Review

13. Januar 2018

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. In this study, the focus is on the application of different estimators. 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 data 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 number of different tree species.

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*

   Thank you very much for the hint to this study. We included it in our discussion, section 4.2 'Choice of Support under Angle Count Sampling'.

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) Concerning merging several years (all except for 2007?) into one factor: As described in section 2.3.1, we already merged the ALSyears 2006 and 2007, and 2012 and 2013. However, as emphasized in Fig.8 and Table 1, the effect of using the ALSyears factor-levels was to allow for higher model accuracies in ALSyears-strata in which the height data showed less noise compared to when merging all or most ALSyear factor-levels. This effect was significantly reduced or even removed when merging the ALSyears-levels except 2007, since this increased the noise in the data and lowered the overall R^2 values as well as the R^2-values within the original ALSyear strata. We show this in the table below, comparing the R^2-values of the final model (as given in the article) and those achieved when merging the factor-levels except '2007' (called the 'merged model'). As depicted in section 'Discussion' line xx-xx, achieving those higher model accuracies within domains that are located in the respective ALSyear-strata are considered to improve the estimation errors in the frame of our upcoming Small Area Estimation analysis. We will particularly investigate this issue in the upcoming article. For the mentioned reasons, we thus decided not to merge any ALSyear factor-levels.

   We added a few sentences in the section 4 'Discussion' where we address and comment on issues 1) and 2)

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*

$ALSyear$	$Area_{ALSyear}$	R^2 final model	R^2 merged model	n
2012	2807	0.61	0.53	408
2011	4361	0.57	0.45	883
2010	4182	0.51	0.43	1171
2009	2100	0.42	0.34	559
2008	2968	0.48	0.41	701
2008_1	2116	0.33	0.33	394
2007	3498	0.46	0.37	418
2003	602	0.27	0.23	529
2002	775	0.44	0.40	314

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) Thank you very much for the hint! We recalculated the entire analysis under the use of circular plots. It turned out that this only had minor effects on all accuracy metrics that were derived in our study. The results were almost identical to the previous ones and did not affect the major findings of the article. It however led to a further improvement of the model accuracy in the ALSyear-strata 2012, 2011 and 2010 (Table 1). We thus decided to switch to the use of circular supports with respect to your comments and changed all respective text-sections, Figures and Tables accordingly. From a purely hypothetical point of view, we think that the choice between rectangular and circular supports did only have minor impacts on the model accuracy, because a) the auxiliary information (ALS canopy height model and tree species classification map) have a rather low spatial resolution (5 x 5 meters); and b) (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.

2) 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. We rephrased the respective text in section 2.4.

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.*

We conducted an Influence Analysis, i.e. leverage / outlier detection, for the 'final model'. We here considered the usually applied criteria of *Leverages* and *Cook's Distance* as amongst others described in Fahrmeir 2013 {Regression: models, methods and applications. *Springer Science & Business Media*. page 164-167}. The critical threshold of $2p/n$ (i.e. twice the average of the hat matrix' diagonal entries), was exceeded by 10% of the observations. However, only 3% of these leverage points were assigned to studentized resiudals > 1 or $<$ -1. Leaving these 3% of points out and recalculating the final model lead to a R^2_{adj} of 0.494 compared to 0.485 when including them in the regression. The Cook's Distance values D_i did not exceed a value of 0.019, and thus were far apart from the often cited critical threshold of $D_i > 0.5$. Based on these findings, we decided not to remove observations from the modelling data set. We added a paragraph commenting on this issue of influential data points in section 3.3 'Final Regression Model'.

Especially if the regression model is used as an internal model for design-based estimations (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. In the design-based framework, removing data points from the modelling/sampling frame is only valid if the respective observations turns out to be truly erroneous. If this is not the case, the removal of outliers or influential data points will increase the model fit, but to the cost of possible bias for the estimates. This is because excluding an observation from the model fit has

in the first instance to be regarded as an interference with the random sampling process. We will comment on this issue in our upcoming study.

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?*

   Although there are indeed many studies where explanatory variables are derived from the ALS raw data, using a rasterized ALS CHM in this kind of studies is up to our knowledge not at all uncommon. In our particular case however, the main reason was that the available ALS data had very low point densities in many parts of the study area (i.e. less than 1 point per m^2). Owing to the low point density it is difficult to extract meaningful statistical descriptors which might significantly improve the model accuracies. This is especially because the 'mean canopy height' as a predictor variable already has the highest predictive power for our plot-based timber volume predictions. We expect that the derivation of this variable from the rasterized representation is well sufficient and that its derivation from the ALS raw data will not increase its explanatory power. We had also tested additional variables derived from the CHM, such as 'height-percentiles' (25%, 50%, 75%) and the 'maximum height value'. None of them added any explanatory power to the model. We are thus convinced that deriving the selected predictor variable directly from the rasterized CHM is justified. Since two years the ALS data acquired by the Geodetic Survey provide substantially higher point densities, which is more promising with respect to deriving ALS echo statistics directly from the point cloud in future studies.

