
SOME NEW TWISTS IN THE ART AND SCIENCE OF IMPUTATION

Albert R. Stage

USDA Forest Service

Imputation is applied for two quite different purposes: to supply missing data to complete a database for subsequent modeling analyses or to support estimates of sub-population totals. Error properties of the imputed values have different effects in these two contexts. We develop a partitioning of the errors of imputation. Two new statistics based on this partitioning facilitate comparison to other means of estimation and of alternative methods of imputation such as k-nn, Most Similar Neighbor (MSN) and choices among their measures of similarity. We apply this partitioning to evaluate an alternative derivation of the canonical-correlation-based weight matrix in MSN using three extensive data sets from western North America.

173

Some new twists in the art and science of imputation.

Albert R. Stage

Rocky Mountain Research Station
Moscow, Idaho

174

Imputation:

- To use what we know about “everywhere” that may be useful, but not very interesting- the X’s,
- To fill in detail that is prohibitive to obtain, except on a sample- the Y’s,
- By finding surrogates based on similarity of the X’s.

Some caveats

- Objective for filling the data base is to impute a value as close to “truth” for each sample unit as if it were examined in great detail for all relevant attributes.
- This is NOT the same as its expected value in estimation.
- The difference (pure error) is of great value to the subsequent analyses.

175

. . . . and one more!

Imputation is not the same as classification—There is no concept of “membership”.

Attributes of Near-neighbor Methods for Imputation

- Measure of similarity
 - Form of distance measure
 - Variables used
 - Weights given to variate differences between sampled (known) and unsampled (unknown) units in distance measure
- Number of sampled units to be imputed to unknown unit.
 - Weights of sampled units if $k \geq 2$

176

Puzzles posed by users:

“I used **k-nn** for a single variable to be predicted—and its error was more than the error from linear regression. Why?”

“If **MSN** is based on a linear model, why aren’t imputations improved by introducing transformations that would improve the fit of the linear model?”

Key question--

- How to choose among analytical methods to attain the best imputations?
 - Pure error (desirable) confounded with “quality” in the usual statistics.
 - Data for the units in reference set not without reproach (measurement error - undesirable).
 - Data distribution problems obscure analytical differences

Topics

- Components of imputation error
- Measures of similarity (a few in particular)
- Alternative MSN similarity measures evaluated using error components
- Transformations of variables to improve resolution

Components of imputation error

- From natural variability
 - Pure error from sources NOT correlated with the selected X's
- From analytical procedures
 - Choice of the X's and Y's
 - Sampling error of unit values (X's and Y's)
 - Distribution of target set in relation to reference set of units
 - Spans of the two sets
 - Density of data points within the reference set
 - Choice of distance function

178

Components of imputation error

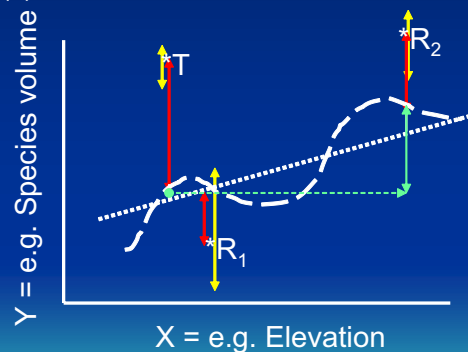
Observation error ε_Y

Pure Error ε_p

Lack of fit $g(X) - B'X = \varepsilon_{L(x)}$

Distance to Reference
 $g(X_T) - g(X_R)$

— — "Truth" $g(X)$
 Linear model $B'X$



If then

- Pure error too large
- Distance component large
- Lack-of-fit large
- Look for additional X's
- Reduce measurement error
- Add data to improve distribution
- Try alternative similarity measures
- Look for linearizing transformations

179

Some notation:

- x_i = observed value of a variable on the i^{th} unit with further identification to assumed sampled status:
 - x_{Ti} used as if a target, but in reference set
 - x_{Rj} in the remaining reference set
- y_i = observed value of a variable on the i^{th} unit (known only for the reference set).
- y_i^* = true value devoid of sampling error

Customary statistic for regression error

$$\text{MSE} = \sum_n (\mathbf{y}_i - \mathbf{B}\mathbf{x}_i)^2 / n$$

$$E[\sum_n [(\mathbf{y}_{Ti} - \mathbf{B}\mathbf{x}_i)^2] / n] = \text{Var}(\epsilon_{Yi}) + \text{Var}(\epsilon_P) + \sum_n [\epsilon_{L(Xi)}^2] / n$$

180

Customary statistic for error of imputation

$$\text{MSIE} = \sum_n (\mathbf{y}_{Ti} - \mathbf{y}_{Rj})^2 / n$$

$$\sum_n [(\mathbf{y}_{Ti} - \mathbf{y}_{Rj})^2] = \sum_n [\mathbf{B}(\mathbf{x}_{Ti} - \mathbf{x}_{Ri})]^2 / n +$$

$$2 \sum_n [\mathbf{B}(\mathbf{x}_{Ti} - \mathbf{x}_{Ri})(\epsilon_{L(XiT)} - \epsilon_{L(XjR)})] + \sum_n [\epsilon_{L(XiT)} - \epsilon_{L(XjR)}]^2 / n + 2 \text{Var}(\epsilon_Y) + 2 \text{Var}(\epsilon_P)$$

$$= \sum_n [\mathbf{B}(\mathbf{x}_{Ti} - \mathbf{x}_{Ri})^2 + 2 \mathbf{B}(\mathbf{x}_{Ti} - \mathbf{x}_{Ri})(\epsilon_{L(XiT)} - \epsilon_{L(XjR)}) - 2 \epsilon_{L(XTi)} \epsilon_{L(XjR)}] / n + 2 \text{Var}(\epsilon_Y) + 2 \text{Var}(\epsilon_P) + 2 \sum_n [\epsilon_{L(X)}^2] / n$$

$\underbrace{\hspace{15em}}_{\approx 2 \text{ MSE}}$

Proposed statistic for pure error

- Defined by differences among observations having identical X's
- Biased by including measurement error
- Procedure:
 - Compute differences among most similar neighbors using Mahalanobis weight matrix
 - Sort by similarity (distance).
 - Compute (MSIE)/2 for fraction of population having near zero distances.

