Answer to Editor

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

Andreas Hill, Alexander Massey

30. Dezember 2018

Feedback from Editor and Reviewer A

Editor:
The editor does not understand why the authors were not able to reproduce the code examples given by the reviewer. Is there a bug, or just a misunderstanding between authors on one side and reviewer + editor on the other? Please check the code and explain.

Reviewer A:
In the revised version of the manuscript (ID 3280), the authors include a new discussion on coverage rates of confidence intervals (p. 25), which, however, appears to only partially embrace the methodological concerns I raised in the earlier review. In particular, the statement that 'the nominal coverage rate of 95% can [be] expected to be met for large sample sizes (e.g., n2=50)' is in conflict with examples 2 and 3 included from my review (n2=80, coverage rate = 85%). I am not convinced by the authors' claim that they could not reproduce the examples. On the other hand, the new discussion and in particular Figure 5 are very helpful in illustrating the issue of coverage rates, and I believe this will for most readers be sufficient to make them aware the related problems. Based on my rather short reading of the manuscript and given that a formal review is not required at this stage (after previous recommendation 'Revisions Required'), I would therefore agree that the manuscript can now be accepted.

Answer

The reviewers code runs without bugs but the simulation was incorrectly formulated. One fundamental problem with the reviewers simulation is that he seems to pose all three examples in the *finite* population context rather than the *infinite* population context. We can see this right at the beginning:

Here, the reviewer sets up a dataframe defining a *finite* population that he then uses in his simulation examples:

```
> # Reviewer's artificial data:
> set.seed(0)
> d <- data.frame(
+           elv = (e<-runif(500, 300, 550)),     # explanatory variable (observed in s1)
+           vol1 = rnorm(e, mean=600-e, sd=100), # response variables (to be observed in s2 only)
+           vol2 = rnorm(e, mean=300, sd=100)*(e<310),
+           resp3 = runif(e) < 0.03)
```

Let's look at the first simulation example to explain where the fundamental problem is:

```
> library(forestinventory)
> # PROBLEM 1: Small sample size (n2=5)
> true_value <- mean(d$vol1)
> cover <- replicate(10000, {
+ d$phase <- 1
+ d$phase[sample(length(d$phase), 5)] <- 2
+ ci <- confint( twophase( vol1 ~ elv, data = d,
+ phase_id = list(phase.col="phase", terrgrid.id=2) ), level=0.95 )
+ ci$ci$ci_lower_g <= true_value & true_value <= ci$ci$ci_upper_g
+ } )
> mean(cover)
```

As obvious, the true population mean (`true_value`) that has to be estimated is here calculated by taking the mean of a *finite* number of response variable values ($N = 500$, i.e. `nrow(d)`). This shows that the reviewer actually defined a finite population for the response variable (`vol1`), and likewise for the auxiliary variable (`elv`). In every iteration of his simulation, he then resamples from a finite set of observed response variable values s_2. **This is a fundamental misunderstanding of the infinite population theory, which was exclusively used to derive all estimators contained in** `forestinventory`. First of all, the true population mean (`true_value`) should be an integral (i.e. spatial mean) over infinite points, as described in section 2.1 in the article, but by using the finite mean, the reviewer has effectively acknowledged that actually $N = 500$ instead of infinity.

Second, we then see that every population unit of his finite population is assigned to the first phase (`d$phase<- 1`). Basically, this means that the auxiliary information is not resampled in every simulation iteration, but the entire population ($N = 500$) of auxiliary information `elv` is used. This indicates that the reviewer intended to apply an *exhaustive* two-phase estimator where the true population mean of the auxiliary information is used. Notice that in the `twophase()`-function call, the reviewer, however, did *not* set the exhaustive parameter as demonstrated in section 3.1 pp. 11-12 in the article, making this a non-exhaustive function call. This means that by simulation design s_1 does not get resampled with every replicate so the variability coming from resampling the first phase is ignored.

