Small Area Variance Estimation for the Siuslaw NF in Oregon and Some Results

S Lin, D.Boes and H.T. Schreuder

***Abstract**—The results of a small area prediction study for the Siuslaw National Forest in Oregon are presented. Predictions were made for total basal area, number of trees and mortality per ha on a 0.85 mile grid using data on a 1.7 mile grid and additional ancillary information from TM. A reliable method of estimating prediction errors for individual plot predictions called the semi-parametric bootstrap is given too. Prediction errors were quite large. Suggestions on how to improve such necessary predictions for small areas are given.*

Introduction

Management agencies need reliable spatial information for decision making. In the past foresters and other land managers cruised or sketch mapped an area usually to decide what is where. Managers were contemptuous of statistical sampling because it might give reliably data on how much was there but not where. Frequent legal challenges changed this. Now interest is in obtaining reliable (defensible) mapped and statistical data together. For example Forest Inventory and Analysis (FIA) of the USDA Forest Service has made this one of the high priority research efforts for their national program. One such area of current research is referred to as small area estimation, basically a model-building approach using statistical data in combination with ancillary data such as from the satellite thematic mapper ™, geographic information systems (GIS), topographic maps and other helpful information.

Small area estimation techniques represent a substantial improvement in terms of quality of data and, especially in defensibility of data based management decisions relative to what used to be done when managers relied on subjective information. As so many new developments, the techniques are oversold as a panacea. This is because estimates generated by such techniques are claimed to have standard errors similar to those for classical sampling. The trouble is the comparison is made for the entire population of interest whereas managers are primarily interested in predictions for much smaller areas such as polygons used as a basis for management. Standard errors for individual predictions can be large, as one would expect, given the variability encountered on the ground in forests.

Considerable work in small area estimation of forest resources is now being done in many parts of the world. Multiple imputation methods (including regression models) and k-nearest neighbor techniques have been proposed for continuous variables. In these techniques, field sample information is extrapolated to the entire population where information on sample locations is input to non-sampled locations by some criteria such as similar TM readings for the sampled and non-sampled locations. In multiple imputations for each unit without sample data, a series of ℓ predictions are made using randomly selected data and an underlying model and database. Then the data sets are analyzed separately and pooled into a final result, usually an average of the results.

Brief Review of Literature

Franco-Lopez (1999) reviews methods for projecting and propagating forest plot and stand information. As he notes, considerable effort has been extended in Nordic countries combining forest monitoring information, remote sensing and geographic information systems (GIS) to develop maps for forest variables such as cover type, stand density and timber volume with emphasis on the k-nearest neighbor technique. He confides that while his results are imprecise for Minnesota, they are comparable to those obtained by other methods in this region.

Methods

An objective in Lin (2003) was to predict mortality, total basal area and number of trees on 1-ha plots at a 0.85 mile grid given data collected on a 1.7 mile grid. These variables were selected for their economic and ecological importance. The sampling design used on the latter grid was the CVS plot design (Max et al. 1996) consisting of a set of a circular 1-ha plot subsampled at 5 locations with subplots of different sizes for different sized trees. Ancillary data used was plot information for plots in the neighborhood of the prediction locations and TM data using bands 1-5 and 7 from TM 5.

Results and Conclusions

In this study: (1) the predictions errors for predictions were derived based on transformed and non-transformed data for non-sampled locations using only field sample plot or subplot information, (2) prediction errors for predictions were derived based on spatial multivariate regression models with distance-related correlation functions, (3) spatial zero-inflated models were used to handle the numerous zeros in the data, (4) the normal approximate and bootstrap-t prediction intervals were compared for the predictors with coefficients based on distance-related correlation functions with and without auxiliary information, and (5) reliable bootstrap methods were developed for different situations.

The semi-parametric bootstrap method is the best method currently available to estimate prediction errors. It works as follows (Lin, 2003 pp 97-98): If we denote $\underline{Y}$ as the vector of the variable of interest, $\underline{\mu}$ as the vector of its mean at the i=1, …,N locations, B the vector of regression coefficients, ε an Nx1 vector of random variables with mean 0 and NxN dispersion matrix σ^2 I, the correlation matrix of Y, I the identity matrix, with H and V the corresponding (0,1) matrices with 1's for the horizontal and vertical neighbors respectively, $C^{(1)}$and $C^{(2)}$the corresponding (0,1) matrices with 1's for corner neighbors in the directions {(1,1), (-1,-1)} and {(1,-1), (-1,1)} respectively and zeroes elsewhere in all matrices. Then suppose the spatial data follows the linear model:

$\underline{Y}=\underline{\mu}+B\varepsilon$ where $\underline{Y}=(Y_1,...,Y_N)'$,$\underline{\mu}=(\mu_1,...,\mu_N)'$,$\underline{\Gamma}=I+\rho_1 H+\rho_2 V+\rho_3 C^{(1)}+\rho_4 C^{(2)}\equiv B'B, \varepsilon\ MVN(\underline{0},\sigma^2 I)$
or more simply:

$$Y_i=\mu+\sum_{j\neq i}^{N}(Y_j-\mu)+\varepsilon_i i=1,...,N$$

where $B^{-1}\equiv I-M$ and $\underline{M}=(m_{ij})$.

