MAPPING FOREST ATTRIBUTES IN THE INTERIOR WEST: COMPARING PREDICTIVE MODELING TOOLS

Gretchen Moisen et al.
(presented by Raymond L. Czaplewski)

USDA Forest Service
Rocky Mountain Research Station

Over the last several years, a number of studies have been conducted in the Interior West to compare the performance of a variety of statistical tools for modeling forest attributes collected on forest inventory ground plots as functions of satellite-based information. These tools include generalized linear and generalized additive models, tree-based methods, multivariate adaptive regression splines, neural networks, kriging, stochastic gradient boosting, hybrid methods, and simple linear models. The See5 and Cubist are the recent tools of choice to map forest attributes across the interior west. They are competitive with other methods tried, inexpensive with little training and start-up time required.

Mapping forest attributes in the Interior West: Comparing predictive modeling tools

Gretchen Moisen, Jock Blackard, Tracey Frescino, Elizabeth Freeman
Interior West Forest Inventory and Analysis Program
UDSA Forest Service, Rocky Mountain Research Station

Mark Finco, Bonnie Ruefenacht
Remote Sensing Applications Center
…and lots of other people….

Overview

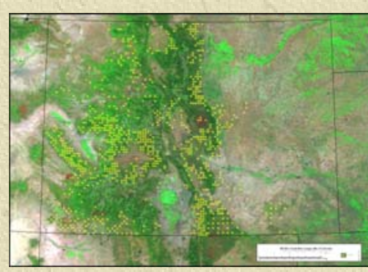
- Process for building maps
- So many tools…so little time
- Model comparisons in the Interior West
- Bottom line

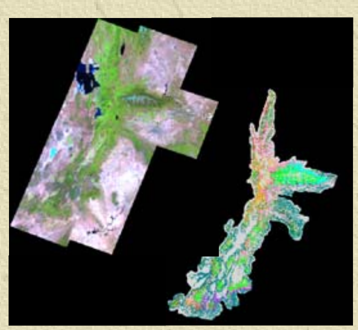
Process:
1. Extract plot-level response variables

Catagorical:
- Forest/nonforest
- Forest type

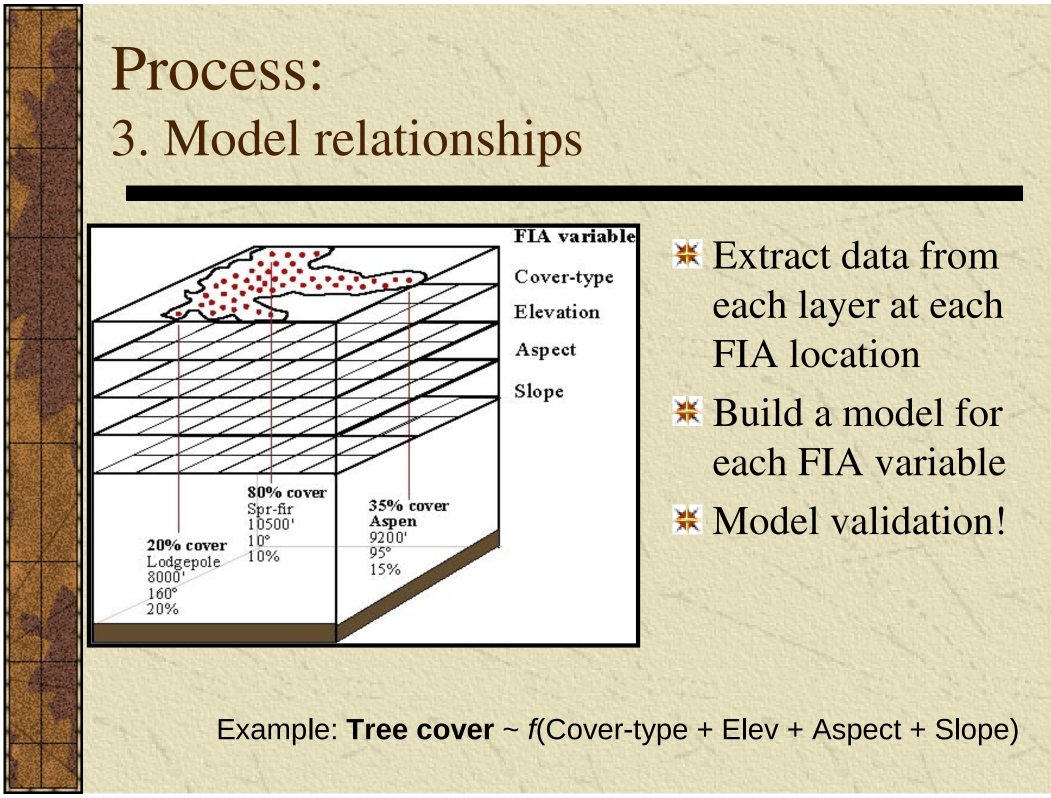
Continuous:
- Basal area
- Biomass
- Crown cover
- Canopy height
- QMD
- Stand age
- TPA
- Volume