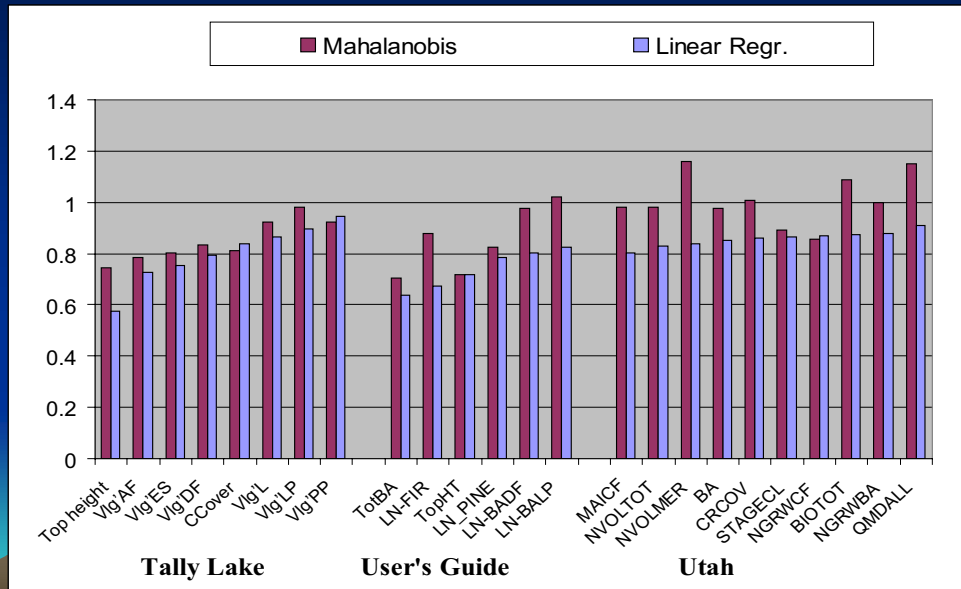
181

Proposed statistic for error of imputation

$E[y_{Ri} - g(x_{Tj})]^2 = E[MSIE]$ minus the variances of pure error and sampling error arising from the Target units.

The required variance component to be deducted is half the mean-square difference for pairs with near zero Mahalanobis distances.

Standard Error of Imputation / Std. Dev.



182

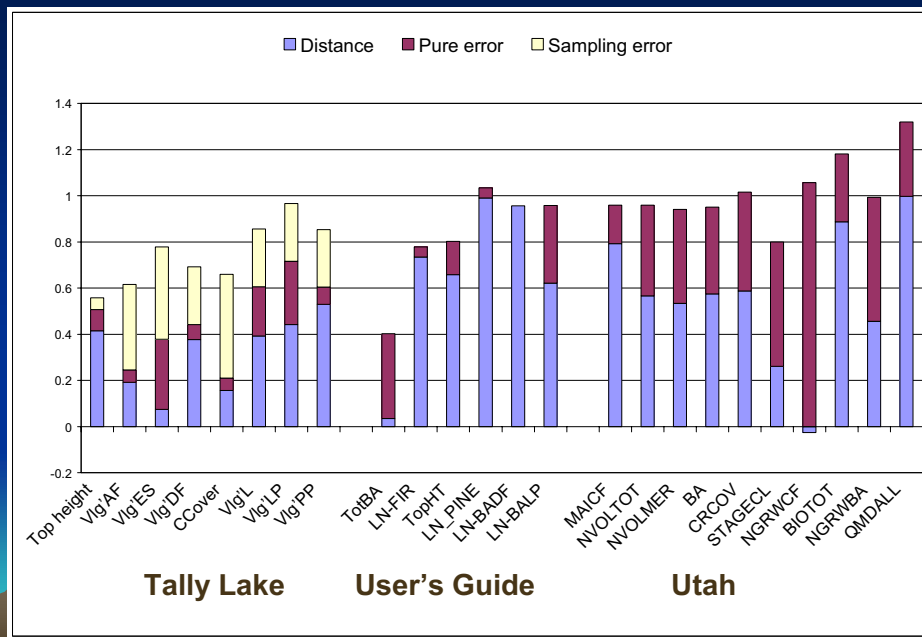
Proposed statistic for distance component

Component of MSIE from separation between Targets and selected Reference units.

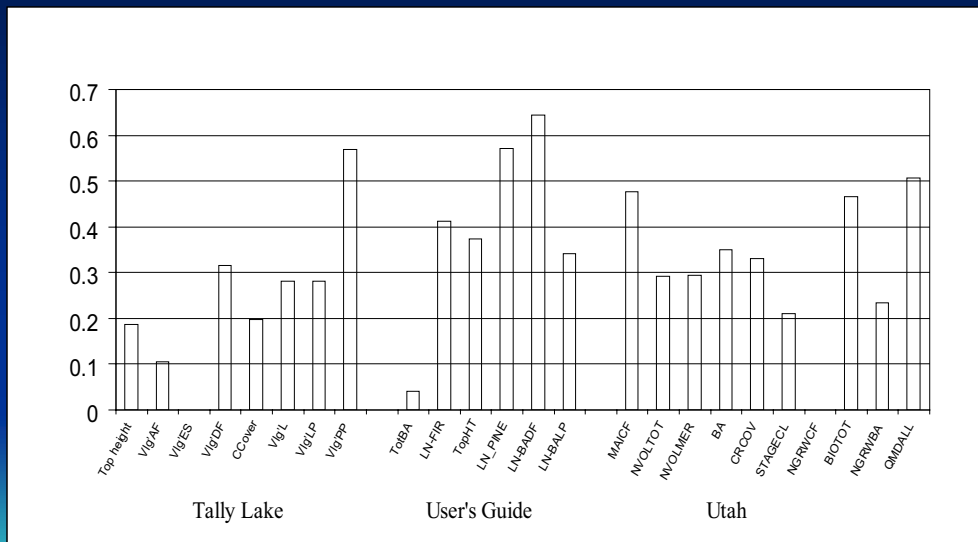
$$\begin{aligned}
 \sum_n [g(x_{Ri}) - g(x_{Tj})]^2/n &= \sum_n [Bx_{Ri} + \epsilon_{L(X_{Ri})} - Bx_{Tj} - \epsilon_{L(X_{Tj})}]^2/n \\
 &= \sum_n [B(x_{Ri} - x_{Tj}) + (\epsilon_{L(X_{Ri})} - \epsilon_{L(X_{Tj})})]^2/n \\
 &= \sum_n [B(x_{Ri} - x_{Tj})^2 + 2 B(x_{Ri} - x_{Tj})(\epsilon_{L(X_{Ri})} - \epsilon_{L(X_{Tj})}) \\
 &\quad + (\epsilon_{L(X_{Ri})} - \epsilon_{L(X_{Tj})})^2]/n \\
 &= MSIE - 2\hat{v}\hat{r}(\epsilon_Y) - 2\hat{v}\hat{r}(\epsilon_P)
 \end{aligned}$$

This error component does not depend on the functional form of the relations of the Y's to the X's .

