Oct. 5, 2006

Partitioning error components for accuracy-assessment of near-neighbor methods of imputation.

Albert R. Stage

and

Nicholas L. Crookston

Albert R. Stage, (Retired) Moscow Forestry Sciences Laboratory, Rocky Mountain Research Station, 1221 S. Main St. Moscow, Idaho 83843 USA. astage@moscow.com: FAX 208 883 2318. Nicholas L. Crookston, Moscow Forestry Sciences Laboratory, Rocky Mountain Research Station, 1221 S. Main St. Moscow, Idaho 83843 USA. ncrookston@fs.fed.us.

Acknowledgement

Interested readers are greatly in debt to one of our anonymous reviewers for insisting in great detail on precise use of wording in this paper. And the authors thank the other reviewer who said the results are so intuitively obvious as to wonder why it had not been done earlier for giving us the incentive to press on with revision. We are convinced that the result is much improved, but any remaining deficiencies are certainly our responsibility.

Abstract: Imputation is applied for two quite different purposes: to supply missing data to complete a data set for subsequent modeling analyses, or to estimate sub-population totals. Error properties of the imputed values have different effects in these two contexts. We partition errors of imputation derived from similar observation units as arising from three sources: observation error, the distribution of observation units with respect to their similarity and pure error given a particular choice of variables known for all observation units. Two new statistics based on this partitioning measure the accuracy of the imputations, facilitating comparison of imputation to alternative methods of estimation such as regression and comparison of alternative methods of imputation generally. Knowing the relative magnitude of the errors arising from these partitions can also guide efficient investment in obtaining additional data. We illustrate this partitioning using three extensive data sets from western North America. Application of this partitioning to compare near-neighbor imputation is illustrated for Mahalanobis- and two canonical correlation-based measures of similarity.

Keywords: Most-similar-neighbor, *k*-nn inference, missing data, landscape modeling.

Introduction

Imputation methods are important tools for completing data sets in which some observation units lack observed values for a portion of their attributes. The objective is to impute a value as close to "truth" for each missing value in the observation unit as if it were examined in great detail for all attributes. Criteria for imputations to support this objective are essentially different from criteria for estimates of population totals. The difference is that pure error, rather than being a nuisance, is of real value for subsequent resource analyses and displays. These analyses are often non-linear optimizations or simulations. For them to be realistic, the structure of the variances and covariances among attributes inherent in the population should be preserved in the data set. Even for display purposes, omission of pure error will cause the range of the displayed values to be contracted. Unfortunately, these inherently useful variances may be combined with variances attributable to the methodology used in the sampling and imputation processes. This mixture complicates choice among analytical methods for imputation. In this report we provide statistics based on a partitioning of the error components which facilitate finding a closer approximation of "truth". We partition imputation errors independently for each variable in the data set although the joint distribution of their error components would be of interest for some applications.

Imputation uses values of variables measured for all observation units ($\boldsymbol{X}$'s) to guide the imputation of values of $\boldsymbol{Y}$'s that are measured only for a sample subset of the observation units (the *Reference* set) to those units for which the $\boldsymbol{Y}$'s are missing (the *Target* set). Both $\boldsymbol{X_i}$ and $\boldsymbol{Y_i}$ may be vectors of attributes for the i^{th} observation unit. Near-neighbor imputation selects units from the reference set to serve as surrogates for members of the target set using a measure of similarity based on the $\boldsymbol{X}$'s. Choice of a particular measure of similarity, in turn, may depend on the relation of the $\boldsymbol{Y}$'s to the $\boldsymbol{X}$'s. Elements of $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$, y_i and x_i, will be subscripted only to identify the i^{th} observation unit. T and R will be used as additional subscripts when it is relevant to indicate that a Reference observation unit is being used as if it were a Target unit (hence a "pseudo-target"). Unit identifying subscripts (i or j) will be omitted when the variables are referred to collectively. Var(·) capitalized will be used for expected values, lower case var(·) for statistics calculated from the data.

Imputation from near-neighbor observations is often used for classification. However, when the "classes" are arbitrary intervals on scales of essentially continuous variables, we argue that the imputation should be based directly on the scales of the underlying continuous variables. If classes are needed for display purposes, the classification algorithm should use the imputed data. We will not consider in this paper errors in classification in which the classes are inherently discrete, requiring the concept of "membership". For discrete classes, other methods for classification such as using a discriminant function may be more appropriate than near-neighbor. For example, classification by a discriminant function may assign different classes to members of a target/reference pair of near neighbors because the discriminating boundary passes between them whereas near-neighbor imputation would assign the target to the same class as the reference member of the pair. However, there is a parallel process of partitioning the error sources in imputation of discrete variables that is beyond the scope of this paper. .

Error properties of estimates derived from imputation differ from those of regression-based estimates because the two methods include a different mix of error components. For example, the reference-set data may not be beyond reproach because of measurement error. These error properties influence how we evaluate quality of the imputations, compare alternative methods for imputation and invest in data collection. Commonly computed statistics that compare imputed values to those of a presumably similar observation unit mask methodological differences in this cloud of variation. We address this problem by partitioning the variation into components that can be estimated from the reference set. Then, new statistics based on this partitioning are presented for assessing the accuracy of imputation methods.

Several questions may be answered using these error components:

1) How does the accuracy of imputed ***Y***'s compare to accuracy of estimates from regression, stratum means or other model-based estimates?

2) How large is the error caused by imputing values to a target unit from reference units where there is substantial difference in their ***X***'s? Is there room for improvement by obtaining additional reference observations to fill gaps in their distribution? How is this error component affected by the choice of a particular measure of similarity?

3) How is the accuracy of imputation affected by the choice of variables and their transformations?

4) What is the effect on imputed values of pooling k reference observations?

5) How do the measurement accuracies compare to components of variation from other sources?

6) Would investments in additional data be more efficient if used to obtain information on variables to be added to the target set (new ***X***'s), to refine the estimates of the ***X***'s already included, or to obtain data on additional units for the reference set?

Resolution of these questions requires quantitative estimates of the sources of imputation error. These estimates can be obtained from the information in the n observation units in the reference data. In analysis of data in the reference-set data, although we will use some of the data as if they were targets, there is no difference in their approximation of "truth", no intrinsic differences between "observed" and "predicted". We are simply describing the properties of differences between members of pairs of observations. When the value to be imputed is a weighted average of k near neighbors, then its error properties are derived from the error properties of the k separate pairs and the weights defined by the particular k-nn procedure.

Bootstrap and cross-validation methods for answering some of these questions have been developed for imputation methods other than near-neighbor (Shao and Sitter 1996) or for classification with *k*-nn (Mullin and Sukthankar 2000). Neither of these papers has addressed the problem of partitioning the errors as to sources. Moeur and Stage (1995) used data-splitting and jackknife methods to evaluate capability of Most Similar Neighbor (MSN) to reproduce the variance and covariance structure of the reference data and to compare error rates to those obtained by stratified sampling and regression. Their analyses of errors also included variation in the coefficients in the measure of similarity caused by sequentially omitting 1.7% of their data as well as the difference between the observed and imputed ***Y*** values for the pair selected by the calculated similarity measure.

Splitting data into "calibration" and "validation" subsets, which was intended to reduce bias in error estimates, introduces a different bias into estimates of imputation errors. The withheld reference observations in sparsely-represented parts of ***X***-space could have supplied imputations for nearby target observations. In the analysis of imputation error, however, those targets will be paired with a more remote reference observation, thereby increasing the **estimated** error. A further disadvantage of the jackknife procedure is that it may increase the estimate of error by increasing the mean-square bias. Targets in the midst of a cloud of reference observations may be paired with an observation from any direction. Targets at the edge of a cloud, however, will likely be paired with a more central point. If there is a trend in the ***Y***'s with distance from the center of the cloud, then the asymmetry of direction to the reference introduces bias in the imputed value. Withholding data increases this bias unnecessarily. The jackknife procedure using a single reference observation as if a target minimizes this bias by using the full range of data (except for the single reference unit). Other methods to reduce this bias in *k*-nn imputation have been evaluated by Malinen (2003).

