A Bayesian Approach To Multisource Forest Area Estimation

Andrew O. Finley[1] **and Sudipto Banerjee**[2]

Abstract.—In efforts such as land use change monitoring, carbon budgeting, and forecasting ecological conditions and timber supply, demand is increasing for regional and national data layers depicting forest cover. These data layers must permit small area estimates of forest and, most importantly, provide associated error estimates. This paper presents a model-based approach for coupling mid-resolution satellite imagery with plot-based forest inventory data to produce estimates of probability of forest and associated error at the pixel level. The proposed Bayesian hierarchical model provides access to each pixel's posterior predictive distribution allowing for a highly flexible analysis of pixel and multipixel areas of interest.

Introduction

Most large forest inventory programs use stratified estimation techniques to couple inventory data collected from field plots and ancillary data, such as satellite imagery, to increase the precision of their estimates. Satellite imagery has provided a useful and cost effective source for deriving the data layers required for stratified estimation (McRoberts *et al.* 2002). These stratified estimation techniques can produce satisfactory estimates and precision for medium to large geographic areas, but they typically fail to satisfy precision expectations for small areas. Obtaining forest area estimates for small areas requires more spatially intensive sampling designs, more and different kinds of ancillary data, and/or methods that extract more information from inexpensive sources of ancillary data. The increased costs associated with more intense sampling and a larger suite of ancillary data often precludes these approaches. Therefore, approaches to make better use of common and affordable satellite imagery merit consideration.

This paper presents a model-based approach that can be used to couple field inventory data from the Forest Inventory and Analysis (FIA) program of the U.S. Department of Agriculture (USDA) Forest Service with mid-resolution satellite imagery to predict pixel-level forest probability with associated error estimates. The Bayesian hierarchical model presented provides access to each pixel's full predictive distribution from which we calculate the desired inferential statistics. Further, when combined with an appropriate area estimator, these individual pixel estimates can provide area and error estimates for arbitrary areas of interest (AOI).

This paper builds a logistic-regression-based Bayesian hierarchical model for incorporating spatial structure then describes methods for parameter estimation of fixed effects and random spatial effects. A procedure for model-based prediction of probability of forest in arbitrary AOIs is described.

Statistical Modeling

Nonspatial Logistic Model

We first outline a basic logistic model that can be used for modeling the forestation. Suppose we have $i = 1, \ldots, n$ subplots. Based on the percent canopy cover and additional forest structure variables, FIA records single condition subplots as either forest or nonforest. We set y_i as the binary variable designating this classification with $y_i = 1$ denoting that subplot i is forested and $y_i = 0$ otherwise. Conditional on the set of independent variables (spectral characteristic for us), say x_i for subplot i,

[1] Research Fellow, University of Minnesota, Department of Forest Resources, 115 Green Hall, 1530 Cleveland Avenue North, St. Paul, MN 55108. E-mail: afinley@stat.umn.edu.

[2] Assistant Professor, University of Minnesota, School of Public Health, Division of Biostatistics, A460 Mayo Building, MMC 303, 420 Delaware Street Southeast, Minneapolis, MN 55455. E-mail: sudiptob@biostat.umn.edu.

we assume that the y_i s follow an independent and identically distributed (i.i.d.) Bernoulli distribution, $y_i \sim Ber(p(\mathbf{x}_i))$ with $P(y_i = 1 \mid \mathbf{x}_i) = p(\mathbf{x}_i)$. The association between the dependent data vector $\mathbf{y} = (y_1, ..., y_n)$, and the $n \times m$ matrix of m spectral independent variables $X = [\mathbf{x}_1^T, ..., \mathbf{x}_n^T]$, where each $\mathbf{x}_i^T$ is the $1 \times m$ vector of spectral characteristics for the i -th point, is modeled through a logistic link regression

$$p(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \theta)}{1 + \exp(\mathbf{x}_i^T \theta)}, \tag{1}$$

where $\theta = (\theta_1, ..., \theta_m)$ is the vector of parameters to be estimated.

