Model-assisted Estimation of Forest Resources with Generalized Additive Models

Jean Opsomer, Jay Breidt, Gretchen Moisen, Göran Kauermann

August 9, 2006
Outline

1. Forest surveys
2. Sampling from spatial domain
3. Model-assisted estimation
4. GAM estimation for forest inventory data
5. Variance estimation for systematic samples
Research Project

- Collaboration between academic and US Forest Service statisticians
- Goal: apply on-going modeling efforts by Forest Service staff to improve efficiency of survey estimators
I. Forest Inventory and Analysis (FIA)

- *Forest Inventory and Analysis* is an annual survey of all forest lands in US.

- Multi-phase survey, including field visits phase with approximately 1 plot/6,000 acres.

- **Expensive**: $68 million in 2004 (nation-wide).
Inference for Surveys?

**Specific Inference**

- expensive, high quality
- targeted to specific application and/or scientific question
- using “custom-built” method (or model) to achieve best possible estimator for particular variable(s)
- willing to defend estimates/inference
Inference for Surveys

Generic Inference

• cheap, reasonable quality, good for many purposes

• using method appropriate for large number of variables

• provide reasonable answers to many possible scientific questions

• validity of estimates resistant to model misspecification; model independent
Survey Estimation

- Classical methods depend only on sampling design (Horvitz-Thompson; Hájek)
- Improved methods are still design-based but take advantage of auxiliary information
  - ratio, regression, post-stratification
  - model-assisted (Särndal et al, 1992)
  - calibration (Deville and Särndal, 1992)
  - nonparametric (Breidt and Opsomer, 2000), nonlinear/generalized (Wu and Sitter, 2001), ...
Current Dataset

- 2.5 million ha ecological region in Utah
- Contains 968 FIA field plots on 5x5km grid
- FIA plots embedded in 24,980 remote sensing locations on 1x1km grid
Current Dataset (2)

Remote Sensing Variables
- Elevation
- Slope
- Aspect
- Location
- Vegetation Index
- TM spectral bands

Field Plot Variables
- Forest/non-forest
- Total wood volume
- Tree basal area
- Biomass
- Percent crown cover
- Mean diameter
- ...

...
Systematic Sampling

- Common in natural resource and other spatial surveys

- Advantages:
  - Simple to implement, intuitive
  - Easy to “nest” within GIS environment
  - Ensures proportional representation of domains
  - Optimal for certain stationary processes
Systematic Sampling (2)

- Disadvantages
  - Inflexible, can miss rare features in region
  - Does not capture spatial relationships at fine scales (modeling)
  - No design-based variance estimator
2. Sampling from Spatial Domain

- Phase I sample $G_1$ is systematic from continuous domain $U \subseteq D = [0, L_1] \times [0, L_2]

- Phase II sample $G_2$ is systematic (discrete) sub-sample of $G_1$

- Conditional on $G_1$, only 25 possible phase II sample
Sampling from Spatial Domain (2)

• Phase I sample $G_1(u)$, with $u = (u_1, u_2)$ uniform random variable on $[0, 1] \times [0, 1]$ and sampling intervals $(\delta_1, \delta_2)$

$$G_1(u) = \{(u_1 + i_1)\delta_1, (u_2 + i_2)\delta_2) : i_1, i_2 = 0, 1, \ldots \}$$

• Phase II sample $G_2(u, d)$, where $d = (d_1, d_2)$ discrete uniform on $[1, 2, \ldots, h_1] \times [1, 2, \ldots, h_2]$

$$G_2(u, d) = \{(u_1 + d_1 + j_1 h_1)\delta_1, (u_2 + d_2 + j_2 h_2)\delta_2) : j_1, j_2 = 0, 1, \ldots \}$$
Population Characteristics

• Interested in estimating finite population total for variable $z(v)$ on $D$

$$\theta_z = \int_{D} z(v) dv$$

• Total $\theta_z$ can be “gridded” into cells $D_{i_1i_2}$

$$\theta_z = \sum_{i_1i_2} \int_{D_{i_1i_2}} z(v) dv$$

$$= \delta_1 \delta_2 \int_{[0,1] \times [0,1]} \sum_{s \in G_1(u)} z(s) du$$
Survey Estimation

• Phase I expansion estimator

\[ \hat{\theta}_1 z(u) = \sum_{s \in G_1(u)} \frac{z(s)}{1/(\delta_1 \delta_2)} \]

(unfeasible for Phase II variables)