All 3 examples given by the reviewer made these same errors so the results clearly can't be considered conclusive on their own merits. This is why we attempted to reformulate the simulation examples correctly in the infinite population context and mimic the examples intended effects. The following is some sample code (similar code was added to the Appendix in the last submission) to demonstrate simulation in the infinite population context. This technique has already applied in Mandallaz (2013a), Mandallaz et al. (2013) and Massey et al. (2015):

```
> library(forestinventory)
> library(rmutil)  # used to calculate the integral in the spatial mean
> #
> #
> set.seed(1)
> #
> # --- function to create density surface: --- #
> target.surface <- function(x0, y0){
+   local.density <- 30 + 13*x0 - 6*y0 - 4*x0^2 + 3*x0*y0 + 2*y0^2 + 6*cos(pi*x0)*sin(pi*y0)
+   local.density
+ }
> #
> # --- True spatial mean for Global Surface Area: --- #
> true_value <- int2(target.surface, a = c(0,0), b=c(3,2))/6 # 39.16667  #true mean
> #
> # --- sample generator for simple two-phase sampling: --- #
> sample.generator<- function(n1, n2){
+
+   realization <- matrix(NA,n1,6)
+   for(i in 1:n1){
+     x0=2*runif(1) # NOTE: The randomness comes from the uniform independent selection of points
+     y0=3*runif(1)
+     realization[i,1] <- target.surface(x0, y0)
+     realization[i,2] <- x0
+     realization[i,3] <- y0
+     realization[i,4] <- x0*x0
+     realization[i,5] <- x0*y0
+     realization[i,6] <- y0*y0
+   }
+   realization <- as.data.frame(realization)
+   names(realization) <- c("response","x","y","xx","xy","yy")
+   realization$phase <- 1
+   realization$phase[sample(nrow(realization), n2)] <- 2
+   realization$response[realization$phase == 1] <- NA
+   realization
+ }
```

As obvious, we take a finite random sample s_2 and s_1 from an infinite number of points (note that the number of possible points, i.e., coordinate pairs `x0, y0`) as input for the density function `target.surface` is infinite). The randomness in the sampling procedure comes from the uniform independent selection of points that are randomly distributed over the density surface defining the distribution of the target variable at every possible point in the plane.

Reviewer Problem #1

```
> # --- Run 10000 simulations with n1 = 500 and n2 = 5 (small sample size n2) --- #
> n1 <- 500
> n2 <- 5
> num_sims <- 10000
> ci.g.logical <- rep(NA, num_sims) # two-phase coverage indicators for g-weight formula
> ci.ext.logical <- rep(NA, num_sims) # two-phase coverage indicators for external formula
> ci.op.logical <- rep(NA, num_sims) # one-phase coverage indicators
> #
> for(j in 1: num_sims){
+ realization <- sample.generator(n1 = n1, n2 = n2)
+ est <- twophase(formula = response ~ y + x + xx, data=realization,
+ phase_id = list(phase.col = "phase", terrgrid.id = 2))
+ est_onephase <- onephase(formula = response ~ 1, data=realization,
+ phase_id = list(phase.col = "phase", terrgrid.id = 2))
+ ci <- confint(est)
+ ci_op <- confint(est_onephase)
+ ci.g.logical[j] <- ci$ci$ci_lower_g <= true_value & true_value <= ci$ci$ci_upper_g
+ ci.op.logical[j] <- ci_op$ci$ci_lower_op <= true_value & true_value <= ci_op$ci$ci_upper_op
+ }
> #
> # results for coverage rates:
> mean(ci.g.logical) # 0.6738, g-weight coverage rate of 2-phase estimator
> mean(ci.op.logical) # 0.9457, coverage rate of one-phase estimator
```

The one-phase estimator has the correct coverage rate even with sample size 5 (it would also have it for $n_2 = 2$) but it will probably have a very large margin of variance (i.e. estimation error). However, the g-weight variance has lower than nominal coverage rates because the contribution of the variance by calculating the beta vector is based on approximations (i.e. a first order taylor approximation for the g-weight) which are not adequate for this sample size. **This particular issue has been addressed in the added paragraph about coverage rates in section 5 "Calculation of confidence intervals", including a graphic, the simulation code in the Appendix as well as a hint for the reader that in such cases of very small sample sizes, one should prefer to apply the small area estimators which ensure the nominal coverage rates even under such limited sample sizes. The reviewer has allready aggreed with this solution.**

