Gradient Nearest Neighbor Imputation for Mapping Forest Attributes and Variability

Kenneth B. Pierce, Jr., and Janet Ohmann

UDSA Forest Service
Pacific Northwest Research Station
Forestry Sciences Laboratory

Our current project consists of three study areas located in Oregon, Washington, and California of around 3 million hectares each. The goal of the project is to map forest structure and composition by combining plot data with spatial data and distributing the plot data spatially using Gradient Nearest Neighbor imputation.

We use a selection of response variables, which we relate to environmental variables using canonical correspondence analysis. The loadings and eigenvalues from the ordination model are then used to locate each sample plot within a multivariate cloud based on their environmental attributes. For the imputation step, each pixel in the target map area is compared based on its environmental attributes to each plot in the training data set. Each plot is then ranked using the Euclidean distance from the pixel's location in gradient space to each plot, with the closest plot getting the highest rank.

For simple single neighbor imputation, the plot ID of the highest-ranking plot is assigned to the target pixel. At the completion of the imputation, each pixel is related to a target plot and any attribute calculated for a plot can be mapped. The selected plot for imputation can also be based on a summary statistic from a suite of plots, for example the five plots with the highest ranks. Imputation methods like kNN impute the summary statistic as the value, such as mean basal area of the five highest-ranking plots.

Imputation provides for an interesting set of accuracy measurements based on the ranking of potential plots. Since the plots are ranked based on gradient similarity, the highest-ranking plots for any pixel provide a sampling region related to that plot and if similar enough can provide a spatial measure of natural variability for each gradient location. That variability can be expressed by the variance of the high-ranking plots. Additionally a measure of similarity is required to assess whether a sample is actually comprised of similar plots. The distance in gradient space to these plots can be compared to the distribution of inter-plot distances to assess similarity. We map “sampling sufficiency” by designating a threshold distance (e.g. 25^{th} centile distance) and count how many plots among the 20 highest-ranking plots fall within this threshold. A plot with 20 close neighbors is considered well sampled and with fewer than 10, poorly sampled.

Imputation is especially useful for retaining covariance between plot attributes and expressing the full range of variability. However, the presence of natural variability within some gradient envelope can cause pixel level accuracy to be low in some instances. Imputation of this sort is probably best used at the small watershed scale or as an accompaniment to more predictive maps to gauge variability from natural and design sources.

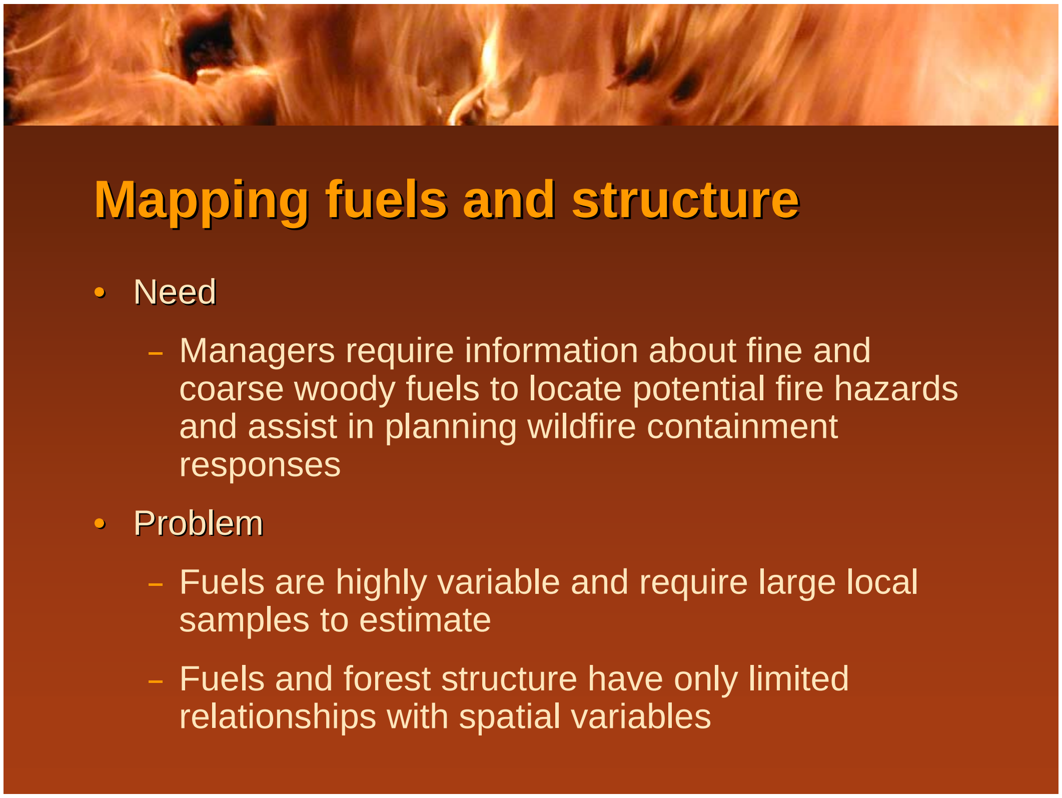
Gradient Nearest Neighbor Imputation for Mapping Forest Attributes and Variability

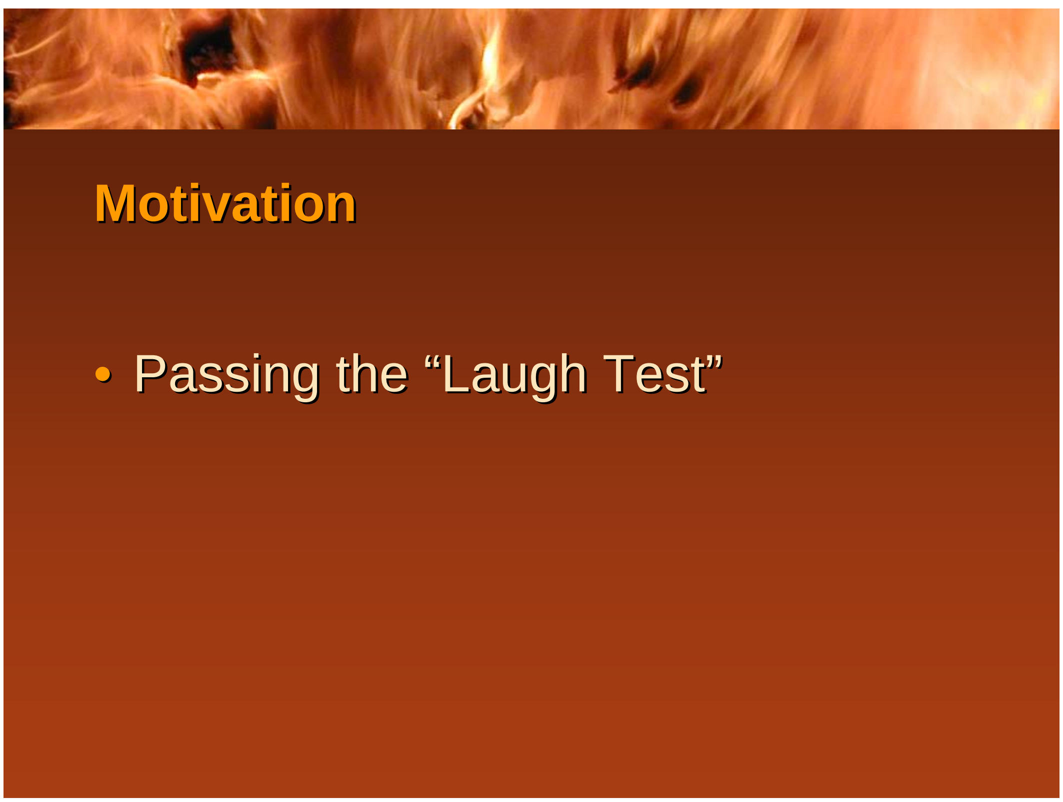
Evaluation of quantitative techniques for deriving National scale data for assessing and mapping risk workshop July 26-28, 2005

Kenneth B. Pierce Jr and Janet Ohmann
Forestry Sciences Lab, PNW Research Station, Corvallis

Objectives/Overview

- Map fuels and vegetation using Gradient Nearest Neighbor (GNN) method
- 3 contrasting ecoregions
- Assess accuracy
- 30 m resolution
- Wa 5.0 million Ha.
- Or 3.1 million Ha.
- Ca 4.6 million Ha.


Mapping fuels and structure
• Need
– Managers require information about fine and coarse woody fuels to locate potential fire hazards and assist in planning wildfire containment responses
• Problem
– Fuels are highly variable and require large local samples to estimate
– Fuels and forest structure have only limited relationships with spatial variables

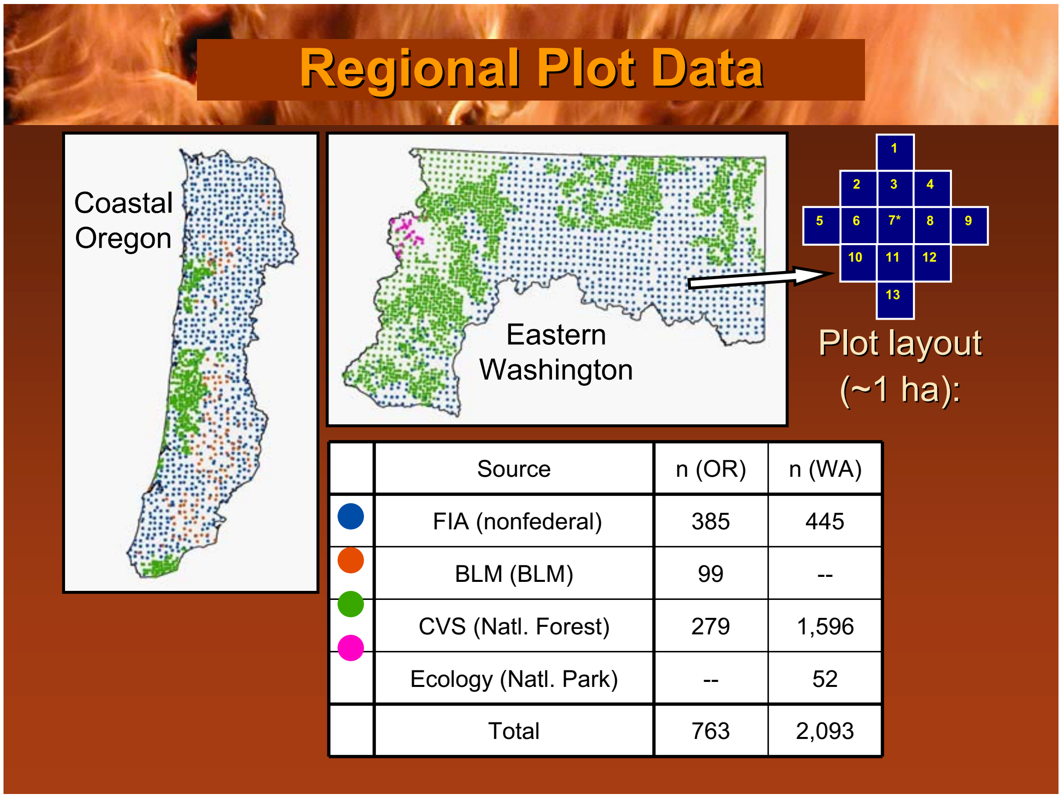

Motivation
• Passing the “Laugh Test”