A statistic commonly used to evaluate imputation error estimates the root-mean-square differences between reference and target observations by withholding each observation unit in the reference set while searching for its similar neighbor in the remainder of the reference set. The term RMSE (root-mean-square error) used for this statistic is unfortunate (e.g. Moeur and Stage 1995, Crookston et al. 2002). The term as used in imputation includes different components of error than the same term used in a regression or sampling context. Therefore, we use the term Mean Squared Difference (MSD) for the statistic describing squared differences in a pair of similar observations. Thus, our partitioning is applicable for evaluating any of the near-neighbor methods of imputation that are judged on the basis of sums of squared errors.

We use the term "distance" for the value produced by the function measuring dissimilarity between the i^{th} and j^{th} pair of observation units. Although Podani (2000) cites more than 60 distance functions, those most widely used for imputation are of the quadratic form:

$$d_{ij}^{\,2} = (\boldsymbol{X}_i - \boldsymbol{X}_j)\,\boldsymbol{W}\,(\boldsymbol{X}_i - \boldsymbol{X}_j)' \qquad (1)$$

where:

$\boldsymbol{X_i}$ is the (1 x p) vector of $\boldsymbol{X}$-variables for the i^{th} target observation unit,

$\boldsymbol{X_j}$ is the (1 x p) vector of $\boldsymbol{X}$-variables for the j^{th} reference observation unit, and

$\boldsymbol{W}$ is a (p x p) symmetric matrix of weights.

If the weight matrix, W, is the diagonal identity matrix, then we have a simple Euclidean distance (squared). As a variation of Euclidean distance, some analysts empirically vary the diagonal elements to improve the imputation. If correlations among the variates are to be considered, then the inverse of their correlation matrix is used for $\boldsymbol{W}$ to produce a Mahalanobis distance—a distance function that plays a key role in estimating the error components. MSN distances are of the same form with $\boldsymbol{W}$ derived from analyses of canonical correlation (Moeur and Stage 1995), canonical regression (Stage and Crookston 2002) or of canonical correspondence (Ohmann and Gregory 2002). With a simple transformation of the $\boldsymbol{X}$'s to $x_i \Big/ \sqrt{\sum_{l=1}^{p} x_{il}^2}$ the quadratic form with identity matrix for $\boldsymbol{W}$ also includes spectral analysis imputation as used by Sohn et al. (1999).

Our following presentation is in four sections: 1) defining error sources in the process of imputation, 2) partitioning MSD into components arising from these sources, 3) presenting some new statistics based on the partitioning relevant to the key questions stated above, and 4) applying these statistics to three extensive data sets.

Components of Error

Variation in imputed values arises from both natural variability of attributes of the ecosystem, and from the measurement and analytical procedures used to describe the ecosystem. While natural variability is useful in analyses requiring the completed data set, variation introduced by measurement and analytical procedures is a nuisance to be reduced.

Imputation error arises from four sources for a given set of $\boldsymbol{X}$ and $\boldsymbol{Y}$ variables:

1) Measurement errors of the $\boldsymbol{Y}$'s in the reference set. These errors are defined as:

$$\varepsilon_{Yj} = y_j - y_j^* \qquad (2)$$

in which the starred variables represent the true, but unknown, values. The $\boldsymbol{\varepsilon_{Yj}}$ are not properties of the ecosystem being described, but rather, properties of the accidents of how we observed it. The measurement errors may arise from using a sample-based estimate as if it were a complete census within the j^{th} unit, from changes during elapsed time since observation, from lack of standardization among different observers or their instruments, or any combination of such causes. These errors often are assumed to be zero (e.g. Moeur and Stage 1995). We now relax that assumption because in some applications, errors from this source have been quite large relative to total error. We assume that the measurement errors can be rendered unbiased and are independent of the true y_j^* and of the observed $\boldsymbol{X}$'s.

2) Pure error. That there exists a relation between the $\boldsymbol{Y}$'s and the $\boldsymbol{X}$'s is a key premise of near-neighbor inference. For a given set of $\boldsymbol{X}$'s the departure of an element of $\boldsymbol{Y}_j^*$ from the underlying true, but unknown, model is termed pure error.

$$\varepsilon_{Pj} = y_j^* - g(\boldsymbol{X}_j) \qquad \mathbf{(3)}$$

Magnitude of the pure error (ε_{Pj}) depends on the particular choice of $\boldsymbol{Y}$- and $\boldsymbol{X}$-variables. By definition, pure error, which arises from effects not associated with the $\boldsymbol{X}$'s is independent of the $\boldsymbol{X}$'s and has zero expectation. Examples of omitted factors are myriad, but would include predicting species composition (the $\boldsymbol{Y}$*'s*) from Landsat spectra (the $\boldsymbol{X}$'s), but omitting elevation as an additional $\boldsymbol{X}$-variable that might improve the imputation.

Not so obvious as a source of pure error would be the effect of lack of accurate registration between the $\boldsymbol{Y}$-variable observation units located on the ground and the paired $\boldsymbol{X}$-variable observation units from a remote sensing platform. In effect, the observed values of $\boldsymbol{X}_j$ from a complete census from the erroneous position are just a differently defined variable for imputation than the $\boldsymbol{X}_j$'s from a properly registered observation unit. Therefore, variation from lack of registration would contribute to pure error that might be reduced by improving registration.

From [2] and [3]

$$y_j = g(\boldsymbol{X}_j) + \varepsilon_{Pj} + \varepsilon_{Yj} \qquad (4)$$

in which the error components include measurement error ($\varepsilon_{\mathrm{Y}j}$), and pure error ($\varepsilon_{\mathrm{P}j}$). Pure error and measurement error are inseparable in many data sets. To estimate pure error alone requires an external estimate of the measurement error. For example, if the observation unit is a spatial polygon represented by the mean of each of the attributes over a number of plots within the polygon, then the estimated variance-of-the-mean would provide the sampling portion of measurement variance to be subtracted to leave pure error.

3) Factors affecting the availability and similarity of reference observation units to serve as surrogates for the target units. This component depends on both the choice of a distance function and on the distribution of observation units in the space spanned by the ***X***-variables. Ideally, all the target data should be within the span of the reference data. The denser the data, the shorter will be the average distance between a target unit and its nearest surrogate in the reference set. And shorter distances usually imply greater similarity. The magnitude of this effect can be appreciated by comparing the distribution of distances to nearest neighbors among the reference data to the distribution of distances from each target observation to its nearest neighbor in the reference set. The distances between the real targets and their near neighbors in the reference set usually would be, on average, shorter than the distances among members of the reference set. Thus, estimated errors based only on the reference set will be biased upward. Effects of the density and range of the data apply to all methods of imputation and are determined by the inventory design.

4) And, finally, the choice of *k,* the number of reference observations and their relative weights in *k*-nn methods of estimating ***Y***'s as a weighted average of *k* near neighbors.

Error analyses we propose are based upon the data in the reference set. Inferences about the error properties of the estimates for the entire population based on these analyses depend on the extent to which the reference set represents the target set. As with inferences about any population parameter, appropriate randomization is a prerequisite to the assumption that the partitioning of error based on the reference set will apply to imputations for the real target set.

Imputation Error Statistics Based on the Reference Set

In the imputation context $\Sigma_i(y_{\mathrm{T}i}-y_{\mathrm{R}j})^2/n$ is the statistic commonly reported as "squared error" based on the *n* observation units in the reference set. We use the term Mean Square Difference (MSD) for it to emphasize that it is not an "error"—rather it is simply a function of the difference between two co-equal observations, neither of which is any more "true" than the other. In this and the expressions to follow, the subscript j identifies the reference observation to be imputed to the i^{th} pseudo-target observation unit. For each observation unit i, the value of j is determined by the minimum of d_{ij}^2 in (1). In k-nn imputation y_{Rj} is replaced by an average of k values of y_m using a weighting rule for the particular flavor of k-nn inference, where m is from the set of indices of the k observations selected as near-neighbors. We will develop the partitioning of error components for $k = 1$ because the notation is much more compact. However, the extension to $k > 1$ introduces no new concepts and will be treated when we discuss the choice of k as an error source.