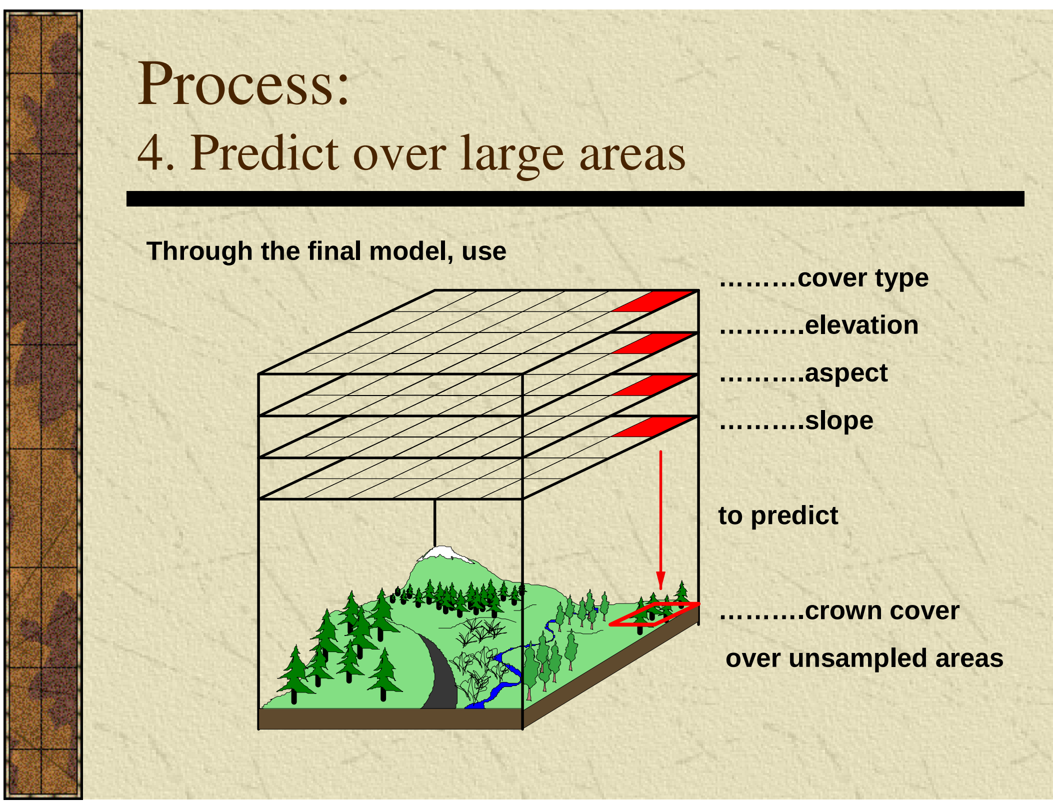
Process:
2. Library of ancillary predictor variables





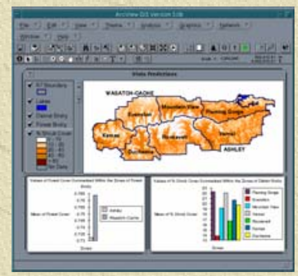
- TM-based products
- MODIS-based products
- Existing cover type maps
- Topographic variables
- Spatial coordinates


