Remote Sensing of Environment 184 (2016) 496–505

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A functional regression model for inventories supported by aerial laser scanner data or photogrammetric point clouds

S. Magnussen [a,*], E. Næsset [b], G. Kändler [c], P. Adler [c], J.P. Renaud [d], T. Gobakken [b]

[a] Canadian Forest Service, Natural Resources Canada, 506 West Burnside Rd., Victoria, BC V8Z 1M5, Canada
[b] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO 1432 Ås, Norway
[c] Forest Research Institute, Wonnhaldestraße 4, Baden-Württemberg 79100, Freiburg, Germany
[d] Office National des Forêts, Département RDI, Pôle de Nancy, 11 Ile-de-Corse, 54000, Nancy, France

ARTICLE INFO

Article history:
Received 12 March 2016
Received in revised form 23 June 2016
Accepted 25 July 2016
Available online 3 August 2016

Keywords:
Airborne laser scanner
Aerial photogrammetry
Forest inventory
Canopy heights
Bootstrap confidence intervals

ABSTRACT

Forest inventories, with a probability sampling of a target variable Y and a potentially very large number of auxiliary variables ($\mathbf{X}$) obtained from an aerial laser scanner or photogrammetry, are faced with the issue of model and variable selection when a model for linking Y to $\mathbf{X}$ is formulated. To bypass this step we propose a generic functional regression model (FRM) for use in both a design- and a model-based framework of inference. We demonstrate applications of FRM with inventory data from France, Germany, and Norway. The generic FRM achieved results that were comparable to those obtained with more traditional approaches based on model and variable selections. The proposed FRM generates interpretable regression coefficients and enables testing of practically relevant hypotheses regarding estimated models.

1. Introduction

An increasing number of forest inventories are now supported by remotely sensed auxiliary variables ($\mathbf{X}$) correlated with the attribute(s) of interest (Y) (Lindgren et al., 2015; Massey et al., 2014; Tomppo, 2006). The benefits of exploiting $\mathbf{X}$ are in the form of an improved accuracy and reduced uncertainty in estimates of the mean or total of Y for an area of interest (Magnussen et al., 2015; Mandallaz, 2014; McRoberts et al., 2006; Saarela et al., 2015). The last decades have witnessed a steady increase in available choices of $\mathbf{X}$ and modelling approaches for linking Y to $\mathbf{X}$ (Brosofske et al., 2014). When Y is a traditional tree size or density attribute, an $\mathbf{X}$ that provides information about canopy heights, canopy porosity, or canopy structure is a candidate for exploration (Brosofske et al., 2014; Gillespie et al., 2004; Koch, 2011; McRoberts and Tomppo, 2007).

The exploration of the utility of $\mathbf{X}$ is context specific. In general, the utility depends on the compatibility of the spatial resolution(s) of $\mathbf{X}$ and Y, temporal synchrony, and numerous technical issues connected to the modalities of observation and recording of data (Holmström and Fransson, 2003; Lovell et al., 2005). The consequence is that a pursuit in search of the 'best' link function has almost become a routine activity in forest inventories supported by auxiliary variables (Alves et al., 2010; Hudak et al., 2012; Maltamo et al., 2009). The notion of 'best' is typically formulated by one or more statistics reflecting how well $\mathbf{X}$ predicts Y (Burnham and Anderson, 2002; Chatfield, 1995; Claeskens and Hjort, 2008). The goodness of fit is commonly assessed by a within-sample cross-validation scheme since a replication of the sample is rarely available or the sample size is too small to afford a split into a test and a validation set.

Although the selection of a link function – whether parametric or non-parametric – may be conducted in compliance with best statistical practices, it remains a fact that the model has been selected based on a fit to a single sample. With a multivariate $\mathbf{X}$ and the typical small sample fraction in a forest inventory, there is a non-trivial risk linked to the 'curse of dimensionality' (Marimont and Shapiro, 1979) of having selected the wrong model due to overfitting and underestimating the uncertainty in both model predictions and estimates of Y given $\mathbf{X}$ (Efron, 2014). This problem is particular relevant with data from airborne laser scanners (ALS) and for data in the form of photogrammetric point clouds — i.e. three dimensional point data captured and processed to $\mathbf{X}$-metrics for an area matching the area of an inventory field plot. Here $\mathbf{X}$ can be any of a large number of metrics, quantiles, and summary statistics extracted from the ensemble of point data observed within a spatial support unit for Y. Examples where the dimension of $\mathbf{X}$ is close to or even exceeds the sample size are not rare.

