



# Modelling Regeneration by Imputation: Some background

Albert R. Stage

West Twin Forestry

Moscow, Idaho

[astage@moscow.com](mailto:astage@moscow.com)



# Model architecture: Growth vs. Yield

- Farnham *et al.* (1986) distinguish between integrating growth processes over time to estimate future states (growth architecture), versus directly estimating the future state as a function of time (yield architecture).
- Regeneration modelling using growth architecture seems too complex—too many poorly understood processes. But see Ek and Monserud (1974).
- Ferguson, *et al.* (1986) and Coates ( ) model regeneration using yield architecture to assemble a “picture” of the outcome.
- Ranneby *et al.* (1987) and Ek, *et al.* (1997) use imputation (also yield architecture) to access relevant samples of the outcomes.



# Model architecture: Process vs. Imputation

- Process modelling requires extensive data, analysis, and programming prior to delivery of finished model.
- Imputation reverses that order:
  - Program the logic for selection of reference sample unit.
  - Analyze distance function using partial data base.
  - Augment data base.



# Imputation Methods

- Stratum-based random sample
- Near Neighbors
  - Most Similar Neighbor ( $k = 1$ )
  - K-near neighbors

All require some measure of similarity



# Attributes of Imputation Methods

- Biological criteria.
  - How to index time?
    - Time since disturbance?
    - Change in overstory?
  - Choice of site and overstory attributes
  - Degree of aggregation
    - Point?
    - Stand?



# Attributes of Imputation Methods

- Statistical criteria
  - Measure of similarity
    - Form of distance function
    - Transformations of Variables
  - Weights in distance measure for variate differences between sampled (known) and unsampled (unknown) units
  - Weights of selected units if  $k \ll 2$

# Similarity Measures

## Method

- Random selection from same stratum
- Proximity in the space of explanatory variables (X's)
- Proximity in the space of predicted variables ( $Y=f\{X\}$ )

## Distance

- Span of stratum "bin"
- Euclidean distance
- Mahalanobis distance
- Canonical-correlation weighted MD



# Sources of imputation error

- All methods
  - Distribution of target data in relation to reference data.
  - Density of reference data.
  - Choice of  $X$  variables and transformations
- MSN Canonical Analysis
  - Choice of  $Y$  variables and transformations



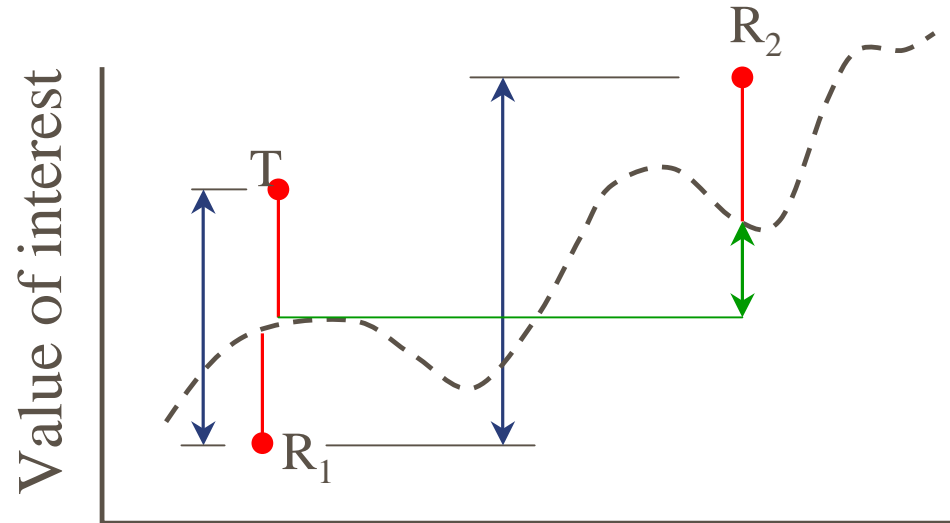
# Calculated RMSE

- Imputed from bin having same ID:
  - Calculated from variance within bin.
- Imputed from near neighbors:
  - Must be calculated from “second nearest neighbor”. So measurement error is included twice.
  - A target observation within the “cloud” will fall between the several near neighbors. Therefore, it will be closer to its nearest neighbor than the distances between that set of near neighbors. (Adding points into the cloud reduces the average distance between points). So distance component of error overestimated.