Components of Error (mahal)



Lack of fit of linear model



Topics

- Components of imputation error
- Measures of similarity (a few in particular)
- Alternative MSN similarity measures evaluated using error components
- Transformations of variables to improve resolution

184

Similarity measures for interval and ratio-scale variables (Podani 2000)

- **Euclidean/Mahalanobis**
- **Chord**
- **Angular**
- **Geodesic**
- Manhattan
- Canberra
- Clark
- Bray-Curtis
- Marczewski-Steinhaus
- 1-Kulczynski
- Pinkham-Pearson
- Gleason
- Ellenberg
- Pandeya
- Chi-square
- 1-Correlation
- 1-similarity ratio
- Kendall difference
- Faith intermediate
- Uppsala coefficient

Similarity measures for binary variables (Podani 2000)

Symmetric for 0/1

- Simple matching
- **Euclidean**
- Rogers-Tanimoto
- Sokal-Sneath
- Anderberg I
- Anderberg II
- Correlation
- Yule I
- Yule II
- Hamann

Asymmetric for 0/1

- Baroni-Urbani-Buser I
- Baroni-Urbani-Buser II
- Russell-Rao
- Faith I
- Faith II
 - *Ignore 0*
- Jaccard
- Sorenson
- **Chord**
- Kulczynski
- Sokal-Sneath II
- Mountford

185

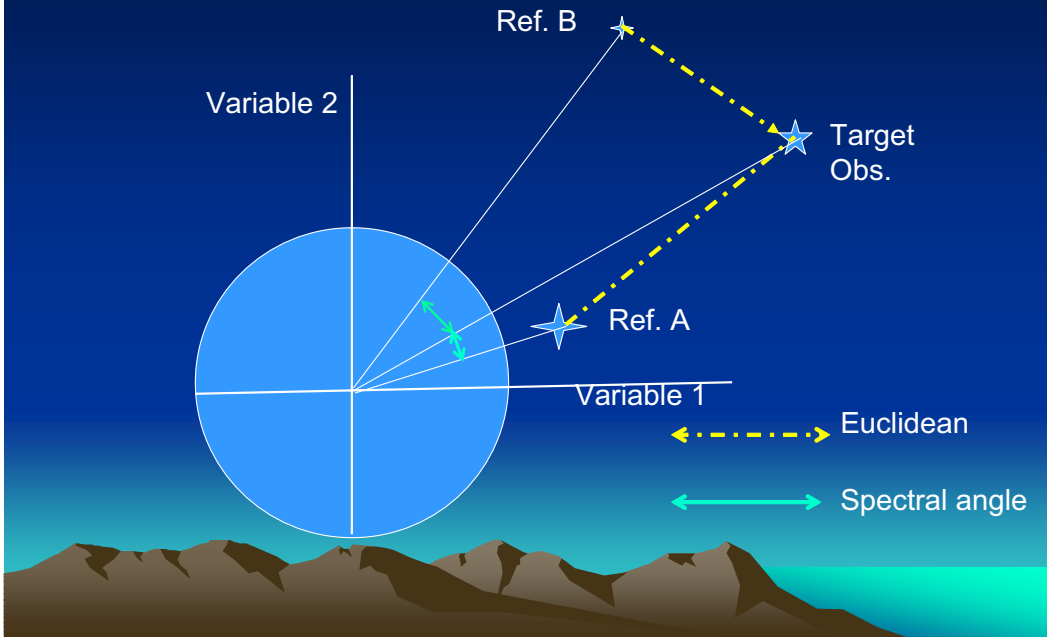
A Disimilarity (Distance) function in matrix notation

$$D_{iu}^2 = \min_i [(X_i - X_u) W (X_i - X_u)']$$

– Where, for

- Euclidean distance: $W = I$ (Identity matrix)
 - All variates get equal weight
- Mahalanobis distance: $W = \Sigma^{-1}$ (Inverse covariance matrix)
 - Correlated variates each get less weight

Euclidean vs. Cosine (Spectral angle)



186

Representing spectral angle as Euclidian distance

$$\cos(a_{ij}) = \frac{\sum_{k=1}^p x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^p x_{ik}^2 \sum_{k=1}^p x_{jk}^2}}$$

Let: $Z_i = X_i / \sqrt{X_i' X_i}$

$$d_{ij}^2 = (Z_i - Z_j)' I (Z_i - Z_j) = 2(1 - \cos(a))$$

Effect of using cosine transformation of TM data on classification accuracy*

Attribute	Untransformed	Cosine trans.
Plant Assoc. Grp. (Oregon) ** (Mahalanobis)	0.340	0.363
Modal Spp. Comp.(Oregon)** (Mahalanobis)	0.276	0.335
Modal Spp. Comp. (Minn.)*** (Euclidean)	0.320	.328

* Kappa statistics **TM data ***TM+ Enhanced data

187

Topics

- Components of imputation error
- Measures of similarity (a few in particular)
- Alternative MSN similarity measures evaluated using error components
- Transformations of variables to improve resolution

A Distance function using the Y information:

$$D^2_{iu} = \min_i [(X_i - X_u) W (X_i - X_u)']$$

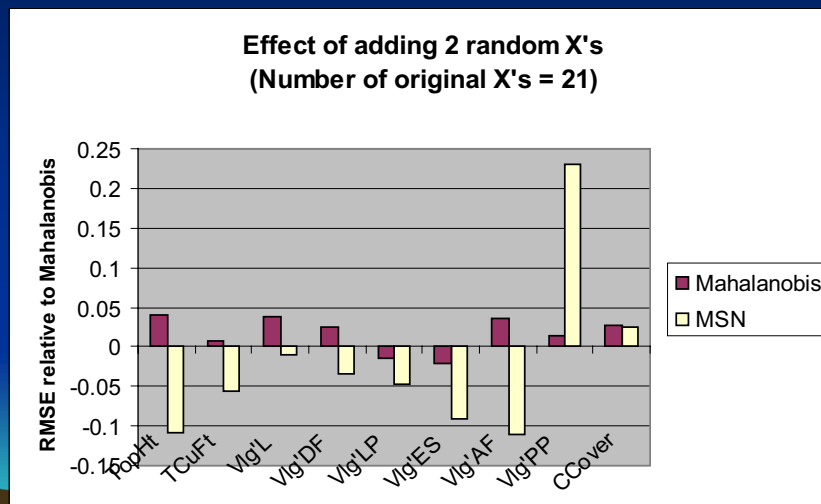
– Where, for

- MSN (1995): $W = \Gamma \Lambda^2 \Gamma'$ with:
 - Γ = matrix of coefficients of canonical variates
 - Λ = diagonal matrix of canonical correlations

188

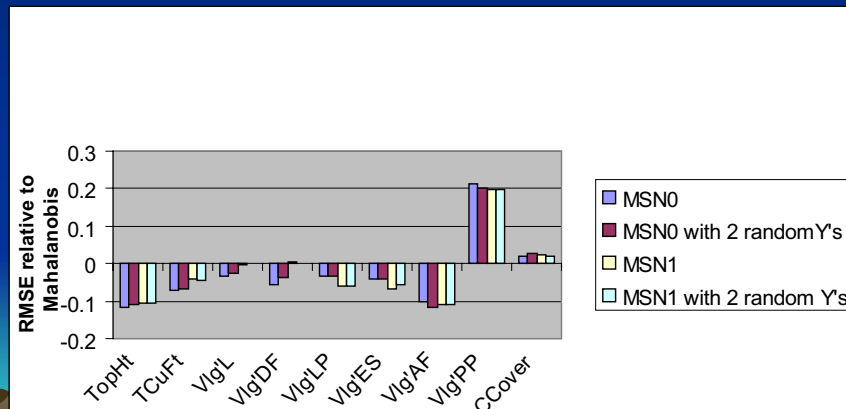
Why Weight with Canonical Analysis?