Letting *Data* denote all the available information, say $\mathbf{y}, X$ above, the likelihood function for the data given the above model is

$$L(\theta; \text{Data}) = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i} (1 - p(\mathbf{x}_i))^{1-y_i} = \prod_{i=1}^{n} \frac{\exp\left(\mathbf{x}_i^T \theta\right)^{y_i}}{1 + \exp\left(\mathbf{x}_i^T \theta\right)}. \tag{2}$$

This yields the corresponding log-likelihood function as

$$\ln\left(L(\theta; \text{Data})\right) = \sum_{i=1}^{n} y_i \left(\mathbf{x}_i^T \theta\right) - \sum_{i=1}^{n} ln(1 + \exp(\mathbf{x}_i^T \theta)). \tag{3}$$

Typically, from equation (3) iterative methods are used to obtain the maximum likelihood estimates of the parameters θ (Amemiya 1985). These methods, however, rely on asymptotic (for large samples) distributional assumptions that are rarely verifiable in practice. Alternatively, we adopt a Bayesian paradigm (Gelman *et al.* 2004) that enables direct probabilistic inference for all the model parameters by first specifying prior distributions for them and subsequently using the likelihood in equation (2) to obtain the posterior distribution. In practice, therefore, if $p(\theta)$is the prior distribution for θ, the posterior distribution of θ is given by

$$p(\theta \mid Data) \propto p(\theta) L(\theta; \text{Dat}$$

Such inference typically proceeds from drawing samples from the posterior distributions of the parameters. Markov chain Monte Carlo (MCMC) integration methods (Gelman *et al.* 2004) provide samples from the full posterior distribution of θ that can subsequently be used for inference.

Logistic Model with Spatial Random Effects

The unexplained residual uncertainty associated with the mean function in equation (1) does not accommodate spatial correlation among subplot observations. Ignoring the spatial structure can impair the precision of predictions. A two-stage hierarchical model allows us to incorporate spatial structure into the basic model. The first stage of the hierarchical model adds spatial random effects to the mean structure in equation (1). If the subplots are spatially referenced (e.g., Easting-Northing or some other coordinate system) as $S = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$, we can envision the response as $y(\mathbf{s}_i) = 1$ or 0 depending on whether the subplot is forested. Within the augmented model, the probability that $y(\mathbf{s}_i) = 1$ depends on spatially-referenced independent variables, $\mathbf{x}(\mathbf{s}_i)$ for subplot $\mathbf{s}_i$, the parameters θ, and the location specific random effects $w(\mathbf{s}_i)$:

$$p(\mathbf{s}_i) = \frac{\exp\left(\mathbf{x}(\mathbf{s}_i)^T \theta + w(\mathbf{s}_i)\right)}{1 + \exp\left(\mathbf{x}(\mathbf{s}_i)^T \theta + w(\mathbf{s}_i)\right)}. \tag{4}$$

In the present context, $\mathbf{s} \in D$ and D defines the surface of interest within $\Re^2$.

The second stage of the hierarchical model is to specify the spatial random effects. Specifically, we presume that $w(\mathbf{s}) \sim GP(0, \sigma^2 \rho(\cdot, \phi))$ is drawn from a Gaussian Process with mean zero and $\rho(d, \phi) = \exp(-\phi d)$ is an exponential correlation function (Banerjee *et al.* 2004). This implies $\mathbf{w} \sim N(0, \sigma^2 R(\phi))$ where $\mathbf{w} = (w(\mathbf{s}_1), .., w(\mathbf{s}_n))^T$ is the $n \times 1$ vector of spatial random effects and R is the $n \times n$ correlation matrix with elements $R_{i,j} = \exp(-\phi d_{i,j})$, where $d_{i,j} = |\mathbf{s}_i - \mathbf{s}_j|$ is the Euclidean distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$. The spatial process is described by the spatial decay parameter ϕ and the spatial effect variance σ^2. Customarily, we describe the effective range d_0 of the spatial process by solving $\exp(-\phi d_0) = 0.05$ (i.e., $d_0 \approx 3/\phi$.

The Priors and Likelihoods

With the addition of the random effects, we have the parameter set $\Omega = (\theta, \sigma^2, \phi, \mathbf{w}')$, with length $m + 2 + n$. A Bayesian analysis requires that we assign appropriate prior distributions $p(\Omega)$ to each parameter. A flat prior can be assigned to θ (i.e., $p(\theta) \propto 1$), a vague Inverse Gamma prior for $\sigma^2 \sim IG(a_\sigma, b_\sigma)$, and Uniform for $\phi \sim U(a_\phi, b_\phi)$. The posterior distribution for Ω becomes

$$P(\Omega \mid Data) \propto P(\phi) P(\theta) P(\sigma^2) P(\mathbf{w} \mid \sigma^2, \phi) \times L(\Omega; \text{Data}), \tag{5}$$

