Characterizing the calibration domain of remote sensing models using convex hulls

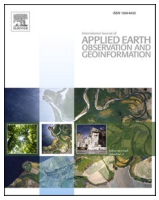
Jean-Pierre Renaud, Ankit Sagar, Pierre Barbillon, Olivier Bouriaud, Christine Deleuze, Cédric Véga

International Journal of Applied Earth Observations and Geoinformation 112 (2022) 102939

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



Characterizing the calibration domain of remote sensing models using convex hulls

J.P. Renaud [a,b,*], A. Sagar [b,c], P. Barbillon [d], O. Bouriaud [e,b], C. Deleuze [a], C. Vega [b]

[a] Office National des Forêts, Département RDI, 8 allée de Longchamp, 54600 Villers lès Nancy, France
[b] Institut National de l'Information Géographique et Forestière (IGN), Laboratoire d'Inventaire Forestier (LIF), 14 rue Girardet, 54000 Nancy, France
[c] UMR Silva, Université de Lorraine, faculté des Sciences et Technologies Campus Aiguillettes, 54506 Vandoeuvre Les Nancy, France
[d] UMR MIA-Paris, Université Paris-Saclay, AgroParisTech, INRAE 16 rue Claude Bernard, 75 231 Paris Cedex 05, France
[e] Ştefan cel Mare University of Suceava, 13 University street, Suceava 720229, Romania

ARTICLE INFO

Keywords:
Calibration domain
Model-based maps
Model transferability
Extrapolation bias
Airborne laser scanning
Sampling effort

ABSTRACT

The ever-increasing availability of remote sensing data allows production of forest attributes maps, which are usually made using model-based approaches. These map products are sensitive to various bias sources, including model extrapolation. To identify, over a case study forest, the proportion of extrapolated predictions, we used a convex hull method applied to the auxiliary data space of an airborne laser scanning (ALS) flight. The impact of different sampling efforts was also evaluated. This was done by iteratively thinning a set of 487 systematic plots using nested sub-grids allowing to divide the sample by two at each level. The analysis were conducted for all alternative samples and evaluated against 56 independent validation plots. Residuals of the extrapolated validation plots were computed and examined as a function of their distance to the model calibration domain. Extrapolation was also characterized for the pixels of the area of interest (AOI) to upscale at population level. Results showed that the proportion of extrapolated pixels greatly reduced with an increasing sampling effort. It reached a plateau (ca. 20% extrapolation) with a sampling intensity of ca. 250-calibration plots. This contrasts with results on model's root mean squared error (RMSE), which reached a plateau at a much lower sampling intensity. This result emphasizes the fact that with a low sampling effort, extrapolation risk remains high, even at a relatively low RMSE. For all attributes examined (i.e., stand density, basal area, and quadratic mean diameter) estimations were generally found to be biased for validation plots that were extrapolated. The method allows an easy identification of map pixels that are out of the calibration domain, making it an interesting tool to evaluate model transferability over an area of interest (AOI). It could also serve to compare "competing" models at a variable selection phase. From a model calibration perspective, it could serve *a posteriori*, to evaluate areas (in the auxiliary space) that merit further sampling efforts to improve model reliability.

1. Introduction

The ever-increasing availability of remote sensing data facilitates the production of forest attributes maps, which are usually made using model-dependant approaches (e.g., McRoberts et al. 2010, Saarela et al. 2015, Magnussen et al. 2016, Stahl et al. 2016, Coops et al. 2021). In such approaches, auxiliary variables are linked to forest attributes through field calibration samples that are used to build models. These calibration samples do not necessarily need to follow a rigorous sampling design, and their purposive selection can sometimes be performed to meet cost effectiveness criteria. According to Magnussen (2015), in simple situations, model-dependant approaches are credible alternatives to the design-based ones usually followed by national forest inventories. Globally, estimates might be considered unbiased when a sufficiently large sample is used for model calibration (Magnussen 2015) and when the model is used within its validity domain. However, locally, for small domains or for conditions not encountered during calibration (i.e., pixels out of the calibration auxiliary space), there is no evidence that model predictions would be reliable or unbiased (Magnussen et al. 2016, Hsu et al. 2020). Even though this crucial role played by the calibration sample is recognised by many authors (e.g., Frazer et al. 2011, Mesgaran et al. 2014, Saarela et al. 2015, Bouvier et al. 2019, Meyer and Pebesma

* Corresponding author at: Office National des Forêts, Département RDI, 8 allée de Longchamp, 54600 Villers lès Nancy, France.
E-mail addresses: jean-pierre.renaud-02@onf.fr (J.P. Renaud), ankit.sagar@ign.fr (A. Sagar), pierre.barbillon@agroparistech.fr (P. Barbillon), obouriaud@usm.ro (O. Bouriaud), christine.deleuze@onf.fr (C. Deleuze), cedric.vega@ign.fr (C. Vega).

https://doi.org/10.1016/j.jag.2022.102939
Received 7 March 2022; Received in revised form 17 July 2022; Accepted 25 July 2022
Available online 28 July 2022

2021), reliability maps of forest attributes are infrequent and, to paraphrase McRoberts (2011), such maps with predictions out of the models' validity domain have a non-negligible risk of being potentially unreliable and rather "pretty pictures".

To minimise model's prediction errors, different strategies have been proposed to efficiently distribute field samples in the auxiliary space of the remote sensing data (e.g. Van Aardt et al. 2006, Pesonen et al. 2009, Hawbaker et al. 2009, Maltamo et al. 2011, Grafström et al. 2014). Some authors have proposed to stratify the ALS data to spread the sampling effort over the whole auxiliary space (Hawbaker et al. 2009, Frazer et al. 2011). Maltamo et al. (2011) also compared different plot selection strategies to optimise precision. Nevertheless, to the best of our knowledge, none have tried to identify the validity domain of models used to predict forest attributes or to identify extrapolated predictions on the resulting maps. Owing to the simplicity of the method (Barber et al. 1996), one motivation of this study was therefore to use convex hulls as a tool for evaluating the quality and the validity domain of models used over an AOI.

Characterizing the validity domain of a regression model is not a new problem, and not restricted to the remote sensing community (Cook 1977; Brooks et al. 1988; Mesgaran et al. 2014; Ebert et al. 2014; Conn et al. 2015; Bouchet et al. 2020, Meyer and Pebesma 2021). In ecology for example, models used outside their calibration range produce frequently unreliable predictions of species distribution for example (Conn et al. 2015). Even though there is no universal definition of extrapolation, some authors have proposed algorithms based on statistical properties of the auxiliary space, such as Cook's or Mahalanobis distances, or dissimilarity indexes to identify extrapolation situations (Cook 1977; Mesgaran et al. 2014; Conn et al. 2015; Bouchet et al. 2020, Meyer and Pebesma 2021), while others have used different forms of convex or concave hulls (Brooks et al. 1988, Ebert et al. 2014; Conn et al. 2015).