For a symmetrical and positive-definite estimator $\hat{\Gamma}$ of Γ, decompose it as:

$\hat{\Gamma} = \hat{B}'\hat{B}$ and define $(\hat{\varepsilon}_1,...,\hat{\varepsilon}_N)\equiv \hat{B}^{-1}(\underline{Y}-\underline{\mu})$ and $\varepsilon_i= \hat{\varepsilon}_i-\sum_{j\neq i}^{N} m_{ij}\hat{\varepsilon}_i/N, i=1,...,N.$

Then the assumed i.i.d. ε_i are bootstrapped.

For each bootstrap sample $\underline{\varepsilon}^*$ calculate $\underline{Y}^*=\hat{\mu}+\hat{B}\underline{\varepsilon}^*$.

Then refit the model to each of the bootstrap samples and predict for each of the samples at the desired locations. The variability between the estimates for the predicted location(s) is then used for the bootstrap variance estimate for that location.

We conclude that prediction models developed are quite unreliable for this data set. For prediction purposes a simple model that assumes a spatial correlation structure without any distributional assumptions worked as well as any other. The specific spatial correlation structure was selected by minimizing an overall mean squared error. Our recommended predictor is a linear combination of measurements at neighboring sites. The coefficients in the linear combinations are functions of the estimated neighboring correlations. These simple predictors have an analytical formula for the prediction errors and these too can be estimated using the estimated correlations. These predictors work for any of the response variables total basal area, number of live trees and mortality. The semiparametric bootstrap method is recommended for computing the reliability of the plot predictions.

The overall results were quite disappointing. There is little spatial dependence among the response variables at the 1.7-mile grid scale and consequently such dependence is of little use in making predictions at non-sampled plots. Also, there is little useful correlation between our response variables and the many available auxiliary variables, including those from satellite imagery (™). Auxiliary data contributed little to improving predictions in our study. The following shows the results for six plots, similar results were found for other plots.

For total basal area we found that certain auxiliary data were useful in reducing prediction errors. To illustrate as to the size of the errors involved for this variable, we predicted the value at each of six sites on the 0.85-mile grid. Predicted total basal area range from about twenty to sixty m^2 per ha. If the sample mean was used as a predictor and independence among sites assumed, the prediction error is estimated to be 24.38. The estimated prediction errors using our simple predictors at the six sites are, respectively, 24.23, 24.22, 24.21, 24.24, 24.11, and 24.12. Clearly there is little improvement over the nominal 24.38. On the other hand, the corresponding estimated prediction errors for the same six sites, when one incorporates aspect and band 4 (representing active vegetation) as auxiliary variables, are, respectively, 18.54, 18.64, 18.53, 18.47, 18.59 and 18.44 showing a nearly twenty-five percent improvement.

For number of live trees/ha, predictions are between above three hundred and five hundred trees per ha. The respective estimated prediction errors at the six sites are 264.81, 264.56, 264.52, 264.70, 264.53, and 264.55 without auxiliary data, and 258.52, 258.14, 258.62, 257.74, 259.87 and 258.25 with auxiliary data. The overall reduction in prediction errors is less than five percent.

For mortality per ha, predictions are between about forty and a hundred and thirty. The nominal estimated prediction errors range from 52.71 to 52.76 without auxiliary data. The estimated prediction errors using auxiliary data ranged from 53.19 to 54.08 so for M, auxiliary data provides no improvement in prediction.

We are not ready for prime time in FIA (except perhaps in some very homogeneous parts of the country), i.e., we should not be publishing results on small area estimates at this time unless quite reliable estimates were generated for the area involved. For successful small area estimation, three conditions need to be met. First of all there should ideally be a good correlation between sampled and non-sampled areas nearby or similar to the ones sampled. This usually requires a much more intensive grid than the 5000 m grid now used by FIA. Secondly, there should be a close relationship between spatial characteristics of a location and plot information and the spatial locations need to be accurate.

Research should be pursued further in this field. We need to use data from an improved resolution remote sensor relative to TM and/or photography as used by the National Resources Inventory (NRI) of the Natural Resources Conservation Service. This improved ancillary information may yield better information for better predictions. Also: there is a need to think about generating estimates emphasizing different sources for different variables and redefinition of some variables could be useful too. For some variables remote sensing alone (including photography) may do better than ground sampling, for others we have to continue to rely on ground sampling primarily.

For management purposes we really want mapped or very detailed local information at least for percent cover, cover types (about 30), and stand structure. Percent cover can probably be obtained as well from remote sensing as from ground sampling. Cover type definitions are actually still quite fluid. In a simple case: some people argue that a mixed pine/hardwood stand has to have at least 40 percent of each, the remainder in the other. Others might argue for a 30/70 percent split. It seems reasonable to use definitions on what remote sensing can objectively give us in that regard. Stand structure on the other hand will probably always have to be obtained from ground sampling. Other examples abound. Large tree mortality and certain types of old growth can probably be best estimated from remote sensing whereas lower vegetation will have to be sampled on the ground usually. In this regard different plots may need to be used for different variables too at the same grid locations.

To obtain reliable predictions today, additional information is required such as that available from improved remote sensors or large- scale photos combined with expertise from local ecologists. Also, at present it is still necessary to correct for location errors with models. Making such corrections requires considerable information on the extent and location of the errors. Hopefully, improvements in GPS-type sensors will allow us to ignore location errors in the future.

References

Lin, J-M. 2003. Small area estimation. PhD dissertation Colorado State University, Ft Collins, CO.

Max, T.A., Schreuder, H.T., Hazard, J.W., Teply, J. and Alegria, J. 1996. The Region 6 Vegetation Inventory and Monitoring System. USDA FS PNW Res. Paper PNW-RP-493. 22 p.