Assessing Risk

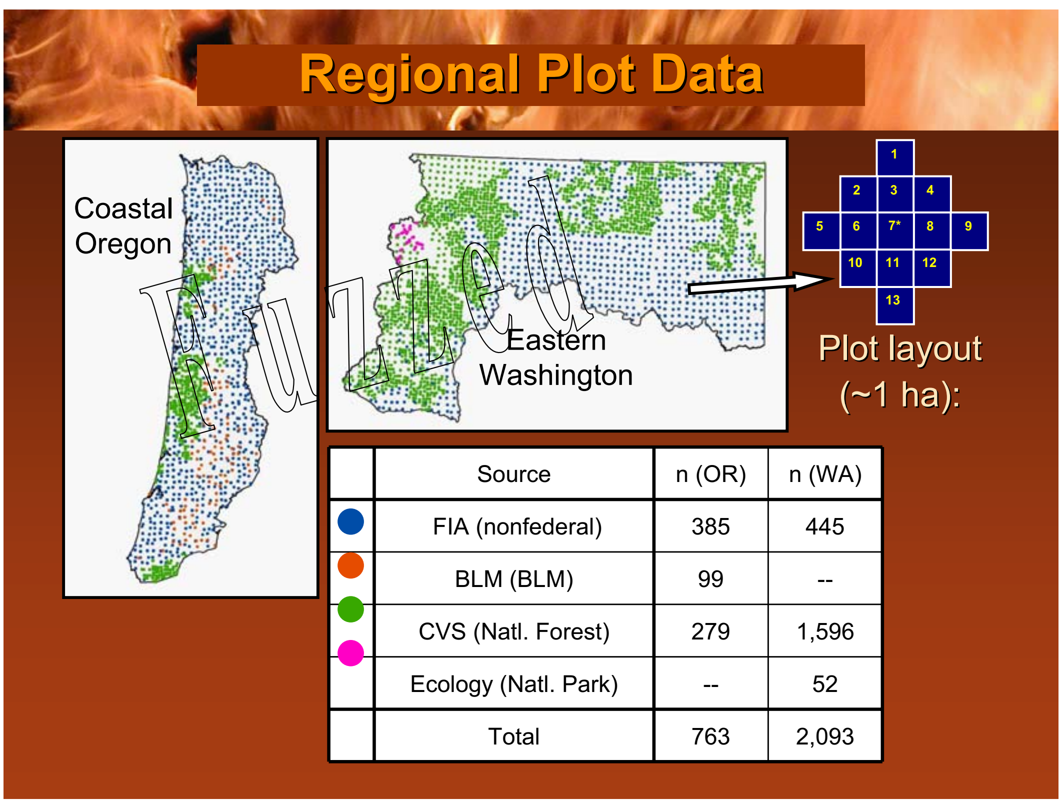
- Identifying factors which influence risk
  - location in a floodplain
  - high fuels
  - proximity to disease
- Uncertainty in data and predictions
  - Decision-making with uncertainty

Major Steps in GNN mapping:

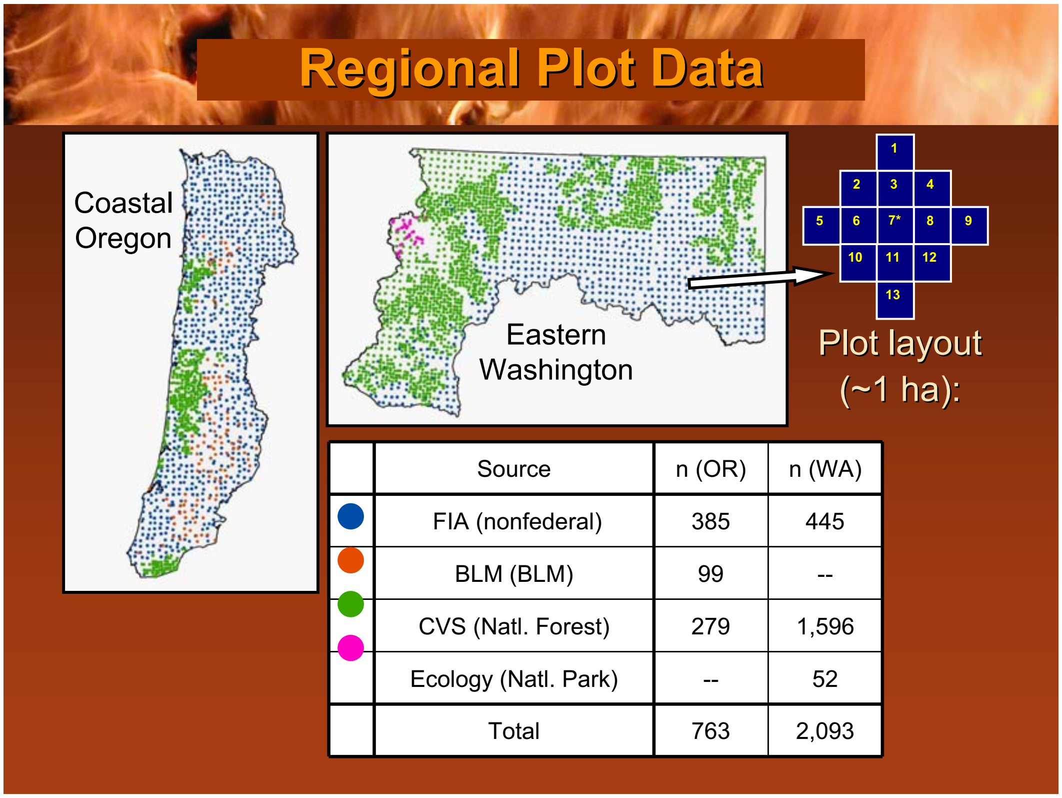
- 1) Data Preparation/Screening
- 2) Statistical Modeling
- 3) Imputation/Map Creation
- 4) Accuracy Assessment
- 5) Applications and Risk Assessment

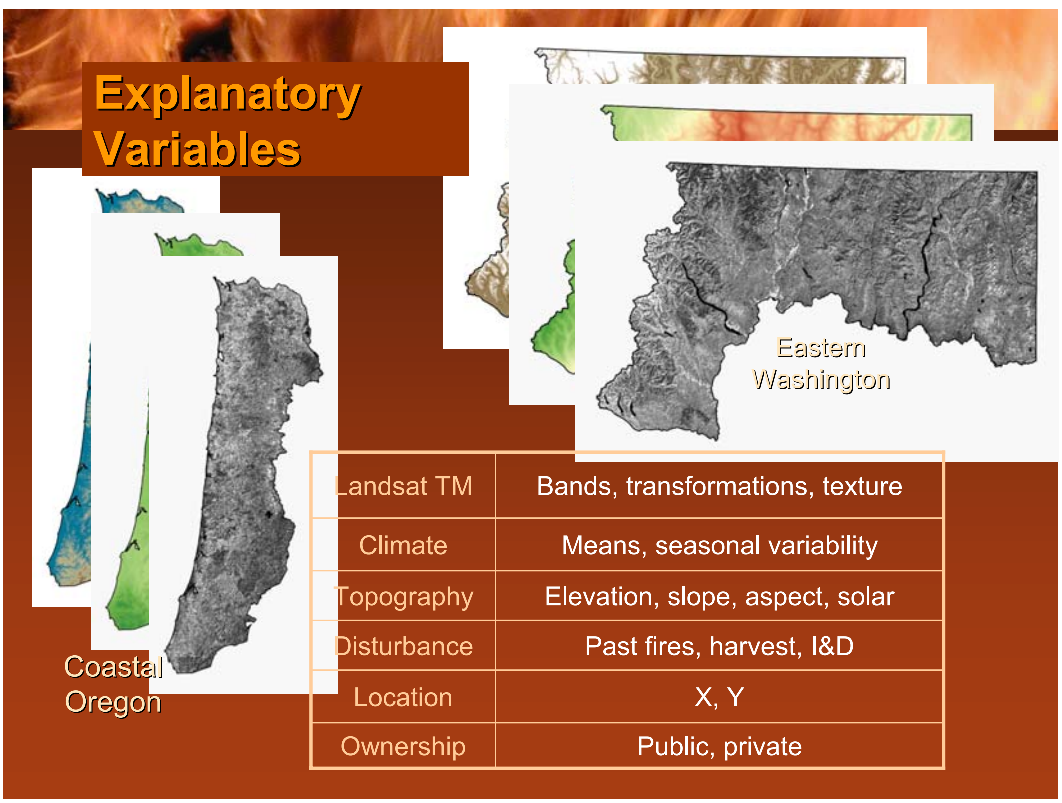


	Source	n (OR)	n (WA)
	FIA (nonfederal)	385	445
	BLM (BLM)	99	--
	CVS (Natl. Forest)	279	1,596
	Ecology (Natl. Park)	--	52
	Total	763	2,093



