Review Round 1

30. Dezember 2017

Reviewer 1:

This relevant and interesting study describes the development of a working/linking model to be used with model-supported estimators. The challenge of the study was the large spatial extent in combination with high resolution auxiliary variables and field data that resulted in severe inconsistencies that had to be handled. The applied approach is statistically rigorous and adequately described; the text is well written and clear.

Comments:

1. *Even though the development of the working/linking model is important, I would have appreciated if it was also presented what the effect of using the model is for estimates. How big is the relative efficiency for estimates in RLP? Consider discussing the impacts of using an internal model (as here) vs. an external model.*

   We are currently writing a follow-up article, where the 'final' regression model presented in this article is in fact used as an internal model for model-assisted small area estimations of standing timber volume within 390 forest management units in RLP. We can already tell that the relative efficiencies ranged between XX and YY (average relative efficiency of VV), and we would very much appreciate having you as a reviewer also for this upcoming article. While the demonstration of a small area double sampling procedure for the German NFI has actually been the underlying overall objective, we decided to publish the findings in two separated articles. Our motivation to address the model building in an own first article as a pre-study has been that the identification of the 'best possible' regression model turned out to be a major issue of the entire study, facing the inconsistencies described in the article. In particular, we came to the conclusion that the three major issues (heterogeneity in the remote sensing data, identification of the optimal support under angle count sampling, incorporating tree species information including the handling of misclassification) have not yet been dealt with in this detail - at least for similar studies in terms of study size and diversity of forest structures and tree species (see Maack et al. {Modelling the standing timber volume of Baden-Württemberg—A large-scale approach using a fusion of Landsat, airborne LiDAR and National Forest Inventory data}). We also considered the findings to be important for a broader range of studies than those of double sampling estimations (e.g. mapping).

   Note: *We could provide the reviewer with some additional details about the rel. efficiency etc.*

2. *Please discuss the results with the study by Kirchhoefer, et al. (2017) Considerations towards a Novel Approach for Integrating Angle-Count Sampling Data in Remote Sensing Based Forest Inventories, Forests*

   Note: *Very good. We shoud discuss ...*

3. *Despite of the large number of observations, the number of 39 explanatory variables is quite large for a linear model. Could you discuss implications of this? Except for 2007, the model parameters do not seem to be all too different (Fig 7). Consider merging several years (all except for 2007?) into one factor as a simple means to reduce the number of model parameters.*

   1) Concerning the number of parameters, we here considered the often cited rule of thumb by Draper and Smith (2014) {Applied regression analysis}, i.e. one should at least have 10 observations per parameter in the regression model in order to avoid the issue of overfitting. For our data set, that would have implied a sample size of at least 390 observations. Since the actual number of observations used for model fitting was 5206 and hence considerably beyond the threshold suggested by Draper and Smith, the number of 39 parameters was not regarded to be critical.

   2) We tested merging several LiDAR years ...

   Note:
   1)Add this argumentation to the article (?).
   2) Check and address no. of parameters within strata! Mention this issue in article and include in answer above. Maybe it turns out that fitting the model on the plot level instead on the cluster level increases the number of observations and is thus to be preferred in our cases (we could include this an argument for this in the article).

4. *The use of square supports is at least uncommon. Basically all studies I am aware of use circular supports (in the case of circular sample plots as here). Therefore, this choice should be justified a bit more or be revised. The reason given that the support should allow for a potential tessellation does not seem to hold because the exact plot location will not allow for an alignment with a potential tessellation grid anyways.*

To my understanding, the area (size) of the support needs to fit to the field data and the tessellation grid, especially if scale-dependent explanatory variables are used. The shape of the support should resemble the field data. I think the model variance is artificially increased by selecting a support that does not fit (i.e. is not circular) with the field data.

1) Rational for rectangular supports: After reconsideration, we became aware that the reason given in the article for using rectangular supports (i.e. theoretical tessellation of the forest area) was indeed not correct. In the *infinite population approach* used by the model-assisted estimators by Mandallaz (and also those proposed by Saborowski), the estimators do in fact not impose any assumptions on the geometry and size of the supports. This is because the explanatory variables are sampled from an infinite population of points defining the continuous distribution (i.e. surface) of the explanatory variable. However, the theoretical tessellation is one prerequisite of model-dependent estimators in order to properly define the population of elements (often referred to as 'pixels' in which the explanatory variables are calculated) used for estimation and is thus not at all uncommon (see e.g. Breidenbach, ...). We rephrased the respective section (now page ... line ...) ... .