- Not degraded by non-informative X's



Why Weight with Canonical Analysis?

- Not affected by non-informative Y's if number of canonical pairs is determined by test of significance on rank.



Distance function of new alternative

Distance is in the r-space spanned by expected values of Canonical Variates conditioned on X

$$\bar{V}^{(r)} = \lambda^{(r)} c^{(r)} X$$

$$\text{Var}\{\bar{V}\} = [I - \Lambda][I - \Lambda]'$$

In Mahalanobis distance, each $\bar{V}^{(r)}$ is weighted by inverse of its standard error about its expectation: $\text{sqrt}(1/(1-\lambda^2))$

New regression alternative:

$$d_{ij}^2 = \min_i (\mathbf{X}_i - \mathbf{X}_j) \Gamma \Lambda [(\mathbf{I} - \Lambda^2)]^{-1} \Lambda \Gamma' (\mathbf{X}_i - \mathbf{X}_j)'$$

Λ is the diagonal matrix of canonical

correlations for $k = \sum_{i=1}^k \lambda_i / \sum_{i=1}^s \lambda_i \geq \text{PROPVAR}$

$$\mathbf{W}^{1/2} = \Gamma \begin{bmatrix} \lambda_1 / \sqrt{1 - \lambda_1} & 0 & 0 & K & 0 \\ 0 & 0 & 0 & K & 0 \\ 0 & 0 & \lambda_k / \sqrt{1 - \lambda_k} & K & 0 \\ M & M & M & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

190

Comparison of MSN Distance Functions

- Moeur and Stage 1995
 - Assumes Y's are "true"
 - Searches for closest linear combination of Y's
 - Set of near neighbors sensitive to lower order canonical correlations
- Stage 2003
 - Assumes Y's include measurement **error**
 - Searches for closest linear combination of **predicted** Y's
 - Set of near neighbors less sensitive to random elements "swept" into lower order canonical corr.

Effect of change:

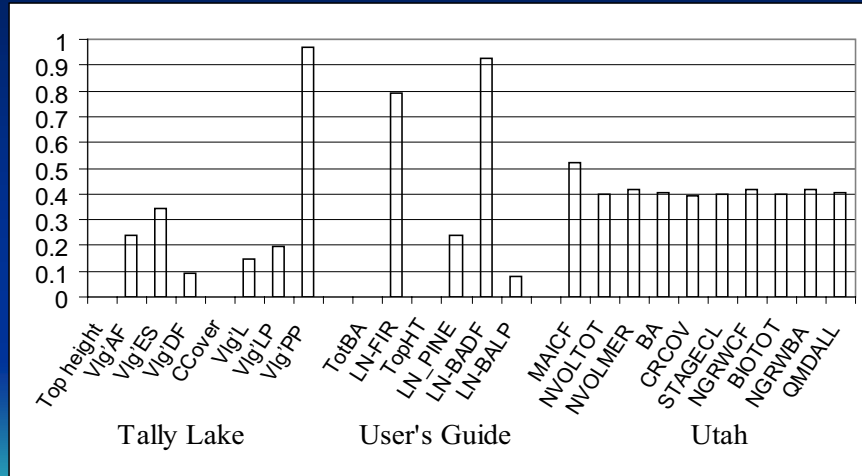
- No change if only first canonical pair is used.
- Regression alternative gives more relative weight to higher correlated pairs.
- Effects on Root-Mean-Square Imputation Error are mixed: e.g. the following three data-sets---

191

Attributes of three data sets

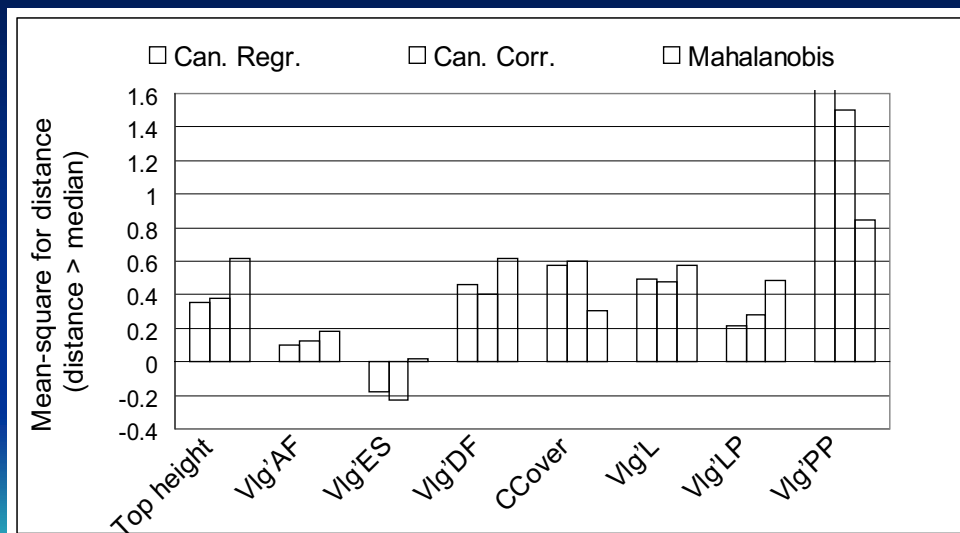
	Utah	Tally Lake	User's Guide
Number of Y's	10	8	6
Number of X's (p)	12	21	12
Number of ref. obs. (n)	1076	847	197
Significant canonical pairs (s)	4	7	5
$n/(s+p \cdot s)$	16.55	5.50	3.03

