Nearest Neighbor Methods for Imputing Missing Data Within and Across Scales

Valerie LeMay
University of British Columbia, Canada
and
H. Temesgen, Oregon State University

Presented at the "Evaluation of quantitative techniques for deriving National scale data for assessing and mapping risk workshop", Denver, CO, July 26-28, 2005

Mapping/Assessment Problem

Measures for all variables of interest and for all scales of interest are not available

Example:

- Forested land, divided into polygons (stands, same age, species, etc.) – complete census based on photos/remote sensing
- Ground data are available for some of the stands
- Wish to "populate" the forested land with detailed information

Imputing Missing Data

Imputation involves estimating missing values for variables of interest

Many methods and variations:

- Univariate (one variable of interest at a time) vs multivariate (all variables of interest simultaneously)
- Single values or means from existing data as estimates for missing values
- Requires probability distribution or can be distribution-free
- Spatial information or variable-space?

Univariate Methods

- Sample means used to impute missing values

e.g all trees with missing heights get average height of 30 m (98 ft), regardless of their diameter

- Generate a random value from a sample estimated distribution
- Use regression or logistic models

E.g. diameter = 50 cm (20 in), predicted height= 30 m (98 ft) Trees of dbh=50 cm without measured heights assigned an estimated height of 30 m.

Issues with Univariate Methods

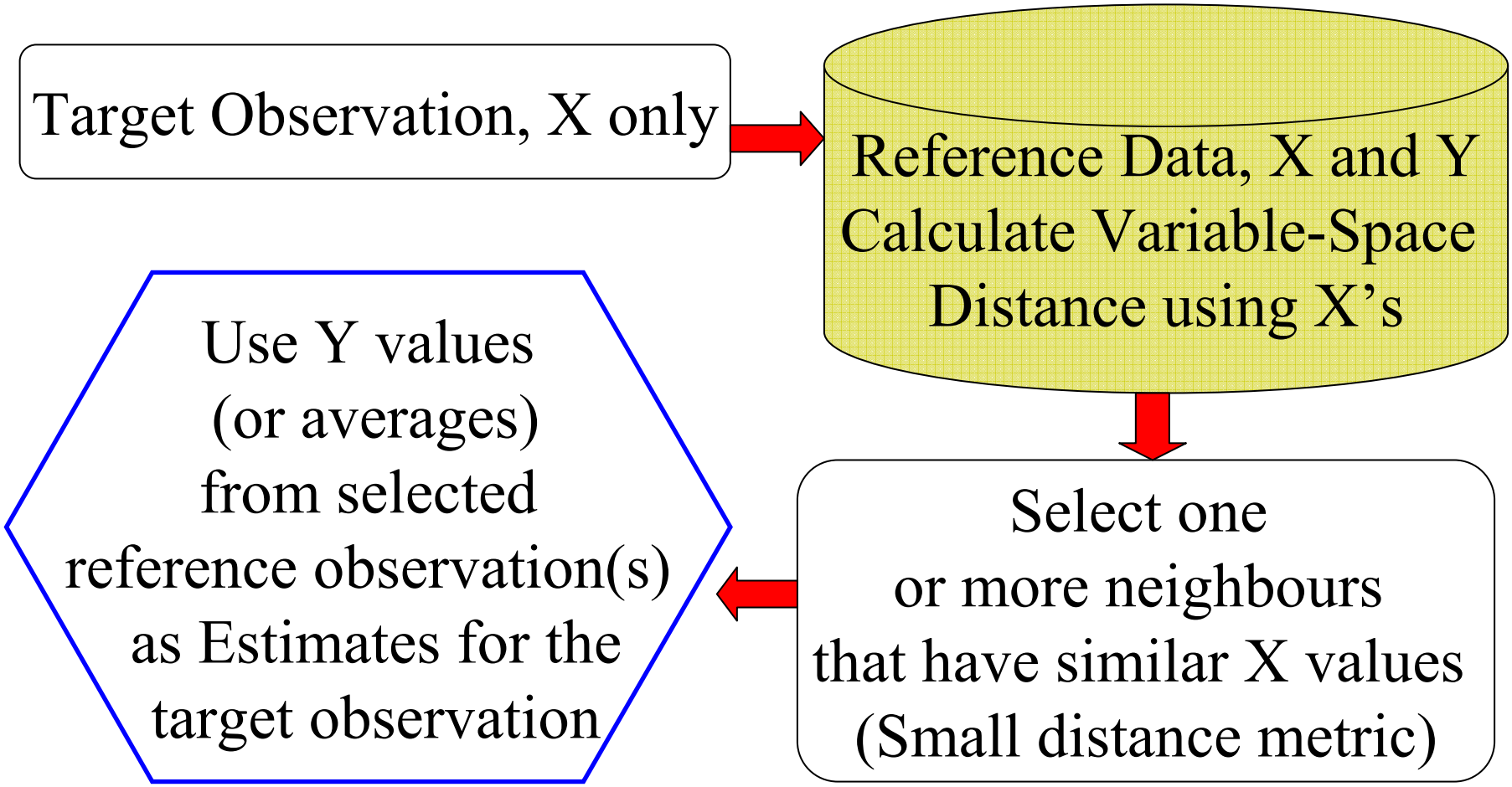
- For means and regression, variables must be ratio or interval scale
- All are unbiased and statistically consistent estimates (if models are correct)
- Only random selection from a probability distribution retains variability (means lowest)
- No assurance of logical consistency across several variables of interest