• Two-phase expansion estimator

\[ \hat{\theta}_2 z(u, d) = \sum_{s \in G_2(u, d)} \frac{z(s)}{1/(\delta_1 \delta_2 h_1 h_2)} \]

• Both unbiased, have exact variance formula
3. Model-Assisted Estimation

- Variables $X(v)$ observed on Phase I can improve precision of survey estimators for Phase II variables

- Model-assisted approach provides convenient framework for incorporating auxiliary information within design-based (generic) inference
Model-Assisted Estimation (2)

1. Assume working model $E_{\xi}(z(v)) = \mu(X(v))$

2. Fit model on $\{z(s), X(s) : s \in G_2(u, d)\}$ to predict $\hat{\mu}(s), s \in G_1(u)$

3. Construct model-assisted estimator

$$\hat{\theta}_{MA,z} = \sum_{s \in G_1(u)} \frac{\hat{\mu}(s)}{1/(\delta_1 \delta_2)} + \sum_{s \in G_2(u,d)} \frac{z(s) - \hat{\mu}(s)}{1/(\delta_1 \delta_2 h_1 h_2)}$$
Properties of Model-Assisted Estimator

- Estimator $\hat{\theta}_{MA,z}$ is approximately design unbiased for large classes of models, with approximate design variance

$$\text{Var}(\hat{\theta}_{MA,z}) \approx \text{Var}(\hat{\theta}_{1z}(u)) + \frac{|D|^2}{n_2^2} \left(1 - \frac{1}{h_1 h_2}\right) E(S^2(u))$$

$$S^2(u) = \frac{1}{h_1 h_2 - 1} \sum_{d_1=1}^{h_1} \sum_{d_2=1}^{h_2} (t_{d_1 d_2}(u) - \bar{t}(u))^2$$

$$t_{d_1 d_2}(u) = \sum_{s \in G_2(u,d)} (z(s) - \hat{\mu}(s))$$
Applying Model-Assisted Estimation

• In typical survey context, many variables of interest instead of single $z(v)$

• Express estimator $\hat{\theta}_{MA,z}$ in the form

$$
\hat{\theta}_{MA,z} = \sum_{s \in G_2(u,d)} w(s) z(s)
$$

(automatic for linear estimators)

• Survey variables “related to” Phase I variables $X(v)$ will benefit from improved efficiency
4. Estimation for Forest Inventory Data

- Forest Service researchers are investigating predictive models for forest characteristics based on remote sensing data.

- Key variable in this survey: FOREST indicator

\[ I_{FOR}(v) \]

- Many other variables not recorded when

\[ I_{FOR}(v) = 0 \]
GAM Variables for FOREST

- (X,Y) coordinates (bivariate)
- ELEV90CU elevation
- TRASP90 aspect (transformed)
- SLP90CU slope
- MRLCOOB5 TM satellite band 5
- NDVI vegetation index (TM)
- NLDC7 vegetation classes (TM)
GAM Model for FOREST

• Model

\[ \mathbb{E}_\xi (I_{\text{FOR}}(v)) \equiv \mu_{\text{FOR}}(v) = g(m_1(x_1(v)) + \ldots + m_6(x_6(v)) + x_7(v)\beta) \]

with \( g(\cdot) \) logistic link and \( x_k(v) \) Phase I variables

• Fitted in S-Plus using \texttt{gam()} with \texttt{lo()} smoothers, to obtain prediction \( \hat{\mu}_{\text{FOR}}(s) \) for \( s \in G_1(u) \)
FOREST Model Components

-lo(Xs, Ys, span = 0.5)

ELEV90CU
s(ELEV90CU, df = 4)

1500 2000 2500 3000 3500

TRASP90
s(TRASP90, df = 4)

-150 -100 -50 0

SLP90CU
s(SLP90CU, df = 4)

0 20 40 60 80

MRLC00B5
s(MRLC00B5, df = 4)

0 50 100 150

NDVI
s(NDVI, df = 4)

-0.2 0.0 0.2 0.4 0.6 0.8

-10 -6 -2 0
Other Phase II Variables

- NVOLTOT: total wood volume (cuft/acre)
- BA: tree basal area (per acre)
- BIOMASS: total wood biomass (ton/acre)
- CRCOV: percent crown cover (%)
- QMDALL: quadratic mean diameter (in)
Modeling Other Variables

Approaches considered:

1. “Classical” model-assisted
2. Model-assisted with FOREST prediction as auxiliary variable
3. Model-assisted with FOREST prediction indicator
We didn’t do...

