Spatial Linear Models: Better Lemon Squeezers for Predicting Potential Forest Productivity

Temesgen[+] H. and J. M. Ver Hoef[++]

[+] Dept. of Forest Engineering, Resources and Management

Oregon State University, Corvallis, Oregon

[++] Alaska Fisheries Science Center, NOAA Fisheries,

Fairbanks, Alaska

Feb. 27, 2015

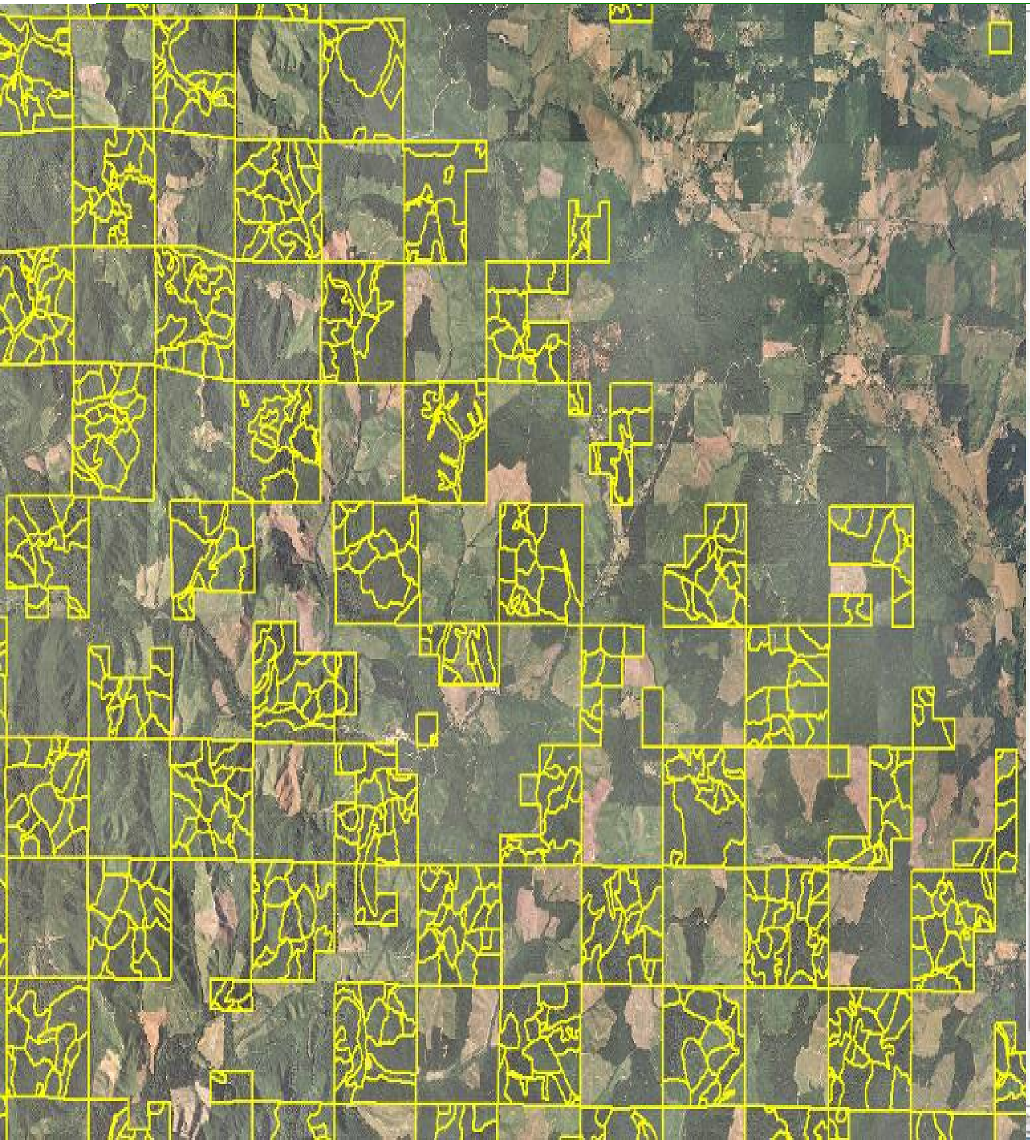
Outline

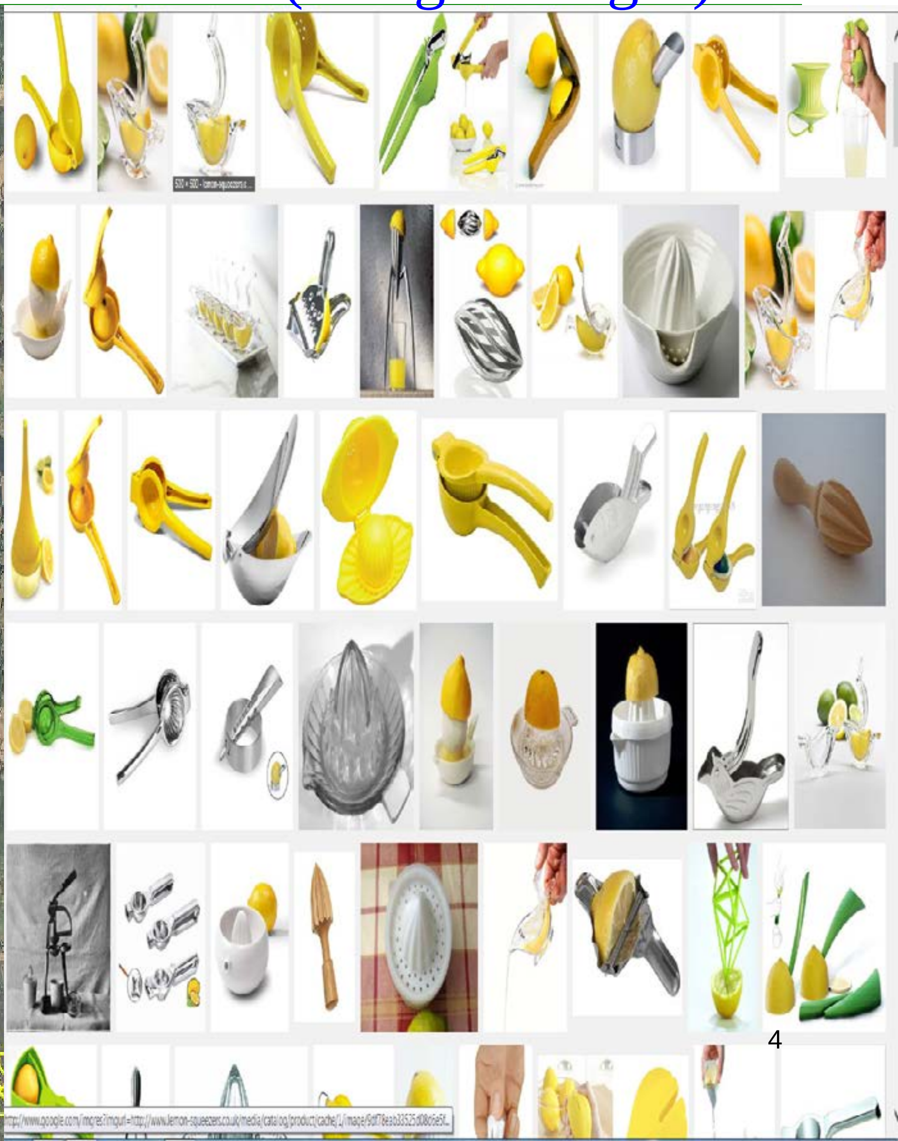
1. Background

1.1. Imputing potential productivity for mapping and for estimating totals

- **Potential mean annual increment (PMAI):** cubic meter volume per hectare per year (**m^3/ha/yr**) at time of culmination
  - indicates the productivity of a forest stand
  - aids in forest planning and assessments
- *1 to 5%* of the land base is sampled for ground variables (Y set, **response**).
- Aerial photos, climate databases, lidar and remote sensing provide complete census of selected auxiliary variables (X set, **covariates**).
- Inference for natural resource planning is improved by "**Populating**" the forested landscape with potential productivity and biomass estimates. **Nearest Neighbor (NN)** methods are commonly used to populate the forest land base.

Lemon Squeezers for Predicting Potential Forest Productivity (Google Images)





Background

Latta, G., Temesgen, H., and T. Barrett. **2009.** Mapping and imputing potential productivity of Pacific Northwest Forests using climate variables. CJFR. 39: 1197-1207.

[Used the Spatial Autoregressive Model (SAR) & localized prediction by using coordinates & selecting neighbors in a circular area (window) around the target unit. Compared MLR, thin plate splines, Most Similar Neighbor Method (MSN), and SAR]

Temesgen, H., G. Latta, and T.M. Barrett. **2011**. Imputing potential productivity of Pacific Northwest forests over space and time. Presented at the International Statistical Institute, 58th Congress, Aug. 21-26, Dublin, Ireland.