In their best practices guide of ALS approaches in forestry, White et al. (2013) illustrated that convex hull could be used to show uncovered forest structures by an ALS model. They showed that an inefficient calibration sample could leave a large part of the model predictions out of the calibration domain. Such a situation is prone to errors, as model transferability could be questioned outside the range of the sampled conditions (Brooks et al. 1988; Conn et al. 2015). Another important aspect about extrapolated predictions is their distance to the calibration data, since a remote prediction is expected to be more prone to bias than one located just beside the calibration domain.

As extrapolated predictions over an AOI are rarely reported in the remote sensing literature and considering that they may produce biased estimates, and locally erroneous maps, a first objective of this study was to use convex hulls to evaluate the proportion of an AOI auxiliary space covered by a model and characterize the residual error associated to the extrapolated predictions. The proportion of extrapolated pixels over the AOI was used as an indicator of model representativity. A secondary objective was to evaluate how sampling intensities affects the degree of extrapolation and how extrapolation distance influence prediction errors.

2. Data and methods

The study area is based on an ALS flight performed in February 2019 in North-eastern France, which covers a forested area of 18,646 ha, with a mean pulse emission density of 16 points per m^2. Within this area, a set of 487 systematic field plots were carried out in the Mouterhouse forest (5,324 ha) during the winter of 2019–2020 (Fig. 1). The Mouterhouse forest consisted of broadleaved, mixed, and coniferous stands (respectively 43 %, 23 % and 34 % on an area basis) that were considered as representative of the whole ALS area. The main species are Scots pine (*Pinus sylvestris*), sessile oak (*Quercus petraea*), beech (*Fagus sylvatica*), Douglas fir (*Pseudotsuga menziesii*) and Norway spruce (*Picea abies*). Diameter (starting at 17.5 cm) and tree species were measured in these calibration plots of 15-meter radius spaced at every ~ 300 m, on a systematic grid. An independent set of 56 plots, located in the same forest area was used as a validation data set.

To evaluate the effect of different sampling efforts, the systematic plot grid was thinned by half several times to obtain coarser grids, down to a minimum of 8 remaining plots, sequentially producing systematic subsamples (Fig. 2). The process follows the approach in used by the French National Forest Inventory and described in Vidal et al. (2007). It consists of a system of nested sub-grids allowing to divide the sample by two at each level. As indicated in Vidal et al. (2007) such an approach preserves the properties of the systematic sample at every grid levels. Furthermore, multiple subsamples of the sampling intensity were used at each thinning iteration, using alternative plot samples of the same grid resolution as replicates (i.e., for example the green and blue dots of the first iteration in Fig. 2). These replicates allowed to obtain estimates of attributes variability at each grid resolution (i.e., for each level of sampling intensity). As the number of candidate replicates grow exponentially at each iteration, a maximum of 40 replicates were used at lower grid resolutions for computational ease. When all replicates of a given grid resolution had not exactly the sample number of plots due to the spatial configuration of the forest, each replicate was trimmed to the minimal value found for that given resolution by randomly removing exceeding plots. This approach was retained since the initial sampling design was systematic and allowed a more even spatial repartition of the sampling effort. As a results, more precise estimates were then available at coarser grid resolution, resulting from the larger number of replicates for these lower sampling efforts.

As shown in Fig. 1, no field plots were available for 71 % of the ALS acquisition area (the orange section). This situation offers the opportunity to examine the impact of model application outside of its "spatial" calibration area. Standard area-based ALS metrics (ABA) were computed using the R package lidR (Roussel and Auty 2021). Metric computation

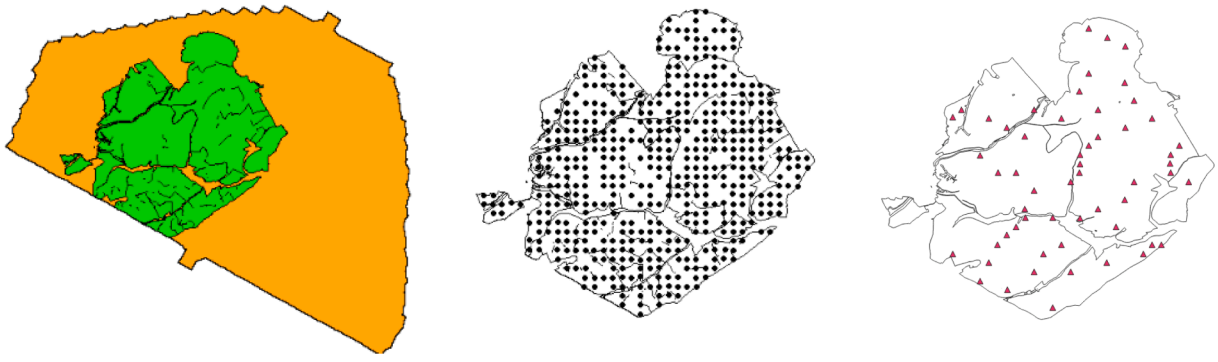


Fig. 1. The ALS flight area in orange, with the Mouterhouse forest in green (left) and for illustration, its initial 487 systematic (middle) and 56 validation (right) plots.