1. “Classical” model-assisted
   \[ E_\xi(z(\mathbf{v})) = m_1(x_1(\mathbf{v})) + \ldots + m_6(x_6(\mathbf{v})) + x_7(\mathbf{v})'\beta \]
   with \( m_1(\cdot), \ldots, m_6(\cdot) \) parametric or nonparametric

2. Model-assisted with FOREST prediction as auxiliary variable
   \[ E_\xi(z(\mathbf{v})) = m_1(x_1(\mathbf{v})) + \ldots + m_6(x_6(\mathbf{v})) + x_7(\mathbf{v})'\beta + \hat{\mu}_{\text{FOR}}(\mathbf{v})\gamma \]
Selected Method

- Construct indicator for FOREST prediction

\[ \hat{I}_{\text{FOR}}(v) = \begin{cases} 
1 & \text{if } \hat{\mu}_{\text{FOR}}(v) \geq \hat{\theta}_{2,\text{FOR}}(u, d)/|D| \\
0 & \text{otherwise} 
\end{cases} \]

and FOREST-interaction Phase I variables

\[ x_{k*F}(v) = \hat{I}_{\text{FOR}}(v)x_k(v) \]

- Working model

\[ E_{\xi}(z(v) \equiv \mu(X(v))) = X_{*F}(v)\beta \]

model predicts 0 whenever \( \hat{I}_{\text{FOR}}(v) = 0 \)
Selected Method (2)

- Estimator constructed as

\[ \hat{\theta}_{\text{MA},z} = \sum_{s \in G_1(u)} \frac{\hat{\mu}(s)}{1/(\delta_1 \delta_2)} + \sum_{s \in G_2(u,d)} \frac{z(s) - \hat{\mu}(s)}{1/(\delta_1 \delta_2 h_1 h_2)} \]

\[ = \sum_{s \in G_2(u,d)} w(s) z(s) \]

- Only approximately linear, due to presence of \( \hat{I}_{\text{FOR}}(v) \) on RHS
Calibration Properties

• Weights $w(s)$ calibrated for Phase I totals of auxiliary variables $x_{k*F}(v)$

$$
\sum_{s \in G_2(u,d)} w(s) x_{k*F}(s) = \sum_{s \in G_1(u)} \frac{x_{k*F}(s)}{1/(\delta_1 \delta_2)}
$$

and approximately for $\hat{\mu}_{FOR}(v)$ and $x_k(v)$

• Estimators for domains $U_h \subseteq D$ improved over simpler estimators, by incorporating forest/non-forest prediction locally
Comparing Estimators

1. **EXP**: expansion estimator $\hat{\theta}_{2Z}(u, d)$

2. **PS**: Model-assisted using NLDC7 categories only (post-stratified, current FIA method)

3. **REG**: Model-assisted with linear model using Phase I variables

4. **GAM/REGI**: Model-assisted with linear model using Phase I variables interacted with GAM forest/non-forest prediction
### Results

<table>
<thead>
<tr>
<th>Study Variable</th>
<th>Estimator</th>
<th>Estimated Mean</th>
<th>Estimated Standard Error</th>
<th>Est. Relative Efficiency of GAM/REGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOREST</td>
<td>EXP</td>
<td>0.51</td>
<td>0.02</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>0.54</td>
<td>0.01</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>0.54</td>
<td>0.01</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>GAM</td>
<td>0.54</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>NVOLTOT</td>
<td>EXP</td>
<td>845.81</td>
<td>44.07</td>
<td>1.79</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>877.41</td>
<td>39.10</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>877.67</td>
<td>35.35</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>853.85</td>
<td>32.98</td>
<td></td>
</tr>
<tr>
<td>BA</td>
<td>EXP</td>
<td>45.19</td>
<td>2.01</td>
<td>1.70</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>47.12</td>
<td>1.77</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>47.29</td>
<td>1.63</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>46.01</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td>BIOMASS</td>
<td>EXP</td>
<td>13.51</td>
<td>0.69</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>14.01</td>
<td>0.60</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>14.00</td>
<td>0.54</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>13.60</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>CRCOV</td>
<td>EXP</td>
<td>21.02</td>
<td>0.86</td>
<td>1.73</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>22.03</td>
<td>0.77</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>22.18</td>
<td>0.68</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>21.64</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>QMDALL</td>
<td>EXP</td>
<td>3.77</td>
<td>0.15</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>3.95</td>
<td>0.14</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>3.96</td>
<td>0.14</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>3.89</td>
<td>0.14</td>
<td></td>
</tr>
</tbody>
</table>
5. Variance Estimation for Systematic Samples