[Compared SAR, Spatial Lag Models, Spatial Durbin Model, Random Forest, and MSN]

*Ver Hoef J. and H. Temesgen. **2013**. A comparison of the spatial linear model to nearest neighbor (k-NN) methods for forestry applications. PLOS ONE. (3):1-11.

[Compared SLM and k-NN theoretically and through simulations (normal, binary, & count simulated data) and using forestry data – PMAI and dry biomass]

*Temesgen, H. and J. Ver Hoef. **2014.** Evaluation of the spatial linear model, Random Forest, and gradient nearest neighbor methods for imputing potential productivity and biomass of the Pacific Northwest forests. Forestry: 6: 1-12 [Evaluated the performance of SLM, GNN, RF, and k-NN, and simulated (normal, Poisson, and lognormal distribution) data]

1.2. Methods Used to Predict Potential Forest Productivity

Imputation is replacing a missing or non-sampled item/unit with another item/unit that has similar characteristics.

A) Nearest Neighbor methods

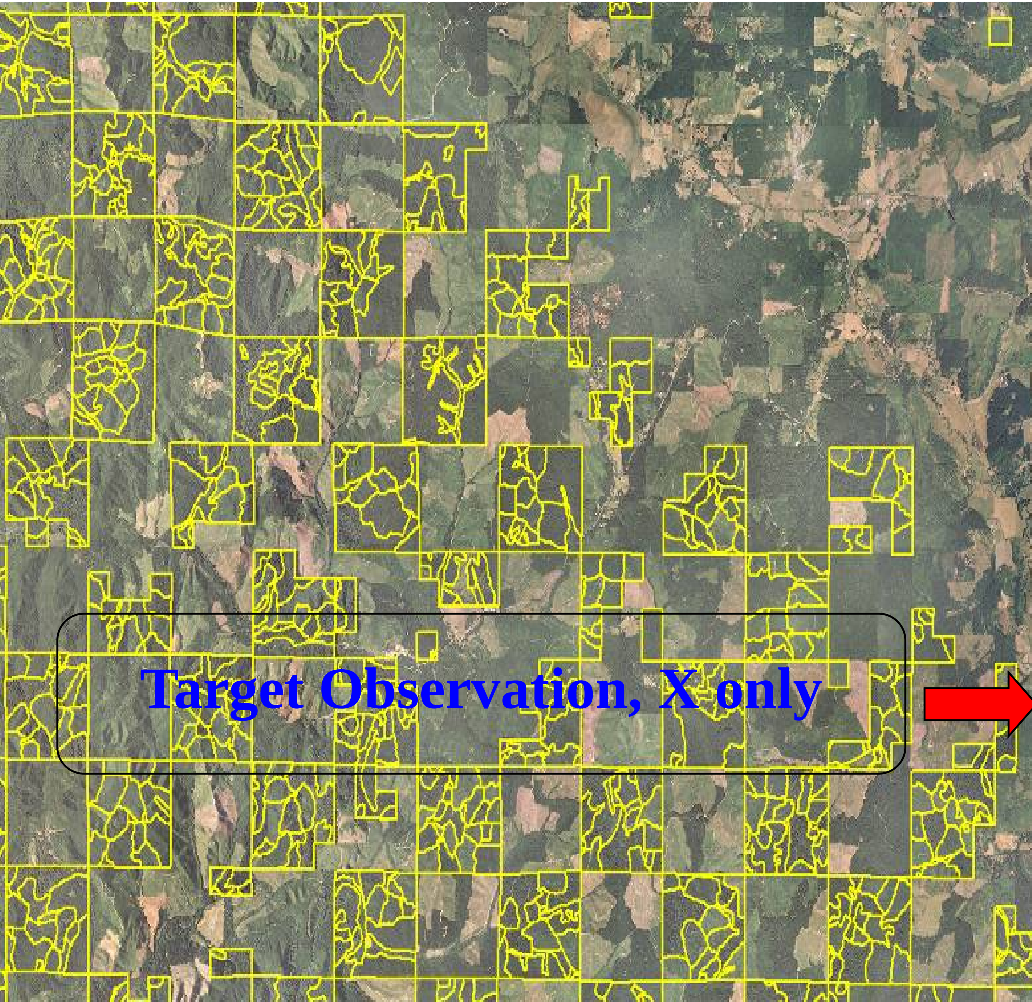
- **A1.** Most Similar Neighbor (MSN, Moeur and Stage 1995)
- **A2.** k-Similar Neighbors (k-NN)
- **A3.** Gradient Nearest Neighbor (GNN) (Ohmann and Gregory)
- **A4.** Random Forest (RF)

B) Likelihood-based approaches

- **B1.** Linear Regression (LM)
- **B2.** Spatial Linear Model (SLM)

A1-A3 NN Methods

NN imputation steps in general (Temesgen et al. 2003)



Sample Data, X and Y
Calculate Variable-Space
Distance using X's

Select one
or more neighbors
that have similar X set values
(Small distance metric)

Use Y set values
(or averages)
from selected
reference observation(s)
as estimates for the
target observation

Pros and cons of NN methods

Imputation methods:

- predictions are within the bounds of biological reality because they are observed in the sample
- reuse existing samples, and distribution free
- maintain logical relationships /dependence structure among response variables (multivariate methods)

E.g., predictive mapping and tree-lists, etc.