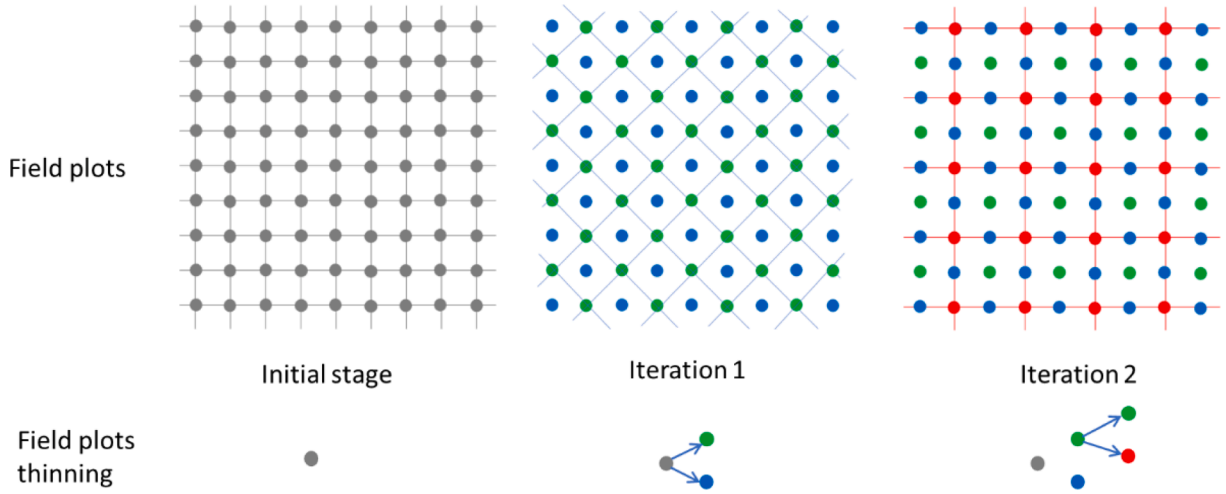


Fig. 2. Illustration of the thinning process from the initial dense grid (left) to the second thinning iteration (right). Field pots are symbolised by dots. At each iteration, every second plot is removed, yielding a thinned grid. At each iteration, all plots of the same spatial resolution were used as replicates of the same sampling effort (e.g., 2 replicates at iteration 1: green and blue dots).

was performed over the field plots, as well as over a raster grid of 30 m resolution for the whole ALS area. A simple working linear regression model (Magnussen et al. 2012, Bouvier et al. 2015) was adjusted to estimate basal area (G), quadratic mean diameter (Dg) and tree density per hectare (N) separately. The retained working model parameters (Eq. (1)) included 3 independent variables: mean (Hmean), and standard deviation (Hsd) of all ALS pulse heights and the average slope (Slope) of the pixels (or of the field plots), computed from the digital terrain model (DTM) using the R package terra (Hijmans 2021). The working model is as follow:

$$(N, G, Dg) \sim \beta_1 H_{mean} + \beta_2 H_{sd} + \beta_3 Slope + \varepsilon \quad (1)$$

where β_i are model parameters obtained for each dependant variables separately *(N, G, Dg)*.

Models' performance was evaluated using root mean square error (RMSE) computed from the validation samples, as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

where n is the number of validation field plots, y_i the observed values of field attributes and $\hat{y}_i$ the predicted values. RMSE was further divided by the validation attribute mean to obtain a relative RMSE.

Using the calibration auxiliary space of the model independent variables, convex hulls were computed for all field calibration configurations associated to the different sampling efforts compared. The "geometry" package (Habel et al. 2019) was used to build convex hulls around the the auxiliary space of the calibration domains. The volume of the convex hull was also extracted from the scaled auxiliary space. Then, validation plots or pixels of the AOI were classified as being located inside or outside the hulls using the "inhull" function of the same package. For the Mouterhouse forest, as well as for the extended ALS area, the proportion of extrapolated pixels were computed. The yaImpute R package (Crookston and Finley 2008), was then used to obtain their Euclidian distances (using scaled X variables) to the nearest calibration plot. The same operation was performed for the validation plots which were also used to compute residuals from the models' predictions (Eq. (3)):

$$residual_i = (y_i - \hat{y}_i) \quad (3)$$

where, y_i is the observed value, $\hat{y}_i$ is predicted value. Means and standard deviations of all these indicators were obtained for each level of sampling intensity by aggregating replicates and were reported in the relevant figures.

3. Results

A conventional way to evaluate the quality of a model adjustment is to examine its validation RMSE. In this study, based on the results of decreasing sampling intensities, it appeared that the relative mean RMSEs tended to reach a plateau when ca. 50 calibration plots were used to build models, regardless of the forest attributes examined (Fig. 3a). This plateau represented relative RMSE values of ca. 18 %, 27 % and 42 % for Dg, G and N respectively. In absolute values, it represented a mean RMSE of 5.9 cm for Dg, 6.5 m^2/ha for G and 117 stems/ha for N.

Mean bias per sampling intensity is shown in Fig. 3b. While for all forest attributes examined no bias was globally observed, a large variability was nevertheless present at small sampling intensities. It showed that in such a situation (e.g., with less than 25 plots to calibrate a model) it was possible to obtain a bias larger than 10 % for a given replicate (Fig. 3b).

As opposed to RMSE, the mean proportion of extrapolated pixels remained relatively high at low sampling intensities (Fig. 3c). For example, with less than 100 calibration plots, more than 25 % of the pixels over the Mouterhouse forest, or 29 % over the whole Lidar area were located outside the auxiliary space defined by the calibration hulls. Only minor differences in proportion of extrapolated pixels between the two AOI examined were observed, slightly more extrapolated pixels (ca. 3 %) were observed within the extended AOI, as opposed to the Mouterhouse forest area. With the full calibration dataset (487 plots) the proportion of extrapolated pixels was less than 18 % (Fig. 3c).

When the sampling effort decreases, a larger number of pixels were found outside the calibration hull and their distances to the nearest calibration plots (in the auxiliary space) tended to increase. For example, with a sampling effort of 487 calibration plots, the mean extrapolation distances observed were respectively 0.6 and 0.8 for the Mouterhouse forest or the extended area respectively, while it was almost twice, i.e., 1.2, when less than 20 plots were used for calibration (Fig. 3d). Greater extrapolation distances to the calibration domain were also consistently observed for the extended lidar area as compared to the Mouterhouse forest, as shown in Fig. 3d. The mean interpolation distance (i.e., distance among the auxiliary space of the calibration plots) was also shown in Fig. 3d (the dashed line). This mean interpolation distance (of the calibration plots) was low (0.3) for the largest sampling effort (487 plots) but increased with at reduced sampling intensities. When less than 50 plots were used to calibrate models, the mean