Note: *We should rephrase the argument for the support geometry in the article and make it more clear why we chose our 'form'.*

2) Effect of circular supports: From a purely hypothetical point of view, we considered the choice between rectangular and circular supports to *not* have impacts on the model accuracy. This is firstly because the auxiliary information (ALS canopy height model and tree species classification map) have a rather low spatial resolution (5 x 5 meters). Secondly - and more importantly - because our analysis already showed that the best possible model fit was not realized using support sizes that most accurately corresponded to the plot individual extents defined by the angle count sampling technique, but turned out to depend more on the spatial resolution and the kind of information of the auxiliary data. We considered this to be an important result of the study which is why we emphasized it in the last paragraph of our discussion.

Note: *We should recalculate everything with circular supports and prove our hypothesis by numbers. (We could start by recalculating the model based on circular plots for the final model data set).*

We checked both (your and our) hypothesis by recalculating the entire analysis using circular plots. It turned out that ...

5. *Was it not necessary to remove outliers or other influential observations from the data set? It sounds almost too good to be true, if that was not necessary.*

   - As stated on page 7 lines ..., we actually did remove the so-called 'zero-plots' (plots with a timber volume of 0 m^3/ha) from the model fit (206 plots). The reason for this was that with an increasing time-lag between the LiDAR acquisition and the BWI3, the ALS mean canopy height showed increasing, in some parts even large discrepancies to the zero-volume plots. For example, we investigated plots with 10 - 27 meters of mean canopy height for plots with a timber volume value of zero (i.e. there must have been a harvest activity in the meantime). However, it turned out that including these zero-volume plots did not have significant leverage on the regression coefficients (they almost stayed exactly the same and thus the same was true for the model accuracy metrics).
   - We emphasize that the objective of the regression model was to predict the outcome of the BWI3 timber volume survey for each sample plot. This means that discrepancies between the response variable value and the corresponding explanatory variable can arise from the remote sensing data due to time-lags, heterogeneity etc., but also from the terrestrial inventory procedure due to non-visible sample trees that should have been recorded (see Ritter et. al {Correcting the nondetection bias of angle count sampling, *Canadian journal of forest research*}) or the applied inventory thresholds that can exclude existing trees from the sample. Hence, a sound justification of removing an influential observation from the data set seems rather demanding. Even if observations turn out to have a leverage effect on the regression coefficients, this does not necessarily justify their removal from the data set. Increasing the model fit by deleting observations does also not ensure a better prediction accuracy for out of sample observations (i.e. unseen data).
   - Especially if the regression model is used as an internal model in the frame of model-assisted, i.e. design-based estimators (which will be the case in our follow-up study), we generally consider removing potential outliers or leverage points an issue that has to be handled with extreme caution. This is because excluding an observation from the model fit automatically means to exclude this observation from the sampling frame. Since the basic assumption here is that the sample has been generated randomly, removing of observations should in the first instance be regarded as an interference with the random sampling process. For example, excluding all zero-plots from the model fit and thus from the sampling frame would maybe yield a better model accuracy, but lead to a severe overestimation of timber volume total for the inventory area.

- Nevertheless, we conducted a mutlivariate outlier / leverage point detection (method). It turned out that ...
  Note: *Perform outlier detection based on mahalanobis distance ...*

6. *It is also somewhat uncommon to derive explanatory variables for CHMs instead from ALS raw data. A lot of information seems to be lost that way. Please justify or revise. Differing pulse densities usually do not have much influence on working models and can easily be considered in the model. This has probably a technical reason?*

   Henning:

7. *Why and how were the ALS raw data thinned before interpolating them to grids? (Consider giving point densities in the more common unit point per m2.)*

   Henning:

8. *Edge correction. Figure 3b suggests that supports were clipped at forest boundaries. 1) There is an additional data set for forest extent of public forests. Could it be described a bit more? Is the model, strictly speaking, only valid for public forests? How were plots on private forests treated? 2) Does clipping of supports fit to the type of edge correction in angle count sampling used in BWI3?*