However, imputation methods are NOT:

- necessarily unbiased

- for Y: match is based on X variables, not Y

- necessarily consistent

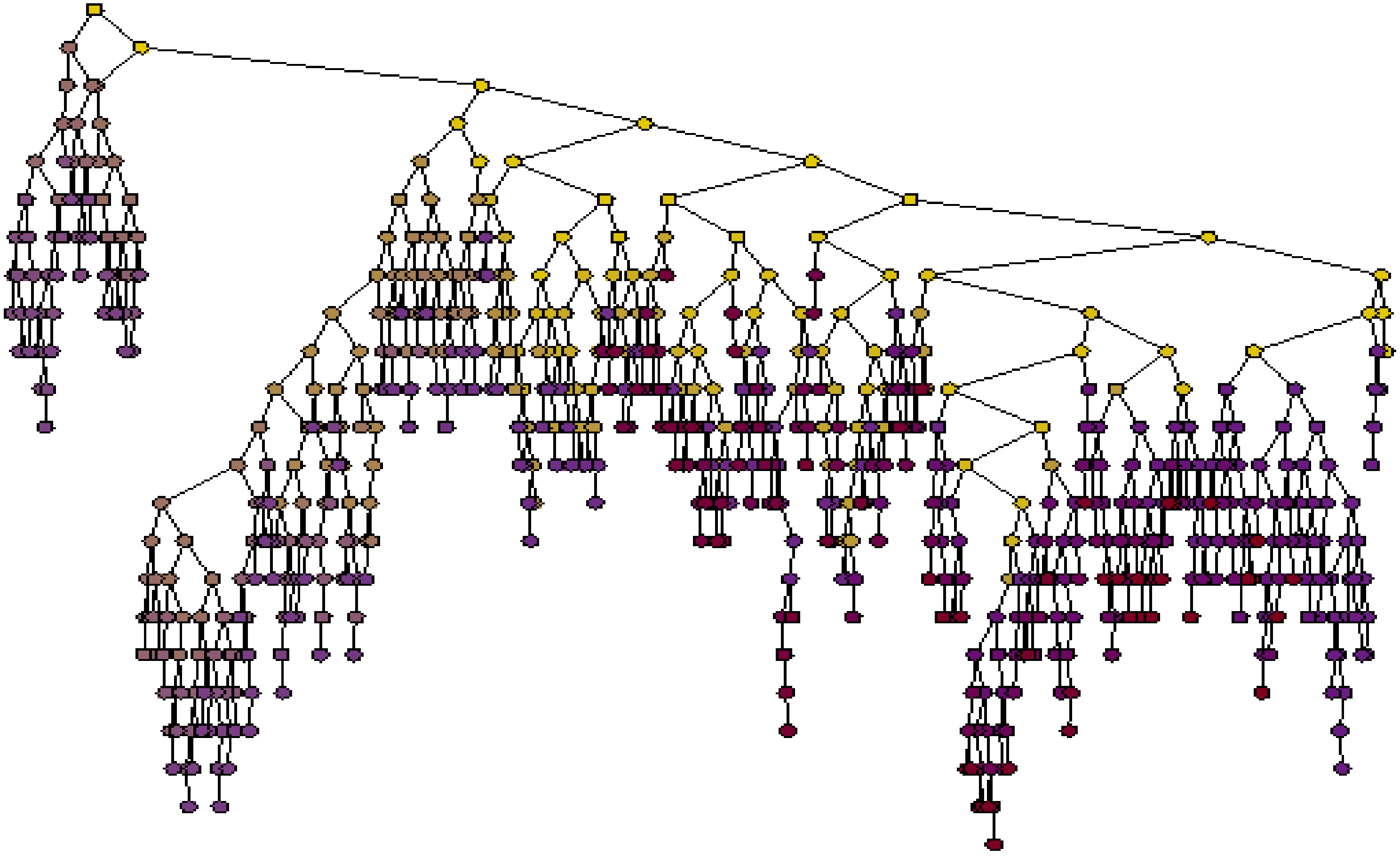
- as sample size increases more likely to find a close match

NN methods lack a good measure of uncertainty. Often global RMSE, from cross-validation, is used for point-wise standard error.

A4. Random Forest

- an ensemble learning method for classification, regression, and other tasks

- *constructs a multitude of decision trees (based on training data) and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees*



A4. Pros of RF (Ensemble) methods

- Very useful for data exploration

- Distribution free

- Work best for classification problems, when they are trained to assign a data point to a class--preferably one of only a few possible classes.

- All variables are assumed to interact (*inefficient if there are variables that have no or weak interactions).*

A4. Cons of RF (Ensemble)

- Can be extremely sensitive to small perturbations in the data: a slight change can result in a drastically different tree (outcomes)

- Lacks a probabilistic framework - unknown confidence intervals, posterior distributions etc.

- Have problems for out-of-sample prediction (*non-smooth*). Can easily overfit. This can be negated by validation methods and pruning, but this is a grey area.

- Poor resolution on data with complex relationships among the variables. At each node, only two possibilities exist. Hence there are some variable relationships that *Decision Trees* just can't learn.

1.2A. Nearest neighbor (NN) Methods: lack a good measure of uncertainty

- Cross validation is often used to compute prediction standard errors.

$$\hat{\xi} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$\hat{y}_i$ is the cross-validadtion prediction of y_i for i=1...,n sample values.

Assuming prediction errors are normally distributed, 90% prediction intervals are formed as $y_j \pm 1.64\xi$ for j=n+1,...,n+m out of sample.

For estimating standard error of a total: $se(\hat{T}) = \hat{\xi}\sqrt{(n+m)m}$

- **Note that the ζ values are constant for all j.**

1.2. Methods Used to Predict Potential Forest Productivity (cont'd)

1.2B. Likelihood-based approaches

Notation for spatial data (after Ver Hoef and Temegsen 2013)

- Let the population of response values be partitioned into those that are:
  - observed $\boldsymbol{y_O} = \{y_i; i = 1, \cdots, n\}$ and
  - unobserved $\boldsymbol{y_U} = \{y_i; i = n+1, \cdots, n+m\}$, and
  - $\boldsymbol{y} = (\boldsymbol{y}'_{\boldsymbol{O}}, \boldsymbol{y}'_{\boldsymbol{U}})'$.

