner et al., 2015). However, although the grasses and shrubs are burnt to ash, the fires do not usually cause permanent land cover changes because most trees survive and the burnt vegetation recovers quickly (Sawadogo et al., 2002). Four fire seasons have been distinguished, the pre-early season in October, the early season in November and December, the late season in January and February, and the post-late season in March and April (Gessner et al., 2015).

### 2.2. Field data

A sampling design proposed by the Land Degradation Surveillance Framework (Vågen et al., 2013) (Fig. 1, right) was used to locate 160 sample plots. Field data were collected between late November 2013 and early February 2014 for the primary purpose of supporting model construction. To cover the range of variation of the variables in the population, the sampling design divided the study area into 16 regular tiles, and in each tile 10 sample plots were randomly selected and measured. The plots were circular, with a radius of 17.84 m and area of 0.1 ha. The mean distance between plot centers was 218.2 m, thus avoiding problems associated with spatial autocorrelation among plot observations. Plot centers were geo-referenced with Global Navigation Satellite System receivers with a real-time accuracy of 60 cm supported by free corrections of Satellite-Based Augmentation Systems based on European Geostationary Navigation Overlay Service.

Plot-level firewood volume ($m^3$/ha) was obtained from fallen or standing deadwood and living trees by selecting the woody material that was not rotten and was usable as fuelwood, and computing the totals per hectare by aggregation. Because specific allometric models for particular tree species volume were not available, a general model for