   With respect to the reasons explained we believe that revising the approach is not justified, considering the trade-off between potential improvements and the required processing efforts (our entire analysis algorithms have been set up and optimized for 'raster'-processing operations in a PostgrSQL database which holds the rasterized CHM as well as the tree species classification map).

   Another more long-term oriented reason for using the rasterized ALS height information was the transferability of the algorithms to the photogrammetric canopy height model (mentioned in section 5 'Conclusion'), which will be updated in much shorter terms as future ALS campaigns, and thus provides a much better temporal alignment with the terrestrial inventories. An example from Norway where such information is used can be found in Breidenbach & Astrup 2012. {Small area estimation of forest attributes in the Norwegian National Forest Inventory}, *European Journal of Forest Research.*

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.)*

   The data was delivered as two separate data sets comprising the Vegetation First Pulse (VEF) and Ground (GRD) points. In order to create a surface model (DSM) in a given raster resolution, the highest point of the combined VEF and GRD data set was identified in each raster cell and saved as a thinned surface point cloud. For the elevation model (DEM), the mean of all GRD points in the cell was calculated, and the result was saved as a thinned ground point cloud. The thinned point clouds were then aggregated to larger tiles and interpolated to raster images using a Delauney interpolation in the Matlab software. The resulting DSM and DEM raster sets were then subtracted from each other to calculate a canopy height model (CHM) in raster format, providing discrete information about the canopy surface height of the entire forest area of RLP in a spatial resolution of 5 meters. The thinning process led to much smaller data sets that could be processed in larger tiles and considerably lowered processing times than the original dense point clouds. Since the data was recorded in leaf-off condition, the original point clouds contained many returns from within the crowns of deciduous trees. The thinned data set also provided the advantage that those measurements did not skew the vegetation height estimate in the final CHM. We rephrased this paragraph in Section 2.3.1 'ALS Canopy Height Model' accordingly to be more clear about this issue. We also changed the units to point per m^2.

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 upcoming 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 xx-xx) 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 in section 2.1 'Study Area'.

3) 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. We added a short note an this in section 2.3.1 'ALS Canopy Height Model', line xx - xx.

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?*

   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 other reviewer and pronounced to be an 'important' part of our work. On suggestion of the other Reviewer to 'put the calibration more popular' in our article, we rephrased and expanded Section 2.3.2 'Calibration'. We here particularly considered your input and tried to be much more clear about the basic idea of calibration in measurement error statistics, and differentiate it from the topic of calibration estimation.

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 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. We thus only added a small comment in this article and will discuss this issue in more detail in the follow-up article.

    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. We added a respective comment in section 3.3 'Interpretation of Final Regression Model' that rarely occurring negative predictions will not have an influence on model-assisted estimates since multiple predictions are averaged over spatial domains.

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 xx). 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 xx-xx).

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.

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 xx-xx 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 xx-xx).

    Note: *Add comment in discussion!*

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. However, we unfortunately had 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 xx-xx accordingly and decided not to mention Packalen et al. 2006 since their approach seemed in fact to differ 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?*

    We rephrased this sentence accordingly. We also rephrased the sentence about how the accuracy assessment was made at the beginning of section 2.5 'Model Building and Evaluation': *'In order to judge the quality of the treespecies variable, the user's accuracy for each classified species category and the overall accuracy of the classification scheme was calculated based on the confusion matrix (Congalton and Green 2008). As reference data, we calculated the actual main plot tree species by applying the respective threshold to the sample trees of each sample plot'.*

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

    We think the Reviewer is referring to P10,L7-32 (section 'Interpretation of Final Regression Model'). In fact, the model interpretation is descibed by a reference to Fig. 7 (first sentence). We do not share the Reviewers opinion that the model interpretation is to lengthy, considering that it is - up to our knowledge - the first study where such detailed tree species information (5 tree species) has been used in a prediction model. We therefore consider providing a detailed interpretation of the model an important part of the article. With respect to the reasons explained we believe that revising the section is not justified.

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

    We added a table on the field measured values within the state and communal forest area at the end of section 2.2 'Terrestrial Inventory Data'. RMSE-values in % have also been calculated for the model accuracies and added in the article. (However, there are no RMSE-values for the field measured values. They are assumed to be error-free.)

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

21. *Fig 4: Please define 'ind'.*

    'ind' is defined now in the caption of Fig. 4. It was also already defined in section 2.4. 'Choice of Support under Angle Count Sampling'.

