Reviews

The R Package forestinventory: design-based global and small area estimations for multi-phase forest inventories

Andreas Hill, Alexander Massey

7. August 2018

Reviewer 1:

The manuscript “The R Package forestinventory: Design-Based Global and Small Area Estimations for Multi-Phase Forest Inventories” by Hill and Massey presents the statistical background and the application of the R package “forestinventory” (Hill, Massey, Mandallaz). This package provides methods for computing estimates and confidence intervals for relevant quantities (such as total timber volume or volume in a subset) in a typical forest inventory setting where precise information about the response is only available at a small sample of locations while explanatory data is available for a wider area.

The package, in connection with the manuscript, is likely to be very useful for a wide range of applications in forest inventories (and possibly other fields), and I strongly recommend publication of the manuscript in JSS once the concerns regarding limitations of the methodology, as described below, have been addressed in a suitable form.

The manuscript itself is carefully and well written with only very minor inconsistencies. It has a clear structure, offering a healthy combination of mathematical details and presentation of the various methods. There is no question as to its usefulness. The formalism is based on that developed by D. Mandallaz in a book and a number of papers following the design-based approach (where stochasticity is tied to the choice of the sampling locations).

The accompanying R package includes more than 7000 lines of code, which in part rely on other packages. Of course, in the context of this review more than a very cursory glance at the implementation (indeed evaluating correctness) is out of the question, so, obviously, any individual results produced by the package, particularly in edge cases, will require double-checking by the user – or trust in the authors. That being said, the code generally appears to be well-written with meaningful comments included throughout. The placement of the various functions in the different files is logical, although the separate handling of one-phase, two-phase and three-phase methods probably involves some duplication. The installation of the package did not cause any trouble, and the R documentation is sound and follows R conventions. The functions tested by the reviewer worked as advertised, but as usual with R involved some time to become familiar with. Specifically, the separate argument phase_id in the fitting functions (cf. page 10 in the manuscript) didn’t seem intuitive when it would presumably be possible to simply define the phases as those rows where the corresponding variables are present (i. e. not NA). In fact, there is some automatic recoding which in part does just that (page 27). Some related discussion is found on page 28.

My main concern is that the limitations of the methodology are nowhere clearly specified in the manuscript. On page 5, it is suggested that a consequence of the design-based approach ïs that the estimation properties of design-based regression estimators (e. g. unbiasedness) typically hold regardless of the model that is chosen."While this is basically true for unbiasedness, unbiasedness by itself means little, and, more important, the statement does not extend to relevant properties such as coverage rates of confidence intervals. Indeed, if s2 contains less than 1-α of the entire population, any true 1-α confidence interval for the population mean necessarily involves distributional assumptions about the population because otherwise any fixed Y(x0), which has probability $> \alpha$ of not being included in the sample, can be anything, so any conclusion about the mean would be wrong with probability $> \alpha$. (Similarly, with the infinite population approach, one needs assumptions on the tails.) Of course, it is very plausible that such issues go away once reasonable assumptions (suitable sample sizes etc.) are met. While this will be obvious in some situations, I expect many potential uses of the package where this is less so, either because users are specifically looking at edge cases in search of interesting phenomena or, more commonly, because there really is very little data available - the typical situation for a forestry undergraduate writing a thesis. The vague invocation of the central limit theorem (page 24) won’t help here. Below I include three numerical examples where the nominal 95% confidence intervals never exceed a coverage of 85%.

As I personally regard solid knowledge of the limitations of a statistical method as a prerequisite of its use in science and in addition believe it would help many potential users of the package, I ask the authors to consider adding some clarification regarding typical situations where the methods can be or should not be applied. For example, showing a few artificial test cases highlighting the most important requirements would probably not be difficult.