Each member of the pairs being averaged in MSD includes stochastic components which do not change whether the observation unit is playing the role of target or reference. Each pair also includes a component determined by the distribution of the $\boldsymbol{X}$'s within the reference set. Thus, the statistics we compute are conditional on distribution of $\boldsymbol{X}$'s in the reference set—and may be used to guide decisions on how or whether to augment that reference set. The stochastic components, pure error and measurement error, are assumed to be drawn from distributions having zero mean and zero covariance. Therefore, for both stochastic error sources:

$$E[\varepsilon_{Pj}] = E[\varepsilon_{Yj}] = 0;\ \mathrm{Var}(\varepsilon_P) = E[\Sigma_j \varepsilon_{Pj}^2/n];\ \mathrm{Var}(\varepsilon_Y) = E[\Sigma_j \varepsilon_{Yj}^2/n];\ E[\varepsilon_{Yi}\varepsilon_{Pj}] = 0. \quad (5)$$

We will define the estimated variances of the stochastic error terms $\mathrm{var}(\varepsilon_P)$ and $\mathrm{var}(\varepsilon_{Yj})$ as the average over the reference set, dividing by n rather than $(n\text{-}p)$ because the error terms are defined relative to true values rather than from a computed mean.

We first introduce the measurement error from (2) into an addend of MSD:

$$(y_{Ti} - y_{Rj})^2 = (y_{Ti}^* + \varepsilon_{Yi} - y_{Rj}^* - \varepsilon_{Yj})^2 \quad (6)$$

Expanding (6) on the starred terms from (3) we have:

$$(y_{Ti} - y_{Rj})^2 = [g(\boldsymbol{X}_{Ti}) + \varepsilon_{Pi} + \varepsilon_{Yi} - g(\boldsymbol{X}_{Rj}) - \varepsilon_{Pj} - \varepsilon_{Yj}]^2 \quad (7)$$

Averaging over the n pseudo target units (y_{Ti}) in (7) assuming ε_{Pj} and ε_{Yj} are independent of each other and using (6), the expectation of MSD becomes:

$$E[\mathrm{MSD}] = E[\Sigma_i(y_{Ti} - y_{Rj})^2/n] = \Sigma_i\,[g(\boldsymbol{X}_{Ti}) - g(\boldsymbol{X}_{Rj})]^2/n + 2\,\mathrm{Var}(\varepsilon_Y) + 2\,\mathrm{Var}(\varepsilon_P) \quad (8)$$

The term: $\Sigma_i\, [g(\boldsymbol{X}_{Ti}) - g(\boldsymbol{X}_{Rj})]^2/n$ in (8) is, therefore, the error component arising from the distance between a pseudo-target point and its selected surrogate reference point. Note that in addition to the distance error component, the other error variances are included twice in MSD.

Estimating pure error and measurement error

In a regression context, sums of squares for pure error plus measurement error can be estimated from differences between the y's for observations having the same $\boldsymbol{X}$'s. The corresponding concept in imputation is for observations separated by zero Mahalanobis distance. Mahalanobis distances are calculated in the space spanned by the normalized, but uncorrelated $\boldsymbol{X}$-variables. The Mahalanobis distance was selected because other distance functions may transform the X's such that the dimension of the space spanned by the transformed $\boldsymbol{X}$'s is of lower dimension than the original $\boldsymbol{X}$-space. Zero distances in the space of reduced dimension would not necessarily indicate that $\boldsymbol{X}_{\mathrm{Ti}}$ for a target unit is identical to the $\boldsymbol{X}_{\mathrm{Rj}}$ for the selected reference unit. We argue that an estimate of the twice the sum of variances of pure error and measurement error can be obtained by averaging the squared differences for some fraction of the units with short Mahalanobis distances. We call this estimate MMSD(0), adding an initial M and the (0) to suggest it is derived from pairs of units with Mahalanobis distances of close to zero. Using (8),

$$E[\mathrm{MMSD}(0)] = 2\ \mathrm{Var}(\varepsilon_{\mathrm{P}}) + 2\ \mathrm{Var}(\varepsilon_{\mathrm{Y}}) + \mathrm{bias} \qquad (9)$$

where the bias equals the amount by which the mean of the squared distance component (as in (8) but averaged over only the observation units with close-to-zero distances) differs from zero. Note that whereas MSD may be derived from any of the many distance functions, MMSD(0) always uses Mahalanobis distance.

The estimate is biased by the average of $[g(X_{Ti}) - g(X_{Rj})]^2$ in MMSD(0). The bias might be reduced by regressing the values of $(y_{Ti} - y_{Rj})^2$ on their distances where the near-neighbor pairings are determined using a Mahalanobis distance function. The intercept of this regression may provide an improved estimate of MMSD(0) by extrapolation to zero distance. However, for some obstreperous $\boldsymbol{Y}$-variables, the squared deviations decline with increasing distance so that the intercept is above the mean. This circumstance indicates that the $\boldsymbol{X}$'s do not measure similarity for those elements of $\boldsymbol{Y}$ or that their stochastic components are heteroskedastic.

Estimating distance component

The distance component depends only on the range and density of the $\boldsymbol{X}$'s and on the measure of similarity used to select the near neighbor(s). Equation (8) showed that E[MSD] is comprised of the distance component, $\Sigma_i\, [\mathrm{g}(\boldsymbol{X}_{\mathrm{T}i}) - \mathrm{g}(\boldsymbol{X}_{\mathrm{R}j})]^2/n$ plus two times the sum of variances of pure error and measurement error. Therefore, the distance component of MSD can be estimated by subtracting twice the components of pure error and measurement error estimated by (9) in the previous section:

$$\boldsymbol{\Sigma_i}[\mathrm{g}(\boldsymbol{X}_{\mathrm{T}i}) - \mathrm{g}(\boldsymbol{X}_{\mathrm{R}j})]^2/n \approx \mathrm{MSD} - \mathrm{MMSD}(0). \tag{10}$$

This error component does not depend on the specific functional form of the relations of the $\boldsymbol{Y}$'s to the $\boldsymbol{X}$'s so any model lack-of-fit is not involved. Therefore, it applies equally to near-neighbor pairing of units without regard for the distance function. Unfortunately, MSD– MMSD(0) is not constrained to be positive if $[\mathrm{g}(\boldsymbol{X}_{\mathrm{T}i}) - \mathrm{g}(\boldsymbol{X}_{\mathrm{R}j})]^2$ decreases with increasing distance.

Using the partitioning to illuminate key questions

We now revisit key questions posed in the introduction, developing some new statistics based on the partitioning to provide answers.