Process:
3. Model relationships
FIA variable
Cover-type
Elevation
Aspect
Slope
80% cover
Spr-fir
10500'
10°
10%
35% cover
Aspen
9200'
95°
15%
20% cover
Lodgepole
8000'
160°
20%
Extract data from each layer at each FIA location
Build a model for each FIA variable
Model validation!
Example: Tree cover ~ f(Cover-type + Elev + Aspect + Slope)


Process:
4. Predict over large areas
Through the final model, use
.........cover type
..........elevation
..........aspect
..........slope
to predict
..........crown cover
over unsampled areas

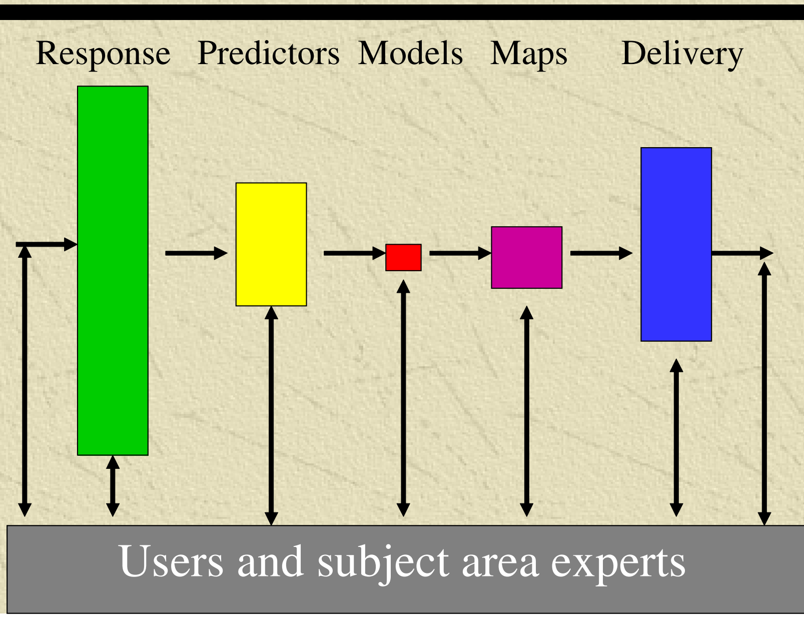
Process:
5. Web-based delivery

- Maps with uncertainty
- Tools to visualize, summarize, query
- Utility to users of map!