where $\mathbf{y} = (y(\mathbf{s}_1), ..., y(\mathbf{s}_n))^T$ is the $n \times 1$ response vector and $L(\mathbf{w}; \mathbf{y}, S)$ is the data likelihood, modified from equation (2) as

$$L(\mathbf{w}; Data) = \prod_{i=1}^{n} P(y(\mathbf{s}_i) = 1 | \mathbf{x}(\mathbf{s}_i), \theta, \mathbf{w}(\mathbf{s}_i))^{y(\mathbf{s}_i)} (1 - P(y(\mathbf{s} = 1 | \mathbf{x}(\mathbf{s}_i), \theta, \mathbf{w}(\mathbf{s}_i)))^{1-y(\mathbf{s}_i)}, \quad (6)$$

with $P(y(\mathbf{s}_i) = 1) = logit^{-1}(\mathbf{x}^T\theta + w(\mathbf{s}_i))$. The Gaussian Process specification implies that the $P(\mathbf{w}|\theta)$ is a multivariate normal $N_n(0, \sigma^2 R(\phi))$. The log-posterior is

$$ln\left(p(\Omega|\mathbf{y})\right) \propto -\left(a_\sigma + 1 + \frac{n}{2}\right)\ln\left(\sigma^2\right) - \frac{b_\sigma}{\sigma^2} - \frac{1}{2\sigma^2}\mathbf{w}^T R^{-1}\mathbf{w} - \frac{1}{2}\ln\left(|R|\right) + \sum_{i=1}^{n} y(\mathbf{s}_i)\left(\mathbf{x}(\mathbf{s}_i)^T\theta + w(\mathbf{s}_i)\right) - \sum_{i=1}^{n}\ln\left(1 + \exp(\mathbf{x}(\mathbf{s}_i)^T\theta + w(\mathbf{s}_i))\right) \quad (7)$$

In the numerical implementation, the prior on ϕ is treated a bit differently. As stated above, its prior is Uniform with the condition

$$p(\phi) = \begin{cases} \frac{1}{b_\phi - a_\phi} & \text{if } \phi \in (a_\phi, b_\phi) \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

In principle, this condition is problematic when the posterior is log-transformed (i.e., $ln(0) = -\infty$); however, this is easily treated in the sampling approach described in the following section.

Posterior Sampling

The Metropolis-Hastings algorithm was used to generate the marginal posterior distribution for each parameter in Ω. Initially, candidate values for the parameters were drawn as a single block from a multivariate normal density. In an attempt to maintain a ~23 percent acceptance rate (Gelman *et al.* 1996), we adjusted the diagonal elements (i.e., the tuning values) of the multivariate normal Σ matrix. In our trials, however, it proved extremely difficult to achieve a reliable acceptance rate that would indicate sufficient mixing. In fact the acceptance rate in our initial trials was typically < 1 percent, while a healthy rate should hover around 23 percent (Gelman *et al.* 2004). Therefore we split Ω into its components, and drew candidate values for θ, σ^2, ϕ, and $\mathbf{w}$ separately. This process required four sequential Metropolis-Hastings steps, where θ and $\mathbf{w}$ were still block updated. In this scheme, we monitored four separate acceptance rates, and generally found much better mixing; this was further improved by specifying the covariance structure among the θ (i.e., off diagonal values) as the dispersion of a multivariate normal proposal for θ.

As noted in the previous section, the Uniform prior on ϕ required that it be treated a bit differently than the other parameters. Specifically, each candidate value of ϕ drawn from the normal proposal density was applied to the conditional statement (8). If the candidate ϕ passed (8), it proceeded through the Metropolis-Hastings iteration; otherwise, the candidate was discarded and subsequent candidates were drawn until the condition was satisfied.

Prediction

Once the samples $\{\Omega^{(k)}\}_{k=1}^{N}$ are obtained from the posterior distribution (i.e., the retained post burn-in sample) $P(\Omega|Data)$, the Bayesian prediction framework is especially simple. The posterior *predictive* distribution we seek is

$$P(y(\mathbf{s}_0) = 1 | \mathbf{y}, \mathbf{x}, \mathbf{x}(\mathbf{s}_0)) = \int P(y(\mathbf{s}_0) = 1 | \Omega, \mathbf{y}, \mathbf{x}(\mathbf{s}_0)) P(\Omega | \mathbf{y}, \mathbf{x}) d \quad (9)$$

