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, μ, 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 (τ) or mean (μ) 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

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 south-eastern 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

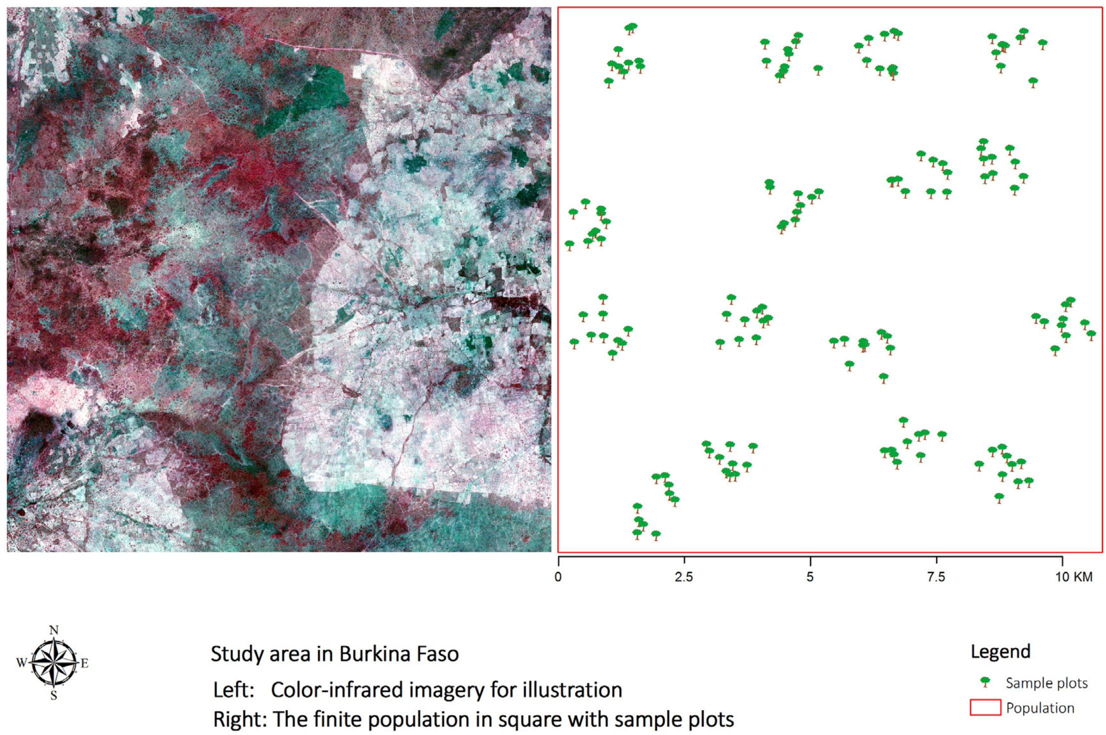

Fig. 1. The population and distribution of sample plots.

dry climates presented by Chave et al. (2005) was applied. Descriptive statistics for the plot measurements are given in Table 1. Altogether 54 tree species were observed of which the most common were *Anogeissus leiocarpa* and *Vitellaria paradoxa*. Most hardwood species except *Parkia biglobosa* and *Sterculia setigera* were included as suitable for firewood; these included *Anogeissus leiocarpa*, *Burkea africana*, *Crossopteryx febrifuga*, *Detarium microcarpum*, *Pterocarpus erinaceus*, *Terminalia avicennioides* and *Vitellaria paradoxa*.

2.3. Remotely sensed data

The sources of remotely sensed data were multispectral images for a single scene from the Landsat 8 Operational Land Imager (OLI) that covers the entire study area. All images were georeferenced to WGS84/UTM Zone 30 N. A time series of Landsat 8 OLI data composed of eight temporally consecutive datasets was collected for the dry season when the field campaign was conducted (Fig. 3). The time interval of 16 days between datasets was a feature of the revisit cycle of Landsat 8. These data with 30-m × 30-m pixels were Provisional Surface Reflectance products (PSR, 2017) obtained free of charge from the U.S. Geological Survey using the online EarthExplorer service (EE, 2017).

3. Methods

3.1. Overview

For modeling purposes, a sample comprised of observations of both the dependent and independent variables is collected for estimating the model parameters, β. A model, $y_i = f(X_i;\beta) + \varepsilon_i$, assumed to adequately represent the underlying relationship between the dependent and independent variables, was used to estimate the superpopulation model where i indexes plots or population units; y_i is the plot-level dependent variable; X_i is a vector of plot-level independent variables; the model parameter estimates, $\hat{\beta}$, are based on the sample; and ε_i is a random residual term following a Gaussian distribution $(0,\sigma_i^2)$. The 0.09-ha pixel size of Landsat 8 was retained because it is similar to the size of the 0.1-ha sample plot. The total number of pixels is thus the population size, M. Because Landsat 8 has the advantage of providing wall-to-wall values, free of charge, a straightforward model-unbiased estimator of μ following this complete enumeration is therefore,

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} f\left(X_i;\hat{\beta}\right), \tag{1}$$

Fig. 2. Controlled fire photos taken in the field, showing scenes on fire (left) and post fire (right).

Table 1
Descriptive statistics for the sample plots.

Attributes	Min	Max	Mean	SD
Tree density (stems/ha)	10	1935	494	401
Mean diameter (cm)	6.4	40	15.2	8.5
Basal area (m^2/ha)	0.2	16.1	5.6	3.6
Firewood volume (m^3/ha)	0	29.1	6.6	6.2

where M is sufficiently large that $E(\varepsilon_i)=0$ can be assumed. For this study, we focus on $\hat{\mu}$ although the total can be readily estimated as $\hat{\tau}=M\hat{\mu}$. The independent variables are selected from each of the Landsat 8 datasets and serve as proxies for the dependent variable. The dependent variable for each population unit can then be predicted using a fitted model. Modeling details are reported in Section 3.4, and methods for selecting the independent variables are reported in Section 3.5.

