Literature Review

Grayson White

Literature Review

Ver Plank et al (2017): Hierarchical Bayesian models for small area estimation of forest variables using LiDAR

This paper proposes three Hierarchical Bayesian (HB) models for small area estimation (SAE) of above ground biomass (AGB). These models are the Fay-Herriot (FH), Fay-Herriot with conditional autoregressive random effects (FHCAR), and FHCAR with smoothed sampling variance. They first introduce the FH model with is defined as:

$$Y_i = \theta_i + \epsilon_i$$

$$\theta_i = x_i \beta + v_i$$

where θ_i is the mean AGB, $\epsilon_i \sim N(0, \sigma_i^2)$, the x_i is a $(p \times 1)$ matrix, β is the $(p \times 1)$ matrix of regression coefficients, and $v_i \sim N(0, \sigma_v^2)$. Note: $1 \leq i \leq m$ where we are indexing over forest stands.

The FHCAR model is very similar, except for the fact that it adds a spatially structured random effect that follows a conditional autoregressive (CAR) prior distribution which allows the authors to take advantage of the spatial autocorrelation seen in the data (stands close together have similar AGB values). For this model, they specify:

$$v = (v_1, \ldots, v_m) \sim N(0, \Sigma(\sigma_v^2, \lambda))$$

where λ is the autocorrelation parameter. They also specify the definition of the covariance matrix:

$$\Sigma(\sigma_v^2, \lambda) = \sigma_v^2 [\lambda R + (1 - \lambda) I]^{-1}$$

The third model they specify is again similar to the first two, but with more: they want to be able to reduce instability in variance estimates in small sample sizes and so instead of saying that we have a fixed and known sampling variance (which is common practice), they specify the FHCAR-SMOOTH model where they define the following:

$$\tilde{\sigma_i^2} = \frac{V_e}{n_i}$$

$$V_e = \frac{\sum_{i=1}^m a_i \sigma_i^2}{\sum_{i=1}^m a_i}$$

where n_i is the number of variable radius plots in stand i.

After the authors introduce these three models, they discuss the priors used in their analysis. They use flat priors for all β's, $\sigma_v^2 \sim InvGamma(\text{shape} = 2, \text{scale} = \sum_{i=1}^m \sigma_i^2 / m)$, and $\lambda \sim Unif(0, 1)$ To sample from the posterior for θ, β, and σ_v^2 the authors used the Gibbs sampler, and for λ they used the Metropolis-Hastings algorithm.

After running their models, they compared their results to Breidenbach et al. (2016) and Mauro et al. (2016) and those author's frequentist results. The HB models had higher σ_v^2 which allows for more realistic inference dealing with the uncertainty associated with estimating above ground biomass.

Finley (2017): Hierarchical Bayesian models for small area estimation of county-level private forest landowner population

NOTE: This paper is *super* similar to the LiDAR paper, I believe that this is a draft of that paper, but without using the LiDAR technology and few extras that got cut out of the final paper. This paper is a draft copy from another journal.

This paper first gives a very nice summary of small area estimation and what kinds of direct estimates are used:

Direct estimates: estimates taken from design-based framework (NWOS survey). Low precision since low response rate. Used for benchmarking against model based approach.

SAE: Combines direct estimates and covariates/explanatory variables to produce better estimates. So, SAE is composed of wo component models: a sampling (direct estimate) and a linking model. "The linking model is a linear model with random effects that relate the small areas of interest with some error."

2.3: Simulation Study. This portion is not in the published paper. Relevant to assesing models: "One iteration in the simulation study produces a set of county-level direct and SAE model estimates by: i) drawing a random probability proportional to size sample from the private forest ownership list sample frame; ii) computing direct estimates (Section 2.2.1); iii) estimating FH and FH- CAR models (Section 2.2.2); iv) evaluating differences between SAE model population estimates and truth. Summarizing results from iv for a large number of iterations allows us to assess precision and bias in SAE model population estimates."

They used population density and total forest area as explanatory variables. When assesing their simulation, the checked bias, "relative bias" = the bias relative to the truth for each county, MSE, RMSE, credible intervals.

Results: explanatory variables were significant in almost all simulations (95% credible interval). The biases for the states were very low with the FH and FHCAR models (less than 1/10th of a percent for Montana and less than 2/10th of a percent for NJ)

Discussion: One thing that they note is that more/better covariates could be useful. I wonder what others we have access to that would be useful? They also note that the FHCAR model did not do a great job, and while its use is to deal with spatial covariates that are not directly included in the model, there was not significant improvement from the FHCAR model.

Breidenbach (2012) (Frequentist methods): Small area estimation of forest attributes in the Norwegian National Forest Inventory

This study's goal is the measure mean forest biomass in a forest of Vestfold County, Norway. The authors compared simple random sampling, generalized regression, and EBLUP estimators. For the simple random sampling estimator, the authors calculate the sample mean, however they note that the MSE is unstable due to the small number of random samples. The next estimator that they discuss is the synthetic regression estimator which is an indirect estimator. This estimator is "synthetic" because it "synthesizes information from sample plots also outside the domain of interest." They estimate the mean with the following:

$$\bar{Y}_{S,i} = \sum_{j=1}^{N_i} \vec{x}_{ij}^T \vec{\beta}$$

For this estimator, they do not calculate the MSE because "the domain-level model bias cannot be considered adequately with the existing MSE estimators, which is why we will not derive MSE estimates for the SRE." To deal with the bias, the authors introduced the generalized regression estimator (GREG) which uses a correction term that accounts for bias given n is large enough.

They next discuss the BLUP estimator. This is a model based estimator rather than a direct or indirect estimator and it combines direct and indirect estimates. The BLUP estimator is very similar to the GREG estimator, except for the factor:

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\epsilon^2/n_i}$$

which handles the weight of the bias correction factor according to model accuracy and n. Finally, the EBLUP estimator uses estimated variances rather than the true value of the variances as in the BLUP model.

The authors found that the MSE of the EBLUP estimator was much lower than that of the SRS given a reliable estimate was possible. Most of the time, the MSE of the EBLUP was smaller than that of the GREG as well. The authors conclude to favor the EBLUP over the two other SAE methods (GREG and SRS).

Finley (2009): Improving the performance of predictive process modeling for large datasets

In this paper, Finley proposes a knot-based predictive process aimed to reduce computation time and preserve "the richness of desired hierarchical spatial modeling specifications in the presence of large datasets." Included in the paper is a section on the application to forest biomass prediction and modeling. When applying this method, they found: "Convergence diagnostics revealed 5000 iterations to be sufficient for initial burn-in and so the remaining 30,000 samples from each chain were used for posterior inference. The 206 knot model required approximately 2 h to complete the MCMC sampling with the 106 and 51 knot models requiring substantially less time to collect the specified number of samples" and concluded that "Ultimately, the predictive process model makes this analysis and subsequent pixel-level prediction trivial for even a common single processor workstation."

I am unsure if this computation barrier will become relevant when doing small area estimation, however if it does this knot-based approach seems extremely relevant to look into and potentially implement.