- Let the index set for the:
  - observed data be $\mathbb{O} = 1, \cdots, n$ and
  - unobserved data be $\mathbb{U} = n+1, \cdots, n+m$.

1.2. Methods Used to Predict Potential Forest Productivity (cont'd)

- We consider two main goals:

  1) point prediction of y_j for $j \in \mathbb{U}$, and

  2) block prediction of the total or average $T = \sum_{i=1}^{n+m} b_i y_i = \boldsymbol{b}'\boldsymbol{y}$,

  where b_i are the weights that define the block objective;

- e.g., if $\{b_i = 1; i = 1, \cdots, (n+m)\}$, then T is a population total, and
  if $\{b_i = 1/(n+m); i = 1, \cdots, (n+m)\}$, then T is a population average.

For all response values, there are covariates contained in a design matrix **X** and the spatial coordinates are contained in matrix S.

1.2. Methods Used to Predict Potential Forest Productivity (cont'd)

- To meet our two goals, we define the linear predictor,

$$\hat{y}_j = \sum_{i \in \mathbb{O}} \lambda_{i,j} y_i = \lambda_j{}' \boldsymbol{y_o}\,, \tag{1}$$

where $j \in \mathbb{U}$.

- A linear block predictor is

$$\hat{T} = \sum_{i \in \mathbb{O}} b_i y_i + \sum_{j \in \mathbb{U}} b_j \hat{y}_j = \omega' \boldsymbol{y_o} \tag{2}$$

1.2. Methods Used to Predict Potential Forest Productivity (cont'd)

❖ Both the SLM and NN methods use distance in various ways.

❖ A general definition: let **A** be a matrix with coordinates in the columns and the *i*th row denoted as a_i. A general distance formula between the *i*th and *j*th rows of **A** is,

$$d(i,j;\mathbf{A},\mathbf{W}) \equiv \sqrt{(\mathbf{a_i}-\mathbf{a_j})'\mathbf{W}(\mathbf{a_i}-\mathbf{a_j})} \qquad (3)$$

where **W** is a weighting matrix.

❖Let $\boldsymbol{x}_i'$ and $\boldsymbol{x}_j'$ be the ith and jth rows of X, respectively. Then a "variable-space distance'' between the ith and jth sites can be computed as d(i,j;**X**,**W**).

❖ Several types of distances are possible (e.g., Mahalanobis, proximity matrix)

1.2. Methods Used to Predict Potential Forest Productivity (cont'd)

- Let **D** be a distance matrix with *i,j*th element $d(i, j; \mathbf{X}, \mathbf{W})$, which can be partitioned as:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{\boldsymbol{O},\boldsymbol{O}} & \mathbf{D}_{\boldsymbol{O},U} \\ \mathbf{D}'_{\boldsymbol{O},U} & \mathbf{D}_{U,U} \end{bmatrix}$$

- Let $\mathbf{d}_j$ be the *j*th column of **D**, $j \in \mathbb{U}$, contained in $\mathbf{D}'_{\boldsymbol{O},\boldsymbol{U}}$;

  i.e., $\mathbf{d}_j = \{\mathbf{D}[i,j]; i = 1,2,\cdots,n; j \in \mathbb{U}\}$.

- If *i* is the index for min $(\mathbf{d}_j)$, then for a first-order nearest neighbor, $\lambda_{i,j} = 1$ in $\hat{y}_j = \sum_{i \in \mathbb{O}} \lambda_{i,j} y_i = \lambda_j{}' \boldsymbol{y_O}$ and all other $\lambda_{l,j} = 0; l \neq i$. This essentially assigns the value of y_i to $\hat{y}_i$ for the ith site that is closest to the jth site in "variable space".

- Let *k* be the index of the *k* nearest sites (smalles values), then $\lambda_{i,j} = 1/k$; takes the average of $\{y_i\}$ from the <u>*k* nearest neighbors in variable space</u>.

1.2B. Spatial Linear Model (SLM)

- Assume only the linear model: $y = X\beta + \varepsilon$

Where **X** is a fixed covariates, β is a random vector of parameters, and ε is a random vector with *var*(ε)=V for some unknown spatial multivarite distribution.

- Unlike NN methods, SLM is a **spatial stochastic model** that allow optimization with respect to bias and square error.

- Let **V** be partitioned as: $\mathbf{var}(\mathbf{y}) = \mathbf{V} = \begin{bmatrix} \mathbf{V}_{\boldsymbol{O},\boldsymbol{O}} & \mathbf{V}_{\boldsymbol{O},\boldsymbol{U}} \\ \mathbf{V}'_{\boldsymbol{O},\boldsymbol{U}} & \boldsymbol{V}_{\boldsymbol{U},\boldsymbol{U}} \end{bmatrix}$

- Let $\mathbf{v}_j$ be the j^{th} column of **V**, $j \in \mathbb{U}$, contained in $\mathbf{V}_{\boldsymbol{O},\boldsymbol{U}}$;
  i.e., $\mathbf{v}_j = \{\mathbf{V}[i,j]; i = 1,2,\cdots,n; j \in \mathbb{U}\}$.

1.2B. Spatial Linear Model (continued)

- The best linear unbiased predictor (BLUP) that minimizes squared-error loss for $\hat{y}_j = \boldsymbol{\lambda}'_j \boldsymbol{y_O}$ is (Cressie 1993 pgs 151-155):

$$\boldsymbol{\lambda}_j' = \left[\mathbf{v}_j + \mathbf{X}_O \mathbf{C}\left(\mathbf{x}_j - \mathbf{X}_O' \mathbf{V}_{O,O}^{-1} \mathbf{v}_j\right)\right]' \mathbf{V}_{O,O}^{-1} \quad (8)$$

where $C = \left(\mathbf{X}_O' \mathbf{V}_{O,O}^{-1} \mathbf{X}_O\right)^{-1}$ with prediction variance of

$$\mathrm{var}\left(\hat{Y}_j - Y_j\right) = \mathbf{V}[j,j] - 2\boldsymbol{\lambda}_j' \mathbf{v}_j + \boldsymbol{\lambda}_j' \mathbf{V}_{O,O} \boldsymbol{\lambda}_j \quad (9)$$

- Notice that β is unknown; the only assumption is the linear model and a known spatial covariance matrix.

- The covariance matrix can be estimated with REML.

2. Comparison of Spatial Linear Models to Nearest Neighbor Methods

Objectives

1. Compare the predictive performances of selected **NN methods** to **SLM** for imputing PMAI for mapping (point) and for estimating totals (block predictions).

