Combining sampling errors and model errors in estimating forest carbon stock changes

Göran Ståhl
Sören Holm
Hans Petersson

Background

- Regression functions often are used in connection with sample surveys to predict, rather than measure, quantities at the level of sampling units
  - Examples range from tree volume and biomass to habitat suitability for different species
- One important case is when carbon stock changes are estimated

Is this a problem?

- For a long time predicted values have been used as measurements and we have not really suffered from this...
  - But we often report – e.g. within NFIs – standard errors that are rather small, as if only sampling errors (and potentially random measurement errors) were involved
  - And we know that the regression models applied are not always very good

Do we exaggerate the goodness of our regression function based estimates?

- Yes, probably!
- But we often argue that:
  - the regression functions really give us 'true values'
  - the regression functions give rise to (unknown) systematic errors, and we only include random errors in our precision measures
- In two-phase sampling – with regression – we *do* include model errors in our precision measures
  - Thus, in some estimators' variances we tend to include model errors and in others we don't...

What if we would include regression model errors?

- We would consider the parameter estimation phase as adding to the uncertainty of our estimates (like in two-phase sampling)
- We would be able to assess the contribution of model errors to overall uncertainty and conclude about trade-offs between model development and application

But reading the textbooks there is no standard theory available for this...?

- Most likely there is something to be found since this should be a general problem in sample surveys!?!
- But the issue is seldom (or never?) explicitly treated in the forestry literature?
  - Discussions along this line in (few) references on error budgets (e.g. Gertner&Köhl)
  - Suggestions for relevant references from the SNS participants would be very welcome!!

Thus the objective of this presentation is to:

- Motivate why the topic is relevant...
- Present an approach for combining sampling and model errors
  - Examples from a simple simulation study

The set-up

- A (large) sample is taken, where regression functions are applied – called S1 (*m* units)
  - For example a National Forest Inventory
- A (small) sample is selected for estimating the regression function (volume, biomass, ...) – called S2 (*n* units)
  - Some researcher generally would do this, rather independent of the NFI...

Cases explored

- Simple random sampling in S1 – either of individual units or as cluster sampling
- The sample S2 is selected independently from S1 in order to estimate the model parameters

Now to some theoretical details...

- Population model
- Estimator of population mean
- Variance
- Variance estimator

  – But we will not go into very much detail with this...

The population model

$$Y(x) = g(x, \alpha, \varepsilon)$$

- x and α are vectors
- The following model is assumed to be known

$$E(Y \mid x) = g(x, \alpha)$$

- With some simplification, this framework should be applicable to all sorts of regression functions

Estimator of population mean

$$\hat{\mu}_Y = \frac{1}{m}\sum_{i=1}^{m} g(x_{i1}, \hat{\alpha})$$

- The regression function is applied to all the *m* units in the sample S1

Variances...

- Obtained by a decomposition approach and application of first order Taylor series approximation
- Model based setting

Variance

$$V(\hat{\mu}_Y) = \frac{1}{m} \cdot \sigma_g^2 + \sum_j^p \sum_k^p Cov_\varepsilon(\hat{\alpha}_j, \hat{\alpha}_k) \cdot E_{S1}(\bar{g}'_j \bar{g}'_k)$$

Variance estimation

$$\hat{V}(\hat{\mu}_Y) = \frac{1}{m} s^2_{\hat{g}} + \sum_{j=1}^{p} \sum_{k=1}^{p} C\hat{o}v_{\varepsilon}(\hat{\alpha}_j, \hat{\alpha}_k) \cdot \hat{\bar{g}}'_j \cdot \hat{\bar{g}}'_k$$

Empirical test (simulation)

- Population model: $y_i = \exp(\alpha + \beta \cdot x_i + \varepsilon_i)$
- For example a simple volume/biomass function, predicting volume based on diameter only
- Some different sample sizes for S1 and S2 were evaluated and the part of the variance emanating from model errors assessed

Some results

Sample size S1	Sample size S2	True variance	Mean estimated variance	Proportion % due to model
50	25	8674	8593	49
50	100	5098	5095	20
100	25	6530	6447	65
100	50	4116	4143	49

Cluster sampling

- The general framework appears to be rather easily extended to the case of cluster sampling (e.g. trees on plots)
  - That is, the functions may be applied to sub-units rather than independent sampling units

Conclusion

- A framework for combining sampling and model errors, when applying regression functions in sample based surveys, have been developed
  - Probably nothing new, although we have not yet found it in the literature
- The model error proportion of variances often is substantial, especially when the sample sizes are large