Proportion of zero observations

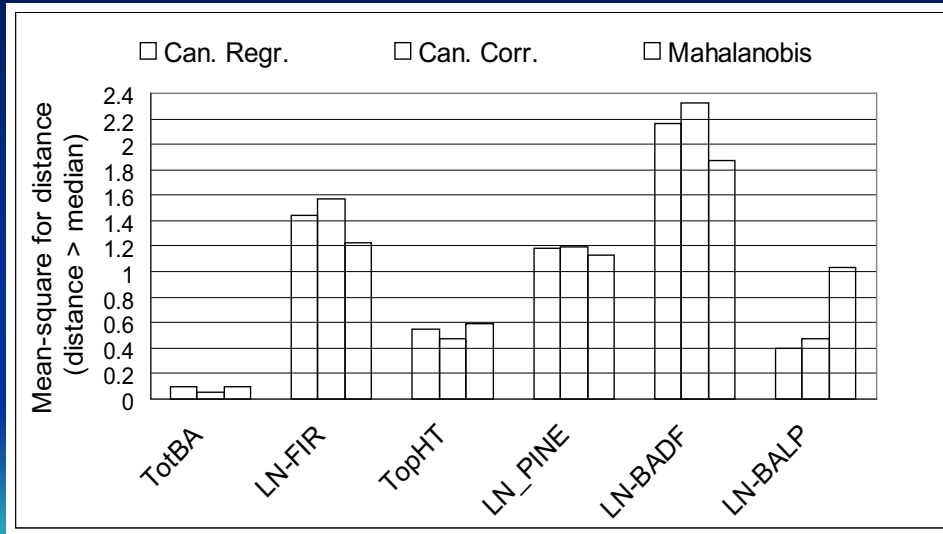


192

Tally Lake

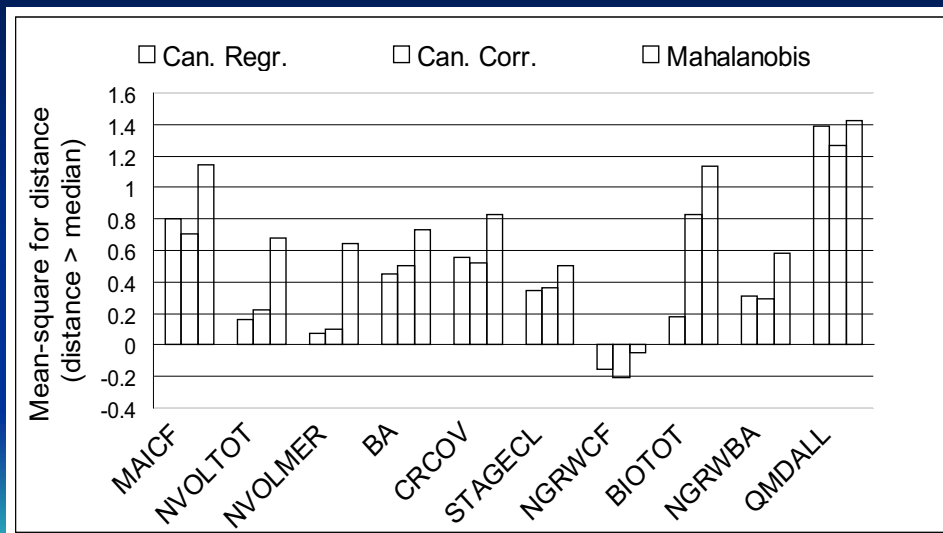


User's Guide



193

Utah



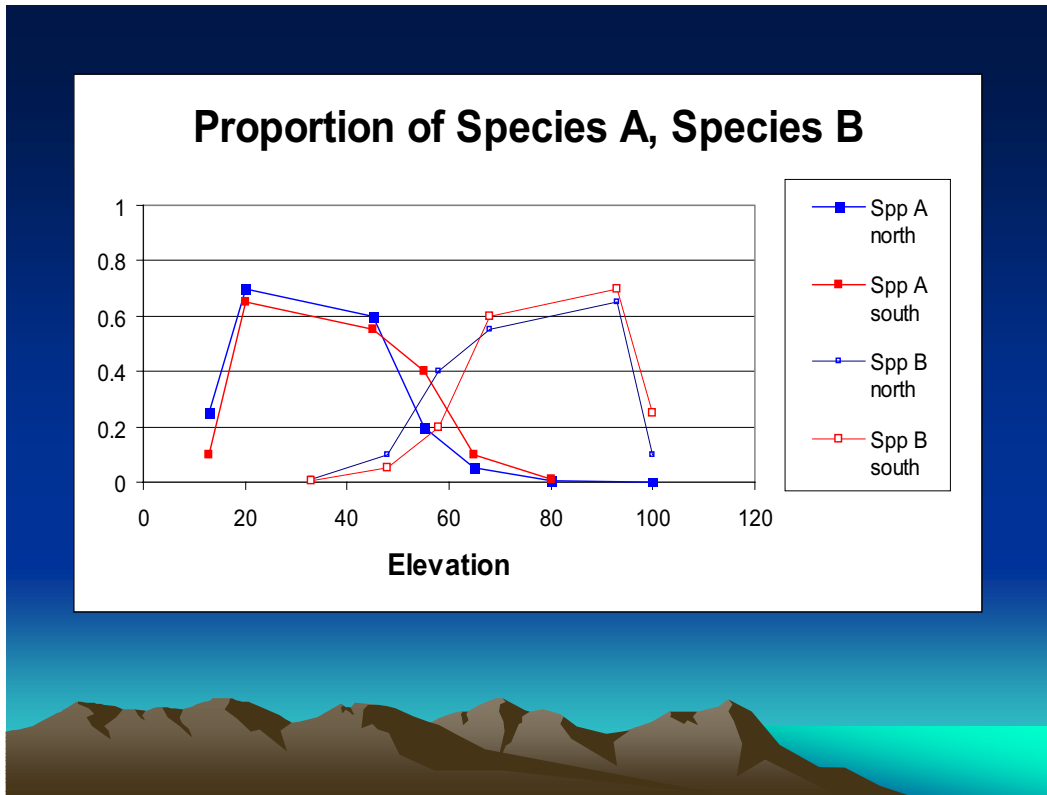
Topics

- Components of imputation error
- Measures of similarity (a few in particular)
- Alternative MSN similarity measures evaluated using error components
- Transformations of variables to improve resolution