http://www.fs.fed.us/rm/ogden/index.html ► Techniques Research

Keeping the parts in perspective

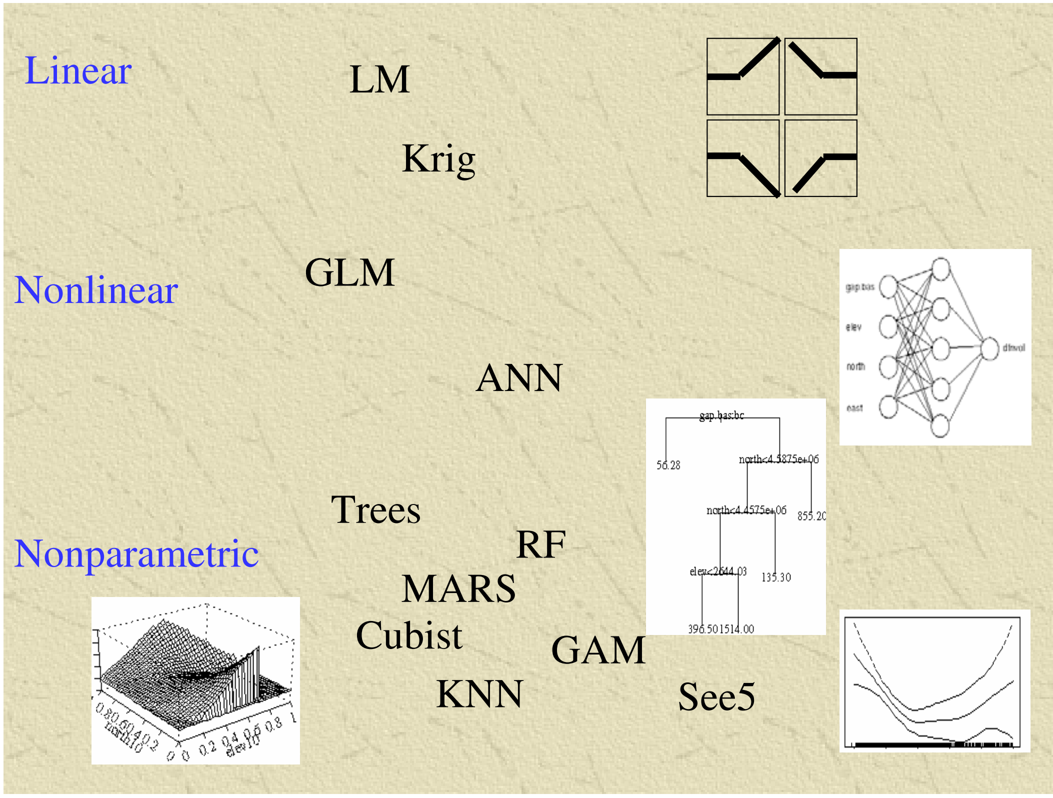


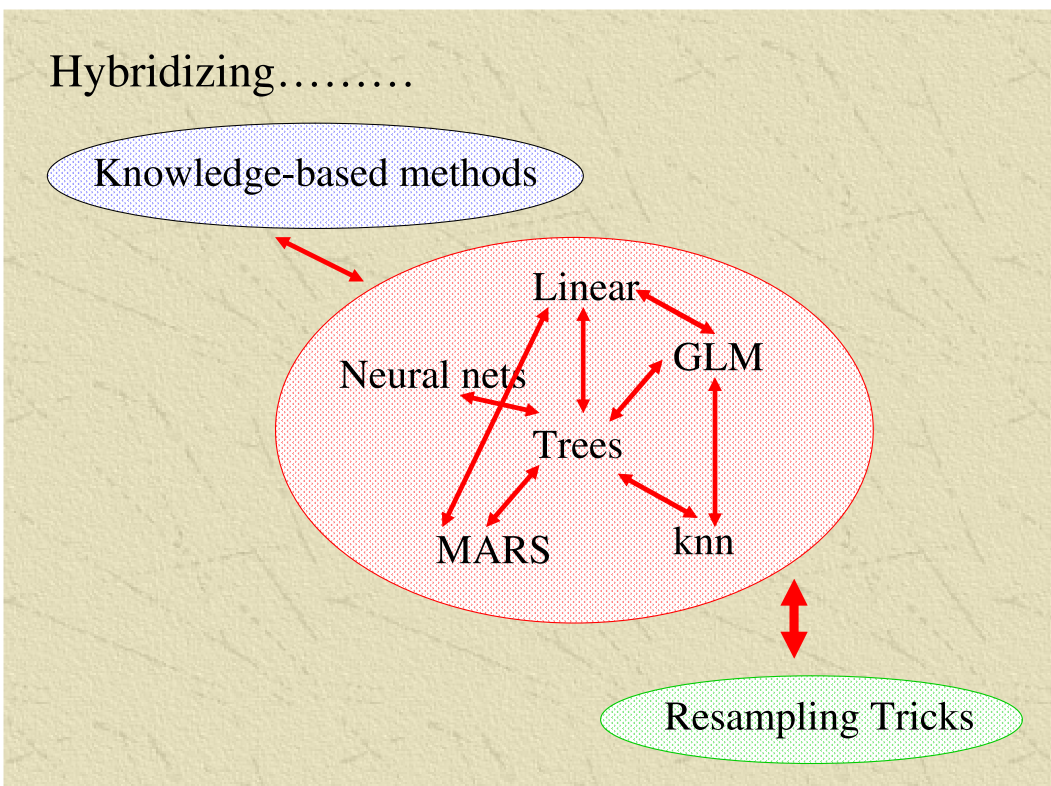
Overview

- Process for building maps
- So many tools…so little time
- Model comparisons in the Interior West
- Bottom line