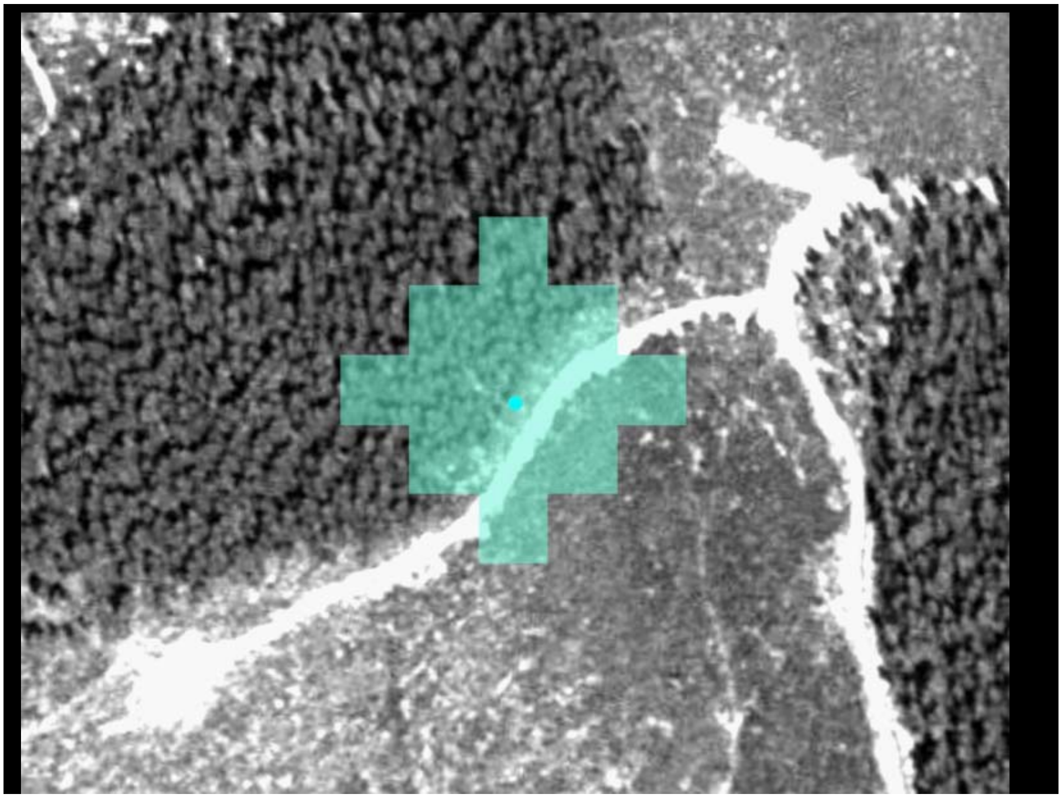
	Source	n (OR)	n (WA)
	FIA (nonfederal)	385	445
	BLM (BLM)	99	--
	CVS (Natl. Forest)	279	1,596
	Ecology (Natl. Park)	--	52
	Total	763	2,093


Regional Plot Data
Coastal Oregon
Eastern Washington
Plot layout (~1 ha):
Source | n (OR) | n (WA)
FIA (nonfederal) | 385 | 445
BLM (BLM) | 99 | --
CVS (Natl. Forest) | 279 | 1,596
Ecology (Natl. Park) | -- | 52
Total | 763 | 2,093


Explanatory Variables
Eastern Washington
Coastal Oregon
Landsat TM | Bands, transformations, texture
Climate | Means, seasonal variability
Topography | Elevation, slope, aspect, solar
Disturbance | Past fires, harvest, I&D
Location | X, Y
Ownership | Public, private

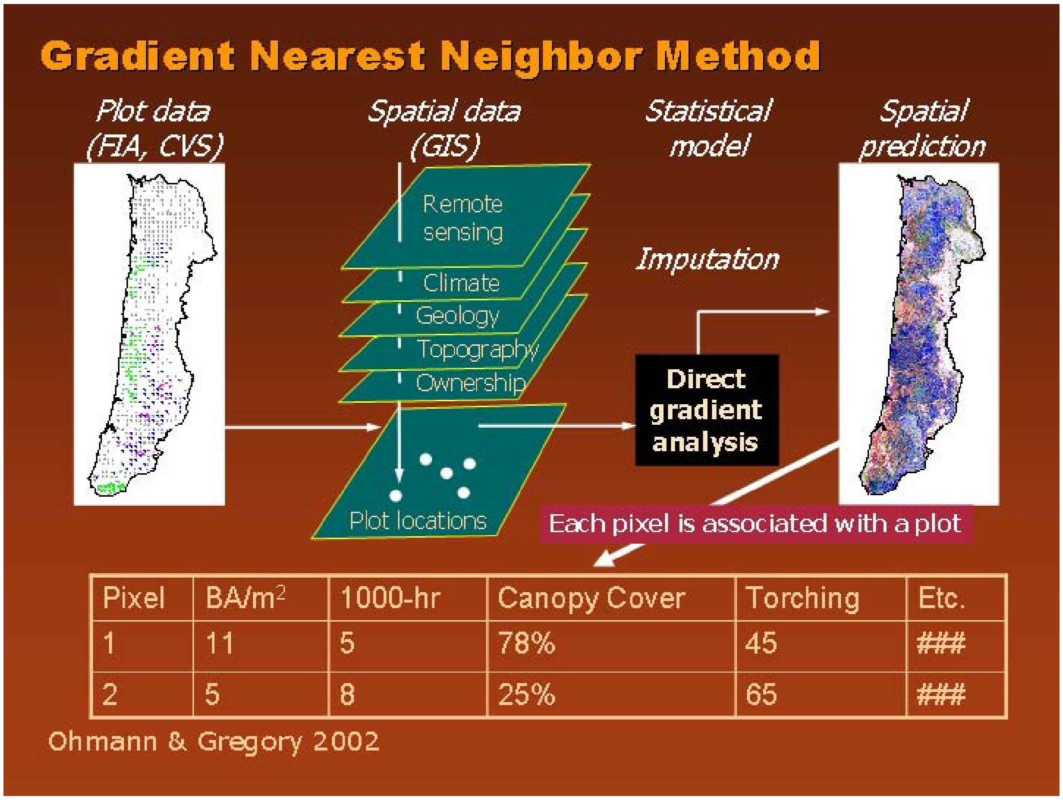
Data Screening

- Included multi-condition plots where minimal spectral variation existed
  - Half oak woodland/Half mixed conifer (included)
  - Half grassland/Half forest (excluded)
  - Half grassland/Half oak (ie sparse or savannah) (included)
- Used within plot spectral variability to help assign heterogeneous plots



Major Steps in GNN mapping:

- 1) Data Preparation/Screening
- 2) Statistical Modeling
- 3) Imputation/Map Creation
- 4) Accuracy Assessment
- 5) Applications and Risk Assessment



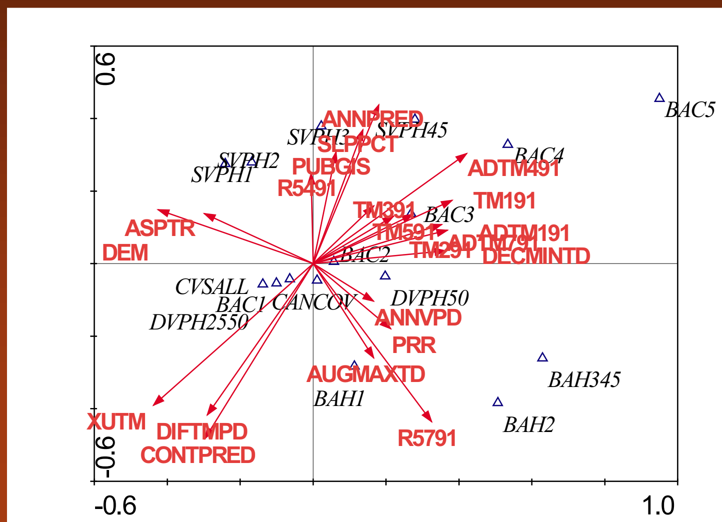
Pixel	BA/m^2	1000-hr	Canopy Cover	Torching	Etc.
1	11	5	78%	45	###
2	5	8	25%	65	###

Statistical Modeling: Canonical Correspondence Analysis

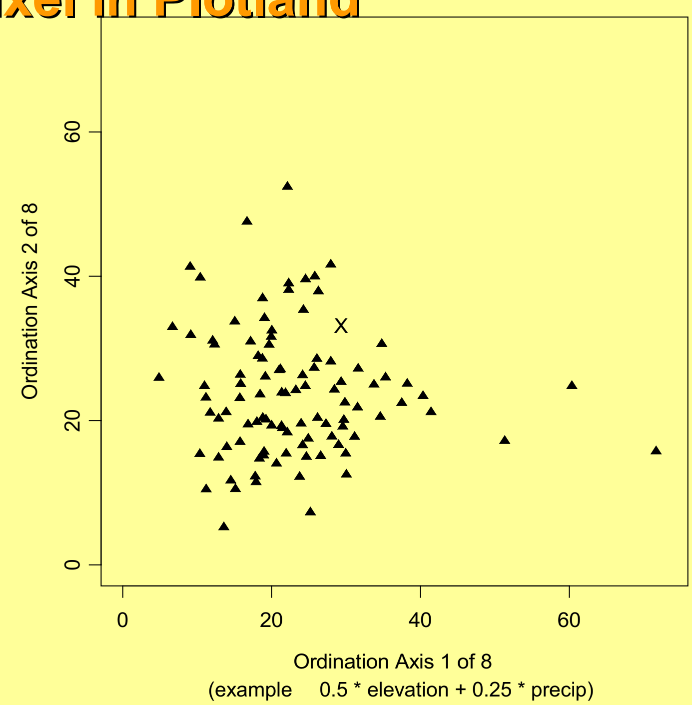
- Multivariate statistical method
  - results in a weight for each spatial variable as to its relationship with the multiple response variables
- Modeling Variables-used as model Y's
  - Structure models (BAC, BAH, STPH, CWD)
  - Species models
- Mapping Variables-retained with plot-map link

Neighbors in Gradient Space

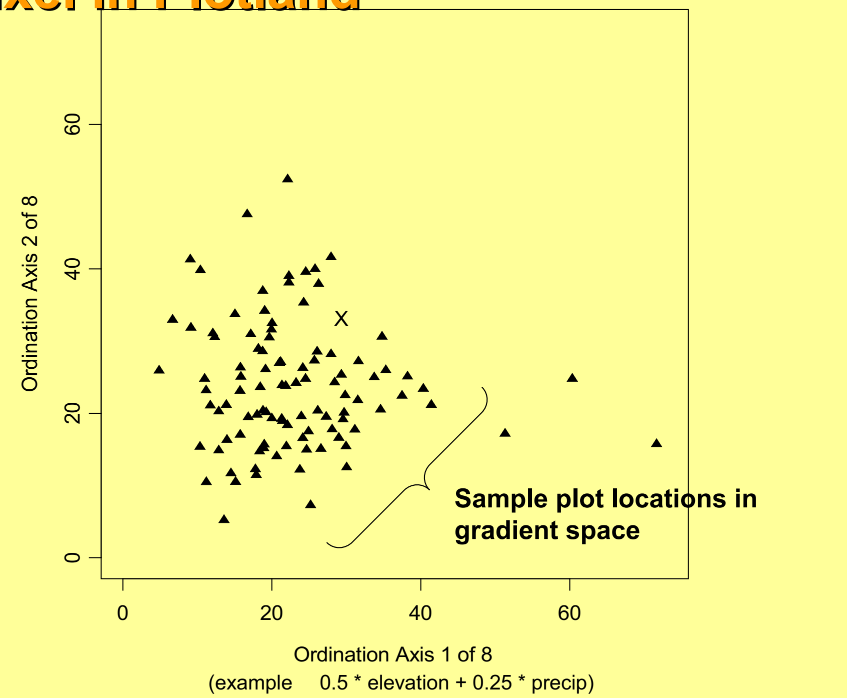
- Direct gradient analysis allows assignment of a multi-dimensional location to each predicted pixel