194

Transforming the Y-variables

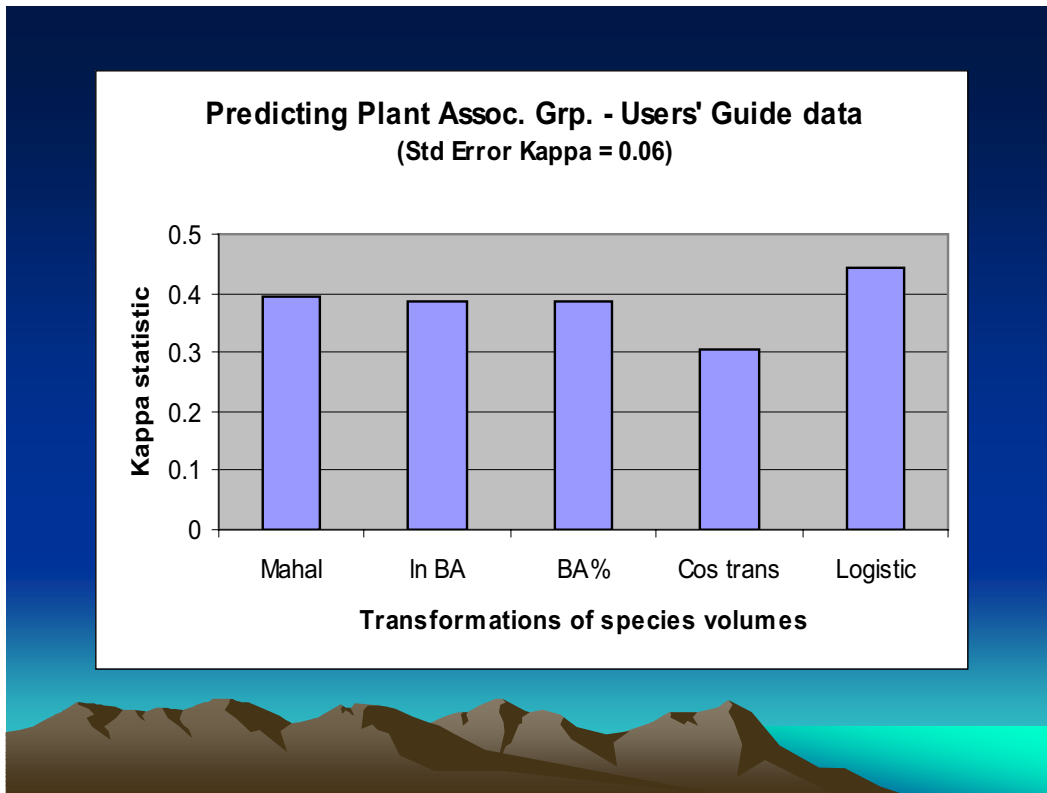
- Variance considerations—want homogeneity
- And a logical functional form for $Y = f(X)$
 - Transformations of species composition
 - Logarithm of species basal area
 - Percent basal area by species
 - Cosine spectral angle
 - Logistic
 - Evaluated by predicting discrete Plant Association Group (PAG), Users' Guide example data (Oregon)



195

Composition transformations:

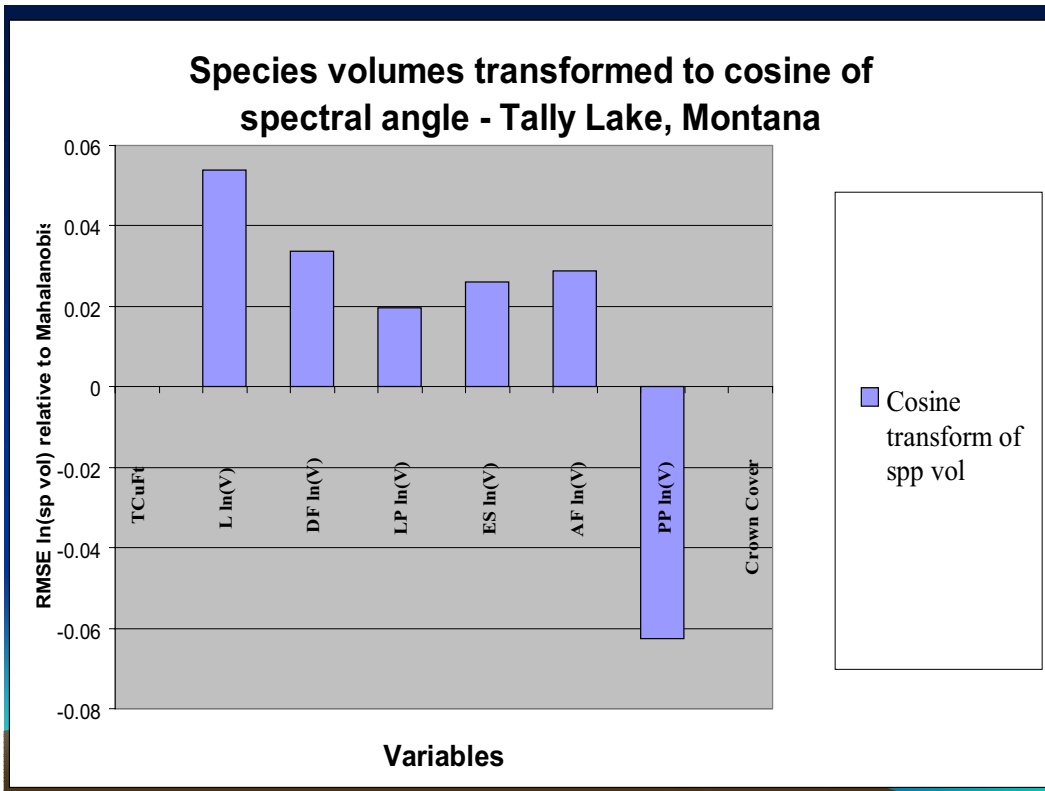
- Logistic:
 - = $\ln[(\text{Total BA} - \text{spp BA})/\text{spp BA}]$
 - = $\ln(\text{Total BA} - \text{spp BA}) - \ln(\text{spp BA})$
 - Represented in MSN by two separate variables.
- Cosine Spectral Angle:
 - = $\text{Spp BA} / \sum (\text{spp BA})^2$