Multivariate Nearest Neighbor Imputation Methods

Data

- Obtain a sample on which X's (auxilary variables) and Y's (variables of interest) are measured [reference data set]
- Can have many Y's
- X's and Y's can be class and/or continuous variables (will affect the methods used)
- On all other observations of the population, measure the X's only [target data set]

Imputation Steps in General



Imputation: Example

For:

Use:

Distance (Similarity) Metrics

- A number of possible metrics
- Distance in variable-space
- Different measures if some are class variables

Squared Euclidean Distance

$$d_{ij}^{\,2} = (X_i - X_j)'(X_i - X_j)$$

X_i = vector of standardized values of the

X variables for the *i*th target observation

X_j = a vector of standardized values of the

X variables for the *j*th reference observation

Most Similar Neighbor Distance = Weighted Euclidean Distance

$$d_{ij}^{\,2} = (X_i - X_j)'W(X_i - X_j)$$

W = weight based on canonical correlation between X and

Y variables using the reference data

Other Distance (Similarity) Measures

- City Block
- Manhattan

} For Class Variables

- Absolute Difference

Variations

Single or Weighting of Many Reference Observations:

- Select one substitute? Or average more than one? Weighted or unweighted average?
- Affects degree of "smoothing" of estimates

Pre-stratification or not?

- E.g., by ecozone? By region?

(Single) Nearest Neighbor (NN)

- Select the closest reference observation (smallest distance)
- Values for all Y variables from the nearest neighbor are the estimates for the target observation
- E.g., Moeur and Stage used NN with their distance metric, Most Similar Neighbour

Tabular Nearest Neighbor

- Stratify reference data into groups
- Calculate variable averages (tables) by group
- Calculate similarity for X variables between a target observation and table averages
- Select the closest table
- Use the table average values for the Y's as the estimates for the target observation

k-Nearest Neighbors (k-NN) and Weighted k-NN

- Select the *k* most similar observations from the reference data
- Average the values for all Y variables from the *k*- nearest neighbors; averages are the estimates for the target observation
- For weighted k-NN, calculate a weighted average of the k-neighbors (e.g., 1/distance as the weight); weighted averages are the estimates for the target observation

Properties: Not Necessarily Unbiased

Over all samples, the mean bias (bias = average difference between observed and estimated value) does not necessarily equal zero for Y or X variables

- For Y: match is based on X variables, not Y
- For X: match may have lowest distance, but not the lowest difference, and compromised among variables

Properties: Bias Example

Target: X1=2 X2=4

Reference 1: X1=0 X2=4 Y1=10 Y2=5

Reference 2: X1=1 X2=3 Y1=7 Y2=4

Ref. 1 better for X2 (squared Euclidean distance of 4)

Ref. 2 better for X1 (squared Euclidean distance of 2)

Properties: Not Necessarily Statistically Consistent

- The average distance between target and match observations tends to decline with increasing sample size (**more likely to find a close match**)
- But mean bias will not necessarily decline with increasing sample size
- Why? Variables that are "hard to find a match for" influence the distance more

  e.g. X1=300 X2=10 Will try to find a match for the extreme X1 value and sacrifice X2.