Reviewer Problem #2 and #3

The 2nd reviewer example involves creating a response variable `vol2`, that is `0` for 98.2% of the values (i.e., for his population) and non-zero where the auxiliary information yields a low value. Presumably this would occur in situations where there is a treeline dependent on elevation and the vast majority of the inventory domain (here 98%) is at elevations where no trees can actually grow. He calls this 'Effect of explanatory variable'. As described in Mandallaz (2014), it is somewhat common in practice to sample information over an entire country, i.e. also at places outside forest even including areas where no trees can grow. This is caused by the fact that in order to calculate totals (by multiplication of the estimated density, e.g. volume per hectare, with the area value of the inventory domain), the area of a country is well known, whereas the area of the forest is often not known precisely. In the mentioned article it is demonstrated that by doing so, one usually inflates the sample by a large number of zeros (i.e. places where the target-value is sampled outside forest). It has been mathematically shown that by this procedure, the R^2 of the regression model is overoptimistic as the zeros are easy to predict. This is e.g. the case if - just as in the reviewer's example - the occurence of zeros depend on a known explanatory variable like the elevation. Furthermore, this procedure is inefficient as it inflates the estimated variance, and a suggested solution in Mandallaz (2014) is to filter out most of the non-forested areas where one can certainly assume that no forest exists (this was consequently done in Hill et al. 2018).
However, the reviewer's example goes beyond that topic by contructing an artifical example where *only 2%* of the entire population (i.e. also appr. 2% of the inventory area) consist of forest/trees. It is extremely hard to imagine that such an inventory would actually be conducted by anyone - unless there is no information at all about the environment where the forest inventory is going to be conducted (i.e.,

wether it is a desert with only few oasis, an alpine mountain side above the tree level, ...). This is hard to imagine, especially in times where aerial photographs and further geodata are availabe and used in the planing process of every forest inventory. Problem 3 is actually the exact same example except that the response variable takes exclusively the value `1` at non-zero plots and these non-zero plots (and thus plots with response variable value=1) are uniformly distributed among the auxiliary variables. This means that now there is per-definition even no causal correlation between the auxiliary variable values and the response value, which arises the question why one should use the auxiliary information in the first place. Nevertheless, such a situation may occur if say the inventorist wanted to count how many trees are in the desert where there are just a handful of uniformly distributed oases, and the auxiliary information does not provide any information about their location. These two examples are also pathological for the one-phase estimator because a significant proportion of the samples drawn will have zeros for ALL of s_2 making the estimate $= 0$ and the confidence interval just a point with no chance of covering the true value. If this occurred in practice the sample would immediately be considered suspect. It is hard to imagine a situation where this would accidentally happen. Furthermore, trying to measure forest volume in areas where 98% of the land has no trees is very difficult to accidentally do. Concluding, **the fact that the estimators did not produce the nominal coverage rates under this artificial setup is clearly not dedicated to a limitation of the estimators, but to conducting a totally pathological inventory that would never come into anyone's mind in practice.**

This being said, it is impossible and has clearly not been our intention to address any imaginable misuse of the implemented estimators in our article. However, we want to emphasize that we already addressed edge-cases that can - from our experience - expected to actually occur in practice in section 6 'Special cases and scenarios'. **We now have also added a short paragraph *'Restricting the inventory domain to forested area'* to section 3.1 in order to address the issue of 'sampling outside forest', as mentioned above.**

Final Remark

Since it's first release in July 2016, the package has been downloaded over 6000 times (1000 downloads within the last 4 month). Considering that the field of forest inventory addresses a rather small user community compared to other R-packages, we think that these numbers support the importance and the added-value of the package to the field. It also underlines the unique characteristic of our package among the existing set of R-Packages. The development of the package as well as the writing of the article was accompanied by many feedbacks and discussions with external experts in the field, and many of their questions and criticisms have been taken up in the article. We are convinced that the revision process helped to further improve the article. After carefully checking the reviewers' questions and concerns, we addressed as much of their requests as possible without altering the general objective of our article.

We hope that the editor can agree with our suggestions, and would be happy if the article can now be published in the *Journal of Statistical Software*.

Yours sincerely,

Andreas Hill & Alexander Massey