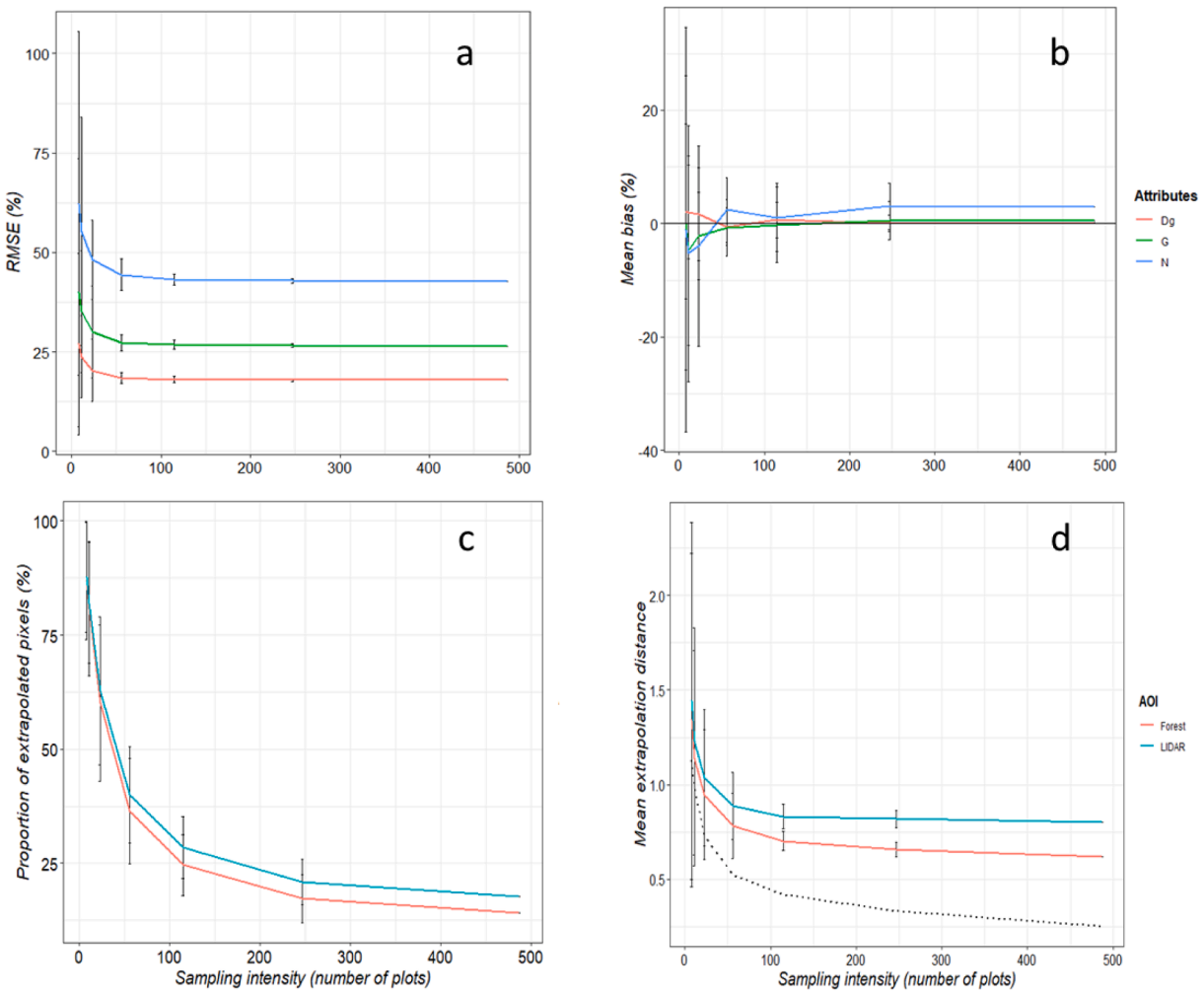


Fig. 3. Impact of sampling intensities on: (a) the relative root means squared errors (RMSE in %) of validation plots for the different forest attributes examined; (b) the mean relative bias (in %); (c) the mean proportion of pixels located outside the calibration hull for each AOI (whole Lidar area or the Moutherhouse forest only); and (d) the mean scaled extrapolation distances for each AOI. (The dashed line is the mean interpolation distance for the calibration plots). In all graphs, error bars represent 2 standard deviations of the replicated sampling intensities.

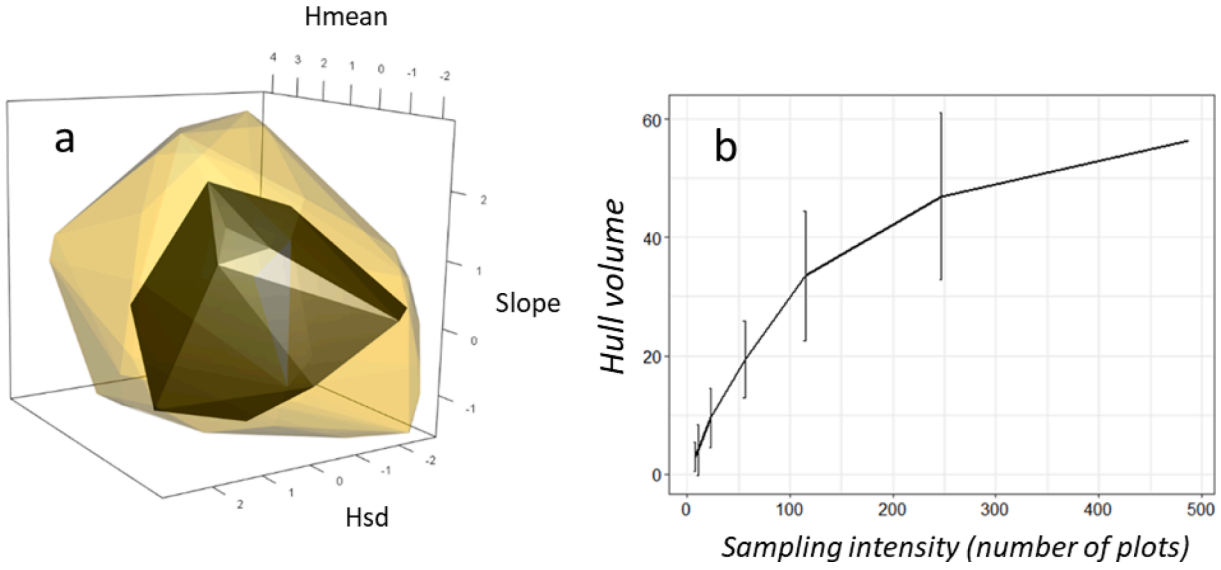


Fig. 4. 3D representation of the convex hull (a) obtained from the model's auxiliary space using 487 (yellow) and 23 calibration plots (black) and (b) the effect of sampling effort on the convex hull volumes. (As all variables are scaled, the hull volume is unitless.).

interpolation distance tended to be larger than 0.5 (Fig. 3d).