For inventory agencies required to produce estimators with design-based properties (Gregoire, 1998), the practice of a statistically guided search for a suitable model based (dependent) on evidence from the sample, constitutes a problem because the model used for inference

* Corresponding author.
E-mail address: steen.magnussen@canada.ca (S. Magnussen).

http://dx.doi.org/10.1016/j.rse.2016.07.035

has to be external to the sample at hand (Gregoire et al., 2016; Särndal et al., 1992, ch. 6.7). That is, the model must be identified prior to observing the sample. For both design and model based inference a sample-based model and variable selection incurs the risk of overfitting (bias) and underreporting of uncertainty (Claeskens and Hjort, 2008).

Is there a way to avoid or at least sharply reduce the inferential challenges associated with a model and variable selection based on sample data? A solution would have to be in the form of a generic parametric model since alternatives (semi- and non-parametric) are sample dependent inasmuch as the model is data-driven, viz. adapting to the data (Breidt and Opsomer, 2009; Breidt et al., 2007; Montanari and Ranalli, 2005).

A generic functional regression model (FRM) (Ramsay and Silverman, 1997) may be a way forward. The basic tenet of a FRM is that Y can be linked to X via an integral $Y = \int \beta_t X_t \mathrm{d}t$ where t is a continuous variable that imposes an ordering of X. The term β_t is an unknown transfer function transforming X_t to Y. In the current context t is canopy height (unit: m) and X_t is the probability density function evaluated at a canopy height of t. Stated as an integral over t, FRM remains an impractical theoretical construct. Unless we impose structures on X_t and β_t we will not be able to estimate a unique continuous transfer function from a sample of n 'discrete' observations of Y. A discretization of t (over a finite set of intervals) and allowing a lack of fit (error) term brings us back to a linear model $Y = \sum_{t=1}^{T} \beta_t X_t + e_t$ where X_t is the relative frequency of t, and the regression coefficients can be estimated by least squares or any other optimization routine that minimizes the variance of the error terms. Depending on the context, the model may include an intercept.

It is often rational to expect a smooth trend in the discretized transfer function β_t across intervals of t. Without constraints, a fitted transfer function may be wiggly, counterintuitive, or outright counter-factual. We should expect β_t to be zero where there is no relationship between Y and X_t, and also that β_t is constant across intervals where there is no statistically significant change in the relationship. James et al. (2009) introduced a new approach to FRM which applies constraints and ideas from variable selection methods to achieve a smooth rational trend. The constraints are applied to derivatives of β_t to satisfy a priori expectations on trends in β_t. They call their method "Functional Linear Regression That is Interpretable" (FLiRTI) to distinguish it from functional regression based on approximations to Eigen-functions with hard-to-interpret results (Ramsay and Silverman, 1997).

The objective of this study is to demonstrate – in the context of ALS and photogrammetry supported forest inventories – the generic aspect and utility of FLiRTI with its potential to eliminate the model and variable selection step. Demonstration examples are from three forest inventories in France, Germany, and Norway. We discuss how FLiRTI facilitates model comparison and hypothesis testing; and we point to other potential forest inventory applications with FLiRTI.

2. Material and methods

2.1. Våler data and sampling design

The study site, the sampling design, and data have been detailed elsewhere (Næsset, 2002; Næsset et al., 2013). Only a brief summary of data and sampling design is given here.

The study area is a boreal forest of 852.6 ha located in the municipality of Våler in south-eastern Norway. The dominant tree species are Norway spruce (*Picea abies*) and Scots pine (*Pinus sylvestris*). The forest was stratified in 1996 to four strata based on: age class; site productivity; and tree species. The strata are: recently regenerated stands (str. 1); young forest (str. 2); mature spruce forest (str. 3); and mature pine forest (str. 4). We report results from strata 2, 3, and 4. The 1999 areas (A) were: $A_2 = 120.9$ ha; $A_3 = 140.4$ ha; and $A_4 = 195.6$ ha. Stratum 1 was excluded because field data from recently regenerated stands was obtained with a protocol different from that followed in strata 2, 3, and 4.