# Sources of error using near neighbors:

- Distance (D) between target (T) and reference (R) observations:  $f\{D_T - D_R\}$
- Measurement error

- True:  $\sigma^2$
- $E(T-R)^2 = 2\sigma^2 + [f\{\Delta D\}]^2$
- $E(R_1 - R_2)^2 = 2\sigma^2 + [f\{\Delta D\}]^2$



Unobservable "truth" - - - -

Distance



# Implications of non-linearity in Canonical Analysis (MSN)

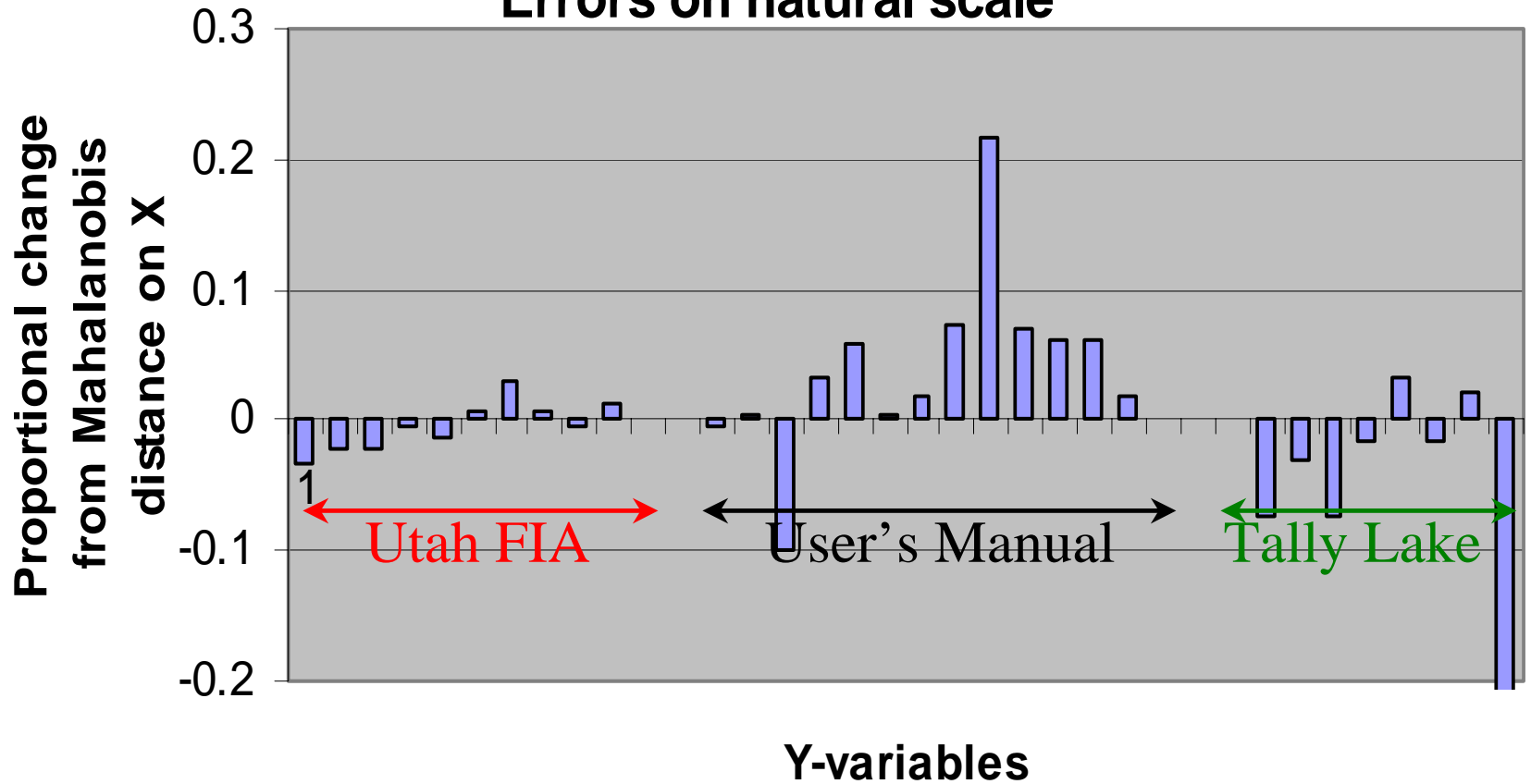
- Imputed value derived from the neighbor, not directly from the model as in regression.
- Neighbor selection may be influenced by curved relations between  $Y$ 's and  $X$ 's .
- Multivariate  $Y$ 's can resolve some indeterminacies from functions having extreme-value points (maxima or minima).