The explored variability of the auxiliary space was also presented in terms of convex hulls volumes (Fig. 4). Fig. 4a presents envelops of the auxiliary space for two sampling intensities: 487 plots (in yellow) and 23 plots (in black). An obvious shrinkage of the volume of the calibration domain was associated to the reduction of the sampling intensity. To illustrate more completely this phenomenon, the volumes of the convex hulls were presented as a function of the sampling intensity in Fig. 4b. From that figure, it appeared that below a sampling effort of 247 plots, a large reduction in the hull volumes was observed. The convex hull volumes were also computed (data not shown) for the pixel's population of the Mouterhouse forest and the extended lidar area. It yielded volumes 2.5 and 2.9 times larger than the convex hull volume obtained with the full set of calibration plots (487 plots). This result suggested that a large part (greater than60 %) of the convex hull volumes of the AOI was not included within the volume of the calibration domain.

For each replicate of the sampling grid resolutions, validation plots were classified as being inside or outside the calibration domain, based on their auxiliary space. The forest attributes observed in each group were aggregated, and their distribution compared (Fig. 5a-c). Results showed, for G and N, that the validation plots located outside of the hulls (InHull = FALSE) were more frequently at the lower end of the distributions compared to interpolated ones (InHull = TRUE) (Fig. 5a and c). For quadratic mean diameter, both distributions were overlapping (Fig. 5b). This translated, for basal area and stand density, into a clear tendency for an underestimation bias associated with an increase in extrapolation distances (Fig. 5d and f). For quadratic mean diameters, a trend toward reduced residuals with increasing extrapolation distance was also observed. Clearly from Fig. 5 (d-f), even for large sampling efforts (green and blue points), the extrapolated plots were globally underestimated for G and N and overestimated for Dg. This result underlined the importance of minimising extrapolation in model predictions and suggested a contribution of the extrapolation distance to bias.

4. Discussion

Our results are emphasizing the interest of convex hulls to identify model predictions that are made outside their calibration range not only for producing potentially more reliable maps, but also to ascertain forest attribute predictions. Even though it is well-known that unreliable predictions may results from such situations (Brooks et al. 1988), the proportion of extrapolated predictions are rarely reported in the remote sensing community involved in forest modelling. In most studies, the main model quality indicators reported are often the RMSE and coefficient of determination (R2). However, as underlined by Persson and Stahl (2020), the use of RMSE has shortcomings, since it represents only a limited facet of the error structure. In a simulation study, Kangas et al. (2016) also showed that RMSE can significantly underestimate the real model uncertainty. Our results showed that RMSE tended to reach a plateau, at sampling intensities where the proportion of extrapolated pixels was not yet stabilized. This result is not due to a curse of dimensionality problem (Sagar et al. 2021) since we purposefully used a simple working model (Magnussen et al. 2012, Bouvier et al. 2015). At low sampling intensities the calibration domain tended to have a reduced convex hull volume (Fig. 4), and the proportion of extrapolated pixels tended clearly to increase. Distances of extrapolated pixels to the calibration domain also tended to increase at low sampling intensities (Fig. 3). This shrinkage of the model's auxiliary space is problematic, since one would expect the calibration domain to cover as much as possible the ranges of the target population to produce reliable predictions (Stage and Crookston 2007). This is particularly critical for models that cannot extrapolate due to their nature (e.g., non-parametric models such as k-NN or random Forest) or for non-linear relationships of

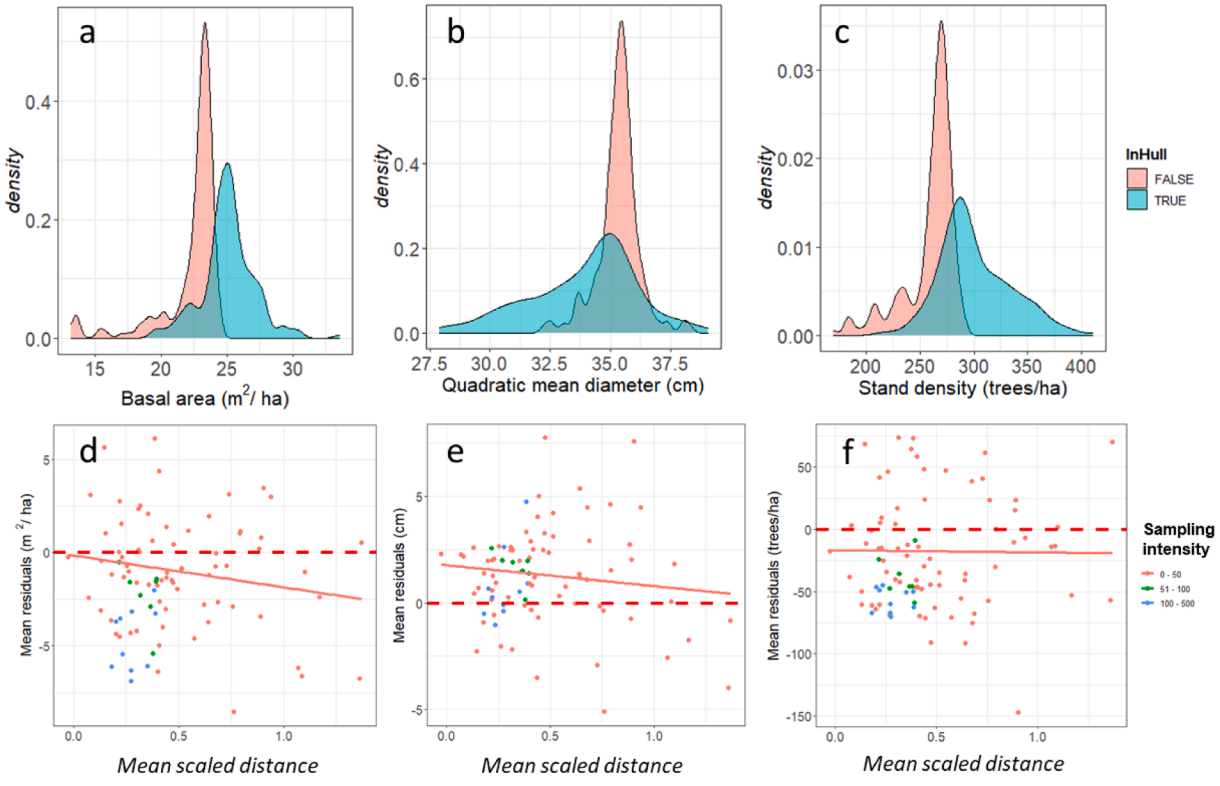


Fig. 5. Distribution of forest attributes of the mean values of the validation plots grouped as inside (InHull = TRUE) or outside (InHull = FALSE) the calibration domain, aggregated over all replicates of sampling intensities (a-c). Mean residuals of the extrapolated plots (d-f, InHull = FALSE) are also given according to their mean distances to the calibration domain. The solid lines in plates d to f represented the regression line. Legend corresponds to classes of sampling intensities. (Attributes are basal area (a, d), quadratic mean diameter (b, e), and tree density per hectare (c, f).

forest attributes over a wider range of the predictors space (Stage and Crookston 2007, Magnussen et al. 2010, Corona et al. 2014). In heterogeneous forests, a low sampling intensity is thus synonymous of a restricted calibration domain and of a potentially low model transferability to unsampled areas (White et al. 2013, Mesgaran et al. 2014, Meyer and Pebesma 2021), especially when empirical models are used.

