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 values for each population unit
- Relies on a probability sample for validity

Model-based inference

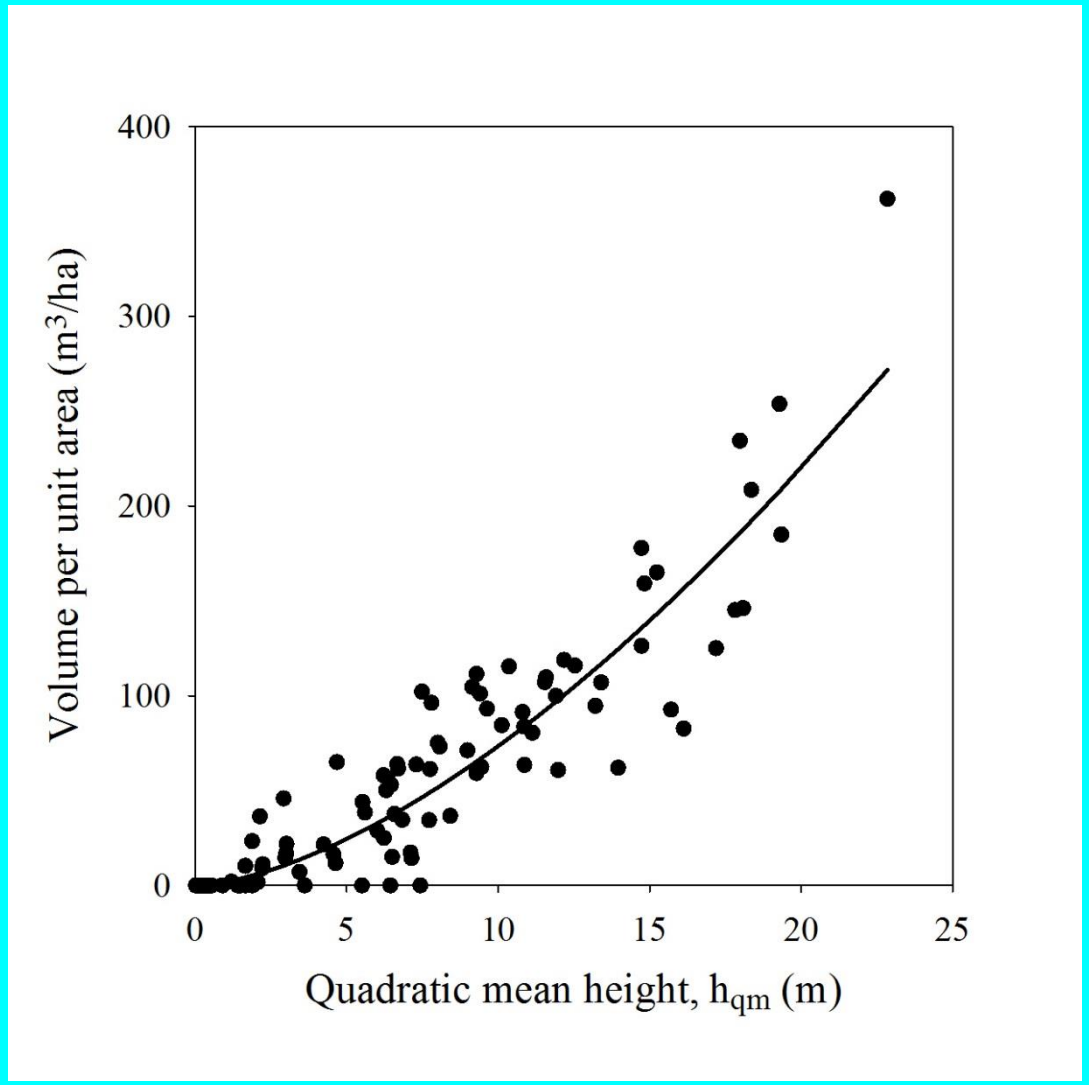
- Assumes a distribution of possible values for each population unit
- Relies on correct model specification for validity