The estimator of the variance of $\hat{\mu}$ is,

$$\hat{V}(\hat{\mu})_{an}=\sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{Cov}\left(\hat{\beta}_j,\hat{\beta}_k\right)\widehat{\bar{f}}'_j\widehat{\bar{f}}'_k+\frac{1}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}\widehat{Cov}(\varepsilon_j,\varepsilon_k), \quad (2)$$

where the subscript *an* denotes this estimator being analytical, distinguished from an empirical estimator reported in Section 3.3; $\widehat{Cov}(\hat{\beta}_j,\hat{\beta}_k)$ is the estimated covariance between the model parameter estimates $\hat{\beta}_j$ and $\hat{\beta}_k$; $\widehat{\bar{f}}'_j=\frac{1}{M}\sum_{i=1}^{M}f'_j(X_i;\hat{\beta})=\frac{1}{M}\sum_{i=1}^{M}\frac{\partial f(X_i;\hat{\beta})}{\partial\hat{\beta}_j}$; and $\widehat{Cov}(\varepsilon_j,\varepsilon_k)$ is the estimated covariance between pairs of residuals. As shown empirically in McRoberts (2006), the second term in Eq. (2) approaches zero as M becomes large and thus in this study assumed to be ignorable in comparison to the first term, which simplifies assessment of the other factors remaining in the first term.

The desire for smaller variances means that the sampling distribution of $\hat{\mu}$ must be concentrated and preferably symmetric and centered on μ. To understand the factors that contribute to small variances, each component of the first term in Eq. (2) must be investigated in turn. The partial derivative calculation relates to two factors, the model and the values of the independent variables for the entire population (x_{app}). The covariance matrix calculation relates to three factors, the model, the values of the independent variables for the sample units (x_{mod}), and the form of the covariance matrix estimator. Fulfillment of the objectives of this study is rooted in the analysis of these factors.

3.2. Effects of temporally external auxiliary data

Relationships between factors related to the model and the independent variables are intermingled. In practice, the model used to predict the attribute for a population unit is constructed using dataset x_{mod} corresponding to imagery acquired at a point in time t_1, whereas the model is often applied using the same variables but to a dataset x_{app} whose values correspond to imagery acquired at another point in time t_2. Although the variables are the same, models of the relationships between the dependent variable and independent variables obtained from x_{mod} may differ from the model of the relationship with independent variables obtained from x_{app}. Therefore, to understand how $\hat{\mu}$ and $\widehat{Var}(\hat{\mu})$ respond to the temporal differences that manifest themselves through the identified factors requires evaluations and comparisons for two scenarios: (1) x_{mod} and x_{app} are obtained from the same dataset, and x_{mod} is a sample of x_{app}; and (2) x_{mod} and x_{app} are obtained from temporally different datasets.

The analyses consisted of firstly constructing a model using x_{mod} obtained from one of the Landsat 8 datasets, and then secondly, applying this model to x_{app} obtained from all the datasets including the dataset from which x_{mod} was obtained. The magnitude of differences in $\hat{\mu}$ and $\widehat{Var}(\hat{\mu})$ at the application stage can be attributed to temporal differences between x_{mod} and x_{app}. Because the effects of obtaining x_{mod} and x_{app} from different dataset are expressed via model predictions, two models were constructed, Model-1 using x_{mod} obtained from Landsat 8 dataset A and Model-2 using x_{mod} obtained from Landsat dataset E. Selecting which two different datasets from the eight available Landsat 8 datasets for x_{mod} is an unimportant factor in regard to analyzing the effects of

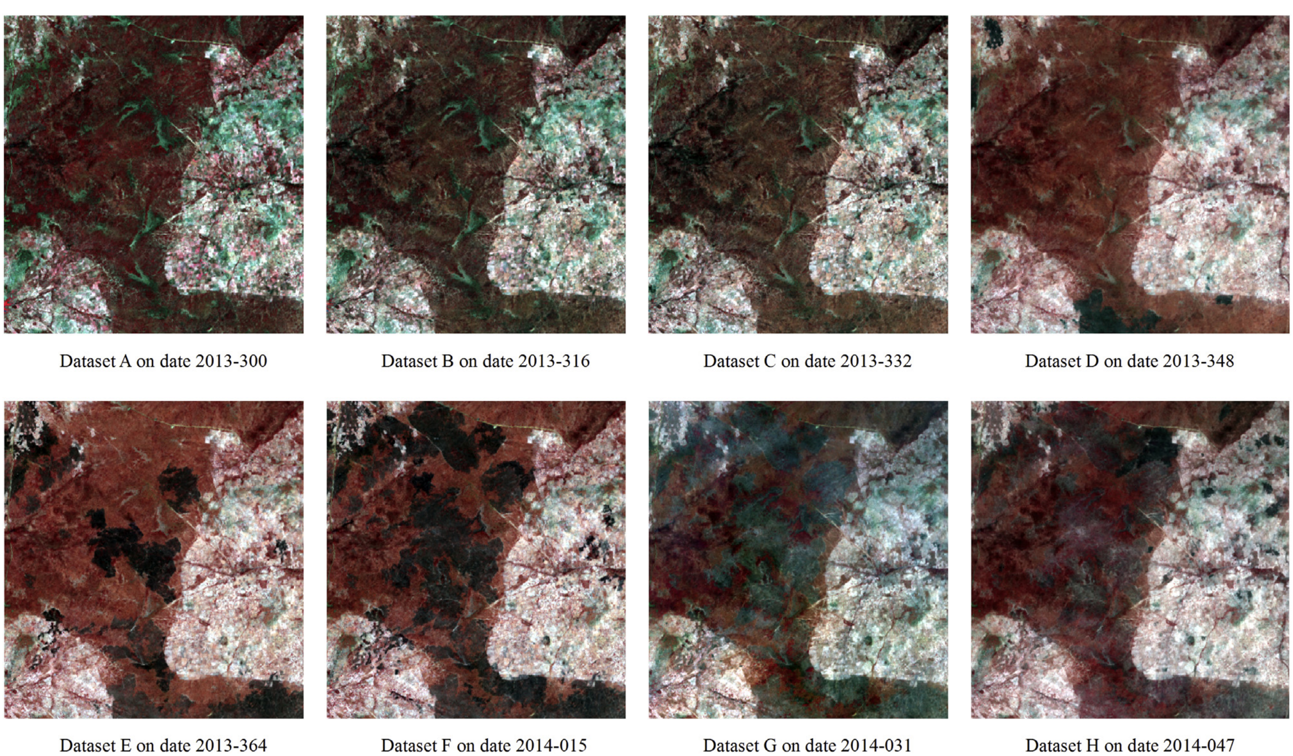

Fig. 3. Landsat 8 OLI time series composed of eight temporally consecutive datasets. The date “2013-300”, for example, stands for the 300th day in the year 2013.

temporally external data. Datasets A and E were selected based on preliminary analyses because their respective models were of the same independent variables and this feature facilitates the comparability of model parameter estimates and population parameter estimation. In this way, the effects of the model as an influencing factor can be assessed. For future reference, describing x_{mod} and x_{app} as having been obtained from the same dataset means that either both x_{mod} and x_{app} were obtained from dataset A for Model-1 or both x_{mod} and x_{app} were obtained from dataset E for Model-2.