Accuracy of imputed values

The fundamental variance statistic in sampling inference compares an estimate with its true value. In our notation that comparison is $y_{Rj} - \mathrm{g}(\boldsymbol{X}_i)$ for k equal to one. Therefore, we propose that the efficacy of the imputation process should be based on a statistic we term the Standard Error of Imputation (SEI).

$$\mathrm{SEI}^2 = \Sigma_i[y_{\mathrm{R}j} - \mathrm{g}(\boldsymbol{X}_{\mathrm{T}i})]^2\,/n \qquad i = 1,\cdots,n \text{ and } j \text{ minimizes } d_{ij}^{\,2} \tag{11}$$

Unfortunately, the addends in the bracket of SEI cannot be computed directly from the data in the reference set because the true value, $\mathrm{g}(\boldsymbol{X}_{\mathrm{T}i})$, is not directly observable. The proposed aggregate statistic (11), however, can be obtained by replacing the "estimate" $y_{\mathrm{R}j}$ in (11) with (4) evaluated for the j^{th} reference unit.

$$\mathrm{SEI}^2 = \Sigma_i[\mathrm{g}(\boldsymbol{X}_{\mathrm{R}j}) + \varepsilon_{\mathrm{P}j} + \varepsilon_{\mathrm{Y}j} - \mathrm{g}(\boldsymbol{X}_{\mathrm{T}i})]^2\,/n \tag{12}$$

Then averaging with the same assumptions of error independence used in deriving (8).

$$\mathrm{E}[\mathrm{SEI}^2] = \mathrm{E}[\Sigma_i[y_{\mathrm{R}j} - \mathrm{g}(\boldsymbol{X}_{\mathrm{T}i})]^2\,/n] = \Sigma_i[\mathrm{g}(\boldsymbol{X}_{\mathrm{R}j}) - \mathrm{g}(\boldsymbol{X}_{\mathrm{T}i})]^2/n + \mathrm{Var}(\varepsilon_{\mathrm{P}}) + \mathrm{Var}(\varepsilon_{\mathrm{Y}}) \tag{13}$$

which differs from MSD (8) by omitting the terms for the variances of pure error and sampling error arising from the target members of (11). If the distance component of MMSD(0) can be assumed to be trivially small when (8) is averaged over only the shorter distances, then:

$$\mathrm{E}[\mathrm{SEI}^2] = \mathrm{E}[y_{\mathrm{R}j} - \mathrm{g}(X_{\mathrm{T}i})]^2 \approx \mathrm{MSD} - \mathrm{MMSD}(0)/2. \tag{14}$$

Imputation compared to estimates using f(X)

The regression model is: $y_j^* = f(X_j) + \varepsilon_j$ where the ε_j includes pure error, and the lack of fit of the assumed model. The regression model could be, but is not limited to the familiar linear parameterization $f(\boldsymbol{X}_j) = \boldsymbol{BX'}$. Alternatively it could be a nonlinear or nonparametric regression model or a collection of means for strata defined by the $\boldsymbol{X}$'s. The true model $y_j^* = g(\boldsymbol{X}_j) + \varepsilon_{Pj}$ differs from the regression model by the lack-of-fit of the regression model:

$$\varepsilon_{L(\boldsymbol{X}j)} = g(\boldsymbol{X}_j) - f(\boldsymbol{X}_j). \tag{15}$$

The error statistic commonly calculated for a regression is the Standard Error of Estimate (SEE) (ignoring the reduction of the divisor by the number of estimated parameters):

$$SEE^2 = \Sigma_j(y_j - f(\boldsymbol{X}_j))^2/n\,. \tag{16}$$

We assume that the lack-of-fit will sum to zero for the particular $\boldsymbol{X}$'s (certain if $f(\boldsymbol{X})$ is fit by least-squares and includes an intercept) in the Reference set.

Then, from (2), (3) and (15):

$$(y_j - f(\boldsymbol{X}_j))^2 = (\varepsilon_{Pj} + \varepsilon_{Yj} + \varepsilon_{L(\boldsymbol{X}j)})^2 \tag{17}$$

The terms for the model lack-of-fit were assumed to be independent of the $\boldsymbol{X}$'s and of ε_{Pj} and ε_{Yj} so $E(SEE^2)$ is the sum of these three sources:

$$E[SEE^2] = E[\Sigma_j(y_j - f(\boldsymbol{X}_j))^2/n] = Var(\varepsilon_P) + Var(\varepsilon_Y) + \Sigma_j[\varepsilon^2_{L(\boldsymbol{X}j)}]/n \tag{18}$$

Comparison of $E[SEI^2]$ in (13) with $E[SEE^2]$ in (18) shows that they differ only by the substitution of the distance component, $\Sigma_i[g(\boldsymbol{X}_{Rj}) - g(\boldsymbol{X}_{Ti})]^2/n$, in imputation error variance for lack of fit, $\Sigma_j(\varepsilon_{L(\boldsymbol{X}j)})^2/n$, in regression estimation error variance.

Rearranging (18) and substituting (9):

$$\Sigma_j[\varepsilon^2_{L(\boldsymbol{X}j)}]/n = E[SEE^2] - E[MMSD(0)/2] \tag{19}$$

The ideal contents for a data set for subsequent analysis would be $\boldsymbol{Y}_j^*$ which would have variance about $g(\boldsymbol{X}j)$ of $Var(\varepsilon_P)$. Unfortunately, the best imputation can do for a given data set is $\boldsymbol{Y}_{Tj}$ which differs from the ideal by inclusion of measurement error variance plus the distance component. Alternatively, regression estimation could supply as estimates $f(\boldsymbol{X}_j)$ plus a random element drawn from a distribution with variance $\mathrm{Var}(\varepsilon_P)$. Using (4) and (15) and the independence of pure error relative to the model lack-of-fit, these estimates would have variance about $g(\boldsymbol{X}j)$ given by:

$$E[\Sigma_j[f(\boldsymbol{X}_j) + \varepsilon_{Pi} - g(\boldsymbol{X}_j)]^2/n] = \Sigma_j(\varepsilon_{L(\boldsymbol{X}j)})^2/n + \mathrm{Var}(\varepsilon_P) = E[SEE^2] - \mathrm{Var}(\varepsilon_Y) \tag{20}$$

which can be estimated by:

$$\Sigma_j[f(\boldsymbol{X}_j) + \varepsilon_{Pj} - g(\boldsymbol{X}_j)]^2/n = SEE^2 - MMSD(0)/2 + \mathrm{var}(\varepsilon_P) = SEE^2 - \mathrm{var}(\varepsilon_Y) \tag{21}$$

Subtracting (21) from (14) the comparison of SEI^2 to (21) becomes:

$$E[y_{Rj} - g(\boldsymbol{X}_{Ti})]^2 - E[\Sigma_j[f(\boldsymbol{X}_j) + \varepsilon_{Pi} - g(\boldsymbol{X}_j)]^2/n] \approx SEI^2 - [SEE^2 - \mathrm{var}(\varepsilon_Y)] \tag{22}$$

which is the same as (13)–(18) plus pure error variance:

$$\Sigma_i[g(\boldsymbol{X}_{Rj}) - g(\boldsymbol{X}_{Ti})]^2/n - [\Sigma_j(\varepsilon_{L(\boldsymbol{X}j)})^2/n + \mathrm{var}(\varepsilon_P)] \tag{23}$$

Thus, the variance of the imputed values would be greater than regression estimated values for each *y* if (22) or equivalently if (23) is greater than zero. However, the regression alternative would not guarantee that the true correlation among the estimated *y*'s within each observation unit would be retained.

Effects of distribution of X's

The second key question concerning distributions of the ***X***'s and alternative measures of similarity is addressed by considering the distance component of MSD: $\Sigma_i[g(\boldsymbol{X}_{Ti}) - g(\boldsymbol{X}_{Rj})]^2/n$. This error component should be made as small as possible either by adding new members to the reference set to reduce average distance between target units and their similar reference unit(s) or by adopting a better measure of similarity or both.

An important consideration in accuracy assessment based only on the reference observation units is the relation between the distribution of the ***X***'s in the target set in relation to that distribution in the reference set. Ideally, the reference set would completely cover the ranges of ***X***-variables of the target set and have an approximately uniform distribution over the range of the combined sets. The distance function being invoked may weight variation of some of the X's heavier than others, thereby stretching and rotating the space spanned by the X's, Therefore, the overall effect the distributions should be compared in terms of the distances between reference unit and the pseudo-target unit of the reference pairs of near neighbors and the distances between the paired reference unit and the real target for which imputations are required.

A statistic sensitive to the merits of alternative distance functions would reduce the influence of pure error and sampling error to focus on $\Sigma_i[g(\boldsymbol{X}_{Ti}) - g(\boldsymbol{X}_{Rj})]^2$. At short distances, the values of $(y_{Ti}-y_{Rj})^2$ are dominated by the pure error plus sampling error. Therefore, a better alternative to MSD calculated as the average over all references is to average only using pairs separated by the longer distances.

