Reviews

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

Andreas Hill, Alexander Massey

14. September 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.

The reviewer makes a valid point that it would be easier from a "plug-and-play"perspective to simply interpret what the phase memberships were via the locations of NAs. However, NAs in covariates may have nothing to do with the sampling design e.g. non-response due to inaccessibility or lidar error. We feel that the current implementation of phase_id forces the user to be more attentive when these situations occur that may interfere with the uniform and independent sampling assumption for the sampling design. To assist the user in understanding how function calls are made we have included several examples for the usage of all estimators in both the documentation and the vignette using real-life examples. It should also be noted that defining the phase_id for all observations with the same (terrestrial) id essentially does the same as the reviewer has described with a warning message given. However we do not recommend this usage for the aforementioned reasons.

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%.

The reviewer is concerned that more attention needs to be given concerning issues in boundary cases with the coverage rates. The given explanation seems to have switched α and $1 - \alpha$ in a few places but made clear that there was a link between too low sampling fractions and coverage rates under the finite population approach. However, the concept of having a sampling fraction too low for the confidence level isn't really applicable in the infinite population approach where the total population is a continuum of points and is thus infinity. We have performed simulations under the infinite population approach and could not reproduce the reviewers examples, but did show that extremely low sample sizes could have an influence. The cited literature includes many simulations and artificial examples demonstrating the distribution of the parameter of interest (the integral of the local density function) is beyond analytical treatment under the sampling schemes but can be well approximated by t-distribution (and consequently the normal in large samples). The reviewer does have a very valid point that discussing coverage rates would improve the paper. For this reason we have expanded the confidence interval section to include some new simulated examples of coverage rates to develop the reader's intuition about when the sample size is too low to produce reliable results.

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").*
   Has been changed accordingly.
2. *On page 27, the element symbol should be a subset symbol..*
   Has been changed accordingly.
3. *On page 32, grisons should probably be capitalized..*
   Has been changed accordingly.
4. *Equations (1), (4), (13), (16), (20) might be better readable without the n2 fractions.*
   Whereas we agree that they might be slightly more readable without the n2 fractions, the decision to include them was based on the fact that this is how the formulas were originally presented in Mandallaz's papers. We feel that the marginal loss in readability is offset by perfect continuity with the original notation.
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..*
   It has been added that the models require intercept terms.

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.

Different versions of this article were written that attempted to shorten it in ways just like the review mentioned. However, we ultimately decided that the purpose of this article is to bridge the gap between likely users

of the software (e.g. forest engineers and technical practitioners) and the mathematical details that might be beyond their skillset. For this reason, we decided to ensure that every function call was connected explicitly to the underlying formula that it applies and a citation is given for those interested in further technical details. We believe this also enables the reader to write their own routines using the functions which can be used to error check our routines. Considering that the current length is not excessively long compared to other JSS papers, we strongly prefer to not to shorten sections merely for the sake of making the article shorter.

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.

Reader 1 suggested that this section in fact be extended to include discussion about coverage rates. We prefer to go in this direction.

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.

This is explained in the section on calculating explanatory variables. It is explictly stated that the exact exhaustive mean vector can only be calculated for explanatory variables based on means (note that polygon layers are means of binary variables). It has now been explicitly added as an example that order statistics are not included in this category with the reviewers example given. It is already explicitly stated that the size of the support should ideally be chosen to achieve a best possible explanatory power of the regression model on page 9 and a citation has been added for the interested reader.

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.

The population doesn't need to be tessellated and can be allowed to overlap or not use all the wall-to-wall map in the infinite population approach. This is already mentioned on page 9. Section 2.1 discusses the Infinite Population Approach. Some brief discussion comparing to the finite is included, but as this package is about software written for the infinite not finite approach, we prefer not to delve too deeply into a topic that makes little practical difference.

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.

This is true, but it is already stated many times that the synthetic estimators will usually have a smaller variance than estimators incorporating the regression model uncertainties, but at the cost of a potential bias making them seem deceptively optimistic (see pages 5, 13 (twice), 14, 16 (this is explained here using a concrete example), 21, 26 and 30)

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.

Removing outliers is more of a common practice in model-dependent estimation, which is beyond the scope of this article. You should never remove outliers in design-based estimation unless you can defend the belief that the resulting (and self-inflicted) non-response is missing at random. The observed value of the the response/explanatory variable is fixed at each x so deleting an observation effectively means the plot at x has been removed from the forest. For temporal mismatches this is pretty much indefensible because the outliers likely represent harvesting after the aerial measurement. We initially included a discussion comparing design-based with model-dependent estimation but we ultimately decided that the paper has more focus if we concentrate only on explaining how to use this software under the design-based infinite population approach. In any case, many citations are currently included that already discuss these issues.