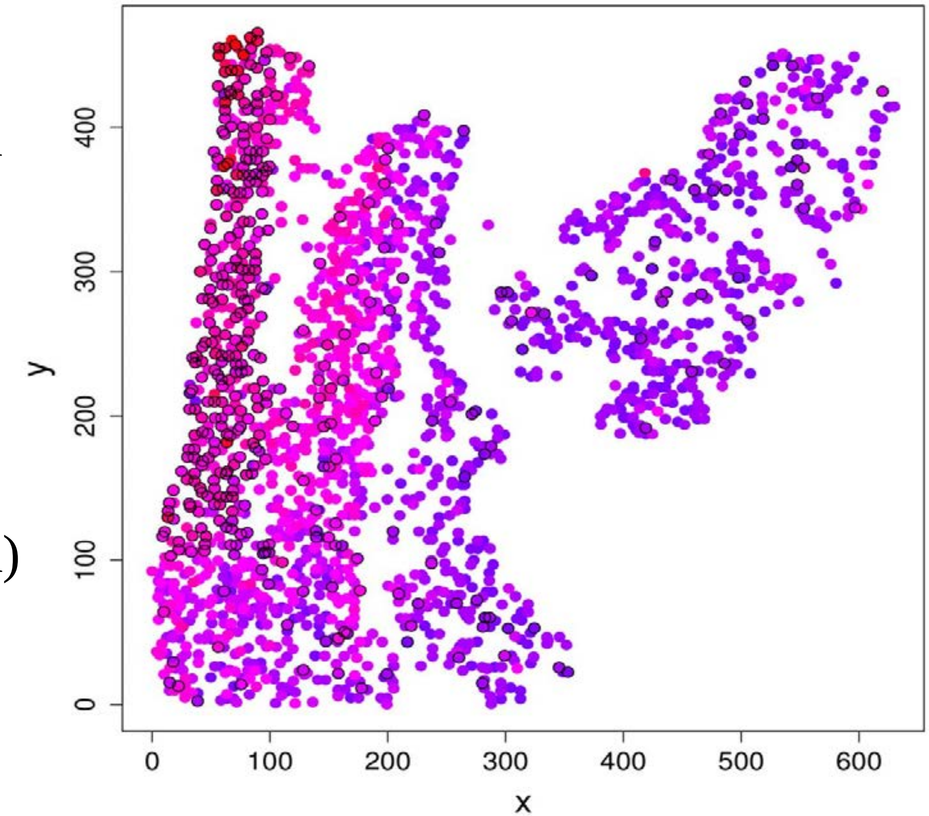
2. Examine the performance of selected NN and SLM under spatially unbalanced sampling.

2.1. Forest productivity data
USFS's National Forest Inventory and Analysis (FIA) Plots

FIA plots in Oregon and Washington (n= 3356)

Response: Maximum potential mean annual increment (m^3/ha/year) (0.2,23.8)

Covariates: 1) Temperature (^{0}c)
2) Precipitation(cm)
3) Elevation (m)
4) Climate Moisture Index (cm)
5) An indicator variable based on western hemlock
6) Ownership



Spatial locations of PMAI variable. The redder shades indicate higher values, and the bluer shades indicate lower values.

2.2. Resampling experiments (500 replications, 885/2471 split) Prediction methods:

MSN1: uses weighted Mahalanobis distance with k = 1. $d_{ij}^{2} = (X_i - X_j)' W (X_i - X_j)$

MSN5: uses weighted Mahalanobis distance with k = 5.

BestNN: uses both Mahalanobis and weighted Mahalanobis distance, and tries k = 1, 2, …, 30, and then chooses the distance matrix and k with the smallest cross-validation RMSPE from the observed data.

RF1: uses proximity matrix with k=1

RF5: uses proximity matrix with k=5

GNN1: uses canonical correspondence analysis on projected ordination of X with k=1

GNN5: uses canonical correspondence analysis on projected ordination of X with k=5

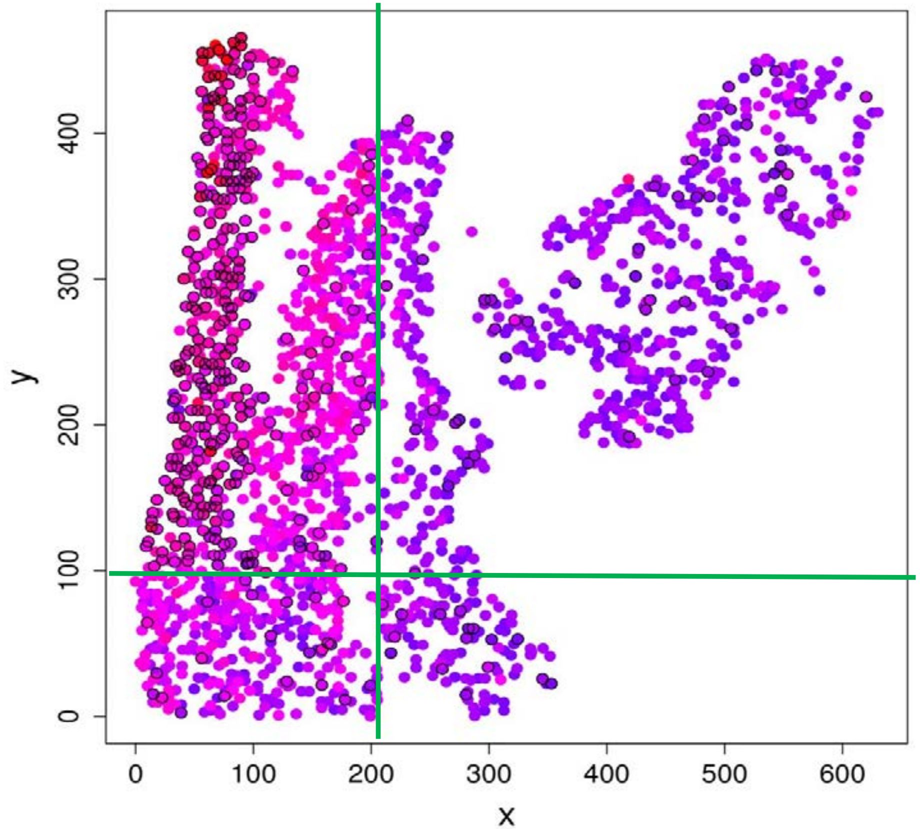
SLM: a spatial linear model using the same covariates as all NN methods as main effects only, with exponential covariance model estimated by REML and FPBK prediction and variance equations (Ver Hoef 2000, 2002).

LM: multiple regression like SLM but assuming all random errors are independent.

Unbalanced sampling (Oregon only)

We preferentially sampled geographically by dividing up the study area into four parts.

From each unbalanced spatial sample, the remaining locations were predicted.



One draw from the unbalanced spatial sample is shown with black circles around the sampled locations.

