Comments on reviews of Chapter 4 of PhD thesis

1 General Remark

I only have Chapter 4 of the PhD thesis and not the paper submitted to remote sensing, so that the numbering of e.g. equations is different.

2 Reviewer 1

1. The reviewer is enthusiastic about the paper.

2. Vectors are always column vectors, i.e $\mathbf{Z}(x)$ and $\mathbf{Z}_c(x)$ are column vectors. When they are used for predictions they have to be transposed in order to get a scalar product written as a matrix multiplication, i.e. $\hat{Y}_c(x) = \mathbf{Z}_c^T(x)\beta$. Of course I do not know what equation 12, 13 refer because of the remark above. In any case, I did not find an error with respect to this issue in the thesis.

3 Reviewer 2

1. I think that you can can make most of the changes suggested (according to the line numbers) in the paper but not necessarily in the Thesis.

2. Of course, not all readers of Remote Sensing are specialists in forest inventory, but most of the later read Remote Sensing!

3. Angle count main text and L 107: optical device based on the apparent angular diameter of a tree. Mention that it is PPS for basal area and give the formula for the radius of the limit circle $R_i = \frac{DBH_i}{\sqrt{k}}$ with $k = 10^4 \sin^2(\alpha)$. The angle count creates problem for the auxiliary information as we know.

4. L 43 Cochran is cited in the bibliography of the thesis but not de Vries, which I think should be. Cite both in the paper.

5. L 43-48. Optimization was not part of the thesis. You could mention the anticipated variance treated in my book. Multi-objective optimization would implies a weighting of the objectives, which is a highly political issue and probably unsolvable in my opinion. Separate investigation are probably simpler and better. Worth emphasizing that the method presented is valid for any response variable which can be used to define a local density.

6. L 454-458: The results are so clear that the conclusions will not be affected. Rerunning for only SRS and PSYNTH with the 321 FR used for PSMALL and EXTPSYNTH might be feasible.

7. **My Question**:
   It is only after having read this review that I realized that all the variances given are based on matrices (i.e. g-weights) and that the external variances have not been used to produce the errors. If you have calculated them it might be good to say how close (hopefully) they are to the more elaborated g-weight versions. If you have not calculated them would be a lot of work to get them?

4 Reviewer 3

This reviewer can be really pedantic and is probably not familiar with forest inventory.

1. Comment 1: the populations are domains (i.e. a continuum of infinitely many points) of the plane F, also known as the Monte Carlo approach: the objective is to estimate the integral $\int_G Y(x)dx, G \subset F$ of a function $Y(x)$, which is essentially the Horwitz-Thomson estimator for a response variable defined for the population of trees. This is the main difference with standard survey sampling. Another important difference is that instead of having a single large sample we have n_2 points x, each one with a sample of trees (terrestrial sample). The sample size n_2 can be large over the entire domain but small over sub-domains, and to small to fit models using auxiliary data available at a much larger number of points n_1. You do not have to borrow strength if you are strong, which is obvious. We do not propose to fit models over extremely large areas with very different forest structures (say all Europe) but to rely on prediction models based on remote sensing data to provide better estimates for smaller management units. Asymptotic results for the variances are, given the complexity of the problem, the only available ones. When are asymptotic results approximately valid? Take for instance $n = 6$ (not a very large number), then the distribution of the mean of $n = 6$ uniformly distributed random variables is practically undistinguishable from the normal distribution, so that in this case $n = 6$ is the asymptotic validity range. On the other hand with extreme values statistics $n = 100$ can be far to small. Some simulations performed on artificial examples (see our CJFR papers) and on a unique data set (zrichberg data with FIESTA, see book of Mandallaz for examples) show that for SA problems $n > 12$ or even 6 is acceptable. All the variance estimates are based on the assumption of i.i.d uniformly distributed points (or cluster origins) for remote sensing followed by SRS selection for terrestrial observations, whereas in practice systematic grids are used.Simulation and past experience wit alternative model-dependent Kriging procedures (far more difficult to implement) show that the variances estimates are usually to pessimist. On the other hand, only sampling errors are take into account and not measurement errors (e.g. on bole volume).

2. Comments 2,3: the situation is even more complicated because we treat a systematic sample (grid) as a random sample of uniformly distributed points (or cluster origins). Consequently all sample sizes (i.e. number of plots or cluster in the forest or sub-domains of it) clusters are random. Treating such a systematic sample as a random will as afore mentioned lead to an over-estimations of the the sampling variance. The resulting variances are conditional on the observed sample sizes. In model-assisted design-based inference models do not have to be true (as in model-dependent survey inference), they only have to be useful. In the external approach (asymptotically valid also for internal models) you must correct the estimator with the residuals. (PSMALL does that and EXTPSYNTH also but in a indirect way). It can happen, by chance, that the PSYNTH is indeed closer to the truth (example in the book). The crucial point is to use remote sensing data and direct estimation (still using auxiliary information) can be unstable because of the large number of parameters in the prediction model. The synthetic estimator can be severely biased (take the model with only the intercept term, it is unbiased globally but not locally, unless the forest looks much the same everywhere, this has been confirmed by simulations with more complex model). In the model-dependent approach it can happen (we have such a case study) that the synthetic estimator is closer

to the known true value than all other estimates (included the sample mean of the observations in the small area). It is not clear what direct estimation mean: a model relying on auxiliary information built especially for the SA under consideration (which can be expected to be better than a global model, provided we have enough observations for model-building) or estimates based only on the sampled response variable in the SA. If you have unlimited resources it is clearly better. National forest inventory are time consuming and expensive and local management want this data to be useful for their own use, so that ideally that do not have to perform local inventories.

3. Minor comments:
   1: ok for design-based regression estimation.
   2: the extended model (extended by the indicator variable of the small area under consideration) is a mathematical trick to ensure zero mean residual over the small area, which enables to derive better variance estimates based on the g-weights estimates adapted from Sarndal's work. The technical details and proofs are available in the references. In (Mandallaz D., Massey A., 2015, e-collection-. Regression and non-parametric estimators for two-phase forest inventories in the design-based Monte-Carlo approach, Appendix C) we also consider a model-calibration approach (according to Wu,C. and Sitter R., JASA (2001, pp.185-193) and translated the results into the Monte Carlo approach and found out that it is asymptotically equivalent to the two-phase regression estimator.