Choice of X's and their transformations

How these decisions affect MSD for a particular variable y depends on the choice of the weight matrix $\boldsymbol{W}$ in (1). If $\boldsymbol{W}$ gives little or no weight to a particular x, then that x is effectively omitted. Conversely an x may be heavily weighted because of its contribution to $g(\boldsymbol{X})$ for other y's. Then, even though a subset of the x's may effectively predict the y under consideration, their contribution will be diluted by differences in the extraneous x's and MSD for that element, y of $\boldsymbol{Y}$ will be dominated by pure error and measurement error to such an extent that $[g(\boldsymbol{X}_{Ti}) - g(\boldsymbol{X}_{Rj})]^2$ may decrease with distance. If it does decrease, then the distance component and model lack-of-fit will be under-estimated.

Transformations in variables are typically invoked to simplify a model such as $y=f(\boldsymbol{X})$ and to render errors more homogeneous. Consideration of (8) and (10) and (17) as estimates of sources of imputation errors from the three sources shows that transformations of the $\boldsymbol{X}$-variables, while modifying the fit of the regression model $y=f(\boldsymbol{X})$, affect MSD only through the distance component, $\Sigma_i[g(\boldsymbol{X}_{Ti}) - g(\boldsymbol{X}_{Rj})]^2/n$, and homogeneity of the pure error component. Transformations affect the distance component through the selection of surrogates, which in turn depend on the choice of the weight matrix $\boldsymbol{W}$. In dense regions of the space spanned by the $\boldsymbol{Xj}$'s of the reference set, the distance component in MSD is small relative to pure error plus measurement error for any choice of near neighbor. On the other hand, where the $\boldsymbol{X}_{Ti}$ are not closely spaced (sparse), their imputations to the $\boldsymbol{X}_{Tj}$ will be few in number, so their effect on MSD will be small. This ambiguity explains a puzzling property of near-neighbor imputation: that it has not appeared to be very sensitive to monotonic transformations of the variables. However, for imputation methods that base $\boldsymbol{W}$ on the relations of the $\boldsymbol{Y}$'s to the $\boldsymbol{X}_j$'s in distance calculations (e.g. MSN), the non-linear components represented by lack-of-fit would change the selection of "near neighbors". The extent of the change would be greatest in pairs of observation units in which model lack-of-fits were of opposite sign.

Choice of k

The partitioning of error provides useful insight concerning the choice of k for imputation using a weighted average of k near neighbors. The obvious effect is that larger k, by averaging over the errors of more reference observations, would seem to reduce the error of the imputed value. However, it is not that simple. Following the same assumptions used in deriving (8) MSD becomes:

$$\mathrm{E}[\Sigma_{\mathrm{i}}[y_{\mathrm{T}i} - \Sigma_m w_{im} \mathrm{y}_{\mathrm{R}m})]^2/n] = \Sigma_i\, [\mathrm{g}(\boldsymbol{X}_{\mathrm{T}i}) - \Sigma_m w_{im}\mathrm{g}(\boldsymbol{X}_{\mathrm{R}m})]^2/n + (1 + \Sigma_{\mathrm{i}}\Sigma_m w_{im}{}^2/n)[\mathrm{Var}(\varepsilon_{\mathrm{Y}}) + \mathrm{Var}(\varepsilon_{\mathrm{P}})] \quad (24)$$

In k-nn imputation $y_{\mathrm{R}j}$ of (8) is replaced by an average of k values of y_m using a weighting rule for the particular flavor of k-nn inference, where m is from the set of indices of the k observations selected as near-neighbors to the i^{th} target and $\boldsymbol{\Sigma_m} w_{im} = 1$. When $w_{im} = 1/k$, the multiplier of the variances in (24) becomes $(1 + 1/k)$. To the extent that it is pure error being reduced, increasing k is counter-productive for the subsequent analysis. Offsetting this effect, measurement error will also be reduced in the same proportion. Hence there is a tradeoff, either lose valuable pure error or reduce undesirable measurement error. The net effect of changing k also depends on the change in $\Sigma_i[\mathrm{g}(\boldsymbol{X}_{\mathrm{T}i}) - \Sigma_m w_{im}\mathrm{g}(\boldsymbol{X}_{\mathrm{R}m})]^2/n$. Whether this component increases or decreases the total error depends on the change of $[\mathrm{g}(\boldsymbol{X}_{\mathrm{T}i}) - \Sigma_m w_{im}\mathrm{g}(\boldsymbol{X}_{\mathrm{R}m})]^2$ for the reference observation being added or omitted by changing k.

Application to Example Data Sets

Three data sets will be used to illustrate the estimation of error components and application of these estimates in evaluating alternative weight matrices. All three use suites of remotely sensed data and data from digital terrain models to impute data from ground-based observations. As examples of real imputation analyses, they illustrate the behavior of the statistics we propose. We do not purport to second-guess the analysis of these data sets, so the definitions of the 69 specific variables in these three data sets are mostly irrelevant to our purposes. Where we do discuss behavior of the partitioning as a consequence of the biological situation, we will define those variables explicitly in the text. Otherwise, readers desiring more detail are directed to the original sources.

The first example uses data used by Moisen and Frescino (2002) obtained by the USDA Forest Service, Rocky Mountain Experiment Station Forest Inventory and Analysis Unit (FIA). The ground-based data (***Y***-variables) are from routine FIA observations for Utah, USA. The ***X***-variables were obtained from LANDSAT and digital terrain data.

The other two data-sets use ground data from inventories of stands defined as polygons. One, from the Deschutes National Forest in Oregon, USA has been used in previously reported analyses by Moeur (2000) and is the example in the MSN User's Guide (Crookston et al. 2002). The third data set is from Tally Lake area in the Helena National Forest in Montana, USA. For these comparisons, the ***Y***-variables will be limited to those measured on continuous scales. These analyses differ from those reported by Stage and Crookston (2002) in that all discrete and a few redundant *y*'s have been omitted to achieve approximately equal numbers of *y*'s in the three examples, and additional *x*'s (transformations of the original variables) have been added. Table 1 summarizes numbers of variables and sample sizes for the three data-sets. Of the three data sets, Users Guide has remarkably fewer observations in relation to the number of unique coefficients in the weight matrix being estimated (last line, Table 1).

The Utah data set differs from the other two in that it contains a notable portion of locations in non-forest although the continuous ***Y***-variables describe forest stand parameters. By contrast, *y*-values of zero in the other two data sets indicate lack of stocking in otherwise forested polygons. Proportion of zeroes in the three data sets are indicated in figure 2.

Table 2 summarizes the structure of the correlations between the canonical vectors for the three data sets. Multivariate regression R^2 of y on ***X*** are listed in col. B of table 2. Correlations between the ***Y***'s and the ***X***'s were lowest in the Utah data because the measurement errors of the ***Y***'s from the FIA plot clusters were larger than in the two data sets based on inventories of stand polygons.

Components of Variance

Data for partitioning variance for the three example data sets are displayed in table 3. Columns A-C contain statistics for each ***y***-variable considered independently of the remaining elements of ***Y***. Columns D-F contains statistics for each *y*-variable, but for pairs of near neighbors selected using a multivariate Mahalanobis distance measure.

Accuracy of imputed values

Standard error of imputation squared (SEI^2) (as a fraction of variance of each variable) of values imputed using a Mahalanobis distance function are shown in figure 3. The error component arising from distance between target and reference: $\Sigma_i[(g(\boldsymbol{X}_{Ti}) - g(\boldsymbol{X}_{Rj}))^2]/n$ as estimated by (10) is shown in figure 3 by the shaded portions of the bars for each y-variable. This figure also shows the combined components of pure error and measurement error as estimated by (9).