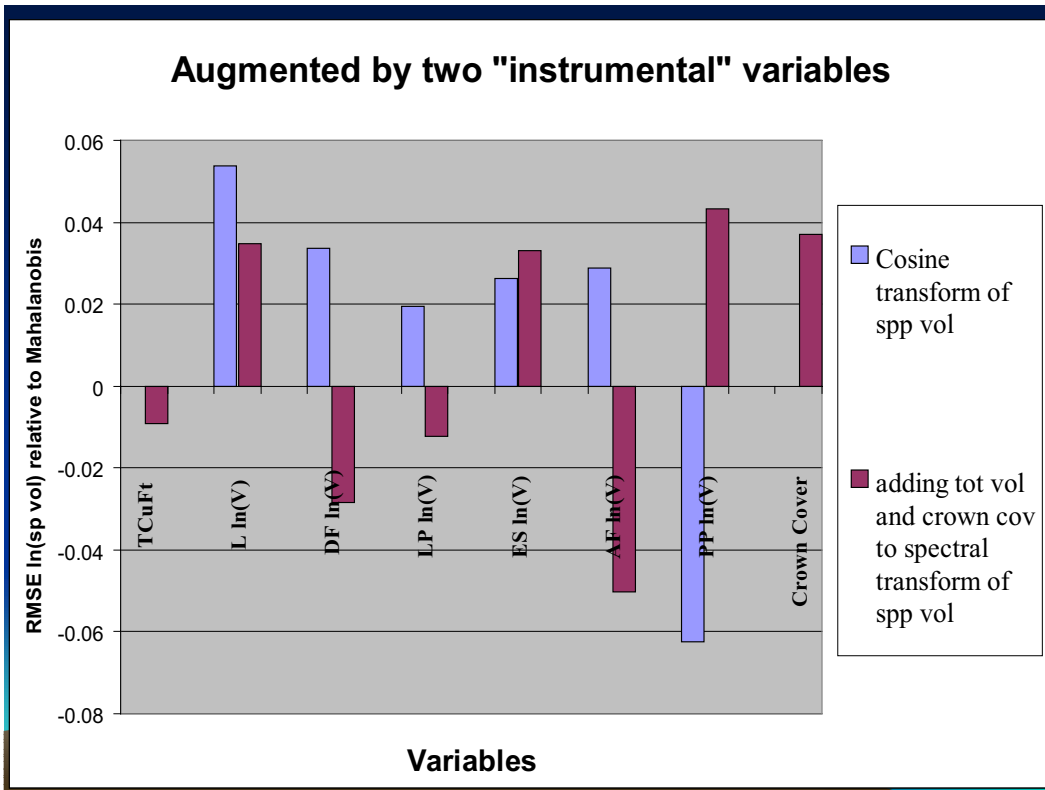
196

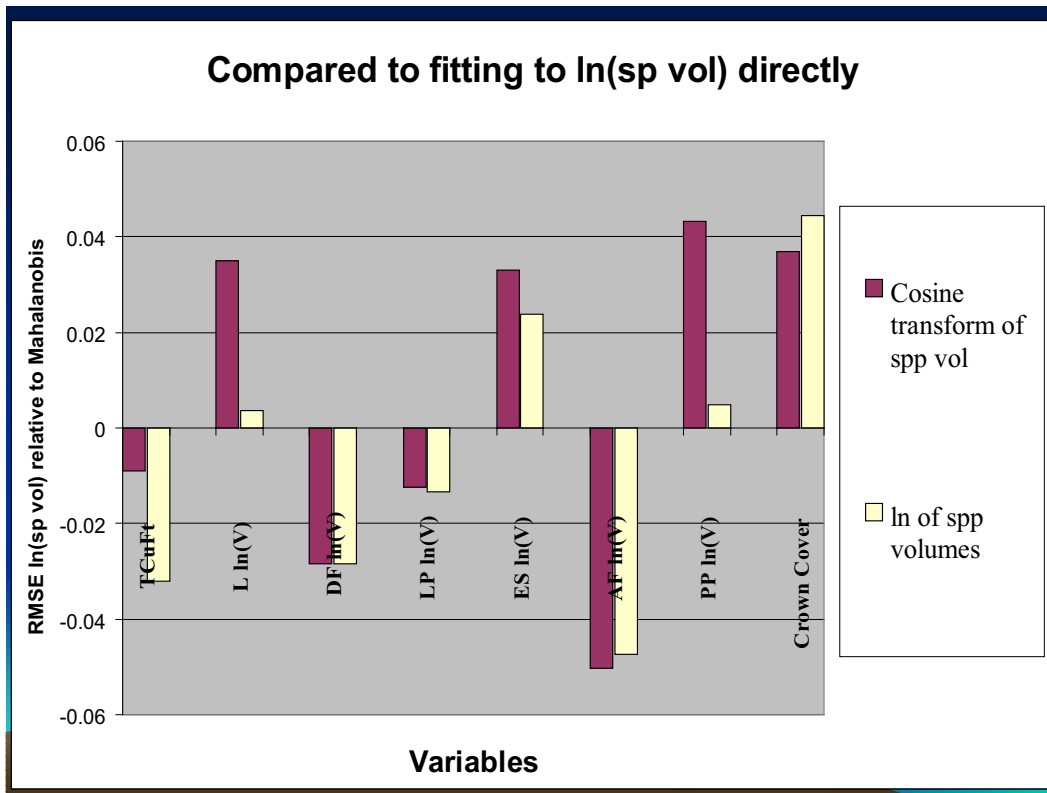
Comparing weights from transformed Y's

- Evaluated by comparison to Mahalanobis weights on all X's
- For imputing $\ln(\text{spp volumes})$