Model-based estimators

$$\hat{\mu}_i = f\left(X_i ; \hat{\beta}\right)$$

$$= \hat{\beta}_1 \cdot X_i^{\hat{\beta}_2}$$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_i$$



Model-based estimators

$$\hat{\mathrm{var}}(\hat{\mu}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} z_i \cdot \hat{V}_{\hat{\beta}} \cdot z_j \qquad \text{where } \hat{V}_{\hat{\beta}} = \left(Z' W Z\right)^{-1}$$

$$\text{and} \quad z_{ij} = \frac{\partial f\left(x_i; \hat{\beta}\right)}{\partial \beta_j} = \begin{cases} j = 1 & x_i^{\hat{\beta}_2} \\ j = 2 & \hat{\beta}_1 \cdot \ln\left(x_i\right) \cdot x_i^{\hat{\beta}_2} \end{cases}$$

- 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

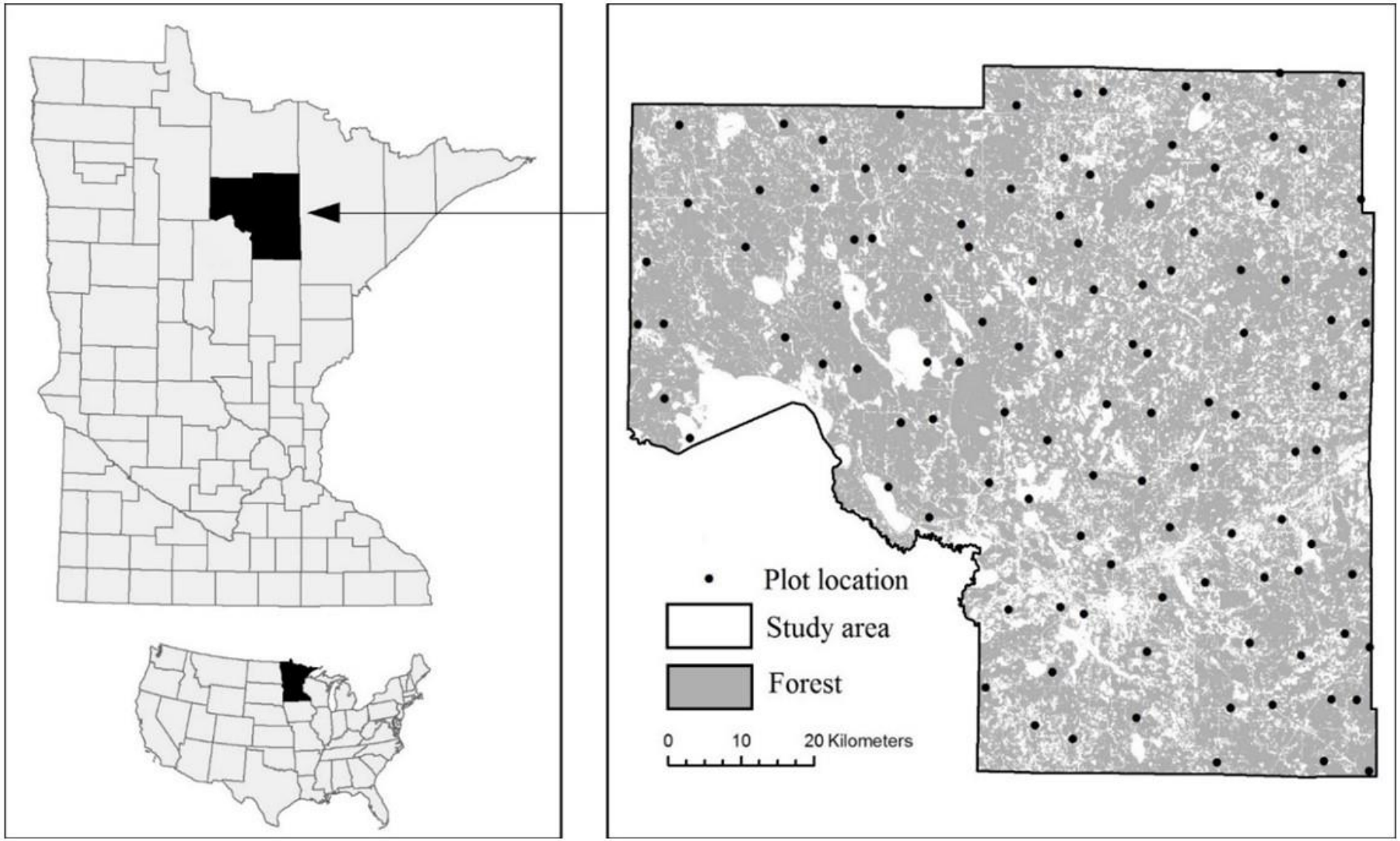
- Forest inventory data for Itasca Co., MN (7,583 km^2, 2,928 mi^2)
  - 115 plots
  - standing live tree stem volume