3.3. Bootstrapping procedure

The incentive for introducing Eq. (2) was to identify the factors that affect the variance in an analytical way. Nevertheless, it is common that some assumptions made during the derivation of the variance estimator or the derivation of a covariance matrix estimator may not be fully satisfied (Saarela et al., 2015). Resampling procedures such as jackknifing and bootstrapping are suitable for such complex applications. Now that the factors were identified, the variance was also assessed empirically using bootstrapping and compared to the variance obtained using an analytical estimator Eq. (2). Bootstrapping also provides an estimate of bias that is otherwise difficult to assess (Davison and Hinkley, 1997). However, it does not identify influential factors as explicitly as does Eq. (2).

Bootstrapping procedures were initially introduced by Efron (1979) and later developed by Efron and Tibshirani (1994). All bootstrap methods are based on the notion of a resample (or bootstrap sample). For regression problems, Efron and Tibshirani (1994) introduced three approaches for constructing resamples, and McRoberts (2011) provides details on various approaches for using bootstrapping with model-based inference. For this study, the *bootstrapping pairs* approach was used. With this approach, a resample of size n is drawn with replacement from the empirical distribution of the pairs, (y_i, X_i), each with probability $1/n$, based on the original sample of size n. The original model calibration sample of 160 plots was resampled with replacement 30,000 times. However, the important issue is not the number of resamples but rather whether the estimates stabilized and converged. For this study, 30,000 repetitions was deemed sufficient because the estimates changed little after 10,000 or 20,000 repetitions. For each resample, the model parameters were estimated, and the resample-specific $\hat{\mu}$ and $\hat{V}(\hat{\mu})_{an}$ were calculated using Eqs. (1) and (2), respectively, and recorded. The empirical bootstrap variance was calculated as,

$$V(\hat{\mu})_{boot} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\mu}^* - \hat{\mu}_b^*)^2, \tag{3}$$

where $\hat{\mu}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b^*$; $\hat{\mu}_b^*$ is the bootstrap estimate from the b-th resample, $b = 1, \ldots, B$, and $B = 30{,}000$. The empirical bootstrap variance was compared with the mean of the analytical variance estimates, $\hat{V}(\hat{\mu}_b^*)_{an}$, for the $B = 30{,}000$ resamples and expressed as

$$\overline{\hat{V}}(\hat{\mu})_{an} = \frac{1}{B}\sum_{b=1}^{B}\hat{V}(\hat{\mu}_b^*)_{an} \tag{4}$$

which correspond closely with $V(\hat{\mu})_{boot}$. An estimate of the bias of $\hat{\mu}$ was computed to assess the deviation as

$$\widehat{Bias}(\hat{\mu}) = \hat{\mu}^* - \hat{\mu} \tag{5}$$

where $\hat{\mu}$ is based on the original sample of 160 plots. Note that $\widehat{Bias}(\hat{\mu})$ pertains only to resampling issues but not to issues related to model misspecification.

3.4. Modeling and covariance matrix estimator

Because the model is an influencing factor, Model-1 and Model-2 were developed using x_{mod} obtained from different Landsat 8 datasets. Formulation of regression models and selection of independent variables can be difficult when the number of independent variables is large. Further, forward, backward, and stepwise variable selection techniques do not work well when correlations among independent variables are large (Harrell, 2001). For each of Model-1 and Model-2, independent variables were selected parsimoniously using the "bootstrap stepAIC" procedure (Rizopoulos, 2009) which integrates bootstrapping to assess the variability in model selection when using a stepwise algorithm based on the Akaike information criterion (Akaike, 1974).

Following preliminary analyses using both linear and nonlinear models, a linear model of the form $y = X\beta + \varepsilon$ was selected where y is the dependent variable; X is the design matrix of independent variables; β is the vector of model parameters; and ε is the vector of residuals assumed to be (multivariate) normally distributed with $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon^T) = \Phi$, a positive definite matrix. To accommodate heteroscedasticity, the model parameters were estimated using weighted least squares (WLS) for which $\hat{\beta} = (X^TWX)^{-1}X^TWy$, and $W = \Phi^{-1}$ is a diagonal matrix with $w_{ii} = 1/\sigma_i^2$ (Carroll and Ruppert, 1988). The diagonal elements of W were estimated using a 4-step procedure introduced in McRoberts et al. (2016b). First, the pairs $(y_i, \hat{y}_i)$ were ordered with respect to $\hat{y}_i$; second, the pairs were aggregated into groups of size 20 each; third, for each group, g, the mean of the predictions, $\overline{\hat{y}}_g$, and the variance, $\hat{\sigma}_g^2$, of the residuals, $\varepsilon_i = y_i - \hat{y}_i$, were calculated; and fourth, the relationship between $\hat{\sigma}_g^2$ and $\overline{\hat{y}}_g$ was modeled as $\hat{\sigma}_g^2 = \lambda \cdot \overline{\hat{y}}_g + \varepsilon_g$, where λ is the parameter to be estimated (Fig. 4).

When $W = \Phi^{-1}$, $Cov(\hat{\beta}) = (X^TWX)^{-1}$, which can be considered an ideal case. Because the structure of Φ is usually unknown, estimating Φ leads to variants of $\widehat{Cov}(\hat{\beta})$. Preliminary analyses revealed differences only in the third significant digit, suggesting the selection of a covariance matrix estimator is a minor factor. Therefore, the following heteroscedasticity consistent estimator, first introduced by MacKinnon