22. *Is Fig 5 really needed?*

    We consider Figure 5 as a very important, if not the most important graphic in our article. It gives the visual overview about all model evaluations carried out under the different paramterization (threshold and support choices, calibration). It thus addresses all 3 objectives of the article, which have been defined at the end of the 'Introduction'. We definitely want to keep the graphic. On suggestion of the other Reviewer, we changed the layout of Fig.5 in order to provide a better visual distinction.

23. *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.*

    The differentiation into two distinct point clouds results from the poor model performance under support size $q100$ for the CHM variables (i.e. the *lower* point cloud). We added this information in Section 3.2 'Effect of Missclassifications'. We also changed the background to white in order to make the dashed lines better visible.

24. *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.*

    We changed the graphic with respect to suggestions of the other Reviewer in order to provide a better visual distinction. We would however very much like to keep all visualized models in the graphic. It was our particular intention to show the differences in model accuracy between these models within the different ALSyears. Visualizing submodel 2 (use of ALSyears as predictor variable) is of importance since one can see the improvement of the model accuracy when using the treespecies predictor variable additional to the ALSyear variable.

25. *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.*

    The order of the Figure 7 and Figure 8 has been corrected. We however would very much like to keep Figure 7 as is, because our intention was to show the different observation point clouds per ALS acquisition year at least once in the article (plotted in the background of each panel). We also wanted to show the reader that the slopes adjusts in the different panels according to each ALS year. We hope the Reviewer is fine with our decision not to revise this Figure.

26. *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.)*

    We revised this notation and changed the R^2-values into percentage points (i.e. 0.5 instead of 50%).

27. *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.)*

    Has been changed accordingly.

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

    We used the article of Breidenbach and Astrup 2012 {Small area estimation of forest attributes in the Norwegian National Forest Inventory}, *European Journal of Forest Research*) as example and added the book-chapters to the book-references of Mandallaz 2008, Gustafson 2003, Caroll et al. 2006 as well as the page range of Draper and Smith 2014 in our article accordingly.

29. *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'.

30. *P5L37: Is the threshold defined?*

    Thank you very much for the hint. We have not been aware that this information was missing. We added this to the paragraph accordingly.

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

    We moved this sentence to the Discussion, section 4.2 'Calibration of Tree Sprecies Map Information'.

32. *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 xx-xx).

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 xx-xx).

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 xx-xx) 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.*

   Values have been changed accordingly.

5. *P4 Line 47 -51: How did you calculate the canopy height model, by subtraction?*

   We forgot to mention this. Yes, the CHM was calculated by subtracting the DEM from the DSM. Based on suggestions of the other Reviewer, we rephrased this section 2.3.1 'ALS Canopy Height Model', and also added the missing information about the CHM calculation.

6. *P4 L1 Column 2: How did you define the forest borders?*

   This is mentioned on page 4, section 2.3.1 'ALS Canopy Height Model': *'The square [...] was previously intersected with the state and communal forest area defined by a polygon mask and thereby corrected for edge effects at the forest border.'*. So referring to your question, the forest borders of the state and communal forests were defined by the mentioned polygon mask. We tried to rephrase the sentence to make this more clear to the reader. We also appended a sentence commenting on the purpose of this method, as requested by the other Reviewer.

7. *P5 Line 34 – 40: Mention here, that you investigate this threshold in your work.*

   Thank you very much for the hint. We have not been aware that this information was missing. We added this to the paragraph accordingly.

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.*

   An advantage for using those explanatory variables in the calibration model was that they also provided explanatory power in the regression model, so they could be used for both, the calibration *and* the regression model. Using these variables thus considerably saved computation time as well as data storage space. We added this comment in Section 'Calibration'. Triggered by your suggestion above, we rewrote the entire Section 'Calibration' in order to be more clear about the technique of calibration in measurement error statistics, and to explain the idea of transferring this method to our misclassification problem in more detail.

9. *P6 Line 46: the title should be "model building and evaluation" instead of model validation*

   Has been changed accordingly.

10. *P7 Line 4-18: What are the values for lidar year, 0 and 1 or N/A and 1 or anything else?*

    Categorical variables are always recoded by a unique combination of '0' and '1'- values in the design matrix of a linear regression model. We however think that the recoding is rather theoretical knowledge which a reader familiar with regression techniques will already have. For the reader, it is just important to see that each factor level (i.e. each tree species and each ALS year) has its own regression coefficient. This leads to the amount of parameters used in a respective linear model.

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*

    We removed the 'n' from the figure as it did in fact not add any valuable information. We also changed the layout likewise Fig. 5 based on your suggestions and think it really improved the graphic. The changes comprised: a) using flexible scales among the tree species groups in order to ßoomïn and make the displayed information better to distinguish. We also placed the legend on the bottom like you suggested and adapted the colour-scheme to support the visual distinction. The background was changed to white.