Properties: May Retain Variability

- Retains the variability of the variables over the population if a single neighbor is used to impute missing values of a target observation
- If many neighbors are selected (k-NN) variation is not retained
  - similar to regression and other models, except that this is multivariate

Properties: Logical Consistency

- Logical consistency across several variables if using one neighbor
  - the combination of variables must exist in the population
- Using averages of many nearest neighbors: some logical inconsistencies may arise

  e.g., volume by species – Ref. 1 has pine and aspen and Ref. 2 (next closest) has larch and spruce. Average will have all four species

Other Properties

- Computationally Intensive: Need similarity between the target observation and each of the reference observations
- Generally, better correlations between the X's and the Y's yield better imputation results
- Multivariate Estimation: can obtain estimates of all the Y variables simultaneously
- Variables of interest can be class or continuous variables or mixed
- Distribution-free

Selecting a Nearest Neighbor: Demonstrations of Issues

Photo 1

Q. 1
Want Coarse Woody Debris and Snags for **Photo 2**

Photo 3

Photo?
X-Variables?

Photo 2

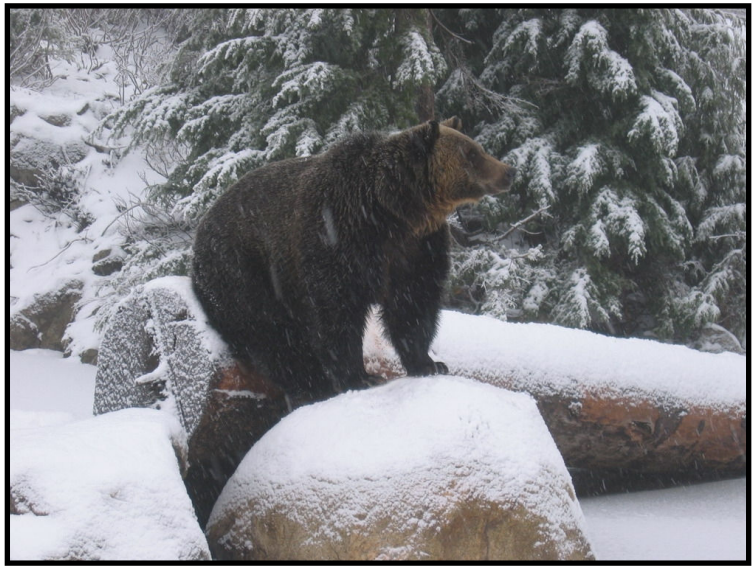


Photo 4 (Yikes!)

Observations

- May be very difficult to obtain the reference data you need
- X-variables matter

Photo 1

Q. 2

Photo 3

Want soil moisture/ nitrogen for **Photo 3** Photo?
X-Variables?