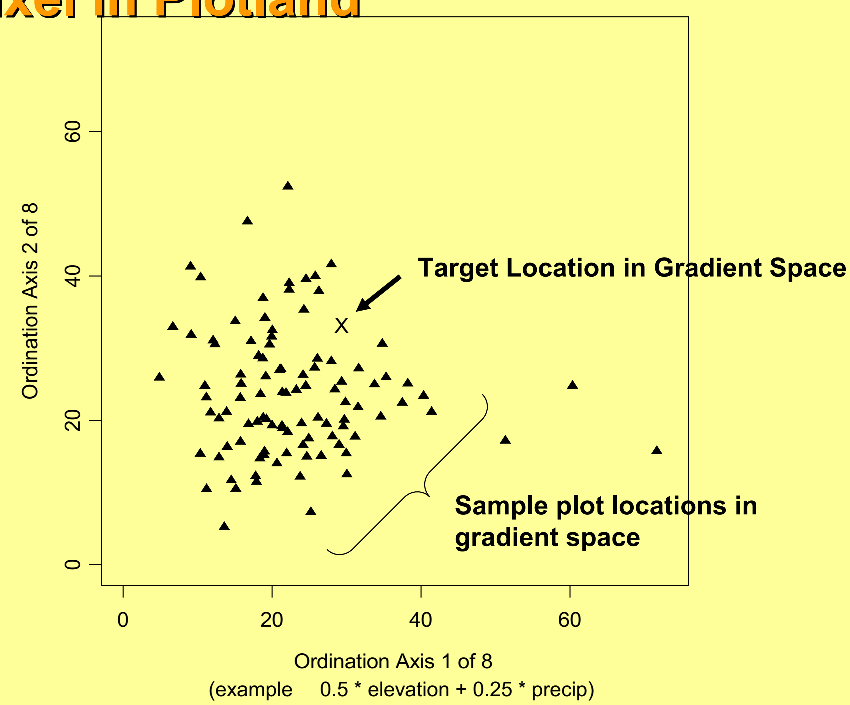
A Pixel in Plotland

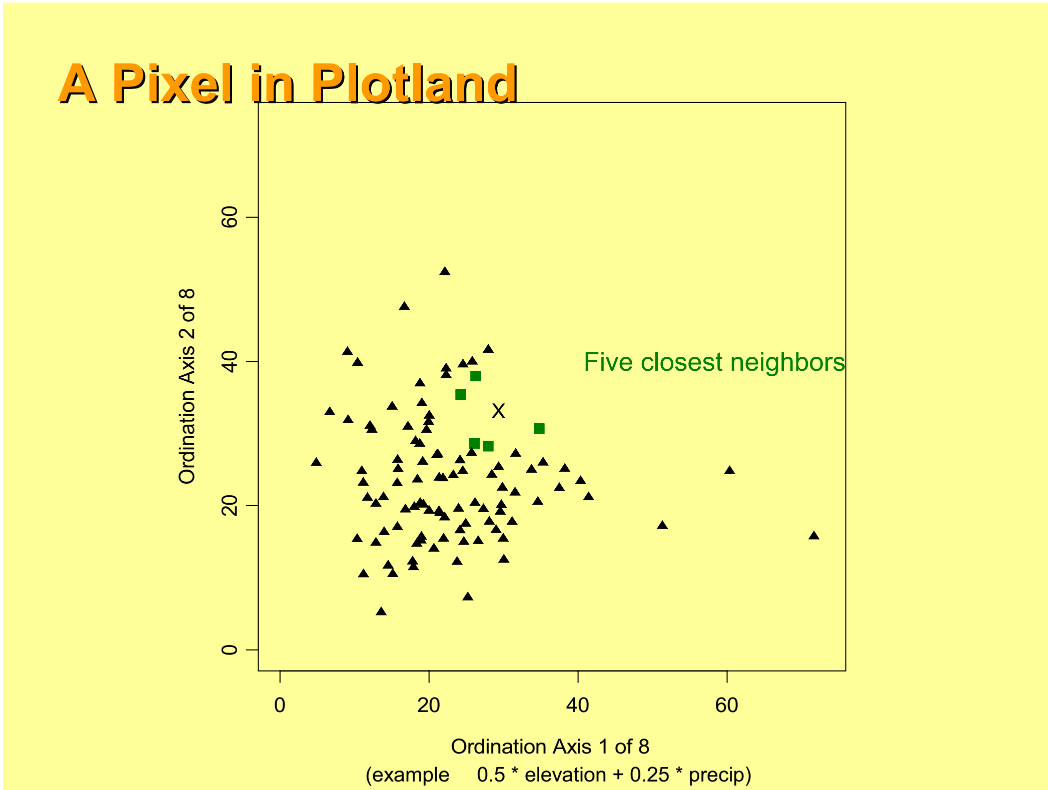


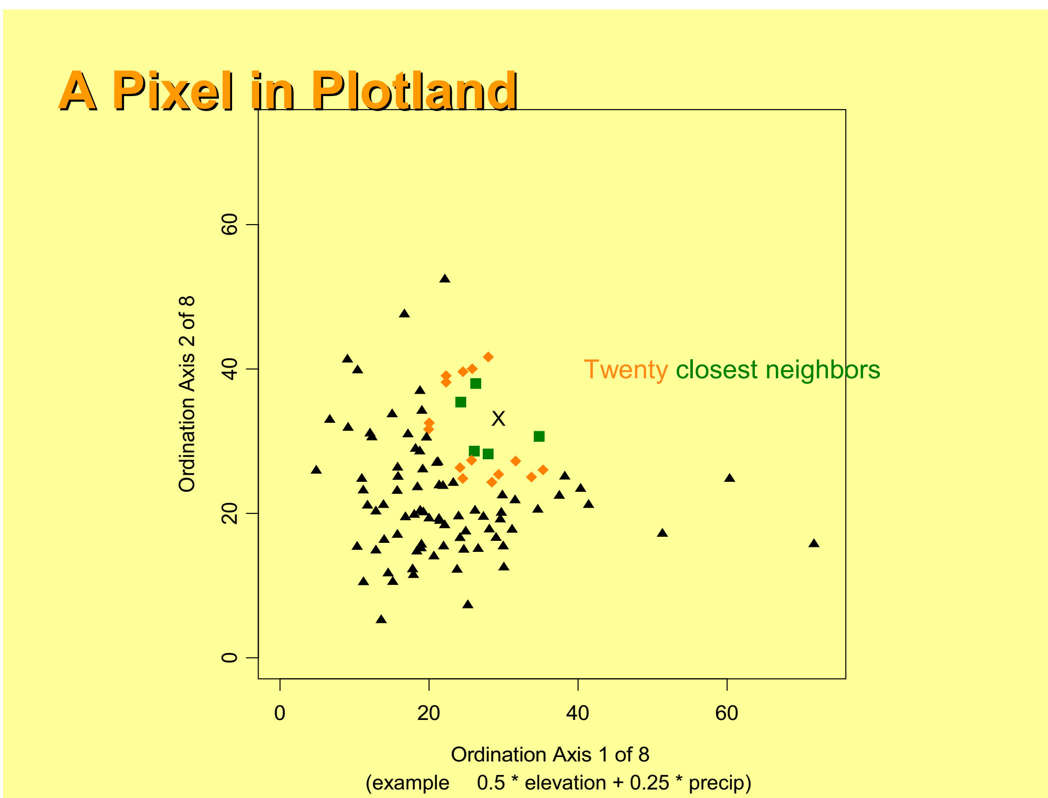
A Pixel in Plotland

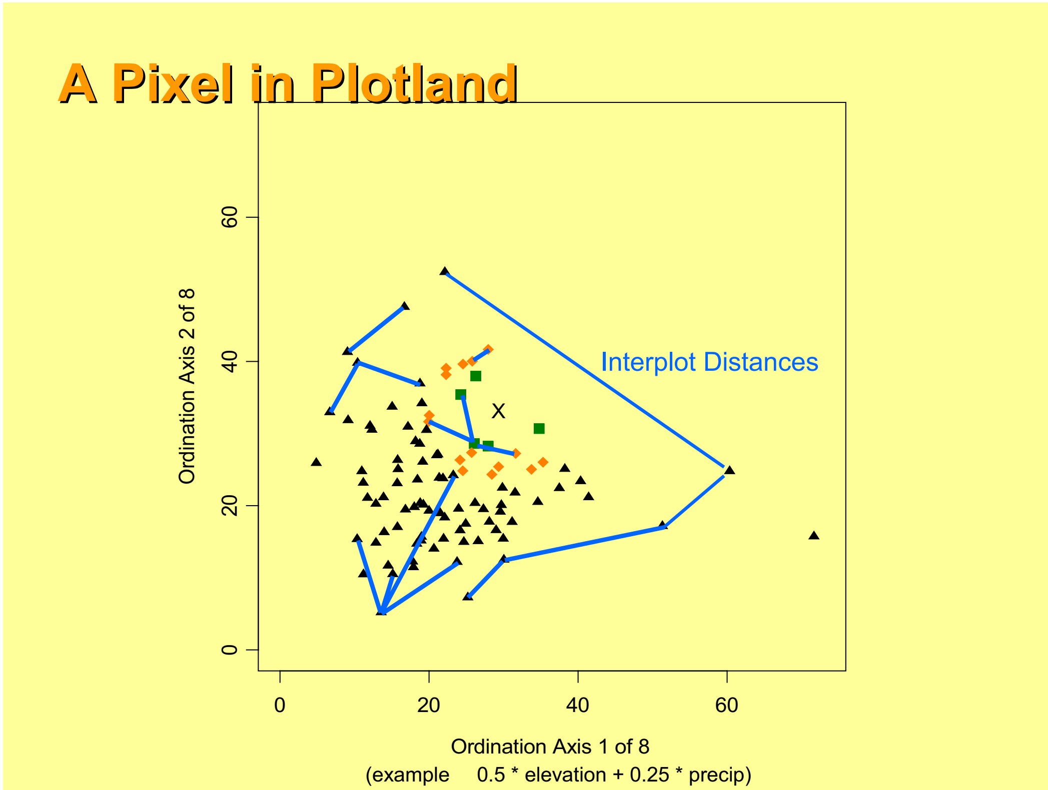


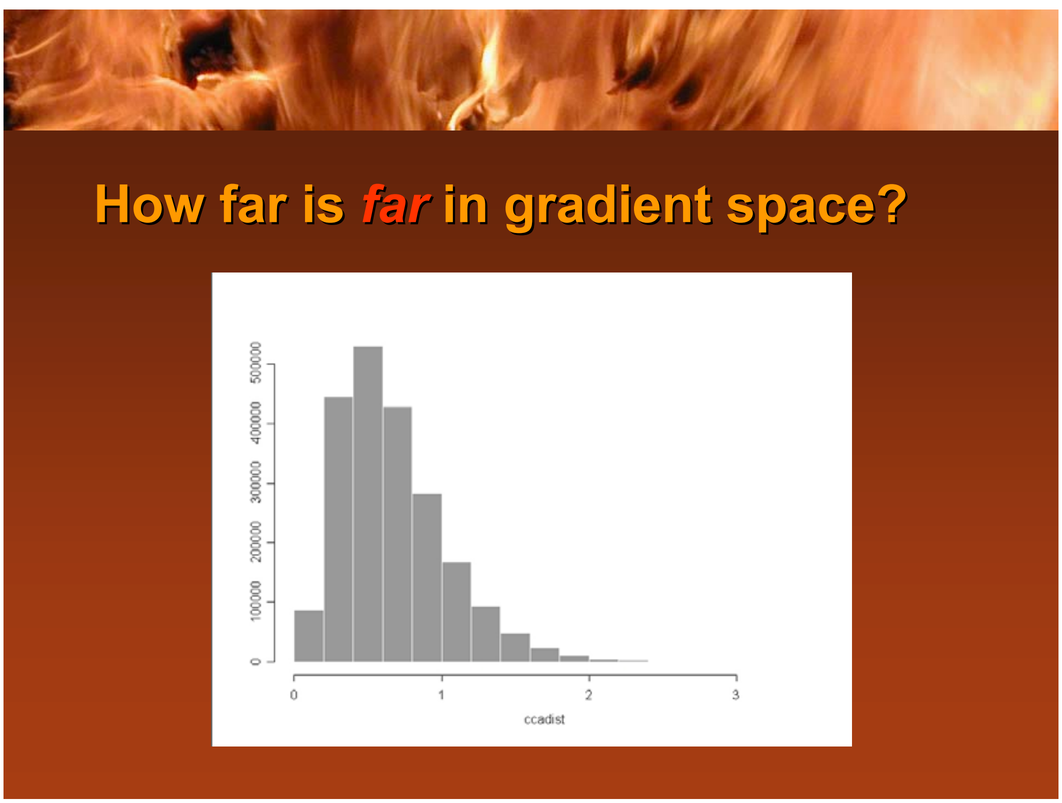
A Pixel in Plotland

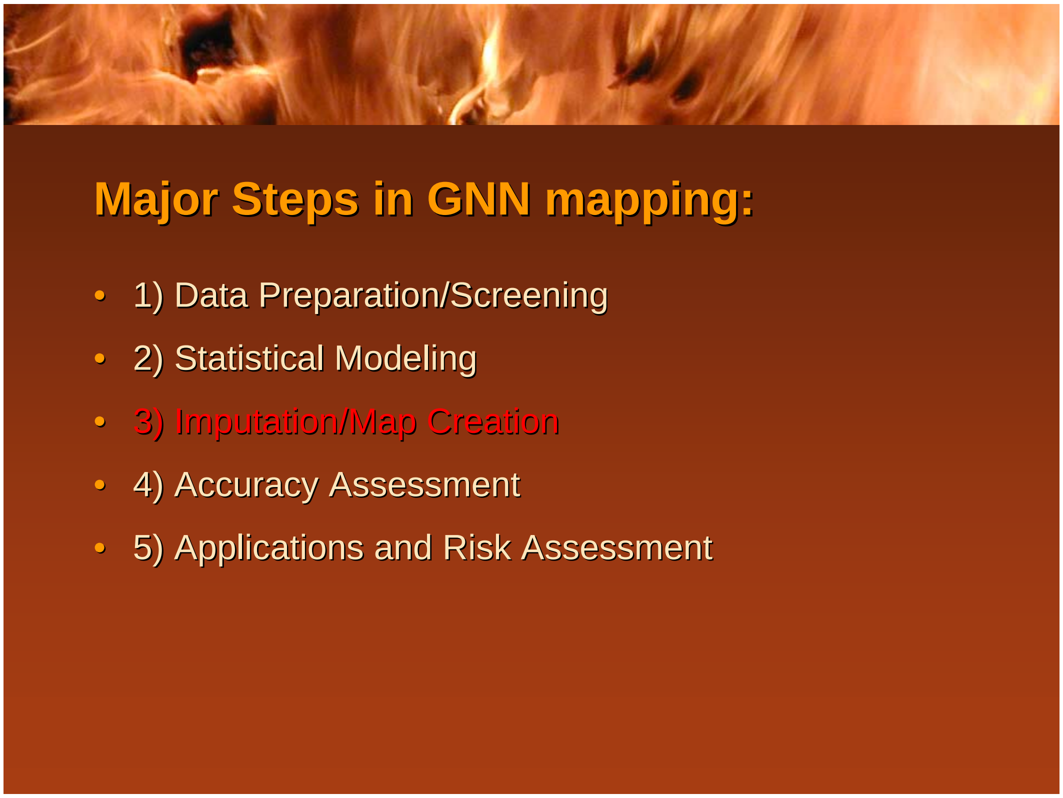


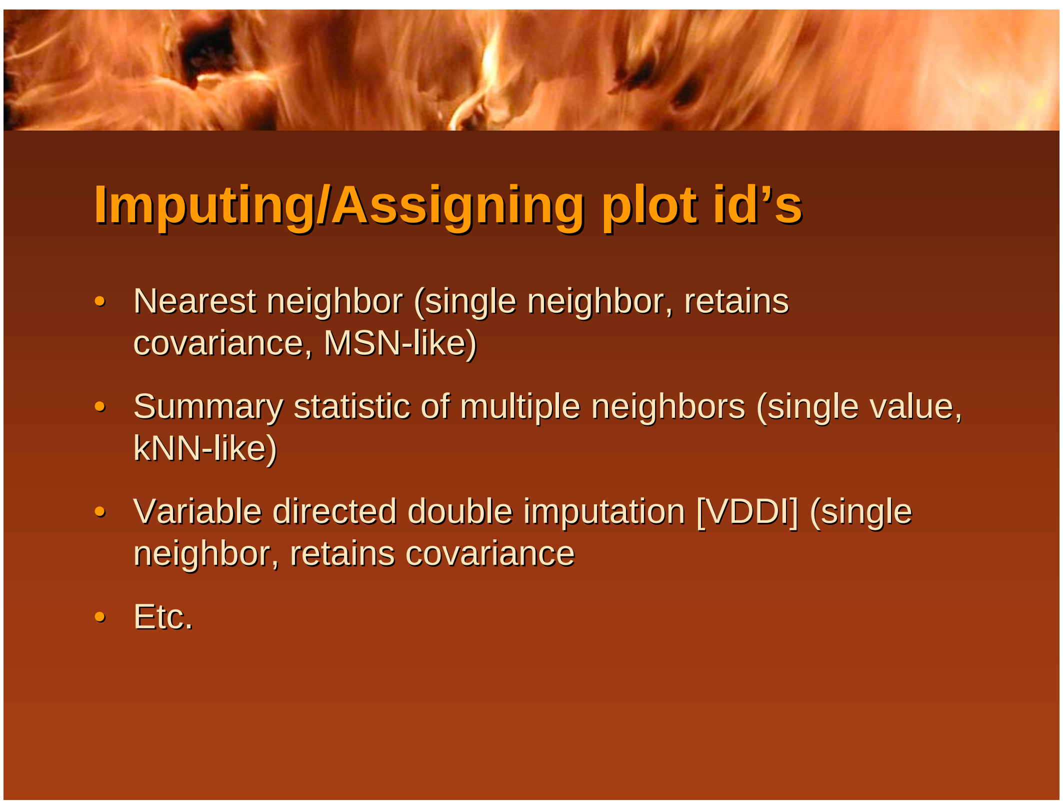

