
COMPARISON OF NEAREST NEIGHBOR METHODS FOR ESTIMATING BASAL AREA AND STEMS PER HECTARE USING AERIAL AUXILIARY VARIABLES

Valerie LeMay, Univ. of BC
H. Temesgen, OSU

Commonly, forested lands are divided into polygons based on forest type. Information for each polygon often includes variables that are measured on aerial photographs (e.g., species composition, height class), and additional variables derived from the aerial attributes using yield or other models (e.g., estimated volume per ha). For detailed information, such as the amount of coarse woody debris, stand structure, or tree-lists (stems per ha by species and diameter), ground sampling of every polygon is usually not possible. However, this information would be useful to represent the current inventory, and as model inputs to project future conditions. In stands that are sampled, detail at lower scales is often also of interest, but this may be available only for some of the sampled stands because of high measurement costs. Also, for recently cut stands, particularly, partially cut stands, estimates of future regeneration are needed as inputs to growth models.

For estimating variables of interest, imputation approaches are used as alternative to regression approaches. Imputation involves substituting, plausible measurements from one or more selected units with similar characteristics to units lacking these measures (Rubin 1987, Ek *et al.* 1997, McRoberts 2001). Data with all variables measured are termed “reference data”, whereas data with some variables missing are termed “target data”. If only one selected unit is used in the substitution, the variability of the missing variables as represented in the reference data will be preserved in the estimates imputed to the target data (Moeur and Stage 1995, Ek *et al.* 1997, Haara *et al.* 1997, Maltamo and Kangas 1998, Moeur 2000, LeMay and Temesgen 2005). This differs from regression approaches where averages, conditional on the values of the predictor variables, are used as the estimates for the missing variables in the target data. Since imputation involves searching the reference dataset for a “match” to the target dataset, this can be computer intensive. However, software has been developed by Moeur and Stage (1995), and updated by Moeur (2000). The most recent version of this software (MSN version 2.12; 2003¹) is quite easy to use.

We have used imputation methods to: 1) estimate regeneration after partial cutting in mixed-species and/or uneven-aged stands (Hassani et al. 2004); 2) estimate tree-lists from aerial variables (forest cover) (LeMay and Temesgen 2001; Temesgen *et al.* 2003); 3) estimate wildlife trees (dead standing, or recently dead and fallen tree (Temesgen and LeMay 2001), and 4) to estimate stand level ground variables from aerial variables (LeMay and Temesgen 2005). We also have used simulation to compare different imputation methods (measures of similarity and numbers of reference plots used in imputation) for different sampling intensities for the reference data (LeMay and Temesgen 2005).

In this presentation, we present background information on imputation methods and demonstrate how these methods are employed to generate tree-lists from aerial attributes for non-sampled polygons, and improve forest inventories, analyses, and management. Examples are then given from our work using data from multi-species and multi-aged stands from southeastern British Columbia.

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

121

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

2



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?

3

122



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.

4

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

5

123

Multivariate Nearest Neighbor Imputation Methods

6

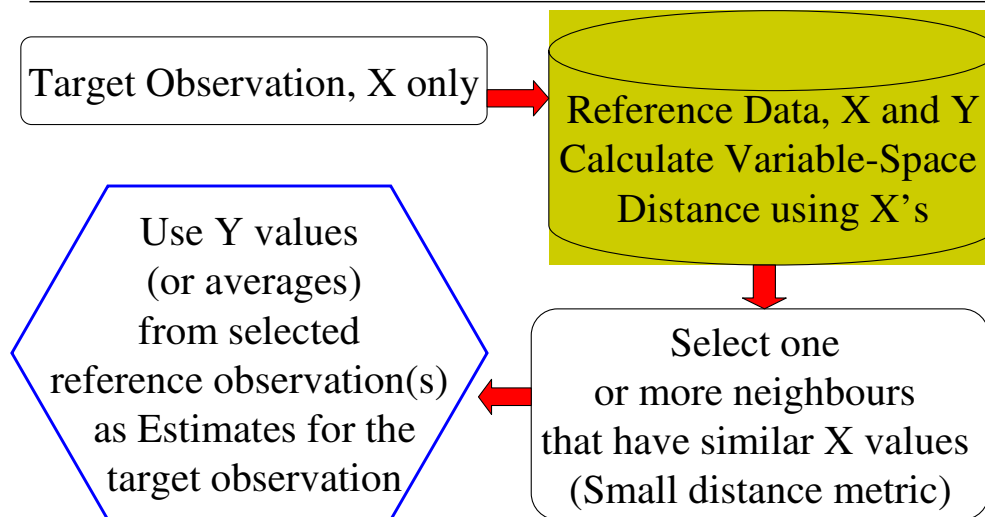
Data

- Obtain a sample on which X's (**auxiliary 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**]