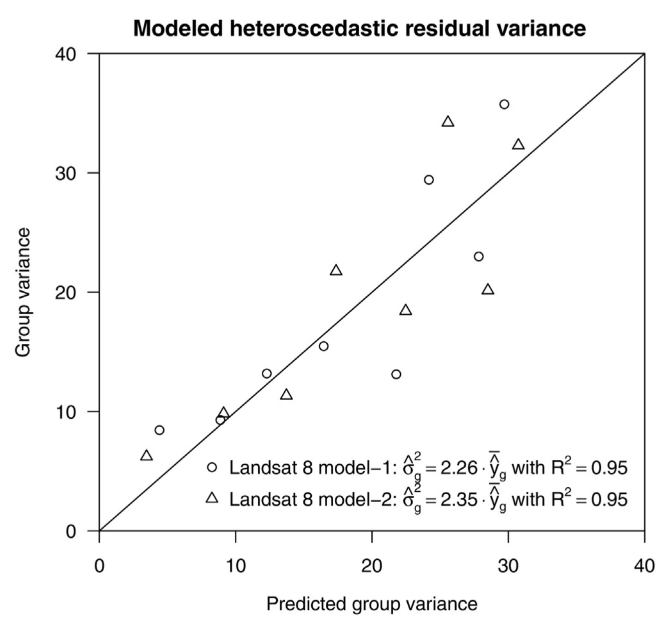

Fig. 4. Modeled heteroscedastic residual variance.

and White (1985), was used,

$$\widehat{Cov}\left(\hat{\beta}\right) = \left(X^T W X\right)^{-1} X^T W \ \mathrm{diag}\left[\frac{e_i^2}{(1-h_i)^2}\right] W^T X \left(X^T W^T X\right)^{-1}, \qquad (6)$$

where $h_i = x_i(X^T WX)^{-1} x_i^T w_i$. Dividing e_i^2 by $(1-h_i)^2$ inflates e_i^2 so that the over-influence of observations with large variances can be adjusted.

In model-based inference, when the calibration and application data are obtained from the same dataset, a common practice is to select the model that satisfies a criterion such as smallest RMSE, greatest R^2, or increasingly the shortest confidence interval for a population parameter (Saarela et al., 2015). However, if the candidate models were from a pool of models developed using different calibration datasets, the selection procedure is more complex. Because Model-1 used independent variables whose values were obtained from Landsat 8 dataset A, and Model-2 used independent variables whose values were obtained from Landsat 8 dataset E, the degree to which measures of prediction accuracy and goodness of fit correspond with the precision of the estimator of the population parameter was assessed. Goodness of fit was assessed with R^2 and prediction accuracy with RMSE based on a leave-one-out cross validation where $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, $RMSE_{\%} = \frac{RMSE}{\bar{y}} \times 100$, n is the number of sample plots; y_i is the field measured firewood volume; $\hat{y}_i$ is the predicted firewood volume; and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

3.5. Independent variables

The independent variables were selected by types from the original spectral bands, the first principal component of the original spectral bands, the textural features of the first principal component, spectral indices and textural features for the respective spectral indices. Principal component analysis is a multivariate method using orthogonal transformation to convert correlated data of multiple dimensions into uncorrelated principal components. Depending on what is being modeled, the first principal component compared with other components usually has the largest possible variance and the greatest power for explaining the variability in the original data. The subsequent components then exhibit gradually less variance and each component in the succeeding position is under a constraint of being orthogonal to the preceding one. Principal components can be resolved as the eigenvectors of a correlation matrix or a covariance matrix. The correlation matrix was used in this study, and only the first principal component was calculated for Landsat 8, using the original spectral bands for each scene.

The spectral indices considered here are described in Table 2. Most of the indices listed have been shown to contribute to statistically significantly improving the quality of fit of stem volume or aboveground biomass models in environmental conditions similar to those for the current study (Hou et al., 2011; Hou et al., 2013; Gasparri et al., 2010; Wu et al., 2013). Textural features were calculated for each first principal component and its respective spectral indices, and these form the first and second-order statistical features, including the mean, variance, homogeneity, contrast, dissimilarity, entropy, angular second moment and correlation, with equations available in Haralick et al. (1973) and Gonzalez (2008). The textures were calculated with window sizes of 3 × 3 pixels for Landsat 8, an offset distance of 1 (averaged over all directions), and a 64 grey level quantization resulting from the tradeoff between the spatial detail preservation and the noise reduction.

Table 2
Summary of the spectral indices. NIR, near infrared band; R, red band; B, blue band; SWIR, short-wave infrared band.

Spectral indices	Formula	References
Enhanced vegetation index (EVI)	2.5(NIR − R) / (NIR + 6R − 7.5B + 1)	Huete et al. (2002)
Generalized Difference Vegetation Index (GDVI)	(NIR^2 − R^2) / (NIR^2 + R^2)	Wu (2014)
Normalized Difference Vegetation Index (NDVI)	(NIR − R) / (NIR + R)	Rouse et al. (1973)
Normalized Difference Water Index (NDWI)	(NIR − SWIR2) / (NIR + SWIR)	Gao (1996)
Specific Leaf Area Vegetation Index (SLAVI)	NIR / (R + SWIR2)	Lymburner et al. (2000)
Simple Ratio (SR)	NIR / R	Birth and McVey (1968)

4. Results

4.1. Models

The models developed using Landsat 8 auxiliary data are summarized in Table 3. Studentized residual graphs and graphs of observations versus predictions are reported in Fig. 5. In general, Model-1 and Model-2 performed similarly, although it is interesting to note that both R^2 and RMSE are smaller for Model-2 than for Model-1.

For each model, only one independent variable, EVI, was selected because of the parsimony criterion, which also circumvented collinearity problems. Although prediction accuracy could be improved by adding more independent variables, our primary interest is comparability between models. For these two models, differences in parameter estimates were attributed to temporal differences in the datasets from which x_{mod} was obtained under the assumption that the relationship between the dependent and independent variables does not change over time. Model-1 indicates that the increase of one unit in EVI causes an increase of 11.07 in the dependent variable, whereas this increase is 29.31 for Model-2. Obviously, the covariance matrix estimates were affected as well, which would further propagate this difference in the calculation of analytical variance estimates.

4.2. Effects on model-based inference

Estimates for cases for which x_{mod} and x_{app} were obtained from the same Landsat dataset are reported in Table 4 and are used for comparison with estimates for cases for which x_{mod} and x_{app} were obtained from temporally different Landsat datasets (Fig. 6). Parameter estimates obtained for Model-1 which used x_{mod} obtained from Landsat dataset A were very similar to estimates obtained for Model-2 which used x_{mod} obtained from Landsat dataset E (Table 4). Further, when the two models were applied to x_{app} obtained from the same dataset as x_{mod}, the model-based estimates of the mean, $\hat{\mu}$, were similar to each other and to the simple random sampling estimate of the mean calculated using only the field observations, 6.6 m^3/ha (Table 1). Differences between estimates of $\overline{\overline{V}}(\hat{\mu})_{an}$ and $V(\hat{\mu})_{boot}$ were in the third significant digit to the right of the decimal point and were so small as to be negligible. $\widehat{Bias}(\hat{\mu})$ did not appear to be a concern, either. Prediction accuracies (Table 3) corresponded closely to variance estimates (Table 4) so that RMSE is useful for screening models.