- No design-based variance estimator for systematic sampling

- In each phase, sample contains only one of all possible grids in population/phase

\[
\hat{\theta}_1(u) = \sum_{s \in G_1(u)} \frac{z(s)}{1/(\delta_1 \delta_2)}
\]

\[
\hat{\theta}_2(u, d) = \sum_{s \in G_2(u, d)} \frac{z(s)}{1/(\delta_1 \delta_2 h_1 h_2)}
\]
Simple Random Sampling Approximation

• Efficiency comparison relied on simple random sampling approximation for variance estimation

• For large numbers of possible samples and stationary populations, approximation is good on average

• Deviations can be significant for individual samples
Simple Random Sampling Approximation (2)

- Stationarity is reasonable for model-assisted estimators (model removes trend)
- But: only 25 possible samples in Phase II
- Approximate variance relies on asymptotic arguments
Alternative Approach to Assess Efficiency Gains

• Ignore Phase I variance component $\text{Var}(\hat{\theta}_{1z}(u))$: identical across all estimators

• Generate “synthetic” population and compute exact Phase II variance over 25 samples
  ▶ avoid both asymptotic and simple random sampling approximations
  ▶ depends on appropriateness of model
Synthetic Population

- Higher order polynomial models fitted to sample data
- Logistic model for FOREST
- Remaining variables fitted only on locations with $I_{FOR}(v) = 1$
- Predict variables for Phase I
- Sample means of variables approximately match those of the original population
<table>
<thead>
<tr>
<th>Simulated Variable</th>
<th>Estimator</th>
<th>Percent Bias of Variance of Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOREST</td>
<td>EXP</td>
<td>12.62</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>9.33</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>24.77</td>
</tr>
<tr>
<td></td>
<td>GAM</td>
<td>-31.01</td>
</tr>
<tr>
<td>NVOLTOT</td>
<td>EXP</td>
<td>-23.71</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>-32.11</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>-43.88</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>-52.71</td>
</tr>
<tr>
<td>BA</td>
<td>EXP</td>
<td>19.35</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>19.81</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>20.97</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>8.93</td>
</tr>
<tr>
<td>BIOMASS</td>
<td>EXP</td>
<td>17.48</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>34.31</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>-1.32</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>-14.61</td>
</tr>
<tr>
<td>CRCOV</td>
<td>EXP</td>
<td>-4.04</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>-20.93</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>-31.38</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>-44.01</td>
</tr>
<tr>
<td>QMDALL</td>
<td>EXP</td>
<td>5.92</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>23.70</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>50.32</td>
</tr>
<tr>
<td></td>
<td>REGI</td>
<td>13.27</td>
</tr>
</tbody>
</table>
## Relative Efficiency

<table>
<thead>
<tr>
<th>Simulated Variable</th>
<th>Estimator</th>
<th>Relative Efficiency of GAM/REGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOREST</td>
<td>EXP</td>
<td>4.51</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>1.92</td>
</tr>
<tr>
<td>NVOLTOT</td>
<td>EXP</td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>1.14</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>1.07</td>
</tr>
<tr>
<td>BA</td>
<td>EXP</td>
<td>2.02</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>1.19</td>
</tr>
<tr>
<td>BIOMASS</td>
<td>EXP</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>1.14</td>
</tr>
<tr>
<td>CRCOV</td>
<td>EXP</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>1.17</td>
</tr>
<tr>
<td>QMDALL</td>
<td>EXP</td>
<td>2.55</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>1.79</td>
</tr>
<tr>
<td></td>
<td>REG</td>
<td>1.20</td>
</tr>
</tbody>
</table>
6. Conclusions: model-assisted estimation

- Model-assisted framework provides flexible approach to incorporate sophisticated models in survey estimation
- Nonparametric models make it possible to capture complex patterns in forest resource data
- Use on-going spatial modeling efforts by forestry researchers to improve tabular data
6. Conclusions: systematic sampling

- Systematic sampling is popular in natural resource surveys, but does not allow for a design-based variance estimator.
- Synthetic population approach provided ad hoc solution in this case.
- On-going research: predict design-based variance under nonparametric model (Li, 2006).
• Almost-final version of this paper available at http://www.public.iastate.edu/~jopsomer/research.html

• Contact info: jopsomer@iastate.edu