Model-based sampling for remote regions

Ronald E. McRoberts
Northern Research Station, U.S. Forest Service
St. Paul, MN

Qi Chen
Department of Geography, University of Hawai‘i
Honolulu, Hawai‘i, USA

Motivation:
• For greenhouse gas inventories, the IPCC good practice guidance specifies that uncertainties in the form of confidence intervals should be reduced “to the degree practicable.”
• Traditional and familiar design-based inference requires probability samples that feature a randomization component.
• Probability sampling is extremely expensive and perhaps logistically not feasible for remote regions such as the Amazon, central Alaska, northern Canada, and Siberia.
• We need an alternative!

Design-based inference
• Assumes only one possible value for each population unit
• Relies on a probability sample for validity

Model-based inference
• Assumes a distribution of possible values for each population unit
• Relies on correct model specification for validity

Model-based estimators
• The variance depends on both the values of the predictor variable and the parameter estimates
• Can minimize the variance by selecting sample units with appropriate values of the predictor variables, given the parameter estimates (probability samples are not necessary)

Sequential sampling
• Sequential sampling entails sampling in stages
• The statistic of interest (e.g., confidence interval width) is evaluated following each stage, and if it satisfies an appropriate criterion, sampling terminates
• Avoids excessive sampling and thereby reduces costs
• Can be used with any form of sampling
Adaptive sampling

- Adaptive sampling designs are multi-stage designs that use data from previous stages to select optimal sampling units for the subsequent stages.
- When little is known about the model form or the parameter estimates, select an initial sample that spans the range of the predictor variable(s).
- Use multiple subsequent stages to refine the parameter estimates and thereby select samples that converge to the optimal sample.
- Can be readily used with sequential sampling.

An illustration

- Forest inventory data for Itasca Co., MN (7,583 km², 2,928 mi²)
  - 115 plots
  - Standing live tree stem volume
- Airborne laser scanning (ALS) data for the entire county
- Fit the model for stem volume versus ALS metrics
- Construct a simulated population by predicting volume for each 13-m x 13-m ALS cell
- Sample from this population

An illustration

- Construct 16 ALS strata based on \( h_q = \):
  - \([0-1), \ [1, 2), \ldots, \ [15, 16), \ [16-\text{max}]\)
- Select an initial sample consisting of one observation from each stratum.
- For subsequent samples, select the four strata that minimize the confidence interval width and select one observation for each of the four strata.
- Terminate sampling as soon as \( \frac{SE\hat{\mu}}{\hat{\mu}} \leq 0.05 \)

Selection of sample units within strata

- Initial sample: select population units within the 16 strata that are closest to reference location.
- Subsequent sample:
  - Option 1: select population units within the 4 strata that are closest to reference location.
  - Option 2: select population units within the first stratum randomly and for the other three strata that are closest to the first selected unit.
<table>
<thead>
<tr>
<th>Field crew base location</th>
<th>Adaptive sequential sampling</th>
<th>Simple random sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean distance (km)</td>
<td>Mean (m$^3$/ha)</td>
</tr>
<tr>
<td>1</td>
<td>48.00</td>
<td>34.38</td>
</tr>
<tr>
<td>2</td>
<td>34.02</td>
<td>34.75</td>
</tr>
<tr>
<td>3</td>
<td>34.02</td>
<td>34.75</td>
</tr>
<tr>
<td>4</td>
<td>34.02</td>
<td>34.75</td>
</tr>
<tr>
<td>5</td>
<td>34.02</td>
<td>34.75</td>
</tr>
<tr>
<td>6</td>
<td>34.02</td>
<td>34.75</td>
</tr>
<tr>
<td>7</td>
<td>34.02</td>
<td>34.75</td>
</tr>
<tr>
<td>8</td>
<td>34.02</td>
<td>34.75</td>
</tr>
<tr>
<td>9</td>
<td>34.02</td>
<td>34.75</td>
</tr>
</tbody>
</table>
## Field crew base location

<table>
<thead>
<tr>
<th></th>
<th>Adaptive sequential sampling</th>
<th>Simple random sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model-based</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>Mean distance (km)</td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>48.89</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>6.48</td>
</tr>
<tr>
<td>3</td>
<td>48</td>
<td>10.32</td>
</tr>
<tr>
<td>4</td>
<td>64</td>
<td>5.52</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
<td>3.80</td>
</tr>
<tr>
<td>6</td>
<td>32</td>
<td>3.21</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>10.32</td>
</tr>
<tr>
<td>8</td>
<td>52</td>
<td>7.50</td>
</tr>
<tr>
<td>9</td>
<td>24</td>
<td>7.07</td>
</tr>
</tbody>
</table>

## Conclusions

- Auxiliary information in the form of ALS metrics contributed to reducing uncertainty
- Model-based, adaptive sampling reduced uncertainties beyond simple random sampling
- Sequential sampling avoided excessive sampling costs
- For remote regions, model-based, adaptive, sequential sampling is a viable alternative if the model relationship is stable over the entire study area