The effects of obtaining x_{mod} and x_{app} from temporally different Landsat datasets are expressed on the full spectrum of estimation, $\hat{\mu}$, $\overline{\overline{V}}(\hat{\mu})_{an}$, $V(\hat{\mu})_{boot}$, and $\widehat{Bias}(\hat{\mu})$. Fig. 6 indicates that selection of a model

Table 3
Summary of the models.

Model	RMSE (m^3/ha)	RMSE$_{\%}$	R^2	Independent variable	Estimate	Std. Error
Landsat 8 Model-1	4.623	58.442	0.604	(Intercept)	−4.872	0.638
				EVI	11.07	0.771
Landsat 8 Model-2	4.562	57.661	0.548	(Intercept)	−9.689	1.159
				EVI	29.306	2.289

All parameter estimates were statistically significantly different from zero at α = 0.01.

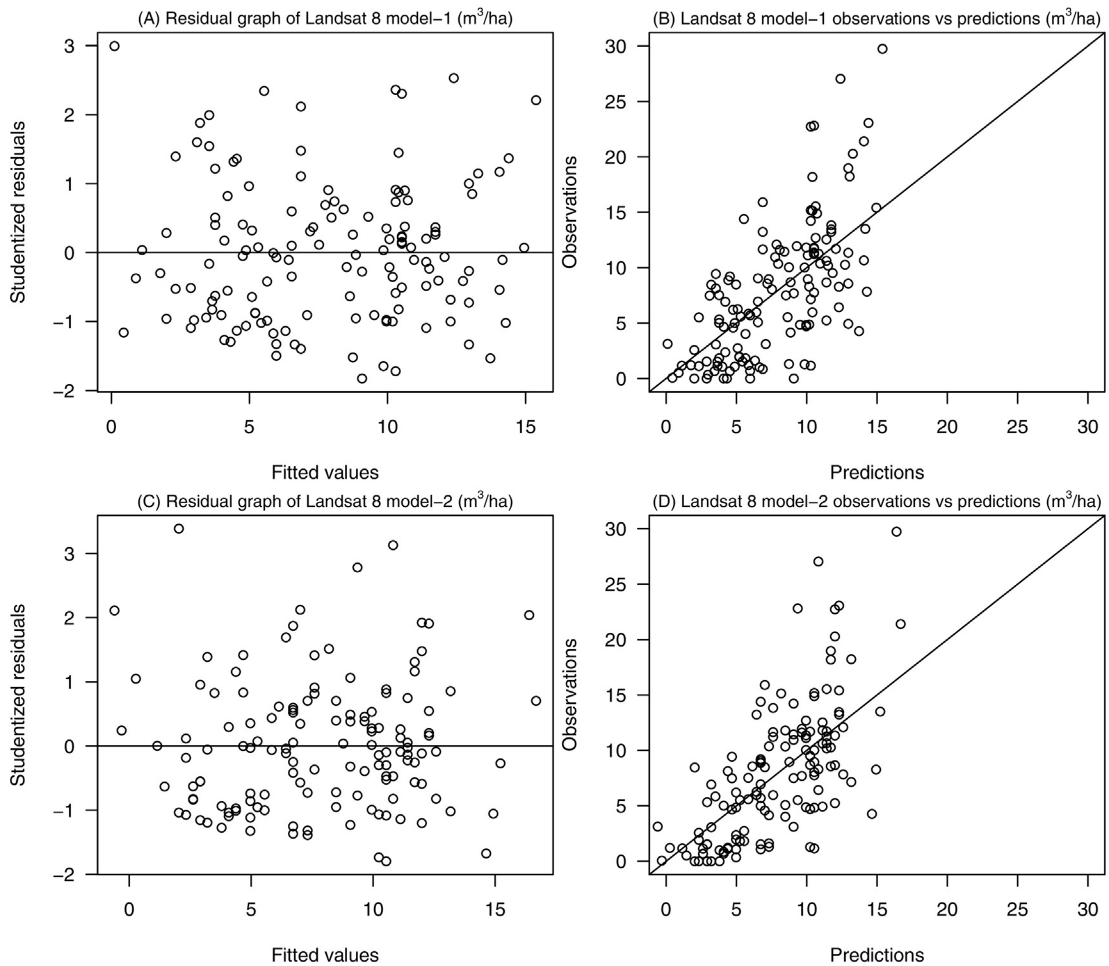

Fig. 5. Studentized residual graphs and the graphs of observations versus predictions.

must be done cautiously, because both the prediction accuracy and the goodness of fit are not indicative of robust population parameter estimation. Model-1 risks under-estimating $\hat{\mu}$ with increasing $\widehat{Bias}(\hat{\mu})$ with increasing temporal externality. In contrast, Model-2 risks over-estimating $\hat{\mu}$ with increasing $\overline{\hat{V}}(\hat{\mu})_{an}$ or $V(\hat{\mu})_{boot}$ with increasing temporal externality. Over- or under-estimation of $\hat{\mu}$ based on models that were originally effective for a non-temporally external situation can be very different for a temporally external situation. Compared to $\overline{\hat{V}}(\hat{\mu})_{an}$ whose calculation is also affected by temporal externality, $V(\hat{\mu})_{boot}$ is more conservative as the temporal difference increases. The general pattern is that the less the temporal externality, the more accurate is the estimation, i.e., the adverse effects of temporally external independent variables are exacerbated over time with respect to a model.

Table 4
Estimates based on x_{mod} and x_{app} obtained from Landsat 8 dataset A for Model-1, and from Landsat 8 dataset E for Model-2, with $\hat{\mu}$ and $\widehat{Bias}(\hat{\mu})$ in m^3/ha.

Source	$\hat{\mu}$	$\overline{\hat{V}}(\hat{\mu})_{an}$	$V(\hat{\mu})_{boot}$	$\widehat{Bias}(\hat{\mu})$
Model-1 applied to Landsat 8 A	7.119	0.131	0.134	0.037
Model-2 applied to Landsat 8 E	6.839	0.120	0.125	0.012

5. Discussion

While many pioneering studies have investigated the effects of model construction, sampling design, estimators and positioning errors (e.g. Breidenbach et al., 2014; McRoberts, 2006; McRoberts, 2010; Saarela et al., 2015; Saarela et al., 2016), this study focused on the response of model-based inference to independent variables selected from temporally different remotely sensed auxiliary datasets. The mathematical form is not explicitly external, but rather only the values of the independent variables, having two models of the same form facilitate comparisons.