2.3. Performance measures

- **RMSPE: Root-mean-squared-prediction error**

$$RMSPE_y = \sqrt{\frac{1}{mR}\sum_{r=1}^{R}\sum_{j=1}^{m}\left(\hat{y}_{j|r} - y_{j|r}\right)^2}$$

for point-wise predictions;

$$RMSPE_T = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{T}_r - T_r\right)^2}$$

for total prediction

where R=no. of resamplings, m=no. of point predictions (2471) per replication r

- **SRB: signed relative bias** $$SRB_k = sign(\tau_k)\sqrt{\frac{\tau_k^2}{MSPE_k - \tau_k^2}}$$

$$\tau_y = \frac{1}{mR}\sum_{r=1}^{R}\sum_{j=1}^{m}\left(\hat{y}_{j|r} - y_{j|r}\right)$$

$$\tau_T = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{T}_r - T_r\right)$$

sign(τ_k) is the sign (positive or negative) of τ_k, and k=y for a point-wise performance measure or k=T for a total performance measure.

2.3. Performance measures over 500 replications

- **PIC90: 90% prediction interval coverage**. For point-wise predictions,

$$PIC90_y = \frac{1}{mR}\sum_{r=1}^{R}\sum_{j=1}^{m} I\left(\left(\hat{y}_{j|r} - 1.645 \cdot \widehat{se}\left(\hat{y}_{j|r}\right)\right) < y_{j|r} \ \& \ y_{j|r} < \left(\hat{y}_{j|r} + 1.645 \cdot \widehat{se}\left(\hat{y}_{j|r}\right)\right)\right)$$

where $\widehat{se}\left(\hat{y}_{j|r}\right)$ is the estimated standard error of $\hat{y}_{j|r}$ for NN methods, and from the square root of variance of ε for the SLM, with covariance parameters estimated by REML.

$$PIC90_T = \frac{1}{R}\sum_{r=1}^{R} I\left(\left(\hat{T}_r - 1.645 \cdot \widehat{se}\left(\hat{T}_r\right)\right) < \text{T\_r} \ \& \text{T\_r} < \left(\hat{T}_r + 1.645 \cdot \widehat{se}\left(\hat{T}_r\right)\right)\right)$$

where $\widehat{se}\left(\hat{T}_r\right)$ is the estimated standard error of $\hat{T}_r$.

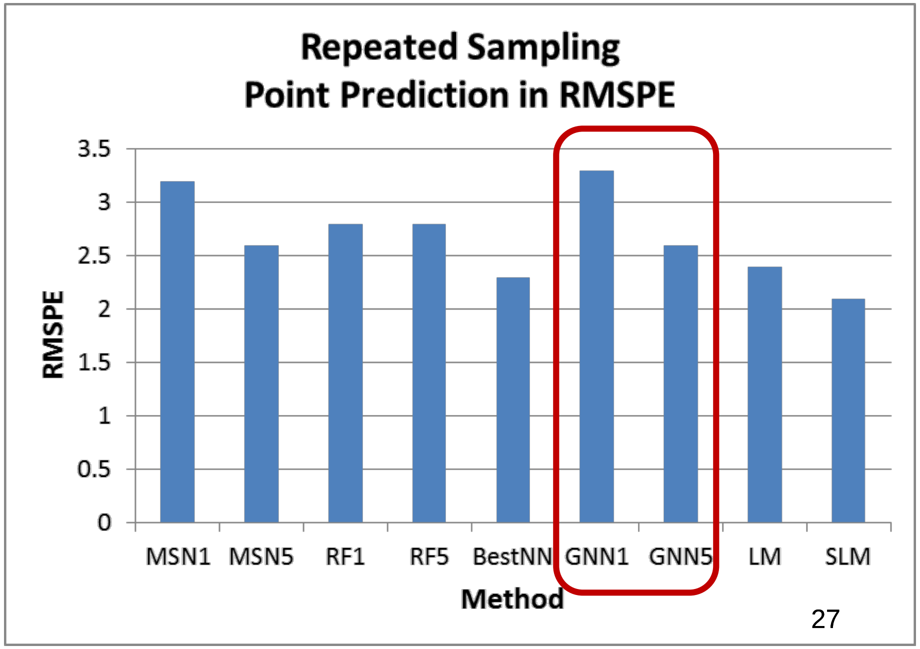
$PIC90_y$ should be near 0.90 if prediction intervals are properly estimated.

3. Results

3. Imputing forest productivity for mapping
Results over 500 resampling

- Point prediction appears unbiased for all methods
- SLM reduced the RMSPE by:
  - 23.4 and 32.8% when compared to RF1 and GNN1.
  - 23.4 and 15.4% when compared to RF5 and GNN5

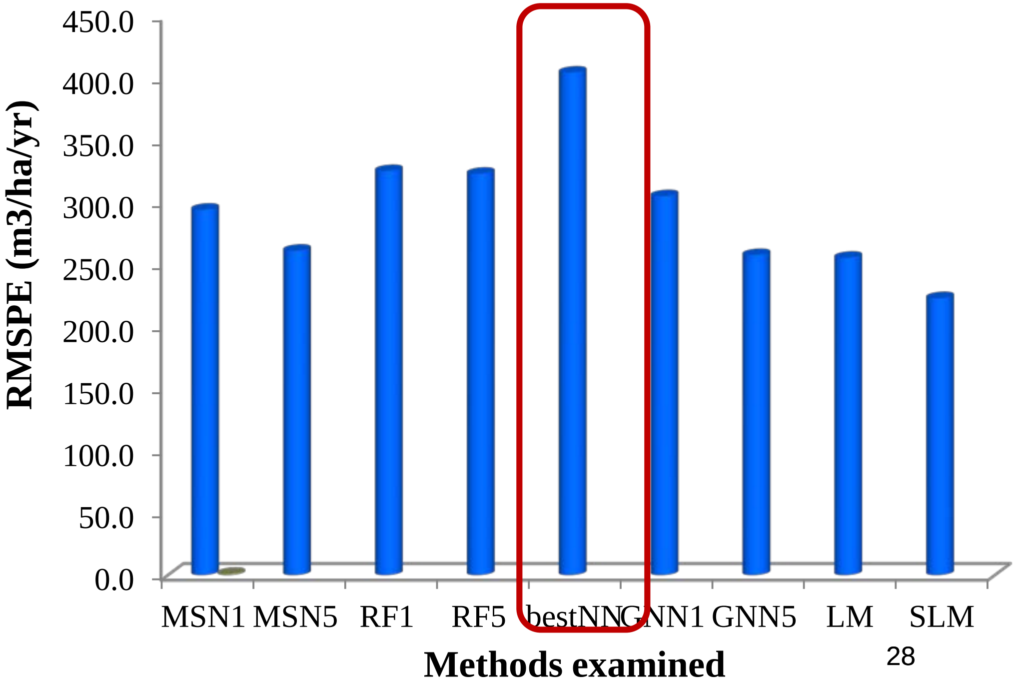
Point Prediction – Repeated Sampling			
Method	RMSPE (m^3/ha/yr)	SRB	PICO90
MSN1	3.2	0.003	0.89
MSN5	2.6	-0.001	0.90
RF1	2.8	0.023	0.85
RF5	2.8	0.024	0.85
BestNN	2.3	0.048	0.90
GNN1	3.3	0.004	0.89
GNN5	2.6	0.004	0.90
LM	2.4	-0.003	0.90
SLM	2.1	-0.002	0.90