A Pixel in Plotland
Five closest neighbors
X
Ordination Axis 2 of 8
0
20
40
60
0
20
40
60
Ordination Axis 1 of 8
(example 0.5 * elevation + 0.25 * precip)

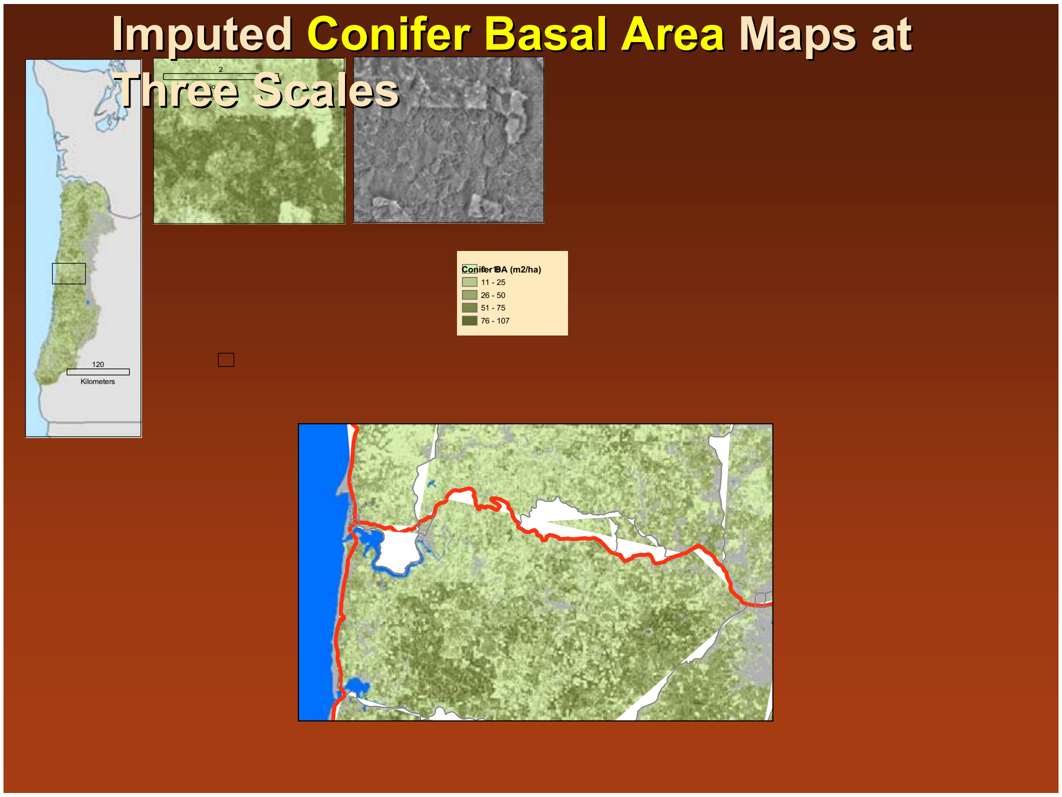

A Pixel in Plotland
Twenty closest neighbors
X
Ordination Axis 2 of 8
0
20
40
60
0
20
40
60
Ordination Axis 1 of 8
(example 0.5 * elevation + 0.25 * precip)