Imputation compared to linear regression

Figure 4 compares the distance component of imputations (plotted as its negative) with the model-lack-of-fit computed as SEE^2-Min(MMSD(0)/2, SEE^2) for $f(\boldsymbol{X}) = \boldsymbol{\beta X'}$. As a corollary of the differences between imputation distance component and regression lack of fit, SEE^2 is almost alsways less than SEI^2. The exceptions to the inequality are Crown Cover (CCover) and logarithm of Pinus ponderosa volume (Vlg'PP) in the Tally Lake data set. We conjecture that the linear regression is just not a very effective model for crown cover, and that the large proportion of zero data for Pinus ponderosa preclude effective prediction of volume. Also, there would be two anomalies leading to negative estimates of lack-of-fit if the minimum of MMSD(0) and SEE^2 were not used: logarithm of Engelmann spruce volume (Vlg'ES) in the Tally Lake data and net growth in cubic feet (NGRWCF) in the Utah data. The larger values of MMSD(0) for these variables are the consequence of squared differences between y_{Ti} and y_{Rj} that decrease with increasing differences in the $\boldsymbol{X}$-variables. As a result, MMSD(0) is larger than SEE^2. We attribute this anomaly to unequal pure error in different regions of the $\boldsymbol{X}$-space. Engelmann spruce in the Tally Lake area occurs bi-modally with elevation—either very common at high elevations, or as sparse stringers in valley bottoms. However, the density in the $\boldsymbol{X}$-space of the observations representing valley bottoms and stands at similar elevations is higher than the density of data representing high elevations. Thus observation pairs with near-zero distances tend to come from low elevations where the sporadic presence of spruce gives large squared differences whereas at high elevations, spruce is more ubiquitous giving smaller differences in volume even at larger separations in $\boldsymbol{X}$-space.

That SEE is almost always less the SEI is not surprising because whereas SEE is a least squares minimization of the model prediction, SEI is not the result of an explicit minimization and includes the pure error and measurement error components. When pure error should be included in estimates for subsequent analyses, the proportion of pure error that might be added to regression lack of fit that would just make (23) equal zero is indicated by the white bars in figure 4. Unfortunately, we lack a direct estimate of measurement error that should be subtracted from SEI, so we can only show the margin from which it would be subtracted.

Effect of distances between X's

The three data sets show differences in the proportions of variance attributable to the Mahalanobis distances between target and reference (figure 3, shaded bar). The low ratio of number of observations compared to number of coefficients to be estimated and large linear model lack-of-fit of the User's Guide data produces a relatively large distance component compared to the Tally Lake data. Utah data show an intermediate level because the effect of the larger number of data relative to the number of coefficients to be estimated is offset by the low correlations between the ***Y***'s and the ***X***'s (table 3) caused by the inclusion of non-forest observations (figure 2).

In the Tally Lake application, average distances from reference observation units to actual target observation units is 2.04 times the average distance from each reference observation unit to its nearest neighbor also in the reference set. Nearly one-third of the targets are farther from their nearest reference than the ninth percentile of the distribution of distances among the references. The significance of this extrapolation might be determined by modeling squared differences for each element of ***Y*** as a function of distance. Such analysis is beyond the scope of this report.

Comparison of alternative distance functions

The difficulty of using MSD to compare alternative distance functions can be appreciated by considering that the influence of pure error plus sampling error would be double that shown in figure 3. Although the absolute value of differences in MSD arising from different distance functions would not change, the relative importance of the differences among the alternative distance functions would be under-estimated.

Figure 5 a,b,c compares three alternative distance functions, the Mahalanobis distance used heretofore in this report, the original canonical-correlation-based distance (CC) of Moeur and Stage (1995), and the newer canonical-regression-based distance (CR) introduced by Stage and Crookston (2002). The panels present both estimated means of $[g(\boldsymbol{X}_{Ti}) - g(\boldsymbol{X}_{Rj})]^2$ based on all data for comparison to means for the 50% of the data separated by the longer distances. Only the Utah data show the alternative similarity measures to rank differently in the full data set than in the reduced data set containing only the 50% longer distances. Also, the Utah data set was the only one to show a distinct advantage to using one or the other of the canonical-based distances over the Mahalanobis distances. And the differences would be even greater if the non-forest data were masked because the Mahalanobis distances did slightly better at matching the zero data. The result seems anomalous because the Utah data had the lowest canonical correlations between ***Y***'s and ***X***'s. However, one of the merits of the canonical approach lies in its capability to ignore ***X***'s that are irrelevant. Moisen and Frescino (2002) found that several of the *x*'s were superfluous. The Mahalanobis distance would have given these variables weights equal to the weights of the useful variables. The other two data sets were obtained after extensive analysis by others that probably had already screened the ***X***'s for utility.

Conclusions

This report concerns the error properties of imputation processes used to fill in a data set by imputing values from a sample of intensively measured observation units to interspersed, less completely measured units. The error statistics for the imputed, continuous-valued variables presented in this report are based on partitioning of the error components into measurement error, error inherent in the particular imputation method and the pure error not associated with the variables measured on all observation units. These statistics can assist in the design of inventories and their analysis with near-neighbor imputation methods. It is now possible to consider the relative gains from reducing measurement error versus increasing the density of the sampled observation units. They also clarify comparisons to other inference methods such as regression or stratum-mean based estimators, and help to choose among alternative weight matrices in similarity measures.

References

Crookston, N.L., M. Moeur, and D. Renner, 2002. Users Guide to the Most Similar Neighbor imputation program Version 2. Gen. Tech. Rep. RMRS-GTR-96. Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 35 p.

Malinen, J. 2003. Locally adaptable non-parametric methods for estimating stand characteristics for wood procurement planning. Silva Fennica 37(1):109-120

Moeur, M. 2000. Extending stand exam data with most similar neighbor inference. P 99-107 *in*: Proceedings of Soc. of Amer. Foresters National Convention; 1999 Sept. 11-15: SAF Pub 00-1.

Moeur, M., and A.R. Stage, 1995. Most Similar Neighbor: An improved sampling inference procedure for natural resource planning. For. Sci. 41:337-359.

Moisen, G. G. and T.S. Frescino. 2002. Comparing five modeling techniques for predicting forest characteristics. Ecol. Modelling 157:209-225.

Mullin, M. and R. Sukthankar. 2000. http://www-cgi..cs.cmu.edu/~rahals/put/icml2000-rahals.pdf (no longer works)

Ohmann, J.L. and M.J. Gregory 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon, U.S.A. Can. J. For. Res. 32:725-741.

Podani, J. 2000. Introduction to the exploration of multivariate biological data. Backhuys, Leiden, The Netherlands.

Shao, J. and R.R. Sitter. 1996. Bootstrap for imputed survey data. J. Amer. Stat. Assoc. 91 (435): 1278-1288.

Sohn, Y., E. Moran, and F. Gurri, 1999. Deforestation in north-central Yucatan (1985-1995): Mapping secondary succession of forest and agricultural land use in Sotuta using the cosine of the angle concept. Photogrammetric Engineering & Remote Sensing 65(8):947-958.

Stage A.R. and N.L. Crookston. 2002. Measuring similarity in nearest neighbor imputation: Some new alternatives. P.91-96. In Symposium on statistics and information technology in forestry. 2002 September 8-12. Virginia Polytechnic Institute and State University. Blacksburg, VA. http://www.forestryubc.ca/prognosis/documents/MSN_StageCrookston.pdf

Table Captions

Table 1. Statistics for three data sets used as examples. Number of coefficients to be estimated in relation to number of samples.

Table 2. Comparison between three example data sets of first four squared canonical correlations between ***Y***'s and ***X***'s.

Table 3. Table 3. Components of variance for three example data sets. Columns C – F are standardized by division by variance in column A Column B and C are for a linear model used as y=f($\boldsymbol{X}$). Columns D-F are obtained with a Mahalanobis distance function.

Table 1. Statistics for three data sets used as examples. Number of coefficients to be estimated in relation to number of samples.