Based on the comparison between Model-1 and Model-2, even when the independent variables are the same, the parameter estimates, $\hat{\beta}$, are different. When these models are applied to estimate population parameters, the effects of the differences in parameter estimates are further influenced by temporal differences between x_{mod} and x_{app}. For x_{mod} and x_{app} obtained from the same Landsat dataset, $\hat{\mu}$, $\overline{\hat{V}}(\hat{\mu})_{an}$, $V(\hat{\mu})_{boot}$ and $\widehat{Bias}(\hat{\mu})$ obtained for Model-1 and Model-2 were similar. Although prediction accuracy depends on the field plot sample, RMSE is indicative of the uncertainty associated with the population parameter estimates and could, therefore, be proxy for the real criterion for screening candidate models which is the confidence interval width for the population parameter estimate. In addition, either the analytical variance estimator or the bootstrapping procedure may be used.

When x_{mod} and x_{app} are obtained from different Landsat datasets, temporal differences in the values of the independent variables

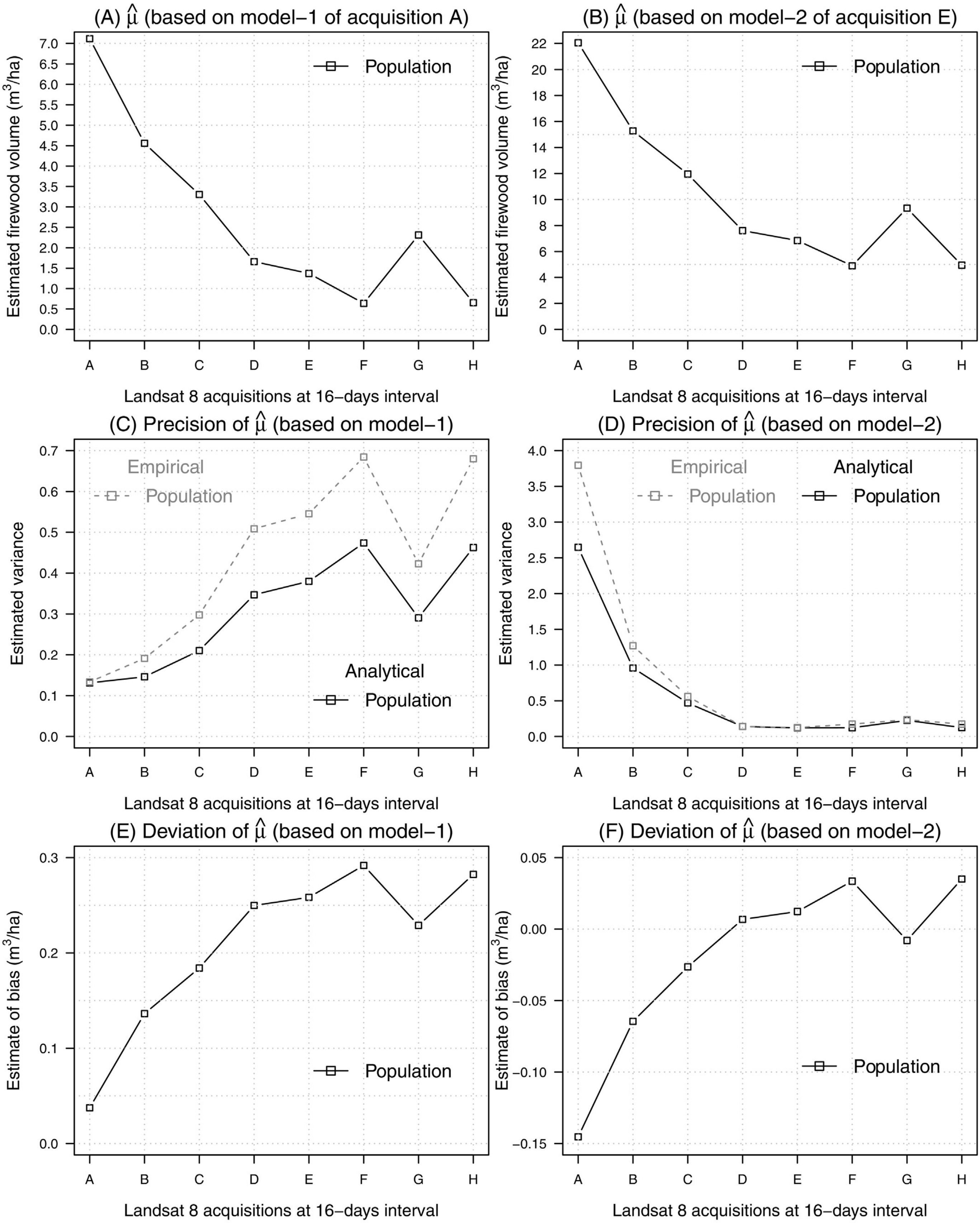

Fig. 6. Response of model-based inference to application of Model-1 and Model-2 to the eight temporally consecutive Landsat 8 datasets at 16-day intervals. The alphabetic codes on the x-axis are defined in Fig. 3.

manifest themselves in the form of enlarged $\overline{\overline{V}}(\hat{\mu})_{an}$, $V(\hat{\mu})_{boot}$, and $\widehat{Bias}(\hat{\mu})$. These effects are expressed in the full spectrum of estimation. Further, prediction accuracy is no longer indicative of the precision of the estimator of the population parameter. Screening of candidate models must be conducted carefully, perhaps requiring an analysis similar to that conducted for this study to prevent misleading estimates. The observed pattern is that, depending on the rate of change of the auxiliary data, the less the temporal difference between x_{mod} and x_{app}, the more credible the estimation. In particular, the adverse effects of temporally external independent variables tend to be exacerbated over time with respect to a model. Further, because temporal differences also affect the analytical variance estimation, bootstrapping variance estimation is recommended.