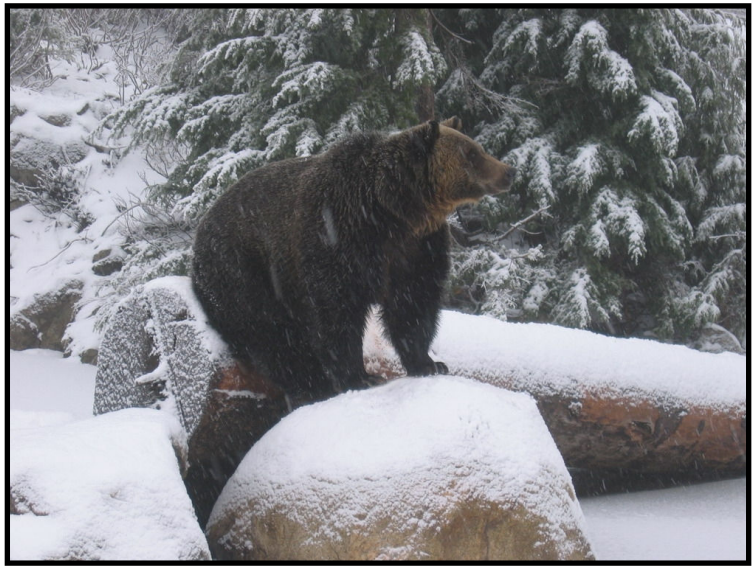


Photo 2

Photo 4

Observations

- Stratifying by location should be considered
- For some variables, time of year when measures are taken are important

Research into Forestry Applications

Examples and Results of Testing Using Simulations

- Tree-lists: X-stand level; Y-tree level
- Regeneration: X-overstory; Y-understory, both at stand-level
- Other Applications:
  - Volume and basal area per ha: X-aerial variables; Y-ground variables both at stand-level (Forest Science Paper)
  - Wildlife Trees: X-stand level; Y-tree level (Conference Proceedings)

Estimating Tree-Lists

- A tree-list (stems per ha by species and diameter) for every polygon would be useful
  - for projecting future stand volume, and
  - for estimating current and future stand structure, as inputs to habitat models
- Can we obtain reasonable estimates of tree lists for non-sampled polygons, based on aerial information?

Data

- 96 polygons were ground-sampled using variable radius plots (Y)
- Up to 9 species in a polygon with a wide diameter range
- Aerial variables (X) were matched to the ground data

Variable Set

Y variables (7):

- basal area/ha
- stems/ha of Douglas fir(D), larch (L), and lodgepole pine (PL)
- Max. dbh of F, L, and PL

X variables (8)

- Percent crown closure
- Average height (m)
- Average age (yrs)
- Site index (m)
- Percents of F, L, and PL by crown closure
- Model estimated volume/ha (stand level model)

Methods:

- SAS 6.12 used to simulate sampling the population (100 replicates)
- Three sampling intensities (20%, 50% and 80%)
- Two imputation methods used: Tabular and Most Similar Nearest Neighbor (NN with MSN Distance)

Correlations Between Ground and Aerial Variables

- Highest for stems per ha of fir (Y) with model estimated volume per ha (X) (about 0.40)
- Lowest for Maximum dbh of larch (Y) with crown closure class (X) (less than 0.01)