where $\mathbf{s}_0$ denotes the location for which the vector $\mathbf{x}(\mathbf{s}_0)$ is known and we wish to predict $y(\mathbf{s}_0)$. Samples from equation (9) are obtained by *composition* sampling: for each $\Omega^{(k)}$ from the posterior sample we simply compute $P(y(\mathbf{s}_0) = 1 | \Omega^{(k)}, \mathbf{y}, \mathbf{x}(\mathbf{s}_0))$ for $k = 1, ...,$ for $k = 1, ..., N$. Programmatically, we first generate a vector of N samples from $\mathbf{s}_0$'s location effect with each element defined by a draw from a normal distribution with mean

$$\varphi_0^{(k)T} R^{-1}(\phi^{(k)}) \mathbf{w}^{(k)} \quad (10)$$

and variance

$$\sigma^{2(k)}\left[1 - \varphi_0^{(k)T} R^{-1}(\phi^{(k)}) \varphi_0^{(k)}\right] \quad (11)$$

where $\varphi_0^{(k)}$ is the $n \times 1$ vector with i-th element given by $\varphi_{0i}^{(k)} = exp(-\phi^{(k)}|\mathbf{s}_0 - \mathbf{s}_i|)$. Then, a vector of probabilities is generated with each element defined by equation (4) replacing $\mathbf{x}(s_i)$ and $w(\mathbf{s}_i)$ with $\mathbf{x}(s_0)$ and $w(\mathbf{s}_0)$. The resulting sample is precisely a sample from the desired predictive distribution in equation (9). A forest probability map can be created using the posterior mean or median (or, for that matter, any other quantile) by simply carrying out the above predictive sampling over a grid of sites. Creating the associated uncertainty map for these predictions is just as simple. For the grid of sites, compute the

uncertainty summary (standard deviation or range) from the predictive sample. We point out that the standard deviations computed from MCMC output are biased as these samples are correlated (Gelman *et al.* 2004). This bias becomes negligible, however, as the size of the MCMC sample becomes large. This sample, being in our control (subject to computational limits), is usually taken large enough and this issue is not serious.

Estimating Multiple Pixel AOI

Once complete posterior distributions are obtained for the pixel-level forest probabilities, interest often turns to obtaining forest area for multipixel AOIs. To be precise, suppose we are interested in a region composed of N_A pixels, say $A = \cup_{i=1}^{N_A}\{\mathbf{s}_i\}$ (perhaps after suitable relabeling of the $\mathbf{s}_i$'s). An estimate of the fraction of the forest area in A is given in terms of the corresponding probabilities at the pixel level by

$$F_A = \frac{1}{N_A}\sum_{i=1}^{N_A} P(Y(\mathbf{s}_i) = 1). \quad (12)$$

Hence, samples $\{F_A^{(k)}\}_{k=1}^{N_{MCMC}}$ from the posterior distribution $p(F_A | Data)$ are immediately obtained from $\{P^{(k)}(Y(\mathbf{s}_i) | Data)\}_{k=1}^{N}$ using equation (12), once the latter are obtained using the methods described in the preceding section. Note that any other functional, such as the total area inside A under forests $\tilde{F}_A = |A| F_A$, where $|A|$ denotes the area of the region A are also immediately accessible to posterior inference.

Summary

This paper described a logistic regression model with hierarchical random spatial effects that can be used to couple field inventory data from the FIA program with mid-resolution satellite imagery to predict pixel-level forest probability with associated error estimates. Once the model's parameters are estimated by the Metropolis-Hastings algorithm, composition sampling is used to delineate each pixel's predictive distribution from which we calculate the desired inferential statistics. These pixel specific predictive distributions of probability forest are then combined to provide forest area and error estimates for multipixel AOIs.

Acknowledgment

The authors thank Dr. Ron E. McRoberts of the USDA Forest Service FIA program for providing the data and critical review. This research was partially funded by the National Aeronautics and Space Administration's Earth System Science Graduate Fellowship program.

Literature Cited

Amemiya, T. 1985. Advanced econometrics. Cambridge, MA: Harvard University Press. 521 p.

Banerjee, S.; Carlin, B.P.; Gelfand, A.E. 2004. Hierarchical modeling and analysis for spatial data. Boca Raton, FL: Chapman and Hall/CRC Press.

Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. 2004. Bayesian data analysis. 2nd ed. London: Chapman and Hall/CRC Press.

McRoberts, R.E.; Wendt, D.G.; Nelson, M.D.; Hansen, M.H. 2002. Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. Remote Sensing of Environment. 81: 36-44.