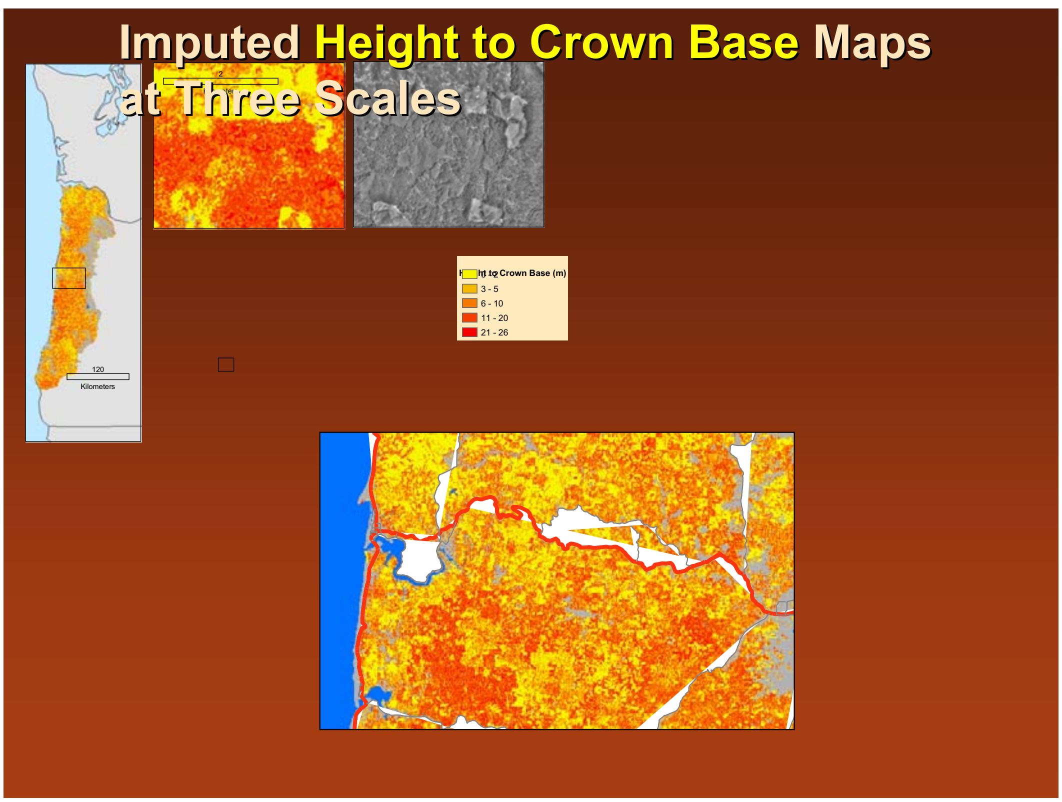

A Pixel in Plotland
Interplot Distances
X
Ordination Axis 2 of 8
0
20
40
60
0
20
40
60
Ordination Axis 1 of 8
(example 0.5 * elevation + 0.25 * precip)

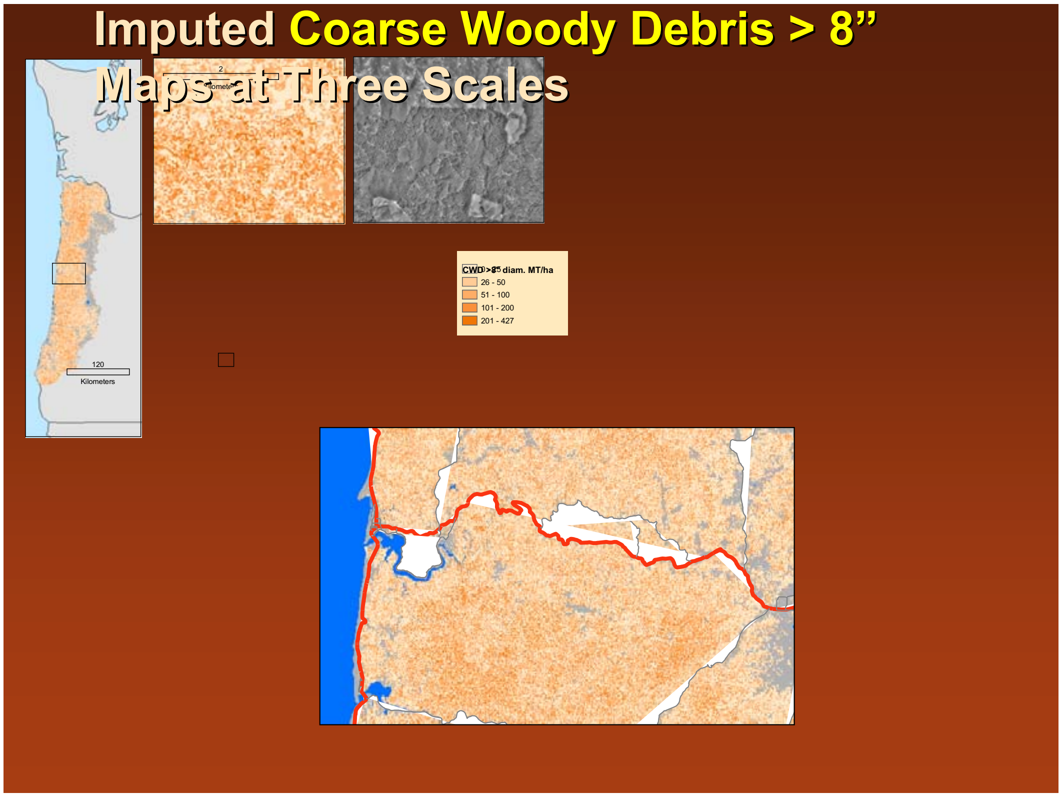

How far is *far* in gradient space?
0
100000
200000
300000
400000
500000
0
1
2
3
ccadist


Major Steps in GNN mapping:
1) Data Preparation/Screening
2) Statistical Modeling
3) Imputation/Map Creation
4) Accuracy Assessment
5) Applications and Risk Assessment


Imputing/Assigning plot id's
Nearest neighbor (single neighbor, retains covariance, MSN-like)
Summary statistic of multiple neighbors (single value, kNN-like)
Variable directed double imputation [VDDI] (single neighbor, retains covariance
Etc.


Imputed Conifer Basal Area Maps at Three Scales
Conifer BA (m2/ha)
11 - 25
26 - 50
51 - 75
76 - 107
120
Kilometers


Imputed Height to Crown Base Maps at Three Scales
Height to Crown Base (m)
3 - 5
6 - 10
11 - 20
21 - 26
120
Kilometers



Major Steps in GNN mapping:

- 1) Data Preparation/Screening
- 2) Statistical Modeling
- 3) Imputation/Map Creation
- 4) Accuracy Assessment
- 5) Applications and Risk Assessment

Uncertainty

“How much better is my knowledge about my study area because I have this map?”

Sources of Uncertainty For Ecological Detectives

- Process Uncertainty/Natural Variability
  - Uncontrollable (often unmeasurable)
    - Natural disturbances
    - Demographic stochasticity
    - Anthropogenic disturbances
- Sampling Uncertainty
  - Not entirely uncontrollable
    - Limited sampling
    - Spatial averaging
    - Temporal sample variation

Hilborn & Mangel 1997

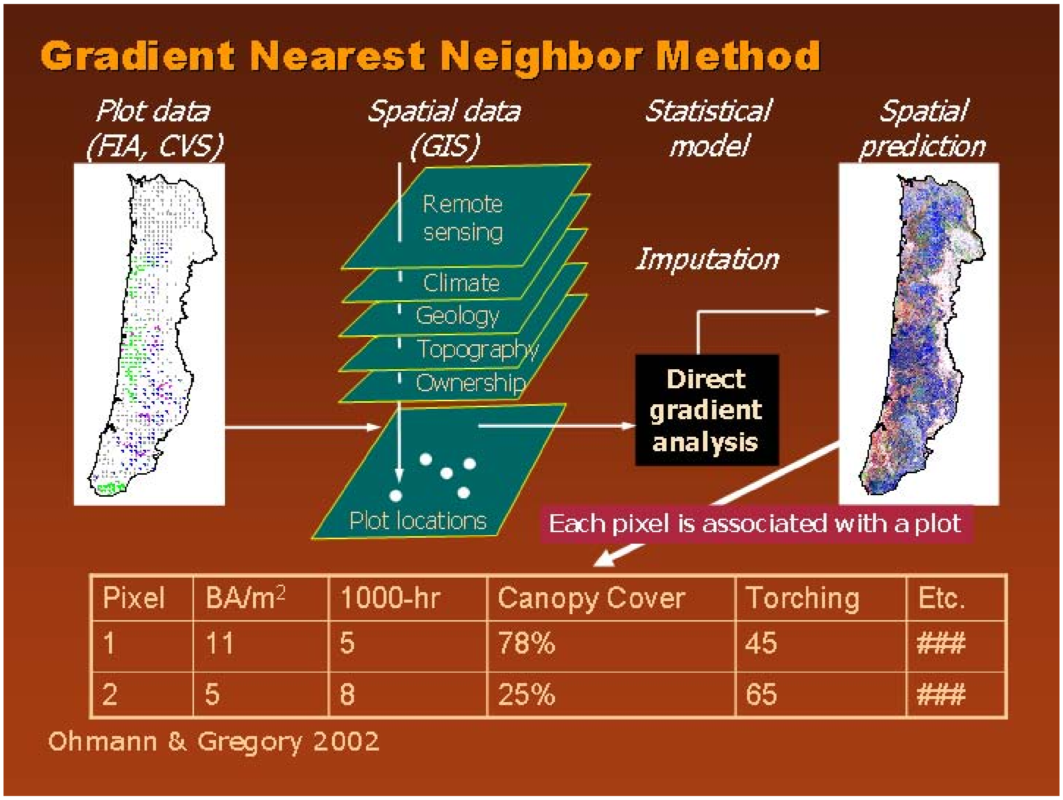
Accuracy assessments

"obsessive transparency"

- Map integral (Value of Map)
  - Confusion matrices/Kappa (local)
  - Correlation statistics (local)
  - Regional histograms (regional)
- Map explicit (Map of Values)
  - Confidence maps (Process)
  - Support (Sampling)