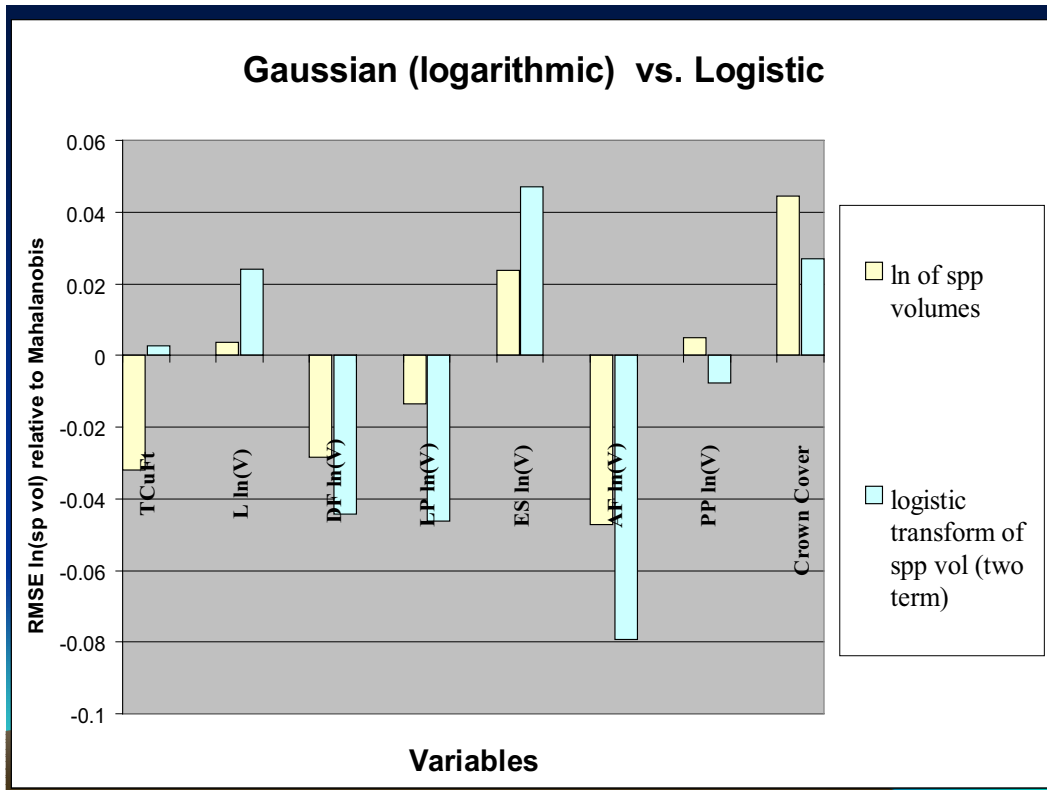


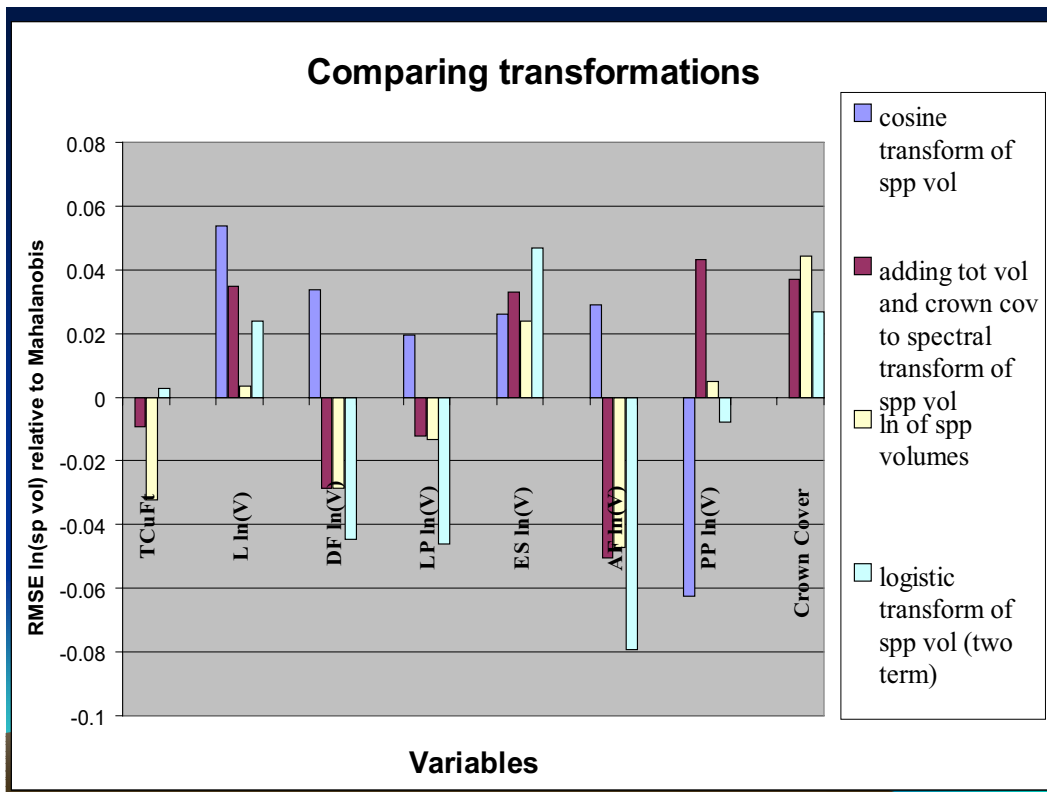
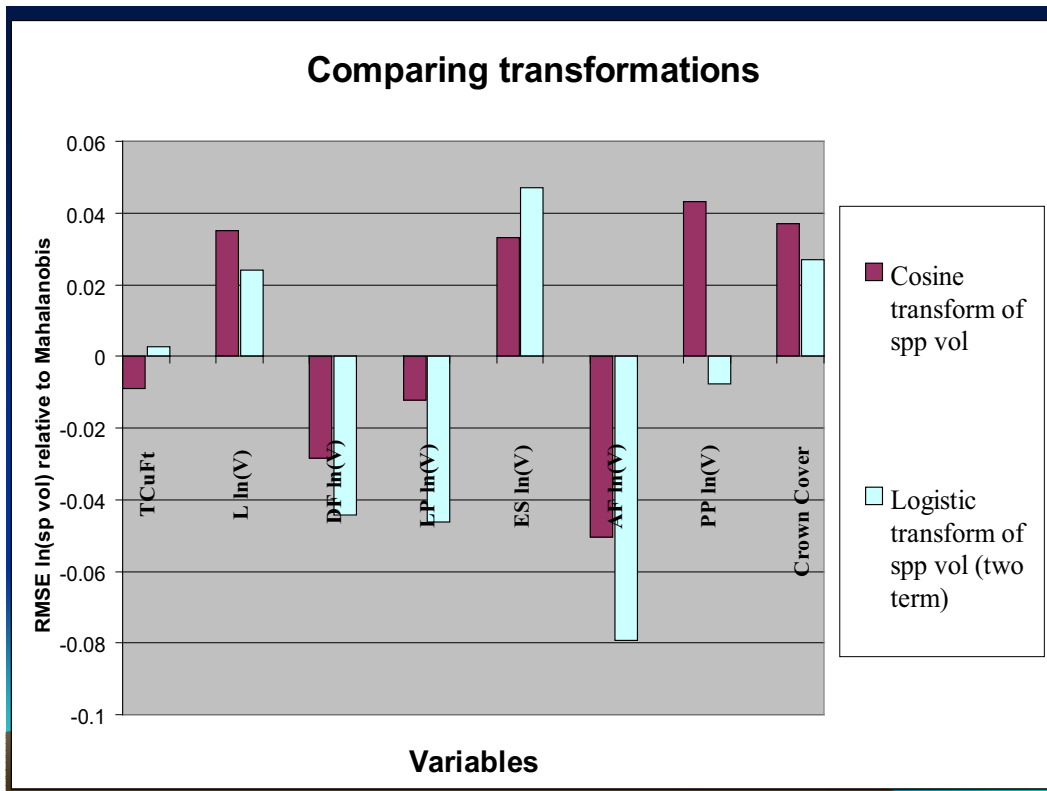
197





198





Implications of transforming

- Imputed value derived from the neighbor, not directly from the model as in regression.
- Neighbor selection may be improved by transforming Y's and X's .
- Multivariate Y's can resolve some indeterminacies from functions having extreme-value points (maxima or minima).

200

MSN Software Now Includes Alternative Distance Functions:

- Both canonical-correlation based distance functions.
- Euclidean distance on normalized X's.
- Mahalanobis distance on normalized X's.
- You supply a weight matrix of your derivation.
- *k*-nearest neighbors identification.

So ??

- Of the many methods available for imputation of attributes, no one alternative is clearly superior for all variables and data sets.

201

“Opportunities” in Imputation

- Samples seldom include extremes, leading to imputed values at the extremes being biased toward the center of the distribution.
 - How should the samples be selected to reduce this bias and increase efficiency?
- Optimization of variate weights may give little or no weight to some variates.
 - How can spatial proximity be introduced when there is little difference among potential candidates?

Software Availability

- E-mail:
ncrookston@fs.fed.us
- On the Web:
<http://forest.moscowfsl.wsu.edu/gems/msn.html>.
- In print:
Crookston, N.L., Moeur, M. and Renner, D.L. 2002.
User's guide to the Most Similar Neighbor
Imputation Program Version 2. Gen. Tech. Rpt.
RMRS-GTR-96. Ogden, UT: USDA Rocky
Mountain Research Station 35p.

202

Transforming X-variables

- To predict discrete classes of modal species composition (MSC) with Euclidean or Mahalanobis distance.
- To predict continuous variables of species composition