First, some alphabet soup

- LM – Plain ol' linear models
- Krig - Kriging
- GLM – Generalized Linear Models
- GAM– Generalized Additive Models
- MARS – Multivariate Adaptive Regression Splines
- CART – Classification and regression trees
- SGB – Stochastic Gradient Boosting
- ANN – Artificial neural networks
- KNN – K-nearest neighbor
- SEE5 and Cubist – Proprietary variations on trees
- RF – Random Forests


Linear
LM
Krig
Nonlinear
GLM
ANN
Trees
RF
Nonparametric
MARS
Cubist
GAM
KNN
See5


Hybridizing………
Knowledge-based methods
Linear
GLM
Neural nets
Trees
MARS
knn
Resampling Tricks

Model Test Using Simulated Data

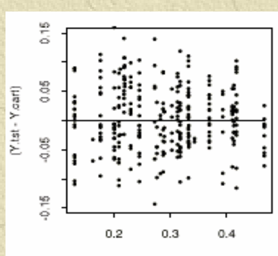
X1,…, X10 ~ Unif(0,1)

Y = 2sin(π*X1*X2) +

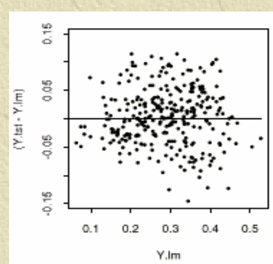
.4(X3-.5)^2 +

.2(X4) + .1(X5)

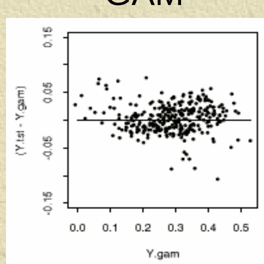
CART



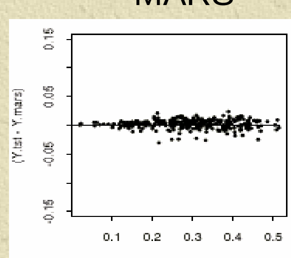
LM



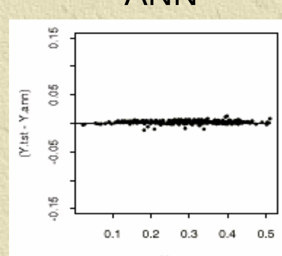
GAM



MARS

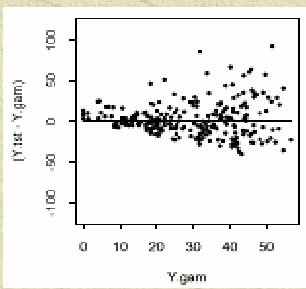


ANN

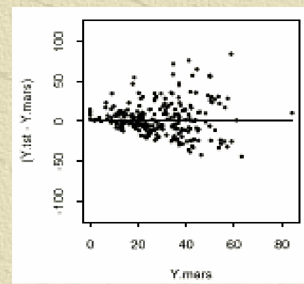


Residual Plots: BIOTOT in UT2

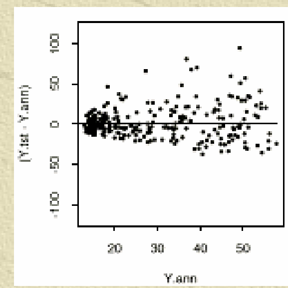
GAM



MARS



ANN



Overview

- Process for building maps
- So many tools…so little time
- Model comparisons in the Interior West
- Bottom line

Comparing: LM, GAM, MARS, CART, ANN

- Responses: Forest type group, biomass, stand age, crown cover, QMD
- Predictors: AVHRR, topography, NLCD
- Area: AZ, UT, MT
- Results: GAMs and MARS were marginally best, sometimes a simple LM worked just fine
- Ref: Moisen and Frescino, 2002

Comparing: CART, SEE5, MARS, and hybrid CART/MARS

- Response: Forest type and tree height
- Predictors: 30 m. multi-date ETM, topography, veg indices
- Area: UT
- Results: hybrid CART/MARS models were marginally best
- Ref: Moisen et al., 2003