- 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


Plot location
Study area
Forest
0
10
20 Kilometers

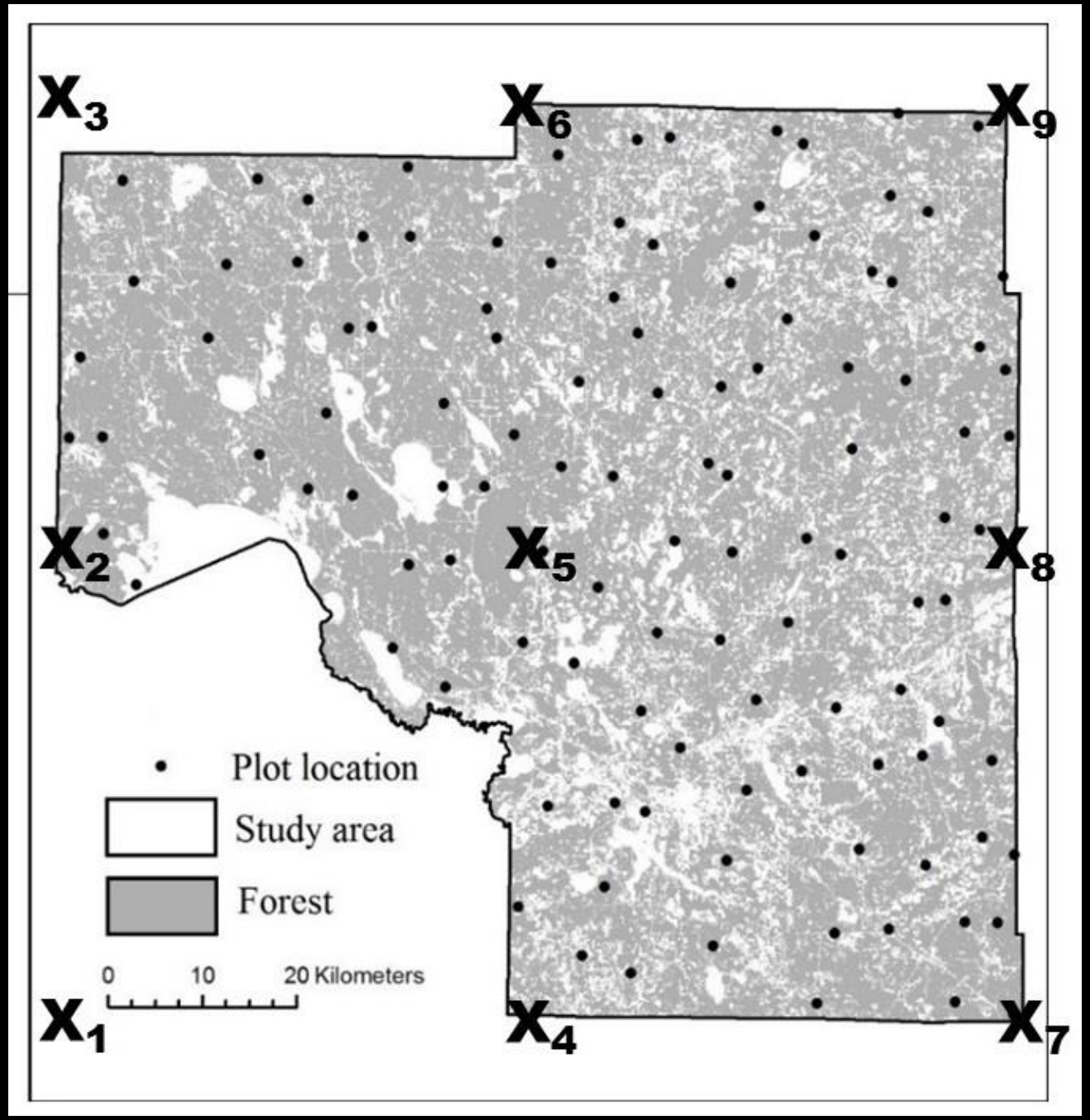
An illustration

- Construct 16 ALS strata based on h_{qm}:
  [0-1), [1, 2), …, [15, 16), [16-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 $\frac{SE\hat{\mu}}{\hat{\mu}} \leq 0.05$