3. Imputing forest productivity for population totals Results over 500 resampling

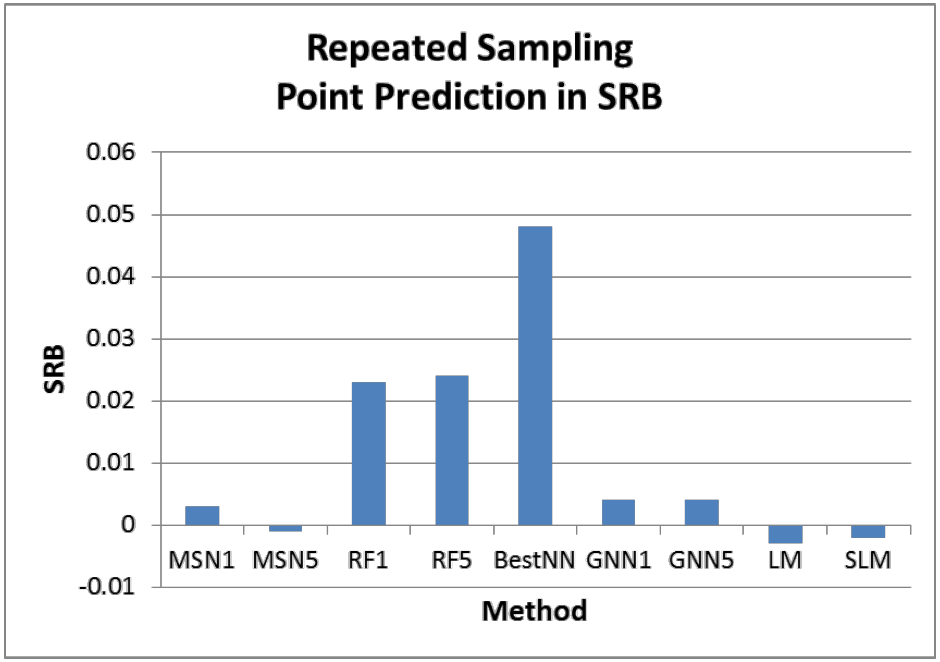
- Total prediction appears biased for RF, BestNN, and GNN.
- SLM reduced the RMSPE by:
  - 31.4 and 26.9% when compared to RF1 and GNN1
  - 31.0 and 13.1% when compared to RF5 and GNN5

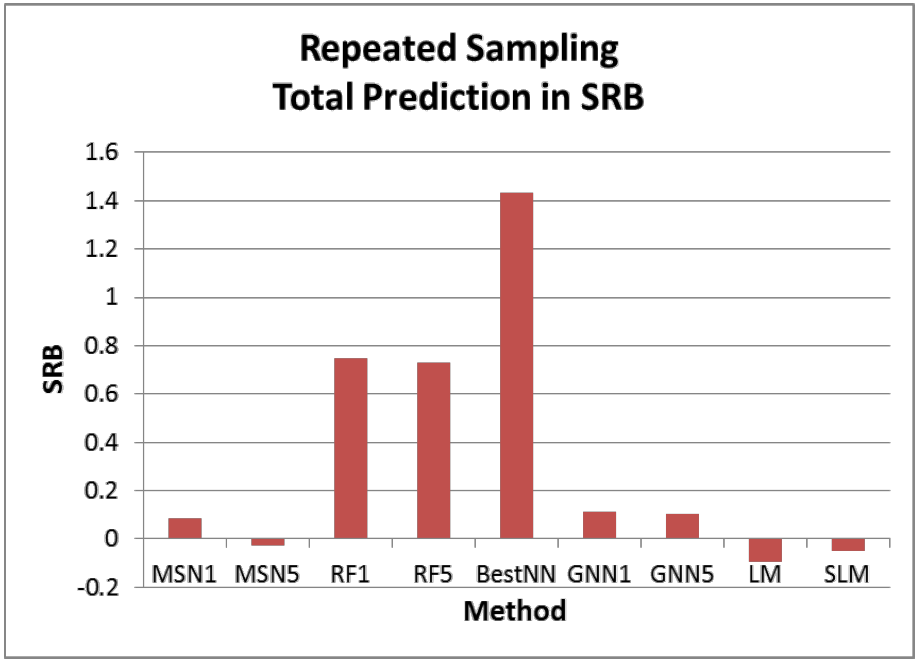
Total Prediction – Repeated Sampling			
Method	**RMSPE**	**SRB**	**PICO90**
MSN1	294.3	0.086	0.93
MSN5	261.3	-0.029	0.91
RF1	325.4	0.746	0.74
RF5	323.2	0.732	0.73
BestNN	404.6	1.434	0.57
GNN1	305.2	0.113	0.93
GNN5	257.8	0.104	0.91
LM	255.6	-0.097	0.90
SLM	223.1	-0.05	0.89



Results over 500 replications of Repeated Sampling Point and Total in SRB

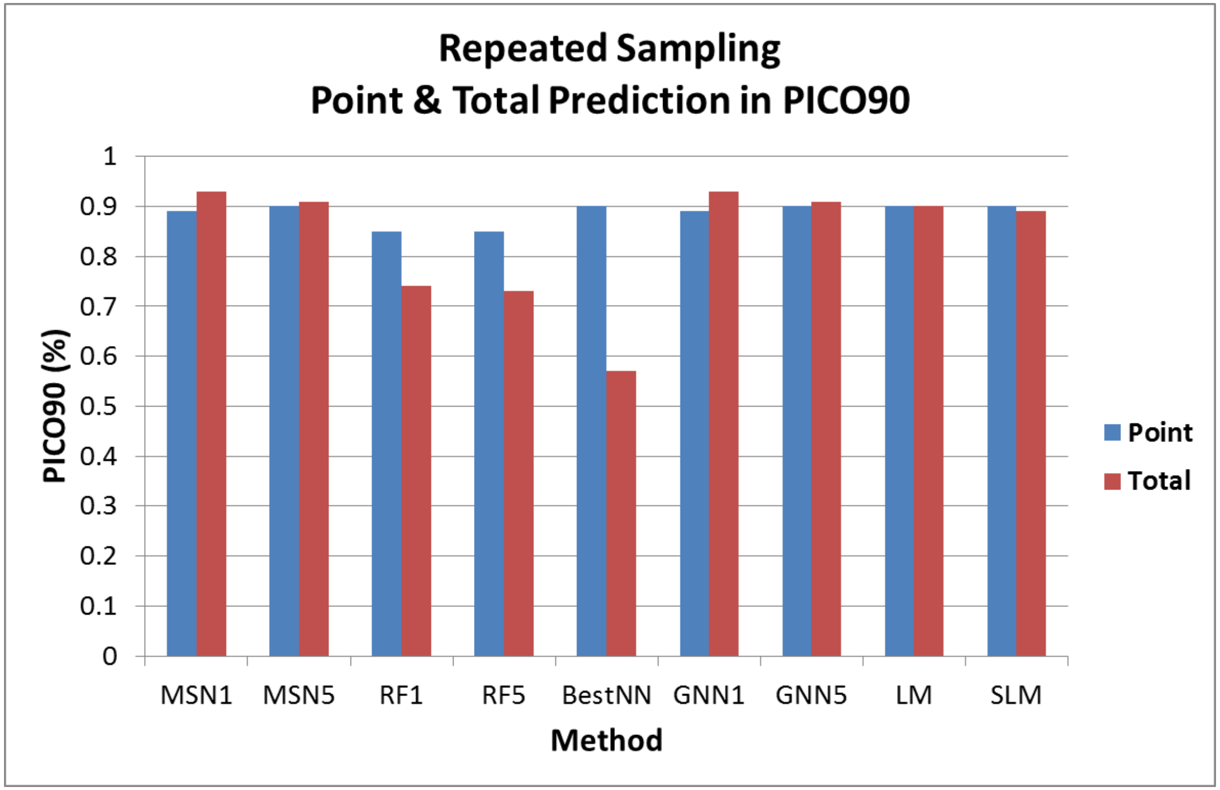
- Total prediction appears biased for RF, BestNN and GNN.
- Except MSN, the NN methods showed positive SRB (over prediction).