A stratified field sampling design with 145 circular 200 m^2 field plots was conducted in 1999 and repeated in the fall and spring of 2010 and 2011. Plots were located on a 150 m × 150 m (str. 2 and 3) or a 150 m × 450 m grid (str. 4). Strata sample sizes (n_j, $j = 2, 3, 4$) were $n_2 = 55$, $n_3 = 58$, and $n_4 = 32$. Stratum sizes N_j are expressed as the integer value of a stratum area divided by the area of a field plot. On each plot, the diameter at breast height (DBH) was recorded for all trees with a DBH ≥ 4 cm. In 81 of the plots with fewer than 15 stems, the corresponding tree heights (HT) were measured with a Vertex hypsometer. On the remaining 64 plots (with >15 stems), tree heights were measured on sample trees selected with equal probability. Overall, HT was measured on three to 43 sample trees per plot with an average of 17.8. Above-ground biomass (AGB) of individual trees was predicted using species-specific allometric models (Marklund, 1988). The target attribute Y is AGB per ha in 1999 and 2010 (AGB_{1999} and AGB_{2010}, unit: Mg ha^{-1}) and was computed from the tree-level AGB estimates and the known plot area.

When the sample plots were visited in 2010 (11), 24 had been regenerated in the intervening years. For these plots, field data were collected in a cluster of four 20 m^2 sub-plots located 5.1 m from the plot center.

ALS data were acquired on 8–9 June 1999 with an Optech ALTM 1210 laser scanner and a flying altitude of approximately 700 m. The pulse repetition frequency of 10 kHz and a scan frequency of 21 Hz resulted in a ground point density of approximately 1.2 m^{-2}. In the 2010 campaign, the ALS data were acquired on July 2nd with an Optech ALTM Gemini laser scanner operated at an altitude of approximately 900 m. The pulse repetition frequency was 100 kHz and the scan frequency was 55 Hz yielding a ground point density of approximately 7.3 m^{-2}. Ground echoes were found and classified using the progressive Triangulated Irregular Network (TIN) densification algorithm (Axelsson, 2000). The first-return heights (CH) were calculated subtracting the respective TIN height from the height associated with a first-return echo.

First return CH-values from each field plot were binned to $T = 41$ height classes with the first class designated to a CH between 0 m and 1.3 m, and the remaining to 40 equal width classes to an upper limit of 31.1 m. A row vector $\mathbf{X}$ of length T of relative class frequencies was computed for each plot. A corresponding census vector was obtained for each stratum. Our choice of T is not entirely arbitrary. According to a formula by Freedman and Diaconis (1981) $T = 41$ is a suitable number of bins for both the Våler and the ONF data; T may exceed the sample size. Note, a fast computation of a census $\mathbf{X}$ is possible as it can be obtained without a tessellation of a stratum to units with the size of a field plot.

2.2. Office national de Forêts (ONF) data and sampling designs

Field data from forests in Aillon and the Vosges were collected in February to April of 2011 (Aillon) and 2013 (the Vosges) as part of the Foresee research project (http://foresee.fcba.fr). Field measurements were performed during the winter period from 94 fixed area plots located in Aillon (49), and the Vosges (45). All but one plot had a radius of 15 m with an area of 706.9 m^2. The exception was a 100 m × 100 m plot in Aillon. Species and diameter at a reference height of 1.3 m (DBH) above ground was, as a rule, recorded with a circumferential tape for all live trees with a DBH ≥ 17.5 cm. For the purpose of this study the field plots were (post) stratified by leading species (dominant by basal area) and with plots considered as a simple random sample (without replacement). The strata included in this study have as leading species: beech (*Fagus sylvatica* L.) and fir (*Abies alba* L.). Total stem volume (VOL) in m^3 ha^{-1} is the target attribute (Y) and it was computed via local volume tariffs (Deleuze et al., 2014).

ALS data were acquired in April (the Vosges) and August (Aillon) of 2011 over a 25.5 km^2 forested area in Aillon and a 1362 km^2 forested area in the Vosges. ALS data for Aillon were captured with a Riegl