X3
X6
X9
X2
X5
X8
X1
X4
X7
Plot location
Study area
Forest
0
10
20 Kilometers

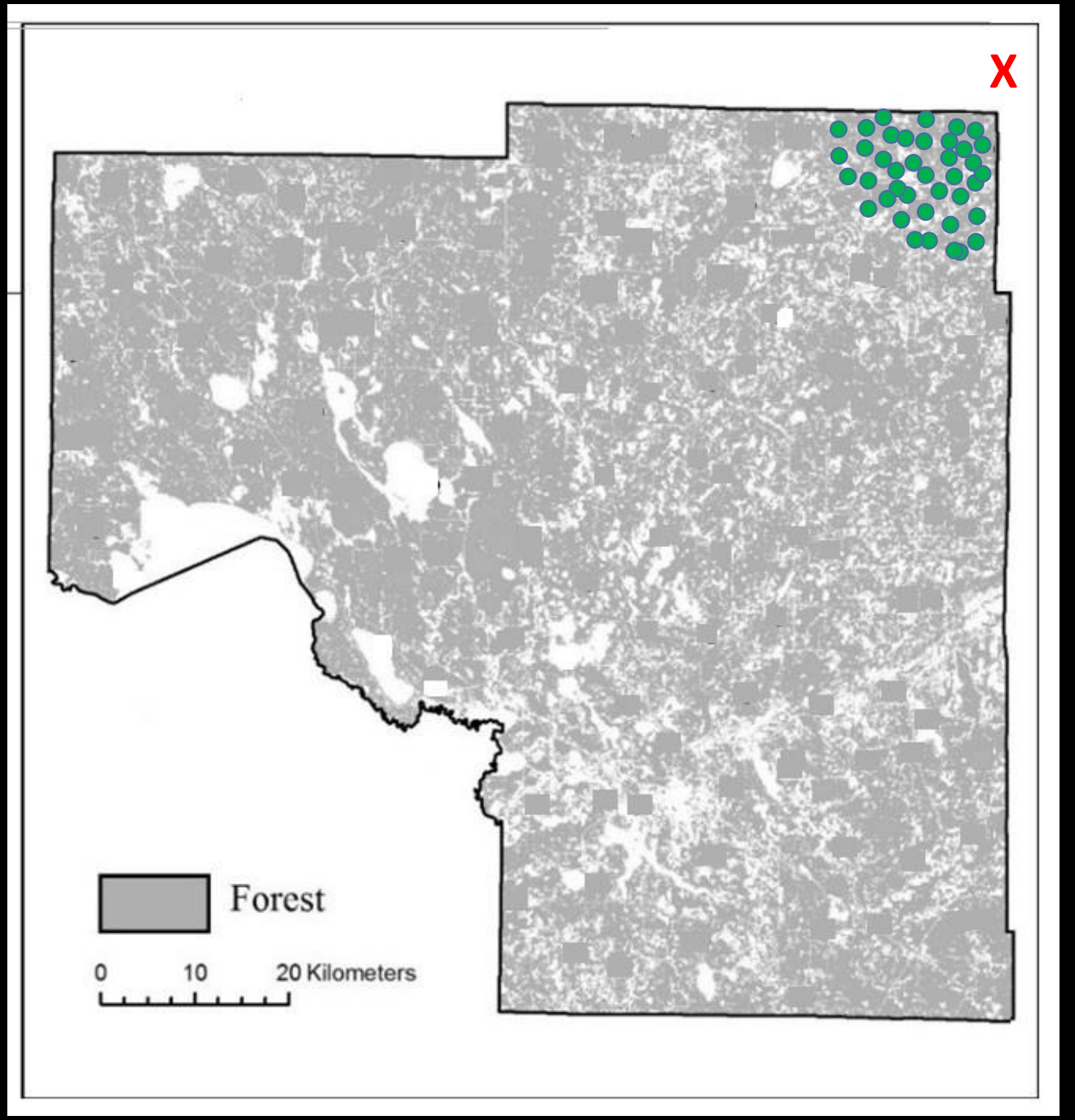
Selection of sample units within strata