Comparing numerous modeling options in See5 and Cubist

- Response: 12 forest attribute variables
- Predictors: 58 MODIS- and other instrument based predictors assembled for FIA/RSAC biomass and forest type modeling efforts
- Area: WY
- Results: recommendations made for production mapping currently underway in Interior West
- Ref: Blackard et al., 2004

Comparing Cubist, Kriging, and Cubist with Kriging

- Response: biomass
- Predictors: 58 MODIS- and other instrument based predictors assembled for FIA/RSAC biomass and forest type modeling efforts
- Area: Interior West
- Results: Cubist far better than Kriging, Kriging Cubist residuals contributed nothing
- Ref: Freeman and Moisen, In review

Comparing GAM, Cubist, See5, and SGB

- Response: 13 species presence, and basal area
- Predictors: ETM, indices, topography
- Area: UT
- Results: SGBs and GAMs performed better
- Ref: Moisen et al., In review

Comparing Cubist, See5, GAMs, MARS, and RF

- Response: biomass
- Predictors: 58 MODIS- and other instrument based predictors assembled for FIA/RSAC biomass and forest type modeling efforts
- Area: AZ, MT, UT
- Results: Nothing did much better than Cubist. RF did well at identifying distinctly different classes
- Ref: Moisen et al., In prep.

Overview

- Process for building maps
- So many tools…so little time
- Model comparisons in the Interior West
- Bottom line

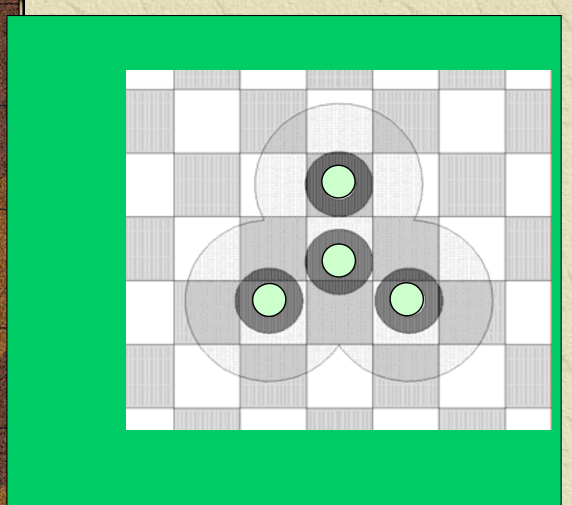
"Production-suitable" criteria

- Allows continuous/discrete predictor variables
- Automatically:
  - selects relevant predictor variables
  - transforms predictor variables
  - determines level and nature of interactions
  - handles missing values
  - protects against overfitting
- Is fast and "push button"

Bottom line

- No miracles yet. The problem is not the statistical model.

What else can we do?... Go for the heart of the problem - Inputs



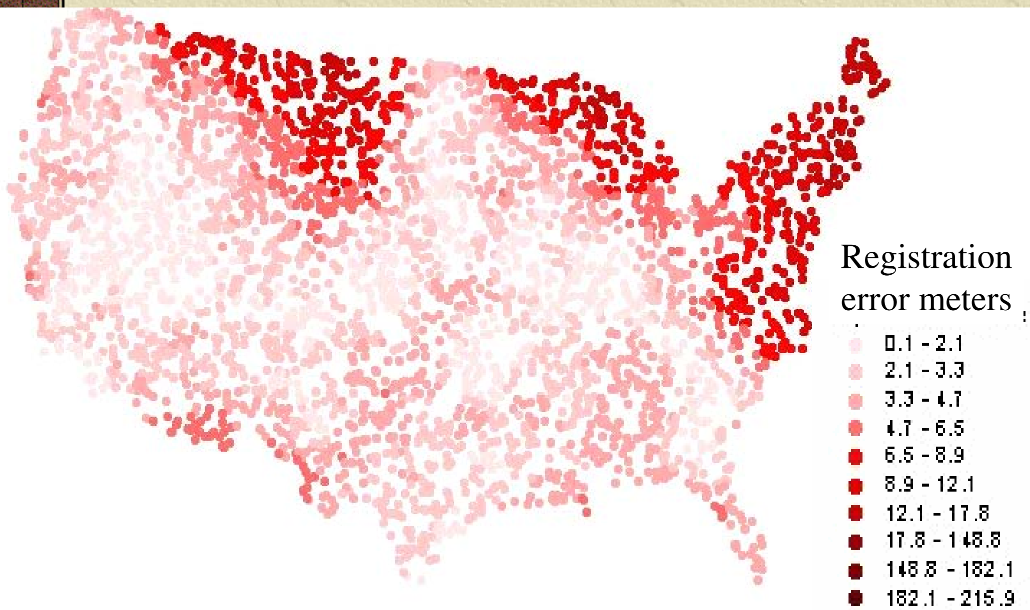
- Scale mismatch
- Plot misregistration
- Pixel alignment
- Date inconsistencies
- Unknown error in response variable
- Missing predictor variables