# Example analyses of distance functions for three data-sets:

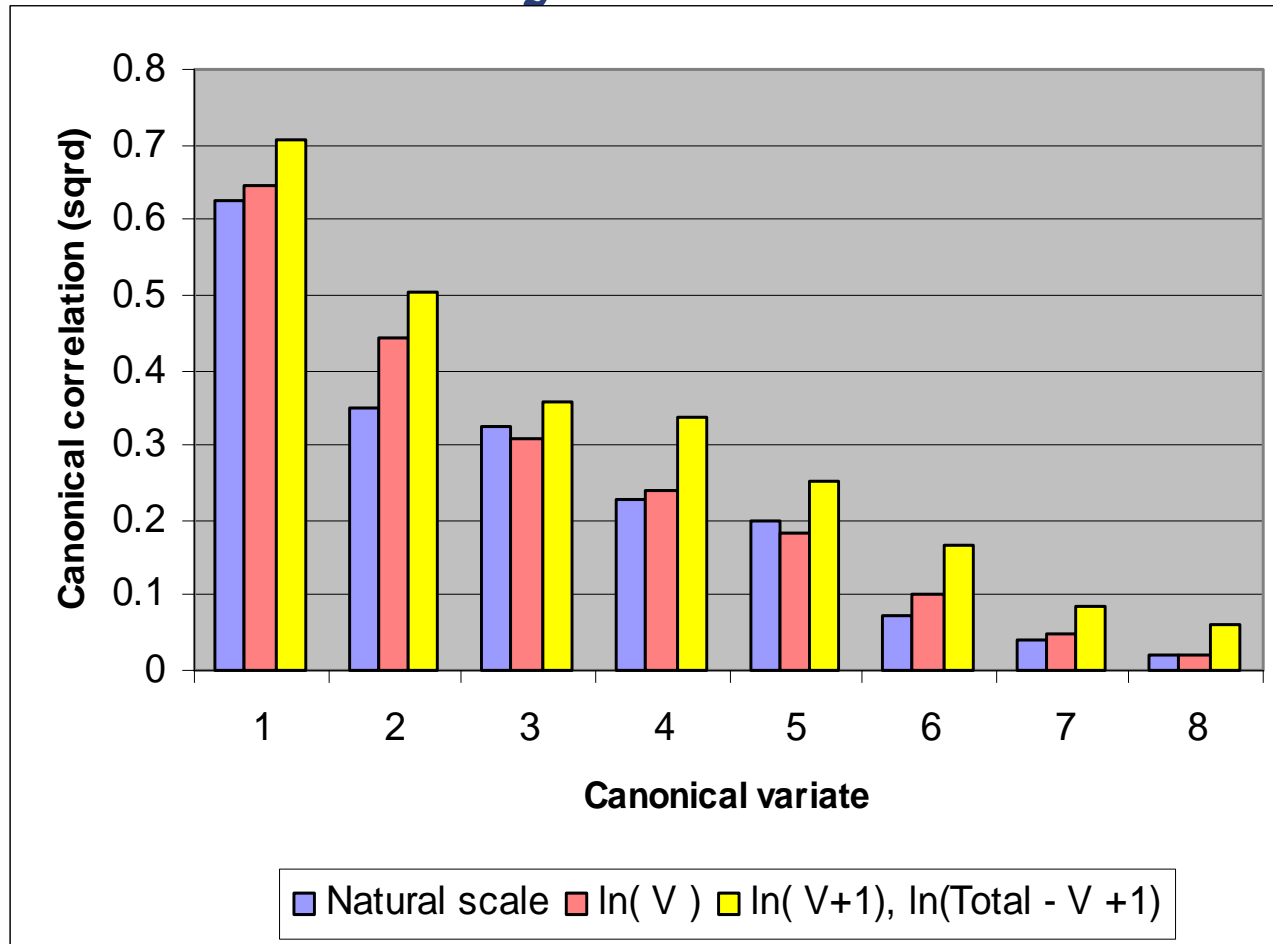
	<b>Utah</b>	<b>Tally Lake</b>	<b>User's Manual</b>
<b>Canonical pairs (s)</b>	<b>9</b>	<b>8</b>	<b>7</b>
<b>Number of Y's</b>	<b>15</b>	<b>8</b>	<b>17</b>
<b>Number of X's (p)</b>	<b>12</b>	<b>20</b>	<b>7</b>
<b>Number of obs. (n)</b>	<b>1076</b>	<b>847</b>	<b>197</b>
<b><math>n/(p*s+s)</math></b>	<b>13.3</b>	<b>5.04</b>	<b>3.52</b>