7

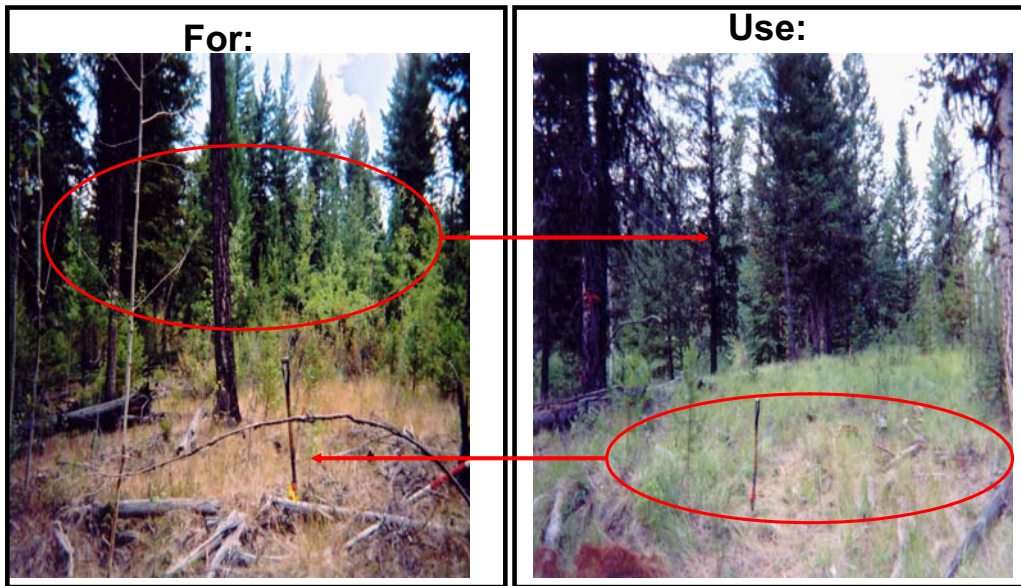
124

Imputation Steps in General



8

Imputation: Example



125

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_j = vector of standardized values of the

X variables for the i th target observation

= a vector of standardized values of the

X variables for the j th reference observation

11

Note: Does not use correlation between X and Y in determining weights

126



Most Similar Neighbor 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

12

Note: Does not use correlation between X and Y in determining weights

Other Distance (Similarity) Measures

- City Block
 - Manhattan
 - Absolute Difference
- } For Class Variables

13

127

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?

14



(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

15

Note: Expect that: 1. MSN better; 2. Largest sampling intensity better; 3. Larger variable set better.

128



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

16

Note: Expect that: 1. MSN better; 2. Largest sampling intensity better; 3. Larger variable set better.

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., $1/\text{distance}$ as the weight); weighted averages are the estimates for the target observation

17

Note: Expect that: 1. MSN better; 2. Largest sampling intensity better; 3. Larger variable set better.

129

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

18

Properties: Bias Example

Target: $X_1=2$ $X_2=4$

Reference 1: $X_1=0$ $X_2=4$ $Y_1=10$ $Y_2=5$

Reference 2: $X_1=1$ $X_2=3$ $Y_1=7$ $Y_2=4$

Ref. 1 better for X_2 (squared Euclidean distance of 4)

Ref. 2 better for X_1 (squared Euclidean distance of 2)

19

130

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. $X_1=300$ $X_2=10$ Will try to find a match for the extreme X_1 value and sacrifice X_2 .

20

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

21

131

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

22

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**

23

132

Selecting a Nearest Neighbor: Demonstrations of Issues

24

Photo 1



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

Photo 3



Photo? X-Variables?





Photo 4 (Yikes!) 25



Note: Photo 4 looks the closest — big log on the ground. But need to measure this site—bear in area? Photo 1—trees too small, but undergrowth more similar? Photo 3—very dry and not much CWD.

133

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



27

Note: Reference Data: Photos 1, 2, 4: poor match (Actually, Photo 3 is from Switzerland, Photos 2 and 4 are BC Coast, and Photo 1 is Alberta pines) Photo 3 looks very dry, compared to Photos 1, 2 and 4. Photo 1 might be the most similar, but hard to tell as this is “measured” in winter – both pine plantations, but Photo 1 is a young plantation.

134



Observations

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

Research into Forestry Applications

29

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)

30

135



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?

31

Note: Have tree-level models such as MGM, Prognosis calibrated to BC, TASS.