250 m pixel = 5.25 ha

30 m pixel = .09 ha

FIA plot ~ .60 ha

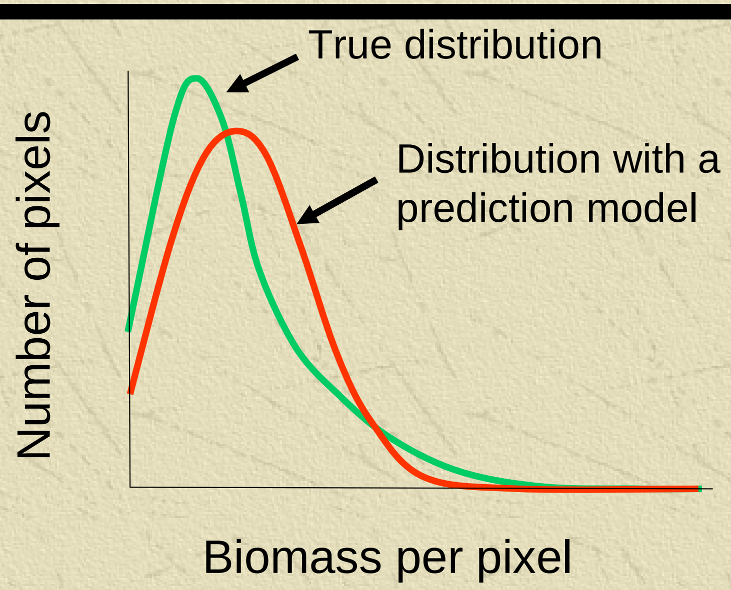
Possible systematic registration errors in GPS/GIS datum conversions



The "best model" depends on how you define errors.

- Is predictions over all pixels unbiased?
- What is the spread of the prediction errors
- Does the distribution of pixel values from the map agree with the true distribution?

The "best model" depends on how you define errors.



Some thoughts

- Modeling tools have to work over ecologically diverse areas and be automated
- Choice of appropriate statistical model is only one small part of the process
- It's inexpensive to change the statistical modeling tool, but not likely to make dramatic improvements in the map
- Real solution to better models is likely to be found working toward compatibility between responses, predictors and the final map product (e.g.,scale mismatch).
- Geographic registration errors between predictor and response variables in training data might be the biggest limiting factor
- The way in which we display the predictions should reflect the real information content (e.g., 3 v. 20 veg types).

Gretchen's picks for production maps

- See5 and Cubist coupled with RSAC's implementation tools have been our most recent tools of choice to map numerous forest attributes across Interior West
- Are they the best for predictive power based on our model comparisons? – No…but they're competitive, and not substantially worse than the others we've tried.
- Most importantly, anyone can put them into production tomorrow, very inexpensively with little training and start-up.

Recommendations

- First, consider if existing map products will meet your needs
- Second, consider simply applying stratum means to enhance an existing map product
- Don't rely on kriging unless sample points are spaced 200m or less and predictor variables weak
- Unless you find a tool that really works miracles, use stuff that's already production-ready