   1) We indicated the reason for this in the Introduction (page 2): 'Our study is embedded in the current implementation of model-assisted regression estimators (Mandallaz, 2013a,b; Mandallaz et al, 2013) for estimating the standing timber volume within the state and communal forest management units over the entire state of Rhineland-Palatinate'. Actually, we use the 'final' regression model presented in this article as an internal model for model-assisted small area estimations of standing timber volume within 390 the state and communal forest management units in RLP (to be presented in our follow-up article).The state and communal forest area thus constitutes the sampling frame on which the regression model identified in this article is subsequently applied. Already restricting the set of sample plots used for modeling in this article provides the advantage that when used as an *internal model* in design-based estimators, the regression model predictions already hold the assumption on the residuals to be zero on average for state and communal forest by construction of OLS technique (see amongst others Mandallaz 2013 {Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Canadian Journal of Forest Research*}). We added this rational to the article (page 4, Line 253-262) to make the reason for restricting to state and communal forest more transparent. We also hope that this incorporates your question regarding the 'validity' of the model. It also has the huge advantage that we can refer to this article for details of the used regression model in our upcoming follow-up article that focuses much more on the application of the double-sampling estimation techniques.

   2) We are not sure what the reviewer refers to as 'public' forest. We added some sentences describing the three forest ownership classes in RLP, i.e. a) state forest, b) communal forest and c) private forest.

   Note: *Add some comments and information in Section 2.2: 1) Kind of forest ownerships*

   2) The clipping of supports at the forest boundary is a means to optimize the coherence between explanatory variables computed at the forest boundary and the corresponding terrestrial response variable, thereby optimizing the model fits for such observations (see Mandallaz et al. 2013 {New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based monte carlo approach with applications to small-area estimation}). In the BWI3 survey, edge correction is applied at the forest border at the individual tree level. This means that sample trees whose inclusion circles are intersected with the forest border are assigned with a corrected (*increased*) counting factor. This method is used to compensate for the fact that part of the trees inclusion circle is outside the forest area. Consequently, the terrestrially determined timber volume value of a sample plot with existing boundary effects would be underestimated if the edge correction was neglected. Now, as obvious from Figure 3b) (upper left support), the ALS mean canopy height will drop to around zero outside the forest area (i.e. beyond the forest border). Including these 'zero' height pixels when calculating the value of the mean canopy height for this plot will severely attenuate the mean canopy height value towards zero (this effect will increase with increasing proportion of the support lying outside the forest border). However, the terrestrially recorded timber volume value has been compensated for the edge effect by increased counting factors for the affected sample trees. Neglecting the boundary correction of the support would thus increase the discrepancy between the value of the explanatory variable (attenuated towards zero) and the terrestrial timber volume value (increased by corrected counting factors). An optimized comparability can thus be realized if we restrict the calculation of the mean canopy height to those height pixels lying within the forest border, thereby avoiding the attenuation towards zero. In our opinion, the proposed clipping-method thus indeed fits well to the edge correction in angle count sampling used in BWI3.

9. *Calibration. Did this study really introduce a calibration technology"? Consider revising. Why is the tree-species model a "calibration model"? Calibration model sounds like the parameters of the original model were adjusted. Calibration is also an estimation technique and could be misunderstood in this context. Would it be a calibration model if a traditional tree species map was used as explanatory variables? Is it not simply a model with some categorical explanatory variables?*

   We think there must be a misunderstanding here. Our proposed methods is indeed a true calibration as known from classical statistical calibration approaches in the sense that 'parameters of the original model were adjusted'. Classical statistical calibration models are used to calibrate an error-prone variable that can cheaply be measured in high quantity on its corresponding exact, i.e. error-free variable whose recording is however very cost intensive. We exactly used this statistical framework to calibrate the estimated main tree species from the classification map (which revealed the quantified misclassification errors shown in Fig.4a) left) on the error-free (i.e. exact) main tree species calculated from the set of sample trees in each terrestrial plot. This calibration of the tree species variable led to an increase in the classification accuracies and more importantly, considerably reduced the effect of the misclassification errors on the regression coefficients and thus increased the model accuracy when using the *calibrated* main plot tree species as a categorical variable in the regression model. Concerning your question, the term 'calibration model' refers to the random forest algorithm that is used to calibrate the error-prone tree species variable on the exactly determined (terrestrially derived) main plot tree species. The regression model using this calibrated categorical variable is indeed 'simply a model with some categorical explanatory variables'. The proposed calibration method was also well received by the second reviewer and pronounced to be an 'important' part of our work. Initiated by your remarks, we rephrased the section to be more clear about the basic idea of classical calibration and differentiate it from the topic of calibration estimation by rephrasing Section 2.3.2 'Calibration' accordingly.