The following is a list of minor issues encountered while reading:

1. *On page 3, n2 is not introduced (e. g. say a "a sample of n2 points, s2").*
2. *On page 27, the element symbol should be a subset symbol..*
3. *On page 32, grisons should probably be capitalized..*
4. *Equations (1), (4), (13), (16), (20) might be better readable without the n2 fractions.*
5. *On page 7, the zero mean residual property requires that the regression model includes a constant term (or equivalent). While the R functions apparently fail when you remove the intercept, this should probably be stated somewhere..*

Reviewer 2:

The authors presented a well-written and interesting paper that will be an important documentation of the well-designed, timely, and highly relevant R package "forestinventory". In the package, the multi-phase, infinite (continuous) population estimators by Prof. Daniel Mandallaz and his group are implemented which will help many analysts to make use of these elegant methods. The package is extremely well documented in the help files with many examples and comes with several data sets. The package also comes with methods for ggplot2 that allow a very elegant comparison of the implemented estimators. To conclude, I'd like to thank the authors for sharing their work – our group is certainly going to use the package a lot.

Comments:

The paper is rather long which can be a good or bad thing. I think that the 2 and 3 phase estimators are so similar that it may be sufficient to present, for example, SAE only for the 2 phase estimator and global estimation only for the 3 phase estimator. The same is true for the exhaustive and non-exhaustive estimators. Describing just once that in all cases both types are possible (as done for cluster sampling), would make the text a lot easier to read. Also section 5 on CIs could be considerably shortened. Needs to be a section on its own right? Sections 6 and 7 are very interesting but the R code under 7.1 could probably be condensed a lot.

The current description of how to obtaining the explanatory variables seems not to fit well into this paper. Explanatory variables have to be calculated by the analyst before using the package (except for an optional boundary adjustment). Also in the interest of making the paper shorter and more general (not only focusing on forest inventories), consider omitting or shorten drastically this part of the document.

Furthermore, I wonder whether the infinite population estimators in combination with the explanatory variables used in the examples are applicable without modification for the exhaustive (wall-to-wall) estimators such as eq. 17a. It may be just a misunderstanding of how to interpret the integral in the equations and the sentence above eq. 2b. However, it sounds like it is suggested to use the full cover of auxiliary information such as lidar height returns and to calculate "metrics" that include order statistics such as 75th percentile and the maximum globally. While this approach is unproblematic for the mean, it results in biased order statistics as they are scale-dependent. For example, the mean of the maximum values for grid-cells (population units) tessellating the area should be calculated rather than the global maximum over the area. The grid cells should have approximately the same size as the sampled units because of the scale-dependency of the explanatory variables but do not necessarily be of the same shape. I am sure this was done correctly when calculating the explanatory variables for the examples. However, this seems to be a crucial point that should be made much clearer; maybe especially so if the authors have a different point of view.

Related to the previous comment is whether the difference between the infinite and finite population approaches is relevant from a practical point of view. At least the estimators may be equivalent (see for example eq. 6a). Furthermore, for some cases as described in the previous comment, the population needs to be tessellated which, in my opinion, practically results in finite populations. I think it would be very helpful to discuss the similarities and differences of the finite and infinite population approach more as most readers familiar with forest inventories are likely to be "indoctrinated" with the finite population approach.

When providing two estimators for the same purpose such as the synthetic or extended synthetic, some (extreme) users might just chose what results in the smallest standard error. Some more guidance or discussion on when to use which estimator may be needed.

Often when using remotely-sensed auxiliary information in forest inventories, due to temporal or spatial mismatch or other errors, analysts have to exclude "outliers" from the sample that would have large leverage in working models. In this case, the zero mean residual property of internal linear models (after eq. 2) does, strictly speaking, not hold. Consider giving some guidance for these very common cases. In the design-based framework, one could of course ignore the issue. However, a lot of efficiency may be lost just due to few observations.

Dealing with sample plots close to borders (split plots) is a common issue in forest inventories. It would be helpful to mention how to handle them in the estimators.