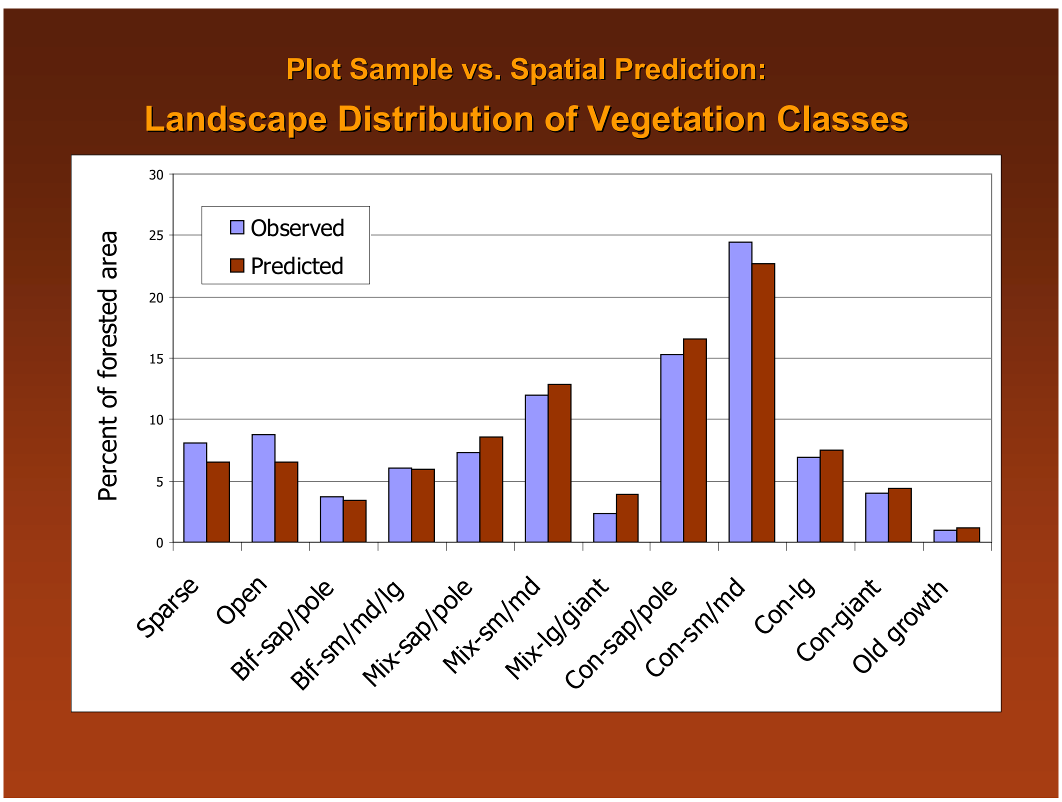
Site-Level Prediction Accuracy for 11 Vegetation Classes

Observed class	Predicted class											% correct	% within one class
	Open	Blf-sap/pol	Blf-sm/md/lg	Mix-sap/pol	Mix-sm/md	Mix-lg/giant	Con-sap/pol	Con-sm/md	Con-lg	Con-giant	Old growth		
Open (<40%)	33	3	2	5	8	0	4	3	1	1	0	55	75
Broadleaf-sap/pole	5	2	1	3	5	0	0	0	0	0	0	13	69
Broadleaf-sm/md/lg	0	2	3	2	18	6	1	0	0	0	0	9	97
Mixed-sap/pole	2	1	0	12	13	1	6	8	1	0	0	27	77
Mixed-sm/md	0	1	4	6	43	4	1	19	4	0	0	52	93
Mixed-lg/giant	0	0	1	0	9	16	1	2	11	4	0	36	93
Conifer-sap/pole	2	0	0	14	9	0	45	23	2	0	0	47	88
Conifer-sm/md	1	0	0	3	27	2	12	95	22	1	0	58	96
Conifer-lg	0	0	0	0	8	1	0	17	31	18	1	41	89
Conifer-giant	0	0	0	0	1	6	0	3	27	36	0	49	95
Old growth	0	0	0	0	1	1	1	1	9	3	0	0	81
% correct	77	22	27	27	30	41	63	56	29	57	0	45	
% within one class	98	89	82	93	77	95	94	90	93	97	100		89


Gradient Nearest Neighbor Method
Plot data (FIA, CVS)
Spatial data (GIS)
Statistical model
Spatial prediction
Remote sensing
Climate
Geology
Topography
Ownership
Plot locations
Imputation
Direct gradient analysis
Each pixel is associated with a plot

Pixel	BA/m2	1000-hr	Canopy Cover	Torching	Etc.
1	11	5	78%	45	###
2	5	8	25%	65	###

Ohmann & Gregory 2002


Plot Sample vs. Spatial Prediction:
Landscape Distribution of Vegetation Classes
Percent of forested area
0
5
10
15
20
25
30
Observed
Predicted
Sparse
Open
Blf-sap/pole
Blf-sm/md/lg
Mix-sap/pole
Mix-sm/md
Mix-lg/giant
Con-sap/pole
Con-sm/md
Con-lg
Con-giant
Old growth

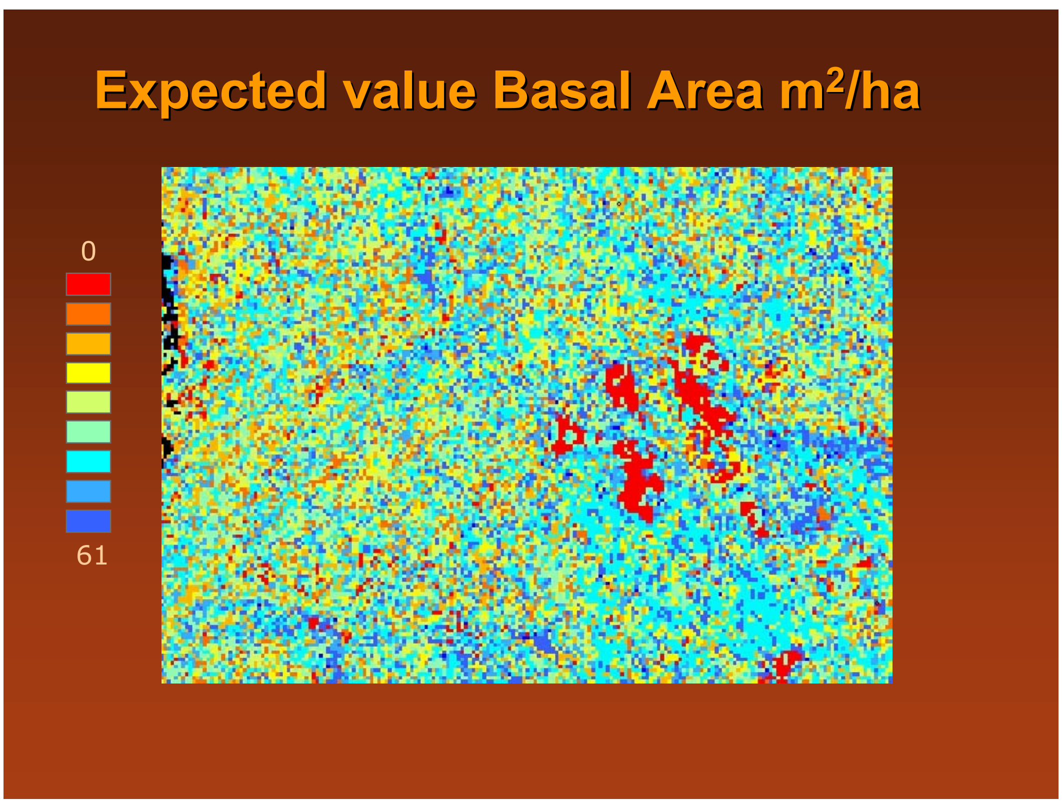
Accuracy assessments

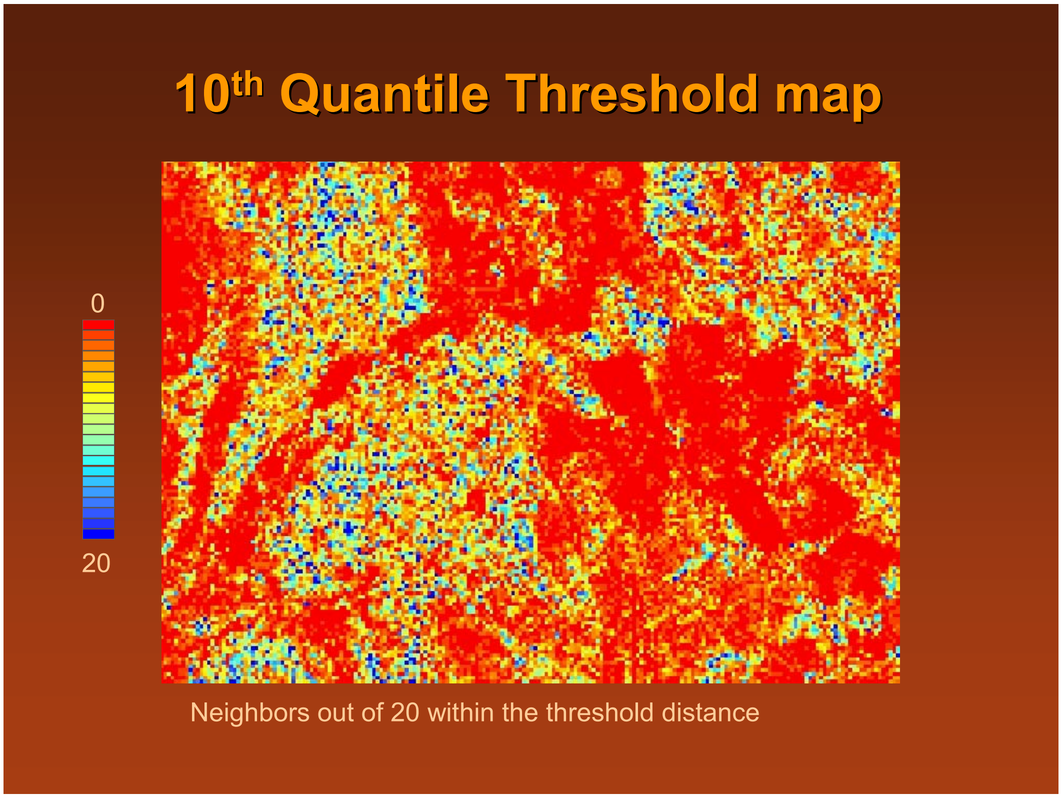
- Map integral (Value of Map)
  - Confusion matrices
  - Kappa statistics
  - Correlation statistics
  - Regional histograms
- Map explicit (Map of Values)
  - Confidence maps (Process)
  - Support (Sampling)