136



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

32

Note: Data supplied by BC MOF

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)

33

Note: More species specific

137

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)

34

Note: Expect that: 1. MSN better; 2. Largest sampling intensity better; 3. Larger variable set better.



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)

35

138



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)

36

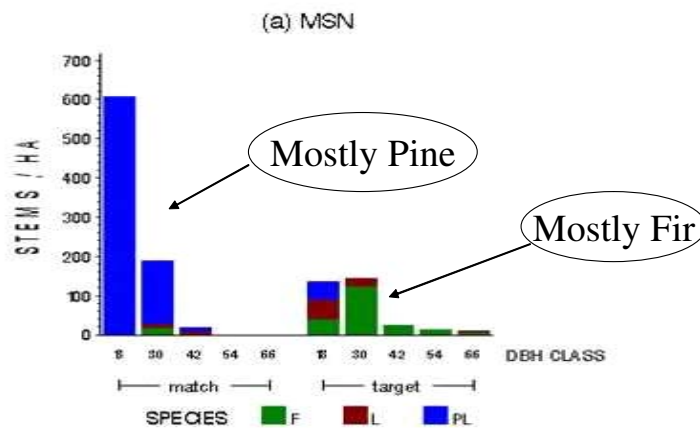
Results Over 100 Replications

- 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

37

139

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



38

Note: Those examples that did not result in the same match polygon for the two methods

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)

39

140

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

40

Imputation Accuracy Over Cells

Match: Presence of regeneration in both the target

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

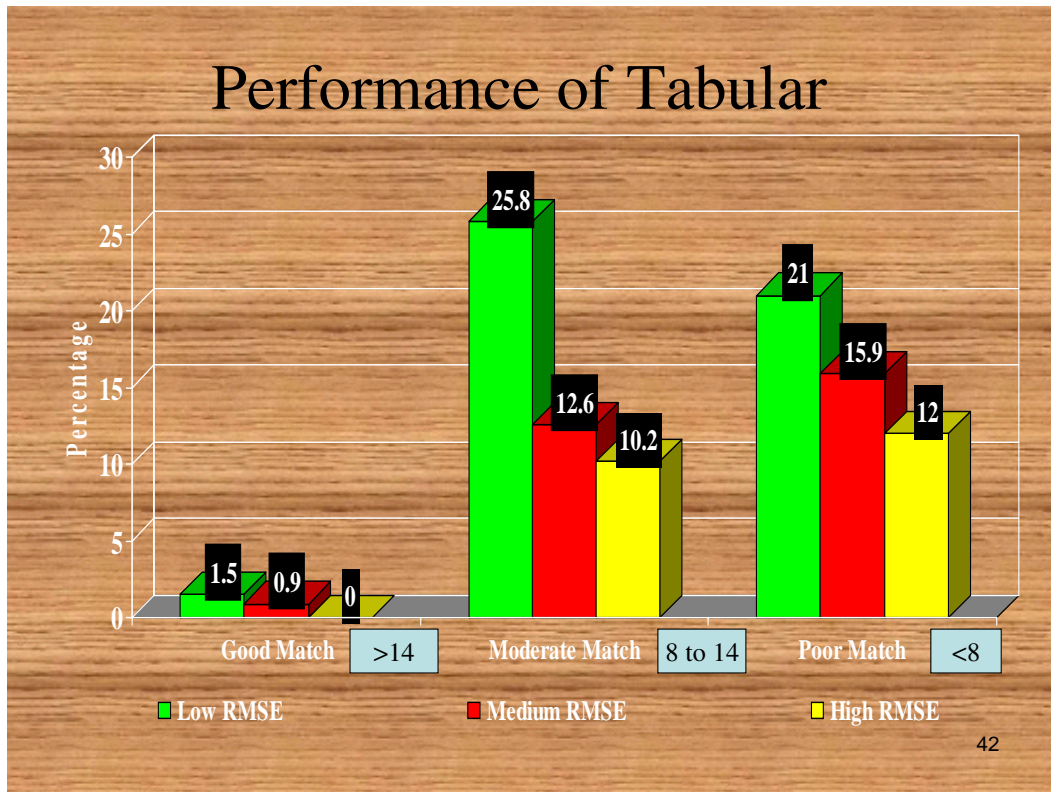
Grouped plots also by root mean squared error