The effects of temporally external auxiliary data may lead to greater loss of precision than other causes such as the selection of modeling technique (parametric or nonparametric or form of models), covariance estimators and variance estimators. The analytical variance estimator risks loss of robustness when partial derivatives in Eq. (2) are based on x_{mod} and x_{app} extracted from temporally different datasets. In comparison, bootstrapping provides conservative variance estimates that lead to larger confidence intervals. Bootstrapping is generally universally applicable, both for parametric models with or without knowing the covariance matrix and for nonparametric imputations. It also provides a bootstrap estimate of bias that is otherwise difficult to assess. In practice, parametric or nonparametric models reported in remote sensing studies can be screened and then used for model-based inference. At hotspot areas where numerous models have been continuously developed, relatively reliable estimates can be obtained by applying respective models to respective auxiliary datasets whose dates are close enough to the dates of the datasets used for model fitting. For these conditions, a temporally consecutive series of estimates will become available. These estimates are not always accurate for forest management and planning, but are effective for revealing temporal trends of the variation of the variable of interest to answer ecological questions.

Sampling that acquires auxiliary data at different times is a pragmatic feature of model-based inference and theoretically allows temporally invariant inferences. However, large correlations between dependent variables and independent variables extracted from respective datasets do not guarantee interchangeability among these datasets. The root cause of the adverse effects on inference is violation of the basic regression assumption that independent variables are stationary and free of errors. In this study, although most trees survived the prescribed burnings, fires may still have affected the spectral signals, especially in the near-infrared spectrum (Koutsias and Pleniou, 2015), and thereby compromised the effectiveness of a model constructed beforehand. Actually, the cause of data deviation may be attributed to multiple factors such as natural or anthropogenic disturbances, climate, phenology, seasonality, clouding or other factors that affect the strength of electromagnetic radiation recorded by remote sensing sensors. It is also common for a sample plot to straddle multiple pixels in which case the mean digital number of these pixels is usually calculated for this plot and used in the calculation of remote sensing variables, as was done in this study. Compared to a perfect alignment, straddling can slightly increase the variance of model parameter estimates, but shall not affect relative comparisons between estimation methods or conclusions considering that these errors are systematic of about similar magnitude (Saarela et al., 2016).

A plausible candidate model may be deceptive because a model with greater prediction accuracy or goodness of fit does not necessarily produce more accurate population parameter estimates or narrower confidence intervals using data other than that used for fitting the model. Reliance on goodness of fit or prediction accuracy requires caution while model comparisons of properties of population estimates are always commended. The reason is related to the timing of the data acquisition, the moment when the natural conditions were captured, because natural conditions are in a perpetual state of flux. The more representative the moment of dataset, the more likely that the model will work with other datasets obtained at different times. Techniques in time series analysis, data assimilation and relative radiometric correction making use of pseudo-invariant features for calibration purposes (Xu et al., 2012) might be applicable for improving the datasets and therefore the robustness of model-based inference.

6. Conclusions

The effects of temporally external auxiliary data on $\hat{\mu}$, $\widehat{Var}(\hat{\mu})$ and $\widehat{Bias}(\hat{\mu})$ were evaluated and compared within the framework of model-based inference. First, for non-temporally external data, four conclusions are relevant: (1) $\hat{\mu}$, $\overline{\overline{V}}(\hat{\mu})_{an}$, $V(\hat{\mu})_{boot}$ and $\widehat{Bias}(\hat{\mu})$ will be similar, regardless of the dates of the datasets from which both the calibration and application data were obtained; (2) selection of an auxiliary dataset for modeling has little effect; (3) prediction accuracy expressed by RMSE is useful for screening candidate models to be used for population inferences; and (4) either the analytical variance estimator or the bootstrapping procedure is satisfactory. Second, for temporally external data, five conclusions are relevant: (1) $\overline{\overline{V}}(\hat{\mu})_{an}$, $V(\hat{\mu})_{boot}$ and $\widehat{Bias}(\hat{\mu})$ are enlarged; (2) over- or under-estimation of $\hat{\mu}$ based on the models that were effective for a non-temporally external situations can be very different for temporally external situations; (3) candidate models should be screened using a criterion related to the real objective, an inference for a population parameter, rather than prediction accuracy or goodness of fit; (4) the adverse effects of temporally external independent variables of the application data tend to be exacerbated over time with respect to a model; and (5) because temporal differences affect the analytical variance calculation, the bootstrapping procedure is recommended.

Acknowledgements

Within the project titled Building Biocarbon and Rural Development in West Africa (BIODEV), work package 1.4 on Sustainable Wood Energy aims to improve the efficiency of firewood use and to develop sustainable energy systems based on wood. This study was supported by BIODEV and the Finnish Ministry of Foreign Affairs. We are grateful to all crew members who carried out the field measurements in Burkina Faso. Last but not least, the comments provided by Professor Crystal Schaaf and anonymous reviewers are sincerely acknowledged.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.
Birth, G.S., McVey, G.R., 1968. Measuring the color of growing turf with a reflectance spectrophotometer. Agron. J. 60, 640–643.
Breidenbach, J., Antón-Fernández, C., Petersson, H., McRoberts, R.E., Astrup, R., 2014. Quantifying the model-related variability of biomass stock and change estimates in the Norwegian National Forest Inventory. For. Sci. 60, 25–33.
Carroll, R.J., Ruppert, D., 1988. Transformation and Weighting in Regression. Chapman and Hall, New York.
Cassel, C.M., Särndal, C.E., Wretman, J.H., 1977. Foundations of Inference in Survey Sampling. Wiley, New York.
Chambers, R., Clark, R., 2012. An Introduction to Model-based Survey Sampling With Applications. Oxford University Press, Oxford.
Chave, J., Andalo, C., Brown, S., Cairns, M., Chambers, J., Eamus, D., Fölster, H., Fromard, F., Higuchi, N., Kira, T., Lescure, J.P., Nelson, B.W., Ogawa, H., Puig, H., Riéra, B., Yamakura, T., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145, 87–99.
Cochran, W.G., 1977. Sampling Techniques. third ed. Wiley, New York.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge.
EE, 2017. USGS EarthExplorer. http://earthexplorer.usgs.gov/ (accessed 17.05.20).
Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. Chapman and Hall, New York.
Gao, B.C., 1996. NDWI-a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257–266.

Gasparri, N.I., Parmuchi, M.G., Bono, J., Karszenbaum, H., Montenegro, C.L., 2010. Assessing multi-temporal Landsat 7 ETM+ images for estimating above-ground biomass in subtropical dry forests of Argentina. J. Arid Environ. 74, 1262–1270.

Gessner, U., Knauer, K., Kuenzer, C., Dech, S., 2015. Land surface phenology in a west African savanna: impact of land use, land cover and fire. In: Kuenzer, C., Dech, S., Wagner, W. (Eds.), Remote Sensing Time Series: Revealing Land Surface Dynamics. Springer, pp. 203–223.