Results over 500 replications of Repeated Sampling Point and Total in PICO90

- RF has poor prediction interval coverage for both point and total predictions.
- Best NN has poor prediction interval coverage for total



3. Results over 500 resampling spatially unbalanced design – point prediction

- Spatially unbalanced design created more bias for NN methods
- SLM remained relatively unbiased, with the smallest RMSPE and valid prediction intervals.

Point Prediction – Spatially unbalanced sampling

Method	RMSPE	SRB	PICO90
MSN1	3.0	0.135	0.912
MSN5	2.5	0.227	0.907
RF1	3.1	0.080	0.905
RF5	2.5	0.104	0.903
BestNN	2.4	0.139	0.9
GNN1	3.3	0.061	0.906
GNN5	2.6	0.082	0.912
LM	2.4	0.159	0.903
SLM	2.1	0.028	0.918

Point Prediction – Balanced Sampling

Method	RMSPE (m^3/ha/yr)	SRB	PICO90
MSN1	3.2	0.003	0.89
MSN5	2.6	-0.001	0.90
RF1	2.8	0.023	0.85
RF5	2.8	0.024	0.85
BestNN	2.3	0.048	0.90
GNN1	3.3	0.004	0.89
GNN5	2.6	0.004	0.90
LM	2.4	-0.003	0.90
SLM	2.1	-0.002	0.90

3. Results over 500 resampling
Spatially unbalanced design – total prediction

- For predicting a total, there are large biases for NN methods.
- The large bias caused the RMSPE for SLM to be much lower than any of the NN methods examined
- Prediction intervals are far from the nominal 90%. SLM was more robust.

Total Prediction – Unbalanced Sampling			
Method	**RMSPE**	**SRB**	**PICO90**
MSN1	637.9	2.635	0.248
MSN5	853.6	4.055	0.11
RF1	457.1	1.418	0.626
RF5	442.2	1.770	0.438
BestNN	576.1	1.651	0.308
GNN1	864.0	1.070	0.72
GNN5	834.3	1.382	0.582
LM	608.1	2.860	0.128
SLM	269.0	0.369	0.92

4. Concluding Remarks

- Re-samplings experiments (and simulations) show that the SLM has smaller eRMSPE with generally less bias and better interval coverage than NN methods.

- For both point and total predictions, the SLM reduced RMSPE from 5% to 67% over NN methods examined.

- Reasons for substantial differences in point and total predictions:
    - SLM localizes the relation between the response variables and covariates in both the geographical and variable space.
    - SLM also accounts for the spatial structure of the data and minimizes prediction error

- SLM was also more robust to spatially unbalanced sampling.

SLM Models: Better Lemon Squeezers (after Ver Hoef and Temesgen 2013)

(Google Images)

- ❖ Theoretical reviews shows that the SLM has the prediction optimality properties, and can be quite robust.

- ❖ SLM provides point-wise prediction standard errors.

- ❖ <u>Unbiased and provides accurate nominal coverage</u>. SLM also accounts for the spatial structure of the data and minimizes prediction error.

- SLM was also more robust to spatially unbalanced sampling.

4. Research Directions

- Mapping probability surfaces for prediction or errors to provide a higher level of confidence in using resource maps.

- Examination of multivariate SLM to preserve the covariance among multiple response variables at new locations.

- Using LiDAR and multispectral imagery to relate PMAI with other covariates (e.g., parent material, soil moisture and soil nutrient regimes) via spatial linear mixed model and Bayesian spatial regression models

- *Choice of SLM or NN method depends* on the number of response variables and objectives. The SLM is a good choice for imputing PMAI (and for imputing ≤ 3 response variables), while NN is suggested when imputing >3 response variables.

5. References

Cressie, N. 1993. Statistics for Spatial Data. New York: John Wiley and Sons. 900 p.

Moeur, M. and A.R. Stage. 1995. Most similar neighbour: an improved sampling inference procedure for natural resource planning. For. Sci. 41, 337-359.

Latta, G., Temesgen, H., and T. Barrett. 2009. Mapping and imputing potential productivity of Pacific Northwest Forests using climate variables. Can. J. For. Res. 39: 1197-1207.

Ohmann, J. and M. Gregory. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. Canadian Journal of Forest Research 32:725-741

Patterson H.D., Thompson R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554.

Temesgen, H. and J. Ver Hoef. 2014. Evaluation of the Spatial Linear Model, Random Forest, and Gradient Nearest Neighbor Methods for Imputing Potential Productivity and Biomass of the Pacific Northwest Forests. Forestry: An Int. J. For. Res. 6: 1-12

Ver Hoef J. and H. Temesgen. 2013. A comparison of the spatial linear model to nearest neighbor (k-NN) methods for forestry applications. PLOS ONE. (3):1-11.

Ver Hoef J.M. 2002. Sampling and geostatistics for spatial data. Ecoscience 9:152–161.

Ver Hoef J.M. 2000. Predicting finite populations from spatially correlated data. In: ASA Proceedings of the Section on Statistics and the Environment. American Statistical Association, 93–98.

Acknowledgments

We thank:

Drs. Greg Latta and Tara Barrett for compiling data and for their insights

Data and funding provided by:

PNW FIA Program, USDA Forest Service, Portland, USA

McIntire-Stennis, USDA