12. *P8 Line 1: Title should be "Calibration" and not only "Effect of Calibration".*

    Has been changed accordingly.

13. *P8 L55-59: Separate it in at least two sentences.*

    We rephrased the section accordingly.

14. *P9 figure5: the figure is hard to differentiate: Use color instead of different markers, put the legend below or above could already help.*

    We changed the layout of Fig.5 (and also Fig.4) based on your suggestions and think it really improved the graphic. The changes comprised: a) using flexible scales among the tree species groups in order to ßoomïn and make the displayed information better to distinguish. We also placed the legend on the bottom like you suggested. We also adapted the colour-scheme to support the visual distinction.

15. *P9 figure 6: Where do the different grey values refer to?*

    We missed to mention in the caption that we used semi-transparent colour for the data points to visualize overlap, and to indicate where most of the calculated model accuracies are located in the plot. We added this information to the caption of Figure 6.

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.*

    We changed the graphic with respect to you suggestions. We also removed the grey background and changed the colour scheme for a better visual distinction.

17. *P10 Table 1: I recommend mentioning RMSE% instead of SSE*

    Additional to $RMSE_{cv}$ or $RMSE$, we now also calculated $RMSE_{cv}[\%]$ and $RMSE[\%]$. These values have respectively been added to the text-sections, the supplementary data tables (Table 2 and 3) as well as Table 1 in the article instead of SSE (requested by Reviewer#2). We also decided to map the $RMSE_{cv}[\%]$ instead of $RMSE_{cv}$ in Fig. 5 (right).

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.*

    We now explicitly mention submodel 3 in the subsection 'Effect of Time-Lags and Heterogeneity in ALS data' (page 10, line xx - xx). With respect to the amount of model parameters, we think that every reader familiar with OLS regression models including interaction terms should well be able to reconstruct the amount of parameters, i.e. regression coefficients, based on a) the model terms given in Table 2 in combination with the extensive and repetitive description of the number factor levels for the *ALSyear*-variable (last sentence in section 2.3.1, Figure 7, Table 1, ...) and for the *treespecies*-variable (section 2.3.2, Figure 4, Figure 7). In our opinion, a further explanation would in fact lead to an extensive repetition of OLS regression theory, which is not the intention of the article. We added a sentence in Table 2 that the ':'-sign indicates interaction terms to support the reconstruction of the amount of parameters and hope that the Reviewer will be agree with our solution.

19. *P12 Line35: This is the wrong reference: Table1 instead of Table2.*
    Has been corrected.

20. *P13 Line 43 Column 2: aerial instead of areal.*
    Typo has been corrected.

Documentation of changes:

1. Triggered by the suggestion of Reviewer#1, we recalculated the entire analysis based on *circular* rather than *rectangular* supports. The results were almost identical to the previous ones and did not affect the major findings of the article. It however led to a further improvement of model accuracy in the ALSyear-strata 2012, 2011 and 2010 (Table 1). We changed the respective text-sections as well as Fig.3(b) accordingly. In addition, the best model we found overall in terms of adjusted R^2 and cross-validated RMSE values (now also given in %) was now the model later used as the final model ($q50$ support for both CHM and *treespecies*-variables). In the previous version, the best model was found with the support settings $q50$ for the CHM-variables and the $q100$ for the *treespecies*-variable.

2. At the suggestion of Reviewer#1, we changed the terminology of LiDAR into ALS.

3. 10-fold cross-validation changed to 5-fold cv (did not change the results and led to much smaller calculation times).

4. At the suggestion of Reviewer#2, we removed the grey background in all graphic-plots.

5. Additional to $RMSE_{cv}$ or $RMSE$, we now also calculated $RMSE_{cv}$[%] and $RMSE$[%]. These values have respectively been added in the supplementary data tables (Table 2 and 3) as well as given in Table 1 in the article instead of SSE (requested by Reviewer#2). We also decided to map the $RMSE_{cv}$[%] instead of $RMSE_{cv}$ in Fig. 5 (right). We also added the RMSE-formulas in section 2.5 'Model Building and Evaluation'.

6. In Table 1, we added the area covered by each ALS acquisition to emphasize for which total area the respective model accuracies have been achieved. (This is particularly relevant when thinking about applying this model in model-dependent or model-assisted small area estimators for domains within these areas).

7. We changed the layout of Fig.4 and Fig.5 based on the suggestions of Reviewer#2. These changes comprised: a) using flexible scales among the tree species groups in order to ßoomïn and make the displayed information better to distinguish. We placed the legend on the bottom like suggested by Reviewer#2. We also adapted the colour-scheme to support the visual distinction.