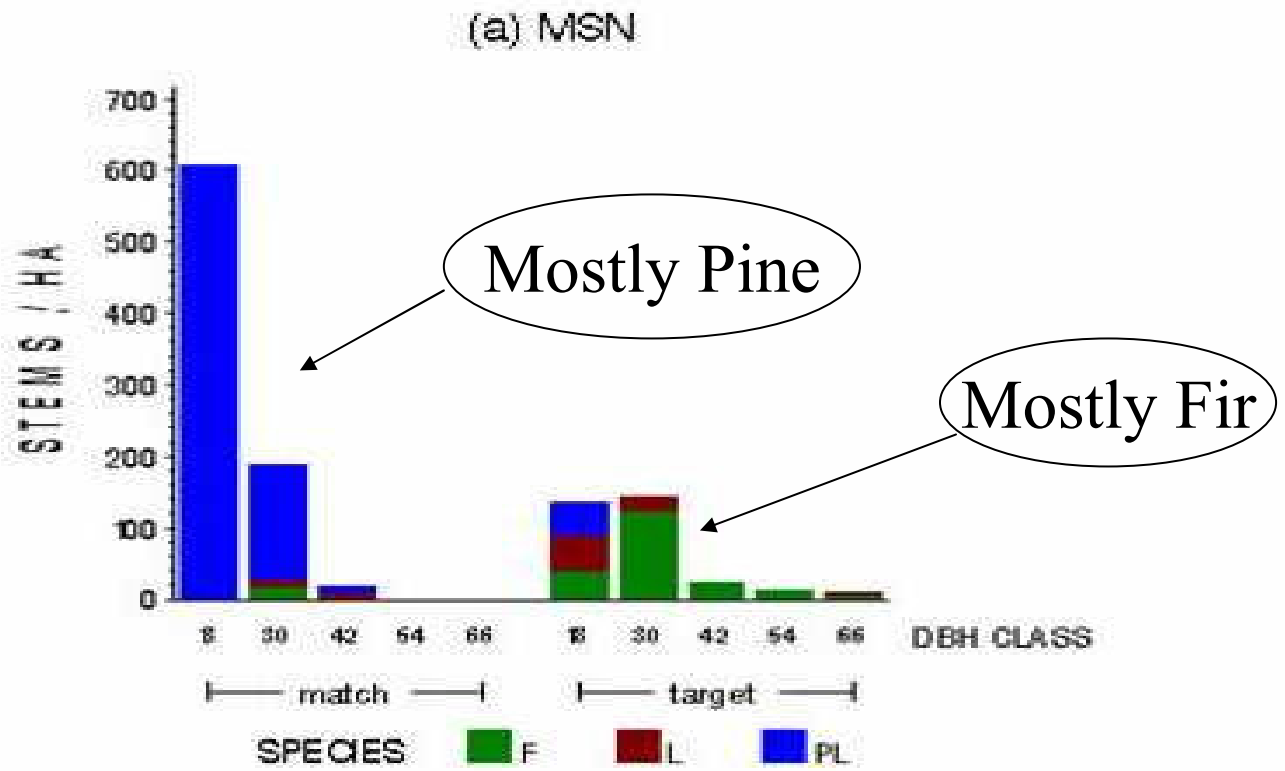
Results Over 100 Replications

- Average correlations between targets measured and imputed variables:
  - For X: Increased as sample size increased
  - For Y: Generally increased with sample size but not for all variables (e.g., decreased for stems/ha larch using MSN)

Results Over 100 Replications

- Mean Bias (average difference) for Y:
  - Generally lower for Tabular than MSN
  - Not declining with increasing sample size
- Mean of Mean Squared Errors for Y:
  - Declined with increasing sample size for most variables
  - MSN and Tabular similar

Example of Target and Match Polygons (80% Sampling Intensity)



Estimating Regeneration Under an Overstory After Partial Cutting

- Stands are multi-species and multi-aged, partially cut; measure overstory variables (X)
- Want to estimate the amount of regeneration (Y) expected to occur following partial cutting
  - Regeneration by 4 species groups by 4 height classes and all very related
- Tabular and MSN (NN with Most Similar Neighbor Distance)

Tabular Imputation: E.g., Dense, Dry (n=18), <6 years after cutting (stems/ha)

Species	Height (cm)				Total
	15-49.9	50-99.9	100-129.2	>130	
Tolerant	3921	1032	454	495	5903
Semi-tol.	2889	949	372	578	4788
Intolerant	1197	41	41	0	1280
Hardwood	454	248	248	743	1692
Total	8462	2270	1115	1816	13663