   Note: *Rephrase Section 2.3.2 'Calibration' to be more clear about the general method of calibration as I formulated it in the answer here.*

10. *The issue of using y-transformed models or not is of high importance. By discussing why g-weight variance is of important, this paper could contribute considerably to the discussion. In addition: How are negative predictions dealt with in practice? Set to 0?*

    1) This is a very very interesting hint and idea for our currently written follow-up article that actually deals with the design-based small area estimations in RLP. In fact, the application of the g-weight variances is a fundamental part of the study and we would love to incorporate your hint in our study.

    2) The purpose of this study was the identification of a best possible regression model to be integrated in the model-assisted estimators in the follow-up study. The objective was however not to apply the model, which is why we have not given any advice how to deal with negative predictions. In Hill et al. (2014){Accuracy assessment of timber volume maps using forest inventory data and LiDAR canopy height models. *Forests*}, we actually proposed to set these values as 'not defined' when it comes to visualizing them in a map. We also emphasized that averaging over a set of multiple predictions in the frame of double sampling estimations will diminish the effect of negative predictions. We added a respective comment in the discussion.

    Note: *Add respective comments in the discussion on a) how often negative predictions occurred for the observed terrestrial data; b) give suggestions on how to treat negative predictions, cite Hill et al. 2014.*

11. *P1,L44 RHS: Is Beaudoin et al really the right reference here? The concept is anyways much older and Næsset, E. (1997) Estimating Timber Volume of Forest Stands Using Airborne Laser Scanner Data. Remote Sensing of Environment should be considered.*

    We revised the given reference of Beaudoin and now refer to the article of Brosofske et al. 2014 {A review of methods for mapping and prediction of inventory attributes for operational forest management. *Forest Science, Society of American Foresters*} who give a very nice and extensive review of applied mapping techniques in forestry, including various references to applications in the Nordic countries (page 1, 'Introduction', Line 41). We also added a sentence particularly emphasizing the long history of timber volume prediction models with reference to Naesset 1997 {Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing of Environment*} (page 1, 'Introduction', Line 48-50).

12. *P2,L15: Does Van Aardt et al. fit into the list of references?*

    We removed Van Aardt from the list of the given references and added Bohlin et al. 2017 {Mapping forest attributes using data from stereophotogrammetry of aerial images and field data from the national forest inventory. *SILVA FENNICA*} as an up-to-date reference for mapping approaches applied in Sweden using data from the national forest inventory.

    Note: *Is there an article about mapping (not only model building) for Norway, maybe from Naesset amongst others?*

13. *P2,L59: It may be of interest around that line that a variable describing deciduous proportion derived from leaf-off ALS data was used by Breidenbach et al. (2008) Mixed-effects models for estimating stand volume by means of small footprint airborne laser scanner data. Photogrammetric Journal of Finland to improve a model for timber volume.*

    Thank you very much for drawing our attention to this article. We rephrased the respective section of page 1 'Introduction' Line 111-118 accordingly and now mention the study of Breidenbach et al. (2008). We also decided to include this article / reference in our discussion, as one of the main findings of the article was that the proposed mixed model approach can yield higher prediction accuracies than simply stratifying according to categorical variables in the frame of OLS regression. We thought that this point fits very well in our case, since we exactly did the latter using the ALS acquisition years as a stratification variable. The reason we sticked to OLS was that the small area estimations (in particular the g-weight variance formulas) of the follow-up study using our regression model are explicitly defined for OLS models. We added a respective comment in the discussion (page XX Line x-x).