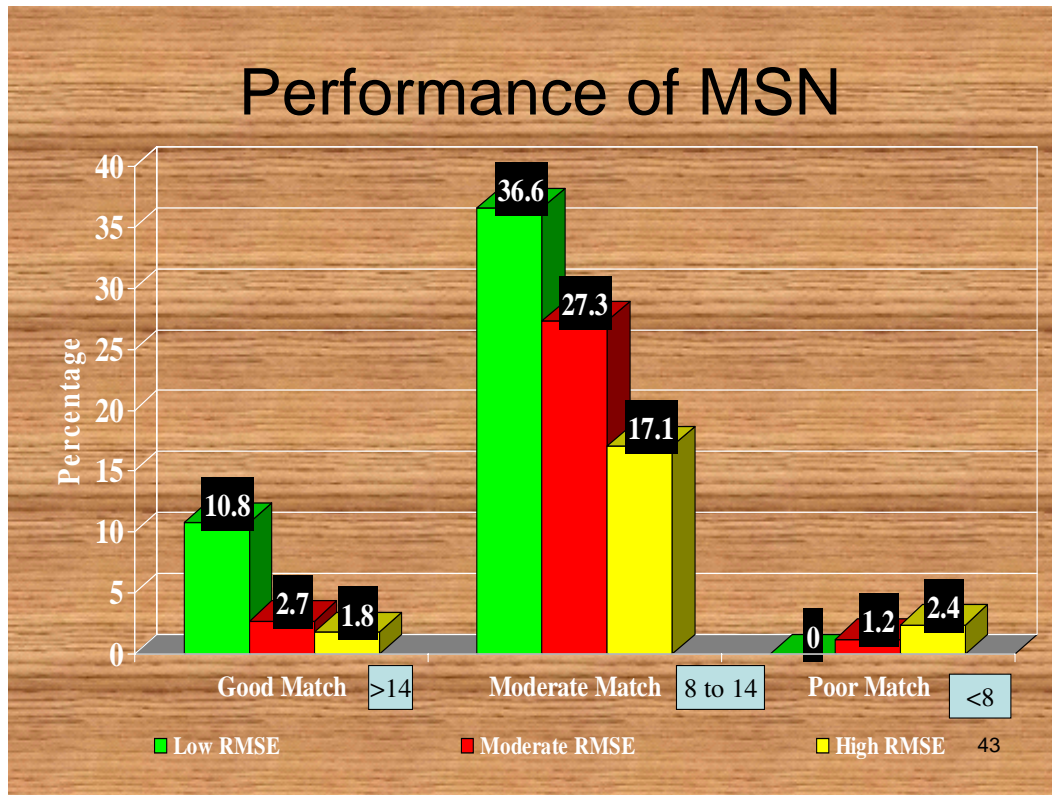
low (<1000 stems per ha, all species)
moderate (1000-2000) high (>2000)

Want Good, Low

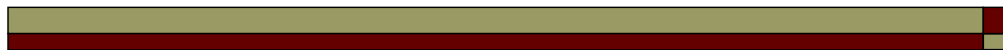
41

141





142



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



45

143

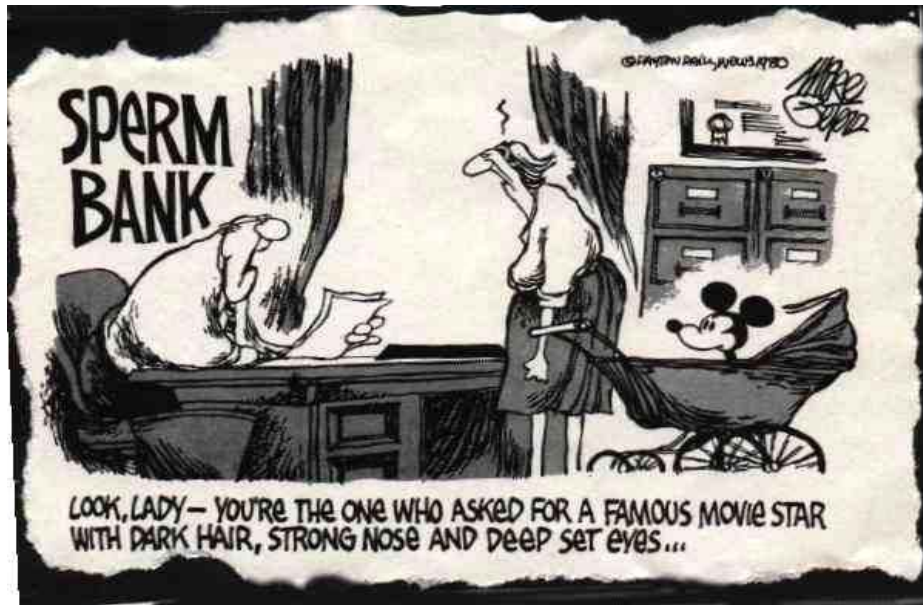
Summary for Imputation Methods

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

46

X-variables matter



47

144

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

48