Imputation Accuracy Over Cells

Match: Presence of regeneration in both the target

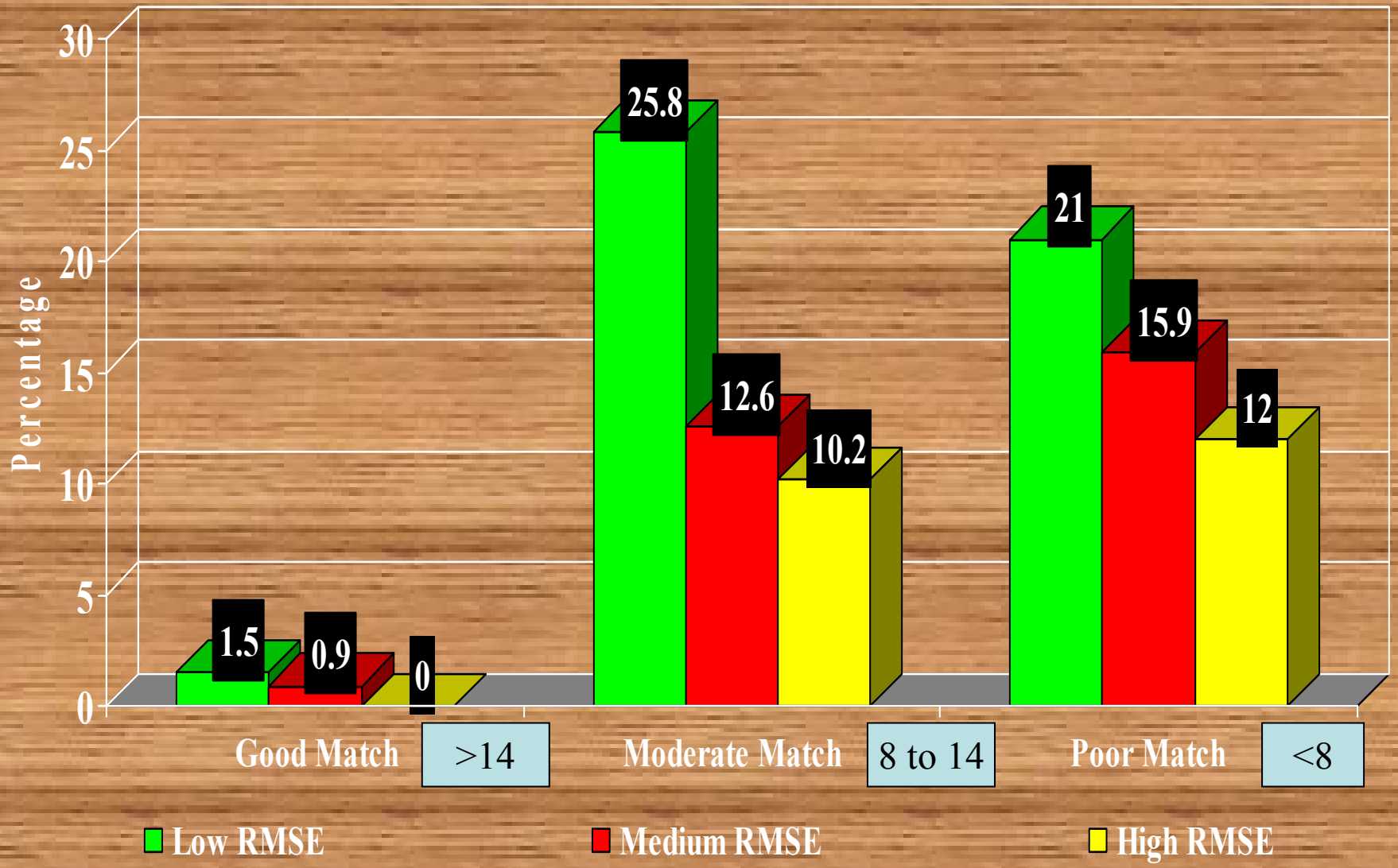
Good (>14 cells matched)	moderate (>8 to 14)	poor (<8)