# RMSE: MSN vs. Mahalanobis distance on X

## Errors on natural scale

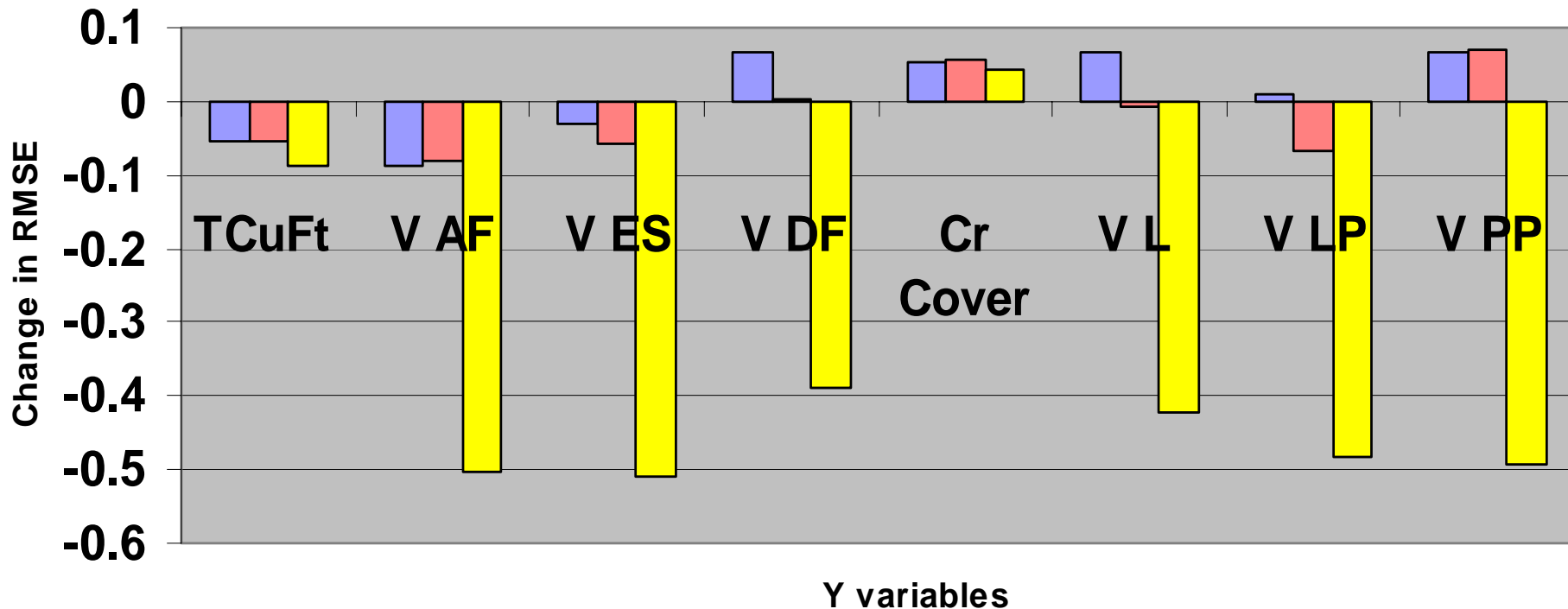


# Canonical analysis of Tally Lake data



# RMSE: MSN vs. Mahalanobis on X's

## Species volumes evaluated on logarithmic scale



- Y's not transformed
- Y's for species volumes transformed to logarithms
- Y's for species volumes transformed to logistic



So ??

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



# “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?



# More Opportunities

- Choice of neighbors can be improved by including covariances of attributes.
  - How to select the best method has not been clearly established.