Of course, the extrapolation problem mentioned above is particularly important in a model-based perspective, or when maps are directly produced from model predictions. Forest managers receiving such maps containing locally extrapolated predictions should be warned, and pixels out of the model's validity domain should be identified as potentially unreliable. However, in a design-based perspective, an estimator frequently used is the generalized regression estimator (GREG). For internal models, GREG is considered as approximately unbiased, regardless of how well the model fits the relationship between the inventory and auxiliary data (Gregoire 1998, Lehtonen et al. 2003, Corona et al. 2014, Stahl et al. 2016, Wojcik et al. 2022). This absence of bias is nevertheless associated with an adequate sample size and the assumption that residuals obtained from the calibration data are representative of the whole population. These conditions thus permit to correct model misspecification (Sarndal et al. 1992, Lehtonen et al. 2003). However, for operational forest management, concerns are frequently associated with small domains where only few, or even no calibration data are available to correct model errors. In GREG, two terms are used: the mean (or total) of model predictions for each pixel of the AOI, and a Horvitz-Thompson estimator of the residuals computed over the calibration plots (Sarndal et al. 1992, Gregoire 1998, Corona et al. 2014, Stahl et al. 2016, Wojcik et al. 2022). In Moutherhouse, predictions that were extrapolated had frequently non-null mean residuals. They concerned mainly young stands with low G and a low number of tallied stems, as shown in Fig. 5, but also occurred in mature plots still having low G and N, but large Dg suggesting that they resulted from thinning operations. Extrapolated predictions were mainly underestimated for G and N and overestimated for Dg (Fig. 5d-f). Even though it is out of the scope of this study, an evaluation of the impact of extrapolations on GREG's estimation of small domains, would certainly be worth further investigations.

We used an initial sampling effort of *ca.* 1 field plots per 8 ha (487 plots). At this intensity, less than 18 % of the pixels from the AOI were found to be outside of the calibration domain. However, this proportion raised drastically when the sampling grid was thinned to a density of 1 plot every 33 ha or more (<115 calibration plots over the AOI). A similar pattern was nevertheless not observed for RMSE, that reached a plateau at a sampling intensity of *ca.* 1 field plot per 70 ha (50 plots) (Fig. 3). When the sampling grid contained less than 50 calibration plots, the proportion of extrapolated pixels exceed 35 % and mean predictions of extrapolated pixels were biased (Fig. 5). Interestingly, a sampling intensity of 1 field plot per 70 ha (50 plots) is also the sample size that allow G to be estimated with a precision of 5 % based on the field sample plots alone (Kangas and Maltamo 2006).

In a Wisconsin forest, Hawbacker et al. (2009) used a much lower sampling intensity (one plot per 700 ha) but stratified their training plots using the ALS predictor space. With this approach, they extended their calibration domain and obtained more accurate results, as compared to a simple random sampling design. This improvement was thus directly related to a better representativity of their calibration domain. In a Norwegian forest, Maltamo et al. (2011) also observed that stratifying field plots based on ALS data improved relative RMSE by 5 to 10 %. That improvement was nevertheless dependant on sampling intensities. With a large sampling effort (at least 1 plot per 340 ha), all sampling designs tested had similar RMSE (Maltamo et al. 2011). We conjecture that the convergence in RMSE past a given sampling intensity is related to the proportions of extrapolated pixels in these studies, which unfortunately were not quantified or reported. Spatially systematic samples have favourable properties because they are spatially balanced (Stevens 1997, Stevens and Olsen 2004), thus avoiding sampling voids over an AOI (Christianson and Kaufman 2016, Meyer and Pebesma 2021). In our case study, a systematic sampling design was used. This sample could be considered as efficient in obtaining a calibration sample covering all aspects of the AOI. But it seems that under a sampling effort of 1 field plot per 70 ha (50 plots), a large amount of the AOI's auxiliary space is left uncovered. This could be related to the heterogeneity or the aggregation structure of the forest. Indeed, spatial regularity, or spatial structures can strongly reduce the efficiency of a systematic sample (Stevens and Olsen 2004), as some aggregated forest structures may be missed bellow a given sampling resolution.

More advanced sampling methods can be implemented to reduce extrapolation. For instance, a sampling based on the cube method could be tested (Grafström et al. 2014). This method aims to improve the spread of the probabilistic samples in the space of some auxiliary variables and could be used to sample from the space of the dependent variables used for model construction. Even though the impact of extrapolation on the prediction's bias hasn't been fully studied or established (Magnussen et al. 2010), it is probable that the informative spreading of the sample could enable to reduce extrapolation situations.

Several authors have shown that NFI data are efficient to train ALS models (Hollaus et al. 2007, Maltamo et al. 2009, Véga et al. 2021). In France, the NFI plot density, over a 5-year period, represents an average sampling effort of 1 plot per 570 ha (Hervé 2016). This intensity over the Moutherhouse forest appears to be low and would probably have led to a large proportion of extrapolated pixels. The use of convex hull in such contexts, could contribute to improved model reliability through identification of area of the predictor space requiring complementary sampling effort or to identify potentially unreliable predictions areas in maps, made for decision-makers or forest managers (Sagar et al. 2021). Certainly, problems associated to low sampling efforts could be more easily diagnosed using the proposed convex hull approach. A drawback though of the method is its large computational requirements in high dimensionalities, that restricts its use to the comparison of parsimonious models. Nevertheless, the question of model transferability at different scales remains a question beyond the scope on the present study.

With an increased sampling intensity, mean extrapolation distances tended to be reduced. Meyer and Pebesma (2021) used such a distance-based criterion to estimate prediction's reliability in random Forest models. They used the mean calibration distance of their auxiliary space (i.e., interpolation distances) to build an index of potentially spurious predictions. From Fig. 5, such a tendency of larger biases with increasing extrapolation distances can be observed, but the relationship is noisy. From a calibration perspective though, a clear impact of the sampling intensities is observed on the mean interpolation distances (Fig. 3), reflecting a denser calibration auxiliary space with larger sampling efforts.