    Note: *Add comment in dsicussion!*

14. *P2,L5-8 RHS: 'One of the rare examples...' I cannot follow here. The way it is described, it is exactly the approach commonly used in the Nordic countries. There must be hundreds of studies one could cite. Probably a misunderstanding?*

    That was indeed not well formulated. Our intention of this section was actually to identify the gap of knowledge that we tried to address by dealing with tree species information in our study. We particularly wanted to point out that - up to our knowledge - there has not been a study that investigated the integration of this amount of tree species categories (i.e. 5 categories: beech, oak, spruce, douglas fir, scots pine) as explanatory variables in a prediction model. We appreciate the various studies made in the Nordic countries and it was not our intention to avoid mentioning them. However, we unfortunately have difficulties to find a lot of studies that use similar approaches than described in our article, i.e. using estimated tree species information (categorical or continuous) as explanatory variables in prediction models which go beyond the distinction of 'coniferous / broadleaf' (maybe because this is a not well investigated issue as depicted in White et al. 2016 and Koch 2010). We rephrased the respective section page 2 'Introduction' Line 112-125 accordingly and decided not to mention Packalen et al. 2006 since their approach seemed in fact differing too much from our approach. If the reviewer cannot agree with the statement, we would very much appreciate if he could give us some references to studies he knows about that are more similar to ours.

15. *P13,L31-37: Why blending ITC into this article? Does the concept of supports fit there at all? Consider removing.*

    The sentence has been deleted from the article.

16. *Consider discussing the model a bit more with other models published in central Europe (maybe based on smaller study sites) and other studies based on NFI data. Why do we see differences?*

17. *P7,L30-31: Ambiguous. Please rephrase. Is dominant species here decided based on field data or the map?*

18. *P7,L7-32: The model interpretation is a bit lengthy and may be best described by a reference to Fig 7.*

19. *P7,L7-32: The model interpretation is a bit lengthy and may be best described by a reference to Fig 7.*

20. *Please add a table on field measured values and relative RMSE values (%).*

21. *Please consider that the Conclusion is not an extension of the Discussion. Consider revising to shorten and to avoid references in the Conclusion*

22. *Fig 4: Please define ïnd".*

23. *Is Fig 5 really needed?*

24. *Figure 6: Could you expand in the text why two distinct clusters are visible? This remained a bit unclear. The dashed lines are hardly visible.*

25. *Figure 8: Consider showing only the simple model and the final model. The others seem to be of little relevance in this context and just make the graphic difficult to read.*

26. *Figure 7: (The fig appears after Fig 8.) The single panels are very small and difficult to read. Consider just showing one panel to exemplify.*

27. *Consider revising the use of percent (%) vs. percentage points. My impression is that the terminology could have been used in the wrong way. (And R2 values are proportions, not percentages.)*

28. *Terminology (p3, l16): Consider using ALS throughout the paper as the correct acronym for airborne laser scanning. (Or lidar which is consistent with acronyms like radar or laser.)*

29. *Please add page range when citing books.*

30. *Eq1: are Y(x) and e(x) defined? Consider presenting table 2 close to eq 1 as the submodel names were undefined until table 2 was presented.*

    The timber volume density Y(x) has been defined on page 4, section 2.2 'Terrestrial Inventory data'.

31. *P5L37: Is the threshold defined?*

32. *P8L17-21, RHS: Discussion?*

33. *Format: Consider submitting a one-column, two line spacing manuscript, such that line numbers are available to all lines.*

    We are sorry for that, but this was done by the editor. The version we prepared for the reviewers did in fact have line numbers on both sides.

Reviewer 2:

It is a very interesting and useful approach and paper. Your research question is well defined, extensively investigated and explained. Regarding your figures please be aware, that printing or digital presentation will be very small. Try to use clear distinguishable colors and markers. I recommend removing the grey background in all figures. Some phrases are quiet long and contain several aspects. Try to shorten your sentences, by separating them according to the different aspects or contents.

Comments:

1. *P3 line 46 Column 2: Did you measure the absolute tree height of every sample tree, or was it a estimation according to the measurements of a sub-sample of tree?*

   In fact, the height is *measured* only for a subset of the sample trees at each plot. The height-values for the remaining sample trees are then *estimated*. The taper functions however use the height-value a sample tree without distinction between 'measured' and 'estimated' as one of the explanatory variables. We rephrased the sentence accordingly (page 3, Line 228-234).