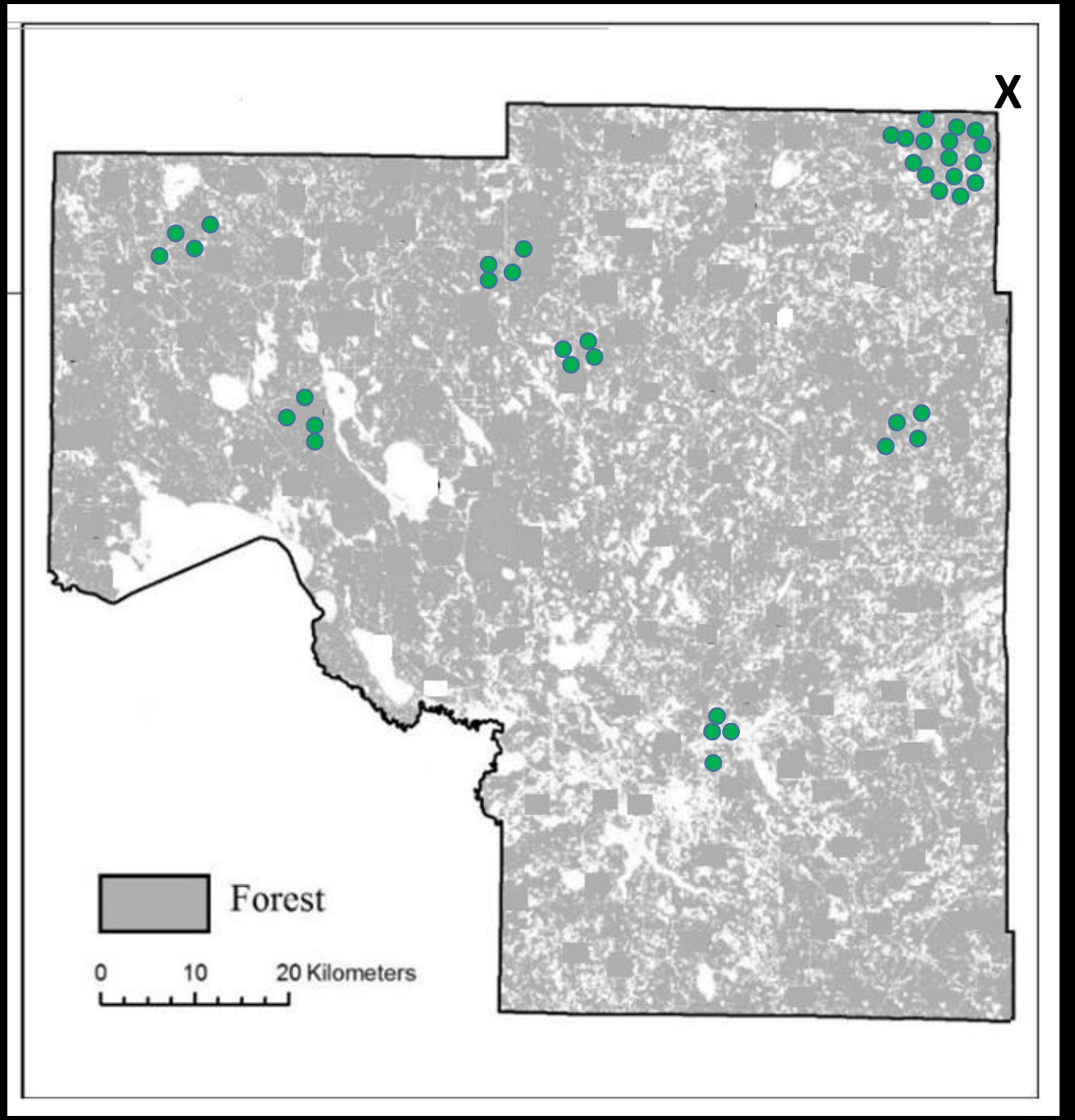
- Initial sample: select population units within the 16 strata that are closest to reference location