	Tally Lake	Users Guide	Utah
Number of Y variables	8	6	10
Number of X's (p)	21	12	12
Number of reference obs. (n)	847	197	1076
Significant canonical pairs (s)	7	5	4
n/(s+p*s)	5.50	3.03	16.55

Table 2. Comparison between three example data sets of first four squared canonical correlations between ***Y***'s and ***X***'s.

Canonical Pair (m)	*Tally Lake*	*User's Guide*	*Utah*
1	0.697	0.686	0.450
2	0.477	0.456	0.153
3	0.325	0.376	0.109
4	0.292	0.244	0.034

Table 3. Components of variance for three example data sets. Columns C – F are standardized by division by variance in column A Column B and C are for a linear model used as y=f($\boldsymbol{X}$). Columns D-F are obtained with a Mahalanobis distance function.

Y-Variable	Total variance of Y-variable in reference set	Multivariate regression R^2	Squared error about regression of single Y SEE^2 1.-B	Mean square between target and nearest reference for all pairs in MSD	Mean square between target and nearest reference for 1/8 of shorter distances (MMSD(0))	Calculated Distance component D –E
	(A)	(B)	(C)	(D)	(E)	(F)
			Tally Lake			
Top height	566.669	0.6713	0.3287	0.6990	0.2837	0.4153
Vlg'AF	8.69601	0.4716	0.5284	1.0380	0.8461	0.1919
Vlg'ES	9.01080	0.4322	0.5678	1.2098	1.4075	-0.1977
Vlg'DF	7.03682	0.3696	0.6304	1.0064	0.6292	0.3772
CCover	222.797	0.2999	0.7001	1.1628	1.0061	0.1567
Vlg'L	6.66466	0.2556	0.7444	1.3189	0.9271	0.3918
Vlg'LP	8.18933	0.1956	0.8044	1.4893	1.0475	0.4418
Vlg'PP	0.71893	0.1076	0.8924	1.1779	0.6486	0.5294
			Users Guide			
TotBA	2822.19	0.5917	0.4083	0.7695	0.735	0.0345
LN-FIR	4.43945	0.5440	0.4560	0.8217	0.087	0.7347
TopHT	292.968	0.4839	0.5161	0.9453	0.287	0.6583
LN_PINE	7.0639	0.3858	0.6142	1.0768	0.087	0.9898
LN-BADF	1.85368	0.3548	0.6452	0.9557	0	0.9557
LN-BALP	3.55926	0.3225	0.6775	1.2927	0.6712	0.6215
			Utah			
MAICF	684.346	0.3567	0.6433	1.1259	0.3334	0.7925
NVOLTOT	2064882.	0.3142	0.6858	1.3522	0.7868	0.5654
NVOLMER	1546287.	0.2976	0.7024	1.3472	0.8143	0.5329
BA	4211.15	0.2736	.07264	1.3271	0.7525	0.5746
CRCOV	779.175	0.2621	0.7379	1.4426	0.8551	0.5876
STAGECL	3746.17	0.2528	0.7472	1.3367	1.0754	0.2613
NGRWCF	905.920	0.2434	0.7566	1.4868	2.1123	-0.6255
BIOTOT	636.018	0.2390	.07610	1.4758	0.5886	0.8872
NGRWBA	0.75238	0.2280	0.7720	1.5302	1.0740	0.4562
QMDALL	19.5868	0.1711	0.8289	1.6427	0.6462	0.9965

Figure Captions

Figure 1. Error components for imputing y_{Rj} (e.g. species volume) to a target observation at x_{Ti} from one of two reference observations in a one dimensional space of $\boldsymbol{X}$ (e.g. elevation). Pure error (ε_{Pi}) is the vertical distance from y_i* to the dashed line g(x). Measurement error (ε_{Yi}) is the vertical distance between y_i* and y_i. Model lack-of-fit ($\varepsilon_{L(Xj)}$) is the vertical separation between the dashed g($\boldsymbol{X}$) and solid f($\boldsymbol{X}$) lines.

Figure 2. Proportion of zero values in example data sets.

Figure 3. Partitioning of relative variance of imputed values (SEI equation (13)) for Mahalanobis distance function. Variables within a data set are ordered from left to right by increasing SEE. Values standardized by division by attribute variance.

Figure 4. Distance error component of imputation (plotted as its negative) compared to lack of fit of a linear regression, and pure error plus measurement error. Clear portion of the bar is amount of error that would be added to lack of fit to make expression (23) equal zero. Stippled bar is remaining portion of pure error plus measurement error. Variables within a data set are ordered from left to right by increasing SEE for a linear regression model. Values standardized by division by attribute variance.

Figure 5a. Tally Lake Comparison of distance components (10) for two canonical-correlation-based distance functions with Mahalanobis distance function. Variables within a data set are ordered from left to right by increasing SEE.

Figure 5b. Users Guide. Comparison of distance components (10) for two canonical-correlation-based distance functions with Mahalanobis distance function. Variables within a data set are ordered from left to right by increasing SEE.

Figure 5c. Utah. Comparison of distance components (10) for two canonical-correlation-based distance functions with Mahalanobis distance function. Variables within a data set are ordered from left to right by increasing SEE.