2. *P3 Line 48 Column 2: Do you have any information about the type and positional accuracy of the used GPS or GNSS-Techniques*

   Yes, we in fact looked at this issue very closely in an individual study of Lambrecht et al. 2017 {A Machine Learning Method for Co-Registration and Individual Tree Matching of Forest Inventory and Airborne Laser Scanning Data. *Remote Sensing*}. The analysis was carried out for a subsample of all sample plots in RLP and indicated that horizontal DGPS errors do not exceed a range of 8 meters for 80% of all plots. We were unfortunately not provided with estimated accuracy metrics from the actual DGPS acquisition software that was used by the field crews. The dataset provided from the authorities only contained information about the number of DGPS measurements used to average the final plot position coordinate (100 measurements at each plot center) and the PDOP-value. Both seemed unsuitable to give a reliable accuracy for the plot positions. The study of Lambrecht et al. is the first study in RLP that have provided an idea about the actual positional accuracy. We added a respective comment in the article (page 3, Line 234-244).

3. *P3 Line 57/58: Why did you restrict to stat and communal forests?*

   We indicated the reason for this in the Introduction (page 2): 'Our study is embedded in the current implementation of model-assisted regression estimators (Mandallaz, 2013a,b; Mandallaz et al, 2013) for estimating the standing timber volume within the state and communal forest management units over the entire state of Rhineland-Palatinate'. Actually, we are currently writing a follow-up article, where the 'final' regression model presented in this article is in fact used as an internal model for model-assisted small area estimations of standing timber volume within 390 the state and communal forest management units in RLP.

The state and communal forest area thus constitutes the sampling frame on which the regression model identified in this article is subsequently applied. Already restricting the set of sample plots used for modeling in this article provides the advantage that when used as an *internal model* in design-based estimators, the regression model predictions already hold the assumption on the residuals to be zero on average for state and communal forest by construction of OLS technique (see amongst others Mandallaz 2013 {Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Canadian Journal of Forest Research*}). We added this rational to the article (page 4, Line 253-262) to make the reason for restricting to state and communal forest more transparent. It also has the huge advantage that we can refer to this article for details of the used regression model in our upcoming follow-up article that focuses much more on the application of the double-sampling estimation techniques.

4. *P4 Line 24-26: In general the point density is given in points per m^2.*
5. *P4 Line 47 -51: How did you calculate the canopy height model, by subtraction?*
6. *P4 L1 Column 2: How did you define the forest borders?*
7. *P5 Line 34 – 40: Mention here, that you investigate this threshold in your work.*
8. *P5 Line 11-36 Column 2: Why did you use these variables and not others? I would recommend putting the explanation of your calibration method more popular in your article. It is an important part of your work.*
9. *P6 Line 46: the title should be "model building and evaluation" instead of model validation*
10. *P7 Line 4-18: What are the values for lidar year, 0 and 1 or N/A and 1 or anything else?*
11. *P7 figure4: the figure is overloaded. Try to use colored lines instead of colors and different markers. Is it really necessary to show both n and threshold? If yes, put them into two separate figures. Or could you explain*

    Note: *remove 'n' from the figure.*

    We reoved the 'n' from the figure.
12. *P8 Line 1: Title should be "Calibration" and not only "Effect of Calibration".*
13. *P8 L55-59: Separate it in at least two sentences.*
14. *P9 figure5: the figure is hard to differentiate: Use color instead of different markers, put the legend below or above could already help.*
15. *P9 figure 6: Where do the different grey values refer to?*
16. *P10 figure 8: I recommend limiting the y-axis between 0.2 and 0.6. A legend is missing and should be added. The long title of the figure contains text, which should be included in the text and not in the title.*
17. *P10 Table 1: I recommend mentioning RMSE% instead of SSE*
18. *P12 Table2: submodel 3 isn't mentioned and explained in the text. It isn't possible to reconstruct how you get your amount of parameters for each model. In addition the amount of parameters isn't discussed in the text. So I recommend either to explain (I would prefer) or to leave it out.*
19. *P12 Line35: This is the wrong reference: Table1 instead of Table2.*
20. *P13 Line 43 Column 2: aerial instead of areal.*