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.

A mention of sample plot boundary adjustments is now included in Section 2.1 with a citation.

Specific comments:

1. *Abstract: "by hand" – change for example to "by field crews".*
   Has been changed accordingly.
2. *Multiphase vs. multi-phase.*
   Multiphase is now used consistently to match package's spelling
3. *Introduction: Consider mentioning the important role of (N)FIs in monitoring carbon change in the context of climate reporting and mitigation..*
   Good point. Has been changed accordingly.
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 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 (independent 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"but how the user wants to estimate forest area is up to them. A citation noting where the reader can find more information on this topic has been included but further discussion was not included as it was outside the scope of this paper.
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..*
   That suggestion only can be done if the plots form a perfect tesselation. Circular units, the most common in NFIs, can't do this. The rest of the sentence actually says "with respect to the forest area". This is true. Points can fill any forested area even if it is a fractal whereas a counter example forest area can always be given for any plot shape you put forward. Extending your tesselation beyond the forest edge is possible, but this increases the weights for boundary cases which likely increases the variance. Many of the cited articles in the infinite population approach already discuss this but the focus for this article is not a comparison of the finite and infinite approaches, but a presentation of a software package written under the infinite approach.
6. *Section 2.4: Consider mentioning already here that the bias correction is not deemed reliable for $n_2G < 6$..*
   $n_2G < 6$ is just a rule of thumb. In Section 2.4, the goal was merely to introduce the small area problem in a general sense for readers that were not familiar. Rules of thumb should not be blindly obeyed though so we thought it best to refer to it when examples are present to give context allow the reader to develop intuition. Further discussion of this rule of thumb including simulation examples are included in the confidence interval section in the context of coverage rates.

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

   The reviewer is correct and this change has been implemented.

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?.*

   The g-weight is the name of a variance estimator for the point estimate in eq.2 as opposed to another possible variance estimator here named the external estimator. The origin of the g-weight name is more obvious in the citation provided in the same paragraph. The g-weight is analogous to a hat matrix and can be extracted from the more compact form that was included in the article. In the spirit of not being too verbose we chose the compact form.

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..*

    This is usually the case, but the estimators are both random and the opposite may sometimes occur. They are asymptotically equivalent though. The point is that one should prefer the g-weight because it attempts to account for sampling variation in the coefficients whereas the external variance completely ignores it. This is explained on page 11 after the example for the global two-phase estimator.

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

    Unbiased point estimates are point estimates produced by unbiased point estimators.

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

    The extended synthetic estimator does not necessarily need to implement an ancova model. That depends on the types of variables that the user has specified. The extended synthetic estimator merely adds an indicator variable for the small area in addition to the other variables that the user specified. This can in fact extend the model to an ancova model if the explanatory variables contain at least one continuous predictor. However, the reason of adding the indicator variable exclusively is to ensure the zero-mean-residual property in both the global domain F and small area G, and not to improve the model fit. We think that linking the indicator variable to the anova or ancova modeling technique might probably cause misunderstanding of the indicator purpose on the side of the readers.

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

    The term is consistent with much of the already published literature included in the citations (e.g., von Lüpke et al. 2012).

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..*

    The reviewers comments refer to the version of the package that is already available on CRAN and has over 5000 downloads. We prefer not to alter the function parameters in order to preserve backwards compatibility. In the future if the usefulness of more general package names becomes apparent a wrapper package could be created.

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

    The *survey* package is in the finite population approach (aka list sampling). It is mentioned on page 2.

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?.*

    It is elegant in the sense that the estimator is unbiased merely by modifying the design matrix. It is not necessary to refit every small area with an indicator, but this is how it was implemented in the package.

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 θ_{s_2} in more detail..*

    It says that "the formulas look similar to the synthetic estimators", because the formulas are exactly same. They merely have a modified design matrix that ensures that the residual correction term is zero by construction and doesn't need to be included. θ_{s_2} is now clearly defined in Eq. 13 and the preceeding paragraph.

18. *Sentence after eq. 17: difficult to follow. Consider rephrasing..*
    Parts of the equation are directly referenced in the sentence to make it easier to follow.
19. *Acknowledgements: grisons - Grisons.*
    Change has been implemented.
20. *Value of twophase() (Package forestinventory version 0.3.1): Z_bar_1G is Z_bar_1?.*
    The reviewer's comment refers to the help documentation of the R-package. The twophase() function can be used to estimate in both the global and small area contexts. Z_bar_1G refers to the small area context whereas Z_bar_1 is the terminology used for global estimation (i.e. when G = F). In the future release of the package we will try to include a short reference to this distinction in the help documentation but given that the package already has over 5000 downloads we prefer not to change the names of the function outputs.