Gonzalez, R.C., 2008. Digital Image Processing. third ed. Prentice Hall, Upper Saddle River, New Jersey.

Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. 28, 1429–1447.

Gregoire, T.G., Valentine, H.T., 2008. Sampling Strategies for Natural Resources and the Environment. Chapman and Hall, Boca Raton, Florida.

Gregoire, T.G., Næsset, E., McRoberts, R.E., Ståhl, G., Andersen, H.E., Gobakken, T., Ene, L., Nelson, R., 2016. Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass. Remote Sens. Environ. 173, 98–108.

Haralick, R.M., Shanmugam, K., Dinsteein, J., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. 3, 610–621.

Harrell, F.E., 2001. Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, New York.

Hou, Z., Xu, Q., Tokola, T., 2011. Use of ALS, Airborne CIR and ALOS AVNIR-2 data for estimating tropical forest attributes in Lao PDR. ISPRS J. Photogramm. 66, 776–786.

Hou, Z., Xu, Q., Nuutinen, T., Tokola, T., 2013. Extraction of remote sensing based forest management units in tropical forest. Remote Sens. Environ. 130, 1–10.

Hou, Z., Xu, Q., Hartikainen, S., Antilla, P., Packalen, T., Maltamo, M., Tokola, T., 2015. Impact of plot size and spatial pattern of forest attributes on sampling efficacy. For. Sci. 61, 847–860.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213.

IPCC, 2006. 2006 IPCC guidelines for national greenhouse gas inventories. http://www.ipcc-nggip.iges.or.jp/public/2006gl/index.html (accessed 17.05.20).

Jonsson, K., Ong, C.K., Odongo, J.C.W., 1999. Influence of scattered Nere and Karite trees on microclimate, soil fertility and millet yield in Burkina Faso. Exp. Agric. 35, 39–53.

Koutsias, N., Pleniou, M., 2015. Comparing the spectral signal of burned surfaces between Landsat 7 ETM+ and Landsat 8 OLI sensors. Int. J. Remote Sens. 36, 3714–3732.

Lymburner, L., Beggs, P.J., Jacobson, C.R., 2000. Estimation of canopy-average surface-specific leaf area using Landsat TM data. Photogramm. Eng. Remote Sens. 66, 183–191.

MacKinnon, J.G., White, H., 1985. Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. J. Econom. 29, 53–57.

McRoberts, R.E., 2006. A model-based approach to estimating forest area. Remote Sens. Environ. 103, 56–66.

McRoberts, R.E., 2010. Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Remote Sens. Environ. 114, 1017–1025.

McRoberts, R.E., 2011. Satellite image-based maps: scientific inference or pretty pictures? Remote Sens. Environ. 115, 715–724.

McRoberts, R.E., Andersen, H.E., Næsset, E., 2014. Using airborne laser scanning data to support forest sample surveys. In: Maltamo, M., Næsset, E., Vauhkonen, J. (Eds.), Forestry Applications of Airborne Laser Scanning. Springer, the Netherlands.

McRoberts, R.E., Næsset, E., Gobakken, T., 2016a. The effects of temporal differences between map and ground data on map-assisted estimates of forest area and biomass. Ann. For. Sci. 73, 839–847.

McRoberts, R.E., Chen, Q., Domke, G.M., Ståhl, G., Saarela, S., Westfall, J.A., 2016b. Hybrid estimators for mean aboveground carbon per unit area. For. Ecol. Manag. 378, 44–56.

MEDD, 2012. Readiness Preparation Plan for REDD. Ministry of the Environment and Sustainable Development, Burkina Faso.

Nicholson, S.E., 2009. A revised picture of the structure of the “monsoon” and land ITCZ over West Africa. Clim. Dyn. 32, 1155–1171.

Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644.

PSR, 2017. Landsat Surface Reflectance High Level Data Products. http://landsat.usgs.gov/CDR_LSR.php (accessed 17.05.20).

Rizopoulos, D., 2009. Bootstrap stepAIC. https://cran.rproject.org/web/packages/bootStepAIC/index.html (accessed 17.05.20).

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERST. NASA Spec. Publ. 351, pp. 309–317.

Royall, R.M., Herson, H.J., 1973a. Robust estimation in finite populations. I. J. Am. Stat. Assoc. 68, 880–889.

Royall, R.M., Herson, H.J., 1973b. Robust estimation in finite populations. II. J. Am. Stat. Assoc. 68, 890–893.

Saarela, S., Schnell, S., Grafström, A., Tuominen, S., Nordkvist, K., Hyyppä, J., Kangas, A., Ståhl, G., 2015. Effects of sample size and model form on the accuracy of model-based estimators of growing stock volume. Can. J. For. Res. 45, 1524–1534.

Saarela, S., Schnell, S., Tuominen, S., Balázs, A., Hyyppä, J., Grafström, A., Ståhl, G., 2016. Effects of positional errors in model-assisted and model-based estimation of growing stock volume. Remote Sens. Environ. 172, 101–108.

Särndal, C.E., 1978. Design-based and model-based inference in survey sampling. Scand. J. Stat. 5, 27–52.

Särndal, C.E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer, New York.

Sawadogo, L., Nygard, R., Pallo, F., 2002. Effects of livestock and prescribed fire on coppice growth after selective cutting of Sudanian savannah in Burkina Faso. Ann. For. Sci. 59, 185–195.

Vågen, T.G., Winowiecki, L., Tamede Desta, L., Tondoh, J., 2013. The Land Degradation Surveillance Framework (LDSF): Field Guide. Center for International Forestry Research http://www1.cifor.org/fileadmin/subsites/sentinellandscapes/document/LDSF_Field_Guide.pdf (accessed 17.05.20).

Valliant, R., Dorfman, A.H., Royall, R.M., 2000. Finite Population Sampling and Inference. Wiley, New York, p. 504.

Wu, W., 2014. The generalized difference vegetation index (GDVI) for dryland characterization. Remote Sens. 6, 1211–1233.

Wu, W., de Pauw, E., Helldén, U., 2013. Assessing woody biomass in African tropical savannahs by multiscale remote sensing. Int. J. Remote Sens. 34, 4525–4549.

Xu, Q., Hou, Z., Tokola, T., 2012. Relative radiometric correction of multi-temporal ALOS AVNIR-2 data for the estimation of forest attributes. ISPRS J. Photogramm. 68, 69–78.