Grouped plots also by root mean squared error

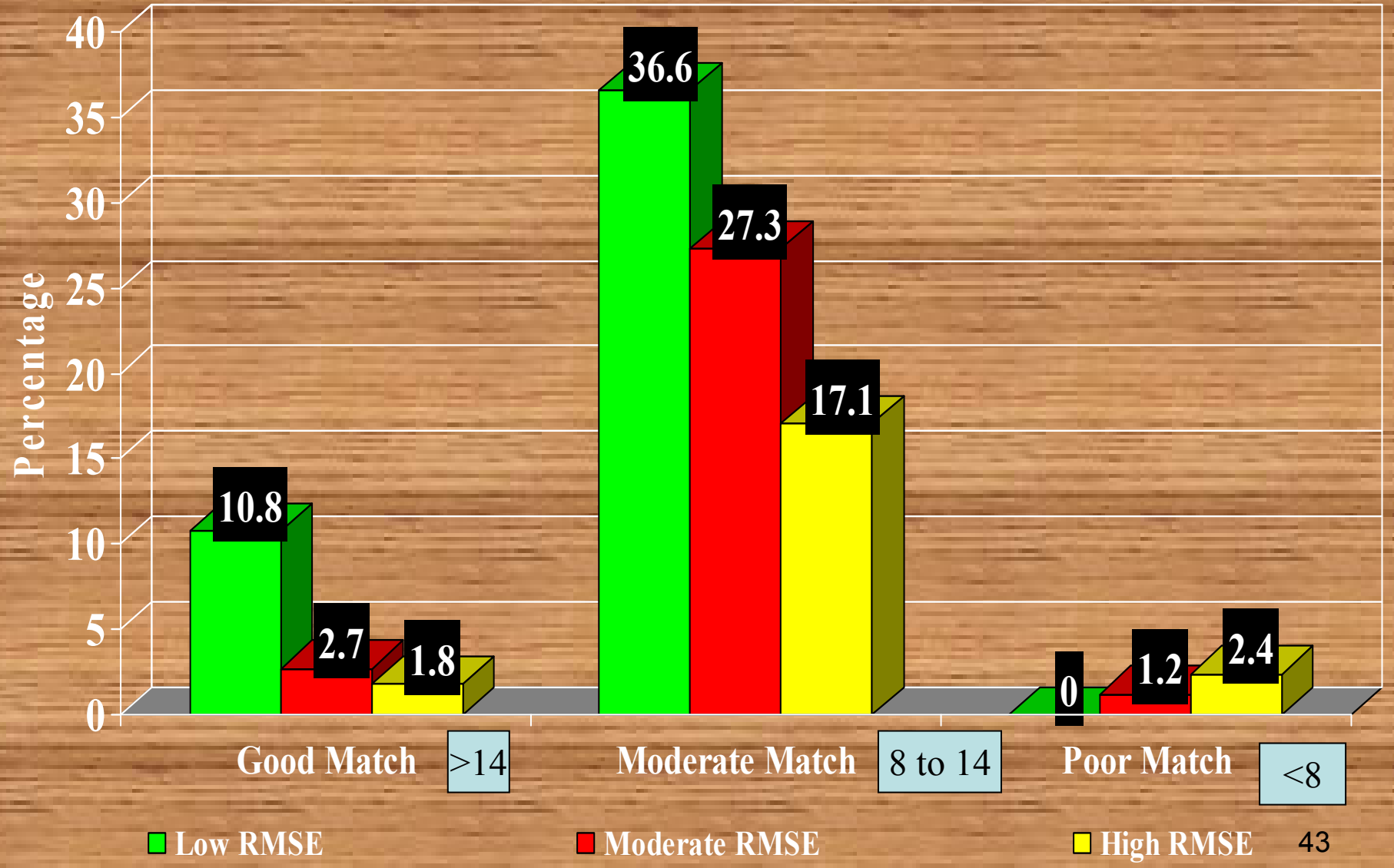
low (<1000 stems per ha, all species)	
moderate (1000-2000)	high (>2000)

Want Good, Low

Performance of Tabular



Performance of MSN



Comparison of Approaches

- Better estimates using MSN
  - MSN uses a single nearest neighbor – variability and logical consistency retained
  - Tabular can be considered “smoothing” (k-NN also is smoothing) – for this problem, too much “smoothing” likely

Summary for Imputation Methods

- Imputation methods are used to fill in missing data for variables of interest across and within scales
  - Can be used to “fill in” data needed for long term monitoring, such as within stand details needed for risk mapping
- Many methods and variations on methods