Specific comments:

1. *Abstract: "by hand" – change for example to "by field crews".*

   Has been changed accordingly.

2. *Multiphase vs. multi-phase.*

   Note: *Andi: There is a slight inconsistency in our article here: Our package says 'multiphase', but here we used 'multi-phase'. However, should we then also change 'two-phase' to 'twophase'?*

3. *Introduction: Consider mentioning the important role of (N)FIs in monitoring carbon change in the context of climate reporting and mitigation..*

4. *Section2.1: Forest area is often itself an important parameter to be estimated. For completeness, it should be mentioned that (or how) it is possible to estimate forest area under the infinite population approach..*

   The estimators presented in our article particularly rely on the assumption that the forest area λ is known. The reason for this is that the local density $Y(x)$ is actually the Horwitz-Thompson

   Note: *Andi: Whereas the forest area $\lambda(F)$ appears in the Horwitz-Thompson estimator, we actually only have to know the forest boundaries to compute the intersection of the circle and the forest boundary. The exact forest area is only needed to compute the total by multiplying, e.g., the sample mean with $\lambda(F)$. The author is, however, right by saying that in most NFIs the forest area is rather estimated than known exactly. A solution, as applied in the Swiss NFI, is to estimate the forest area by multiplying the known country area by the percentage of (randomly and uniformly distributed) sample points falling within the forest (decision made by field crews) -> see Mandallaz 2014: Note on the estimation of totals. So actually, its no big deal "to do it in the infinite population approach". However, in the case of multiphase sampling, also s_1 or s_0 points can no longer be restricted to the forest, which can lead to loss of efficiency (see also Mandallaz 2014).*

5. *Section2.1: "population in the finite approach ... not well defined". Can this really be generalized in this way? If not, consider deleting. Think about estimating forest area where every finite population unit is covered by a certain proportion of forest..*

6. *Section 2.4: Consider mentioning already here that the bias correction is not deemed reliable for $n_2 G < 6$..*

7. *S2.5: systematic grids are not used for reducing travel costs but rather because of practicalities around obtaining a spatially balanced sample..*

8. *Fig 3: Meaning of the asterisk *?.*

   Has been corrected.

9. *After eq. 2: Where comes the term g-weight from? No "g" in any of the equations around eq. 2?.*

10. *After eq. 6: "variance ... external model ... usually slightly smaller than" internal... In the examples for the global estimators further down it is vice versa. Consider rephrasing..*

11. *After eqs. 12, 25 and possibly elsewhere "...unbiased point estimates..." Can an estimate be biased or is it estimators that are biased?.*

12. *Section6.1: Ancova model: That means the model of the extended synthetic estimator is also of type ancova. Worth mentioning?.*

13. *To me, the term "estimation error" seems a bit unconventional..*

14. *I think the title of the package is a bit narrow, because it suggests that it could only be used for forest inventories (FIs). FIs were the background and examples for the estimators. However, the methods are generally applicable for all kinds of infinite and finite populations that are sampled by a probability design. The package name may be difficult to change (are package aliases allowed on CRAN?) but at least the function parameters could be general. terrgrid.id in the phase_id list could for example be replaced by a general term. In this list, consider using the underscore as a consistent separator..*

15. *Some of the estimators are also implemented in the well-known "survey" package which could be mentioned..*

16. *Around eq. 13. The extended estimator is described as being "elegant". However, it seems to be necessary to fit the model for each small area with a change in the indicator variable. How does this fit?.*

17. *After eq. 14. How can there be a residual correction term in a synthetic estimator? There may not even be observations available from within small area G. Please describe and define the term $\theta_s 2$ in more detail..*

18. *Sentence after eq. 17: difficult to follow. Consider rephrasing..*

19. *Acknowledgements: grisons - Grisons.*

20. *Value of twophase() (Package forestinventory version 0.3.1): $Z_b ar_1 G$ is $Z_b ar_1$?.*