Convex hull could also serve at a model selection stage. Comparing convex hull volumes of competing models with the the ones of the AOI could serve as indicators of models' extrapolation potential. The lowest the hull volume ratio, the lowest the extrapolation risk. In addition to the volume ratio, the extrapolation distance could serve as an indicator of outliers. Models with extrapolation distance largely above their interpolation counterpart should be considered with care. The impact of models' predictions out of their validity domain on maps is certainly an aspect that worth further investigation and this study is a first step toward the use of this interesting tool.

5. Conclusions

This study presented the use of convex hulls as a multivariate tool allowing to identify model's calibration domains and their transferability over an AOI. It showed that the achieved RMSE can hide extrapolation issues, with risks of local bias. Our results also suggest that addressing the extrapolation issue requires a higher sampling intensity than the one considered for achieving a given RMSE target. The use of convex hulls allows to produce maps with identified predictions that are

made outside model's calibration domains. From an operational forest management point of view, knowing where predictions are out of there validity domain could represent a valuable outcome of this approach. Convex hulls could also help to compare "competing" models at a variable selection phase, provided that a limited number of variables are constituting the auxiliary space, due to the large computational requirement of the convex hull method in high dimensions. From a model calibration perspective, it could also serve *a posteriori*, to evaluate areas (in the auxiliary space) that merit further sampling efforts to improve calibration or evaluate prediction bias. Finally, further studies are required to examine the possibility of correcting map bias using the extrapolation distances.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We acknowledge the Laboratoire Interdisciplinaire des Environnements Continentaux (CNRS et l'Université de Lorraine) for supplying the ALS data and the Agence Territoriale de Mulhouse (ONF) for the field data. We thank Steen Magnussen for enlightening discussions. We thank Alain Munoz, Flavien Lamiche, and Vincet Pérez (ONF) for their help in preparing and supplying data.

Funding

ONF Département RDI is supported by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (ANR-11-LABX-0002-01, Lab of Excellence ARBRE). Ankit Sagar received the financial support of the French PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE, through the project Impact DeepSurf.. Olivier Bouriaud acknowledges support from the project funded Ministry of Research, Innovation and Digitalization within Program 1 - Development of national research and development system, Subprogram 1.2 - Institutional Performance - RDI excellence funding projects, under contract no. 10PFE/2021.

References

Barber, C.B., Dobkin, D.P., Huhdanpaa, H., 1996. The quickhull algorithm for convex hulls. ACM Trans Math Softw 22, 469–483. https://doi.org/10.1145/235815.235821.

Bouchet, P.J., Miller, D.L., Roberts, J.J., Mannocci, L., Harris, C.M., Thomas, L., 2020. dsmextra: Extrapolation assessment tools for density surface models. Methods Ecol. Evol. 11, 1464–1469. https://doi.org/10.1111/2041-210X.13469.

Bouvier, M., Durrieu, S., Fournier, R.A., Renaud, J.P., 2015. Generalizing predictive models of forest inventory attributes using an area-based approach with airborne LiDAR data. Remote Sens. Environ. 156, 322–334. https://doi.org/10.1016/j.rse.2014.10.004.

Bouvier, M., Durrieu, S., Fournier, R.A., Saint-Geours, N., Guyon, D., Grau, E., Boissieu, F., 2019. Influence of Sampling Design Parameters on Biomass Predictions Derived from Airborne LiDAR Data. Can J Remote Sens 45, 650–672. https://doi.org/10.1080/07038992.2019.1669013.

Brooks, D.G., Carroll, S.S., Verdini, W.A., 1988. Characterizing the Domain of a Regression Model. Am. Stat. 42, 187–190. https://doi.org/10.2307/2684998.

Conn, P.B., Johnson, D.S., Boveng, P.L., 2015. On Extrapolating Past the Range of Observed Data When Making Statistical Predictions in Ecology. PLoS ONE 10, e0141416. https://doi.org/10.1371/journal.pone.0141416.

Cook, R.D., 1977. Detection of Influential Observation in Linear Regression. Technometrics 19, 15–18. https://doi.org/10.2307/1268249.

Coops, N.C., Tompalski, P., Goodbody, T.R.H., Queinnec, M., Luther, J.E., Bolton, D.K., White, J.C., Wulder, M.A., van Lier, O.R., Hermosilla, T., 2021. Modelling lidar-derived estimates of forest attributes over space and time: A review of approaches and future trends. Remote Sens. Environ. 260, 112477 https://doi.org/10.1016/j.rse.2021.112477.

Corona, P., Fattorini, L., Franceschi, S., Scrinzi, G., Torresan, C., 2014. Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: model-based, design-based, and hybrid perspectives. Can. J. For. Res. 44, 1303–1311. https://doi.org/10.1139/cjfr-2014-0203.

Crookston, N.L., Finley, A.O., 2008. yaImpute : An R Package for k NN Imputation. J. Stat. Softw. 23. https://doi.org/10.18637/jss.v023.i10.

Ebert, T., Belz, J., Nelles, O., 2014. Interpolation and extrapolation: Comparison of definitions and survey of algorithms for convex and concave hulls. In: 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 310–314.

Frazer, G.W., Magnussen, S., Wulder, M.A., Niemann, K.O., 2011. Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. Remote Sens. Environ. 115, 636–649. https://doi.org/10.1016/j.rse.2010.10.008.

Grafström, A., Saarela, S., Ene, L.T., 2014. Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. Can. J. For. Res. 44, 1156–1164. https://doi.org/10.1139/cjfr-2014-0202.

Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. 28, 1429–1447. https://doi.org/10.1139/x98-166.

Habel, K., Grasman, R., Gramacy, R.B., Mozharovskyi, P., Sterratt, D.C., 2019. geometry: Mesh Generation and Surface Tessellation. R package version (4), 5. https://CRAN.R-project.org/package=geometry.

Hawbaker, T.J., Keuler, N.S., Lesak, A.A., Gobakken, T., Contrucci, K., Radeloff, V.C., 2009. Improved estimates of forest vegetation structure and biomass with a LiDAR-optimized sampling design. J Geophys Res G. Biogeosciences 114, 11 p. https://doi.org/10.1029/2008JG000870.

Hervé, J.C., 2016. France. In: Vidal, C., Alberdi, I.A., Hernández Mateo, L., Redmond, J.J. (Eds.), National Forest Inventories: Assessment of Wood Availability and Use. Springer International Publishing, Cham, pp. 385–404.