- 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


X
Forest
0
10
20 Kilometers


X
Forest
0
10
20 Kilometers

Field crew base location	Adaptive sequential sampling			
	n	Mean distance (km)	Mean (m^3/ha)	SE (m^3/ha)
1	60	48.89	33.38	1.55
2	68	2 6.48	35.45	1.69
3	48	40.32	33.38	1.65
4	64	25.52	36.47	1.77
5	60	13.80	35.62	1.70
6	32	3 3.21	38.82	1.82
7	44	40.32	35.04	1.75
8	52	37.50	32.03	1.57
9	24	57.07	34.08	1.67

Field crew base location	Adaptive sequential sampling				Simple random sampling		
					Model-based		
	n	Mean distance (km)	Mean (m^3/ha)	SE (m^3/ha)	Mean distance (km)	Mean (m^3/ha)	SE (m^3/ha)
1	60	48.89	33.38	1.55	89.03	34.81	2.09
2	68	2 6.48	35.45	1.69	67.81	34.85	1.98
3	48	40.32	33.38	1.65	79.99	34.59	2.36
4	64	25.52	36.47	1.77	64.53	34.90	2.01
5	60	13.80	35.62	1.70	37.41	34.84	2.09
6	32	3 3.21	38.82	1.82	55.55	34.67	2.87
7	44	40.32	35.04	1.75	77.92	34.69	2.45
8	52	37.50	32.03	1.57	55.92	34.76	2.24
9	24	57.07	34.08	1.67	72.68	34.81	3.34

Field crew base location	Adaptive sequential sampling				Simple random sampling				
					Model-based			SRS	
	n	Mean distance (km)	Mean (m^3/ha)	SE (m^3/ha)	Mean distance (km)	Mean (m^3/ha)	SE (m^3/ha)	Mean (m^3/ha)	SE (m^3/ha)
1	60	48.89	33.38	1.55	89.03	34.81	2.09	34.85	4.87
2	68	2 6.48	35.45	1.69	67.81	34.85	1.98	34.98	4.57
3	48	40.32	33.38	1.65	79.99	34.59	2.36	34.56	5.39
4	64	25.52	36.47	1.77	64.53	34.90	2.01	34.82	4.70
5	60	13.80	35.62	1.70	37.41	34.84	2.09	34.58	4.81
6	32	3 3.21	38.82	1.82	55.55	34.67	2.87	34.66	6.53
7	44	40.32	35.04	1.75	77.92	34.69	2.45	34.50	5.59
8	52	37.50	32.03	1.57	55.92	34.76	2.24	35.00	5.19
9	24	57.07	34.08	1.67	72.68	34.81	3.34	34.70	7.54

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