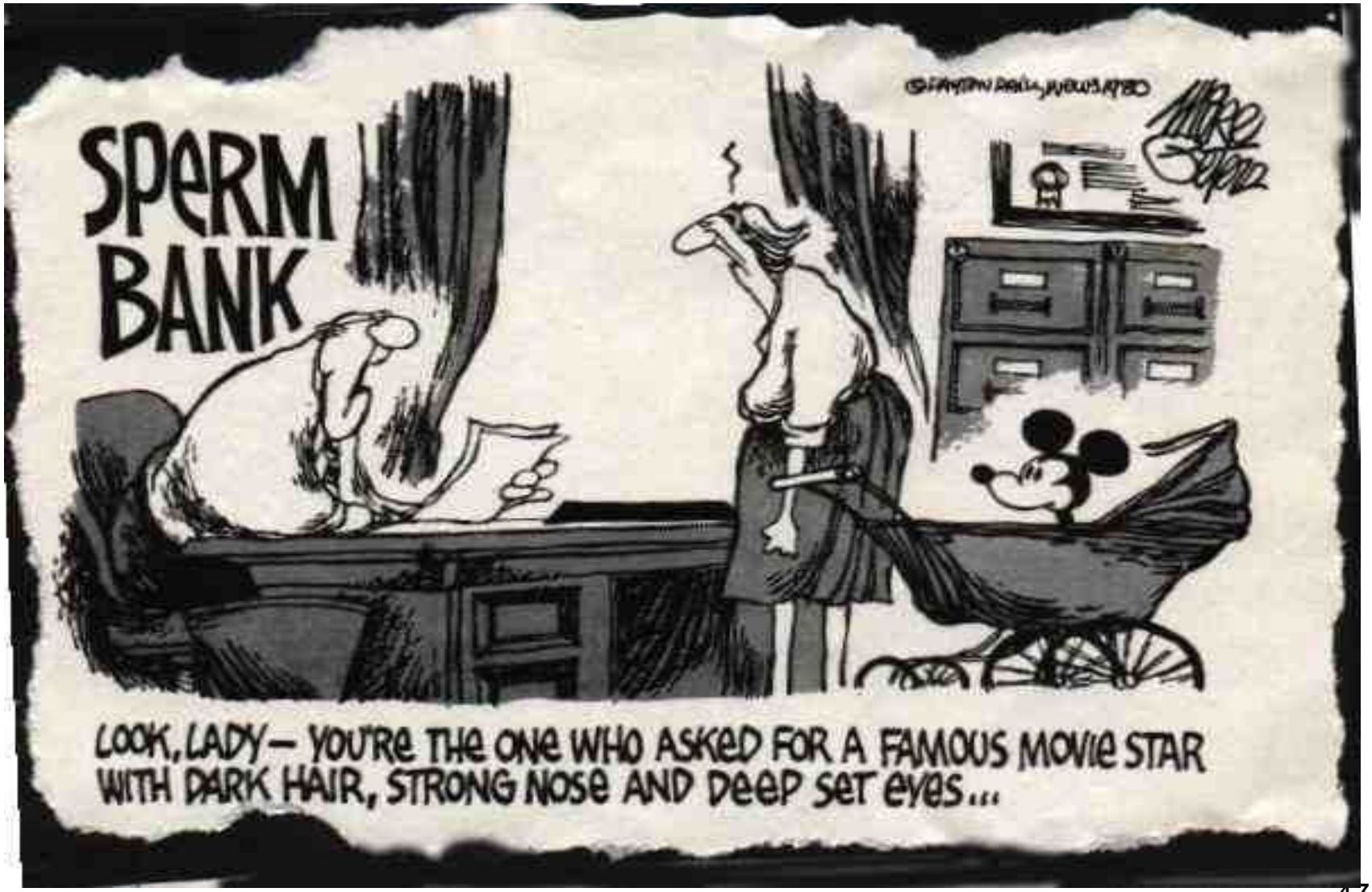
Summary for Imputation Methods

Nearest neighbor methods

- are multivariate and distribution-free
- can retain logical consistency and variation
- can be used for class or continuous or mixed variables of interest

- Degree of "smoothing" – from single nearest neighbor to k-NN to Tabular – can adversely affect accuracy of results
- Need a "good" set of reference data, with auxiliary variables that are well related to variables of interest

X-variables matter



Websites and Acknowledgements

Articles:

www.forestry.ubc.ca/Prognosis

www.forestry.ubc.ca/biometrics

NN Software (website given on the Abstract also):

forest.moscowfsl.wsu.edu/gems/msn.html

Thank you to the organizers for inviting us to present at this workshop. Funding for this research was provided by Forest Renewal BC, NSERC, and Forestry Investment Initiative