Hijmans, R.J., 2021. terra: Spatial Data Analysis. R package version 1.2-10. https://CRAN.R-project.org/package=terra.

Hollaus, M., Wagner, W., Maier, B., Schadauer, K., 2007. Airborne laser scanning of forest stem volume in a mountainous environment. Sensors. 7, 1559–1577.

Hsu, Y.-H., Chen, Y., Yang, T.-R., Kershaw, J.A., Ducey, M.J., 2020. Sample strategies for bias correction of regional LiDAR-assisted forest inventory Estimates on small woodlots. Annals of Forest Science 77, 75. https://doi.org/10.1007/s13595-020-00976-8.

Kangas, A., Maltamo, M., 2006. Forest Inventory: Methodology and Applications. Springer Science & Business Media.

Kangas, A., Myllymäki, M., Gobakken, T., Næsset, E., 2016. Model-assisted forest inventory with parametric, semiparametric, and nonparametric models. Can. J. For. Res. 46 (6), 855–868.

Lehtonen, R., Sarndal, C.-E., Veijanen, A., 2003. The effect of model choice in estimation for domains, including small domains. Survey Methodology, Statistique Canada 29, 33–44.

Magnussen, S., 2015. Arguments for a model-dependent inference? Forestry: An International Journal of Forest Research 88, 317–325. https://doi.org/10.1093/forestry/cpv002.

Magnussen, S., Tomppo, E., McRoberts, R.E., 2010. A model-assisted k-nearest neighbour approach to remove extrapolation bias. Scand. J. For. Res. 25, 174–184. https://doi.org/10.1080/02827581003667348.

Magnussen, S., Næsset, E., Gobakken, T., Frazer, G., 2012. A fine-scale model for area-based predictions of tree-size-related attributes derived from LiDAR canopy heights. Scand. J. For. Res. 27, 312–322. https://doi.org/10.1080/02827581.2011.624116.

Magnussen, S., Frazer, G., Penner, M., 2016. Alternative mean-squared error estimators for synthetic estimators of domain means. Journal of Applied Statistics 43, 2550–2573.

Maltamo, M., Packalén, P., Suvanto, A., Korhonen, K.T., Mehtätalo, L., Hyvönen, P., 2009. Combining ALS and NFI training data for forest management planning: a case study in Kuortane, Western Finland. Eur. J. For. Res. 128, 305–317.

Maltamo, M., Bollandsas, O.M., Næsset, E., Gobakken, T., Packalen, P., 2011. Different plot selection strategies for field training data in ALS-assisted forest inventory. Forestry 84 (1), 23–31. https://doi.org/10.1093/forestry/cpq039.

McRoberts, R.E., 2011. Satellite image-based maps: Scientific inference or pretty pictures? Remote Sens. Environ. 115, 715–724. https://doi.org/10.1016/j.rse.2010.10.013.

McRoberts, R.E., Cohen, W.B., Næsset, E., Stehman, S.V., Tomppo, E.O., 2010. Using remotely sensed data to construct and assess forest attribute maps and related spatial products. Scand. J. For. Res. 25, 340–367. https://doi.org/10.1080/02827581.2010.497496.

Mesgaran, M.B., Cousens, R.D., Webber, B.L., 2014. Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. Divers. Distrib. 20, 1147–1159. https://doi.org/10.1111/ddi.12209.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650.

Persson, H.J., Ståhl, G., 2020. Characterizing Uncertainty in Forest Remote Sensing Studies. Remote Sensing 12, 505. https://doi.org/10.3390/rs12030505.

Pesonen, A., Leino, O., Maltamo, M., Kangas, A., 2009. The comparison of fi eld sampling methods and the use of airborne laser scanning as auxiliary data for assessing coarse woody debris. For. Ecol. Manage. 257, 1532–1541.

Roussel, J.R., Auty, D., 2021. Airborne LiDAR Data Manipulation and Visualization for Forestry Applications. R package version 3 (1), 1. https://cran.r-project.org/package=lidR.

Saarela, S., Schnell, S., Grafstrom, A., Tuominen, S., Hyyppa, J., Nordkvist, K., Kangas, A., Stahl, 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 (11), 1524–1534. https://doi.org/10.1139/cjfr-2015-0077.

Sagar, A., Véga, C., Piedallu, C., Bouriaud, O., Renaud, J.-P., 2021. High resolution mapping of forest resources and prediction uncertainty using multisource inventory approach. In: In ProceedIngs of the SilviLaser Conference 2021, pp. 219–221.

Sarndal, C.-E., Swensson, B., Wretman, J.H., 1992. Model assisted survey sampling. Springer, New York.

Stage, A.R., Crookston, N.L., 2007. Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. Forest Science 53, 62–72.

Stahl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E., Gregoire, T.G., 2016. Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. Forest Ecosystems 3, 5. https://doi.org/10.1186/s40663-016-0064-9.

Stevens Jr, D.L., 1997. Variable Density Grid-Based Sampling Designs for Continuous Spatial Populations. Environmetrics 8, 167–195. https://doi.org/10.1002/(SICI)1099-095X(199705)8:3<167::AID-ENV239>3.0.CO;2-D.

Stevens, D.L., Olsen, A.R., 2004. Spatially Balanced Sampling of Natural Resources. J. Am. Stat. Assoc. 99, 262–278. https://doi.org/10.1198/016214504000000250.

van Aardt, J., Wynne, R.H., Oberwal, R.G., 2006. Forest volume and biomass estimation using small-footprint lidar distributional parameters on a per-segment basis. For. Sci. 52, 636–649.

Vidal, C., Belouard, T., Herve, J.-C., Robert, N., Wolsack, J. 2007. A new flexible forest inventory in France. In: McRoberts, Ronald E., Reams, Gregory A., Van Deusen, Paul C., McWilliams, William H. (Eds.), Proceedings of the seventh annual forest inventory and analysis symposium; October 3-6, 2005; Portland, ME. Gen. Tech. Rep. WO-77. Washington, DC: U.S. Department of Agriculture, Forest Service: 67-73.

White, J.C., Wulder, M.A., Vastaranta, M., Coops, N.C., Pitt, D., Woods, M., 2013. A best practice guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. The Forestry Chronicle 89 (6), 722–723.

Wojcik, O.C., Olson, S.D., Nguyen, P.-H.V., McConville, K.S., Moisen, G.G., Frescino, T.S., 2022. GREGORY: A Modified Generalized Regression Estimator Approach to Estimating Forest Attributes in the Interior Western US. Frontiers in Forests and Global Change 4.