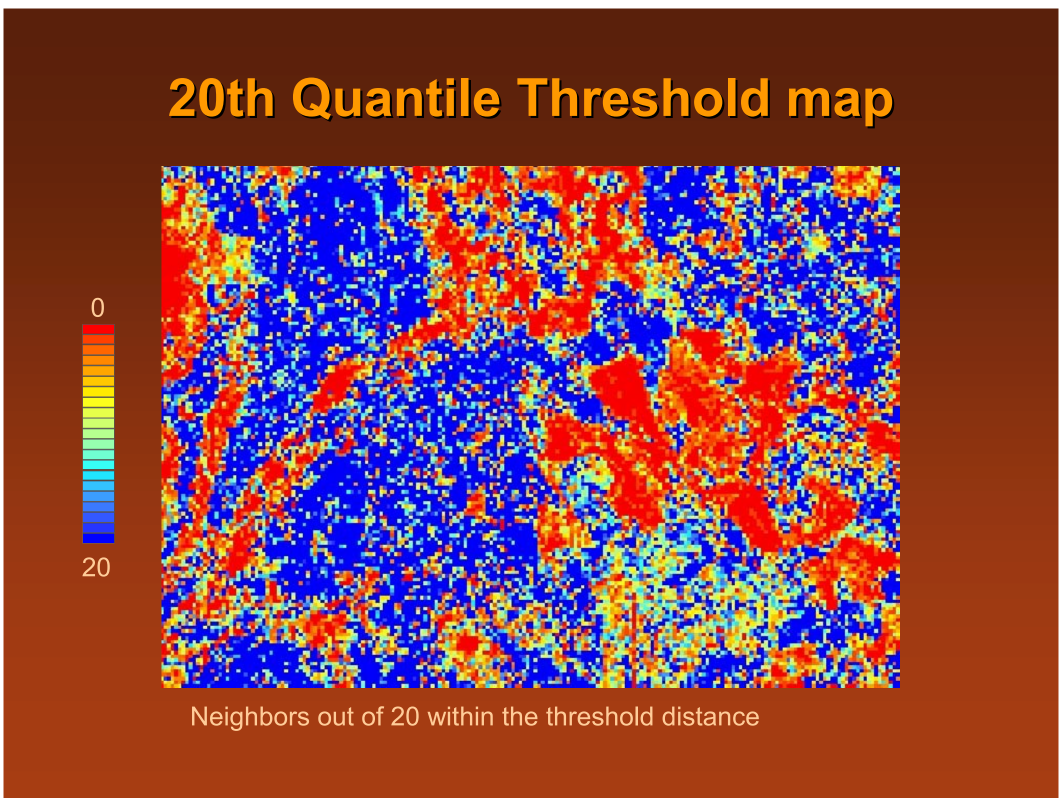
Overview of maps

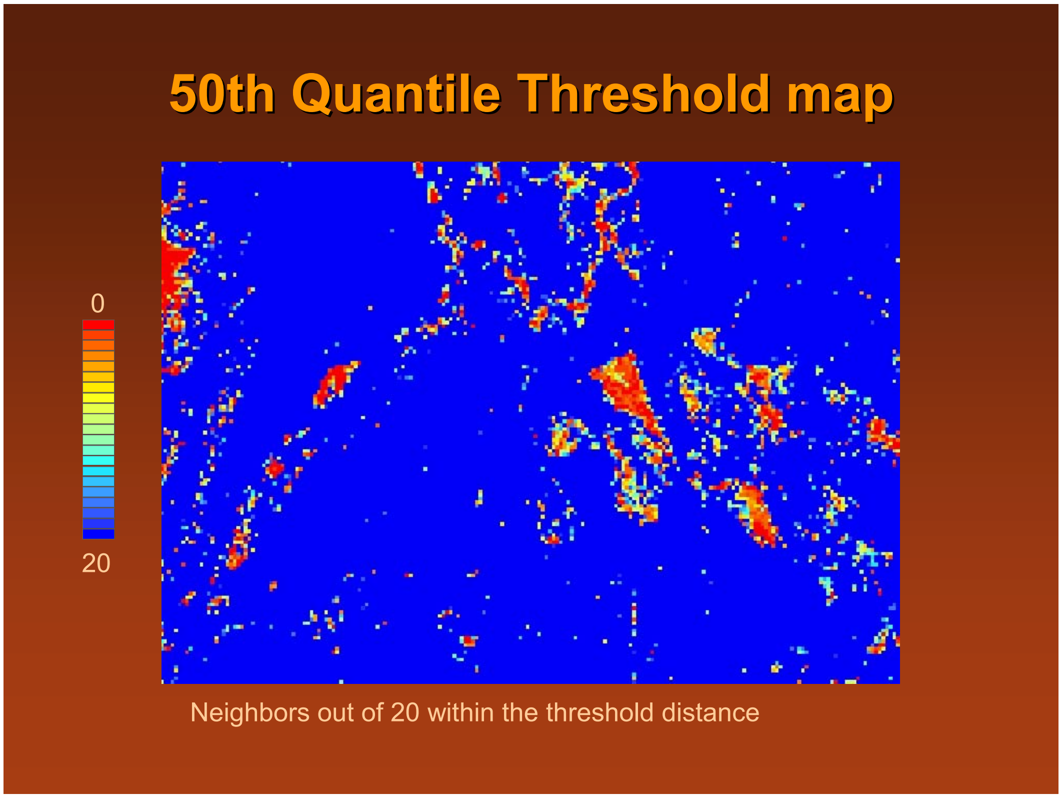
- Vegetation map
  - the predicted value
- Neighbor Count map
  - a measure of sampling sufficiency for a specific ecological location
- Natural Variability map
  - the variability in response at the most similar locations

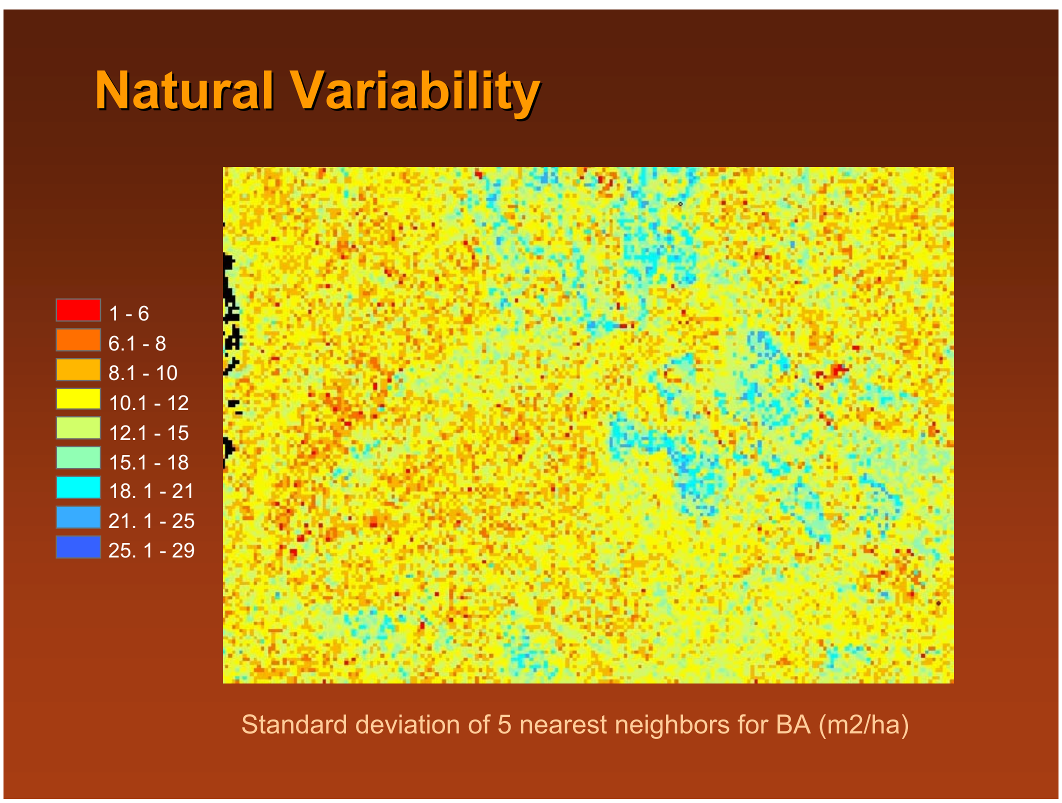

4m Aerial Photo


Expected value Basal Area m2/ha
0
61


10th Quantile Threshold map
0
20
Neighbors out of 20 within the threshold distance


20th Quantile Threshold map
0
20
Neighbors out of 20 within the threshold distance


50th Quantile Threshold map
0
20
Neighbors out of 20 within the threshold distance


Natural Variability
1 - 6
6.1 - 8
8.1 - 10
10.1 - 12
12.1 - 15
15.1 - 18
18. 1 - 21
21. 1 - 25
25. 1 - 29
Standard deviation of 5 nearest neighbors for BA (m2/ha)

Map Usage

- Expected value map
  - Normal map product
- Neighbor count map
  - Map of sampling sufficiency
- Standard deviation map
  - Natural variability within sampled condition
- Combo map

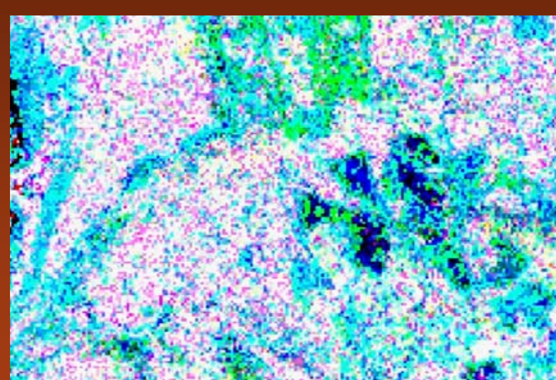
Tri-map plot

All three maps combined with different colored layers

Individual layers
Blue = many neighbors
Red = high basal area
Green = high variability

Combined Indices

Purple = high support and value



White = high support, value and variability

GNN Imputation benefits

- Recaptures most of initial variation
- Maintains multi-attribute covariance
- Provides detailed spatial data for post-mapping classification, analysis and modeling
- Spatial mapping of natural variability
- Ability to assess sampling sufficiency

GNN Imputation Limitations

- Map values are constrained to those at sampled locations
- Natural variability reduces point-level prediction accuracy
- Requires complete spatial coverages and plot inventories which simultaneously measure all attributes of interest

National Scale Applicability

- Practicality
  - Based on pre-existing inventory data and nationally available spatial data
- Robustness
  - Highly robust for analysis since final product retains all input data
  - Sensitive to completeness of model data
- Reproducible
  - All GNN code is open source. Execution and analysis require GIS and statistical software

National Scale Applicability

- Feasibility
  - Simple to implement once plot and spatial data are acquired
  - Could be piggy backed onto other analyses with proper data such as Landfire
- Difficulty
  - May require expertise in database development, GIS, multivariate analysis and remote sensing
- Transparency
  - Crystal (as much so as any multivariate output can get)

Conclusions

Single Neighbor Imputation mapping can facilitate:

1) Mapping natural variability spatially
2) Estimating variability within plant associations and locations
3) Assessing sampling sufficiency
4) Selecting new sample locations
5) Assess expectations from data/reduce uncertainty

Thanks