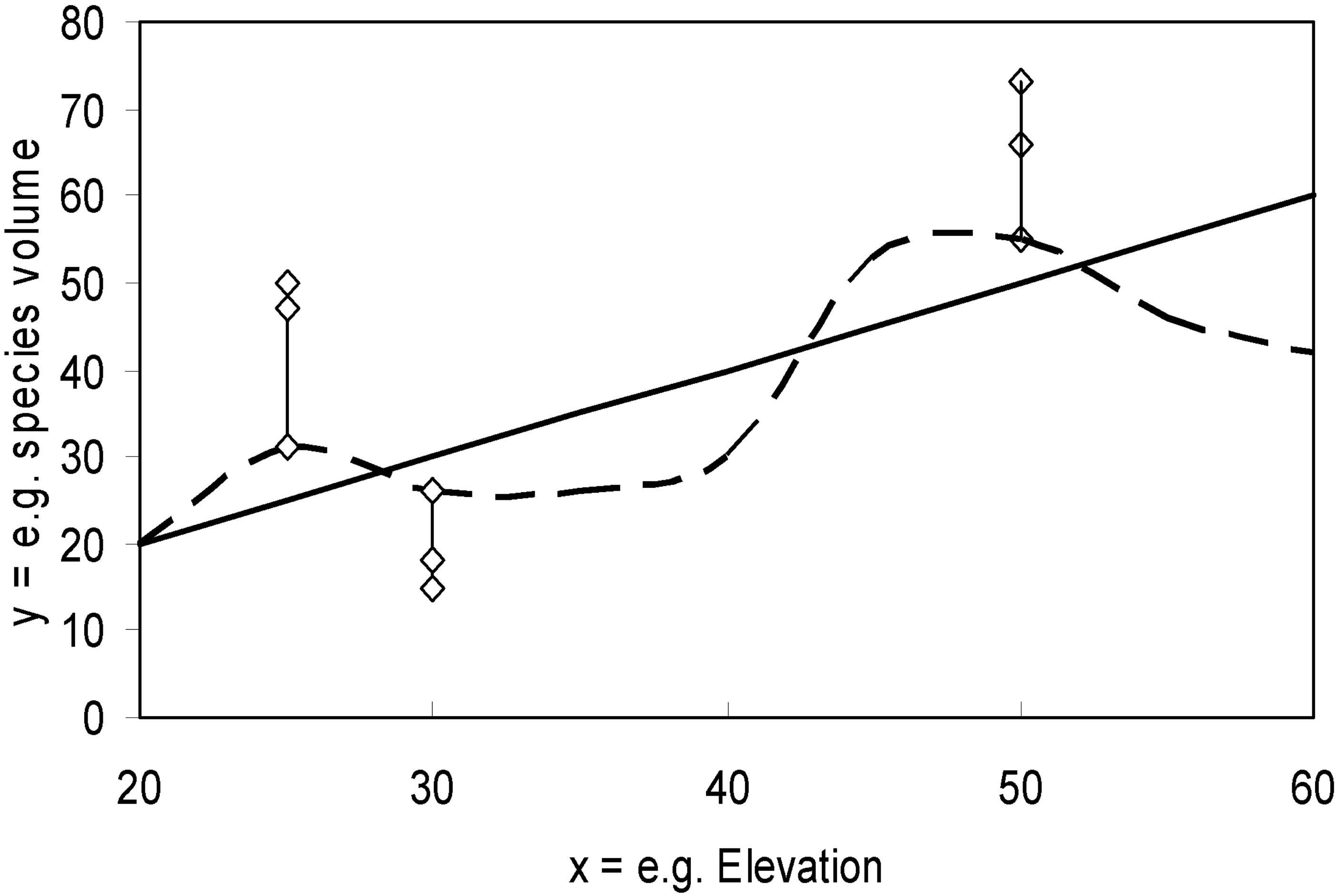
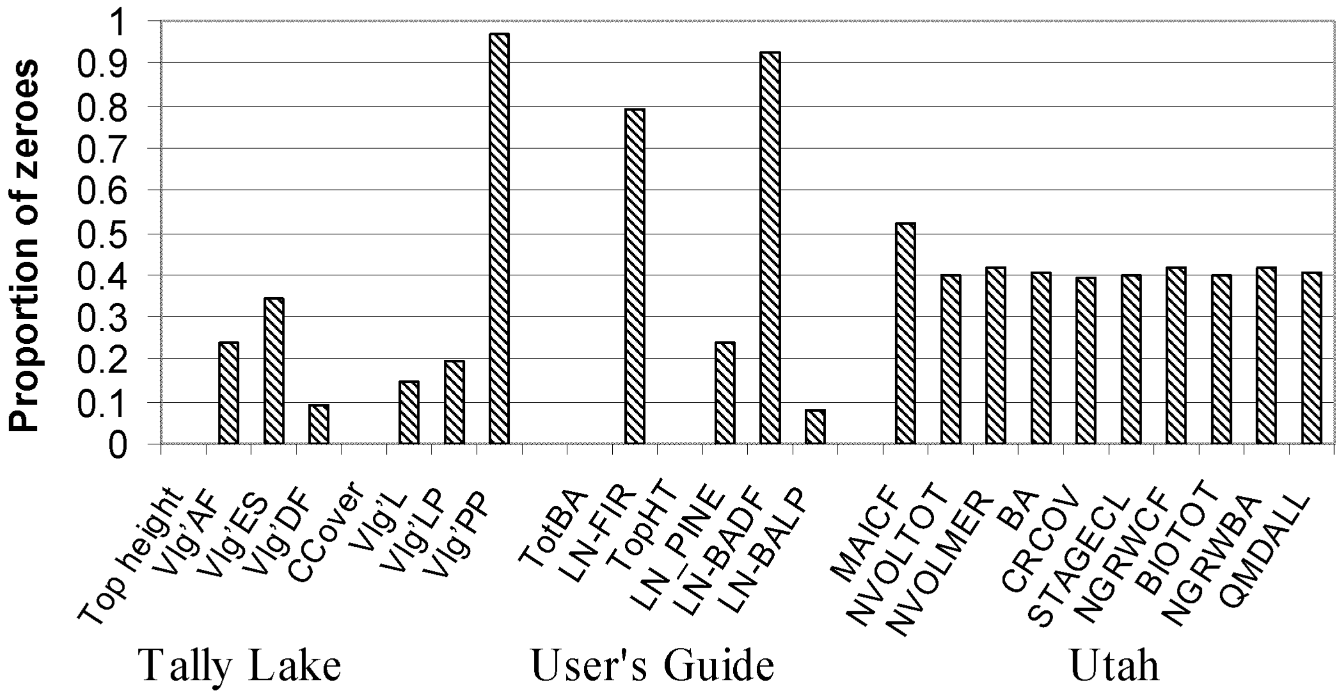
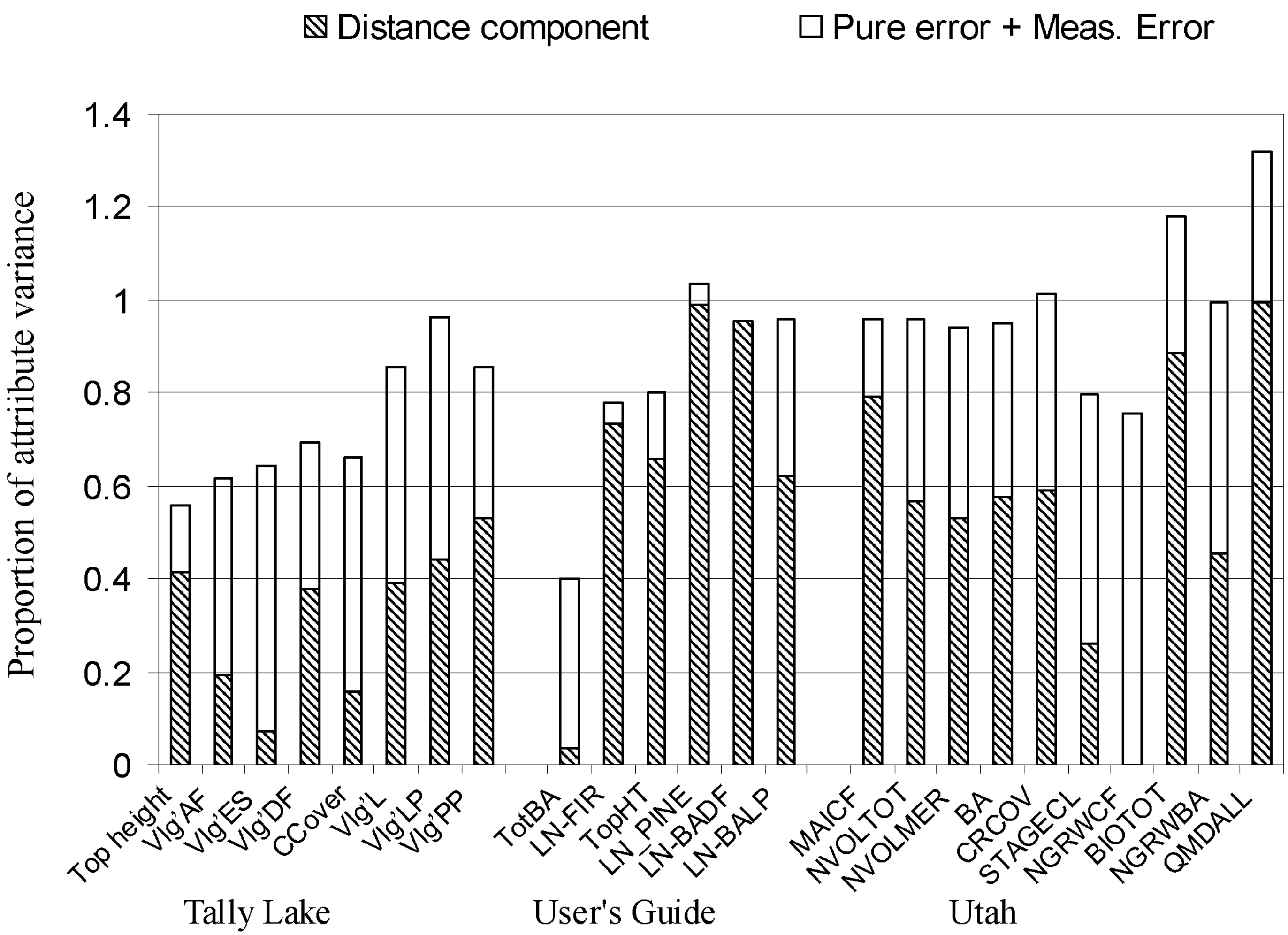

80
70
60
50
40
30
20
10
0
y = e.g. species volume
20
30
40
50
60
x = e.g. Elevation

Fig. 1



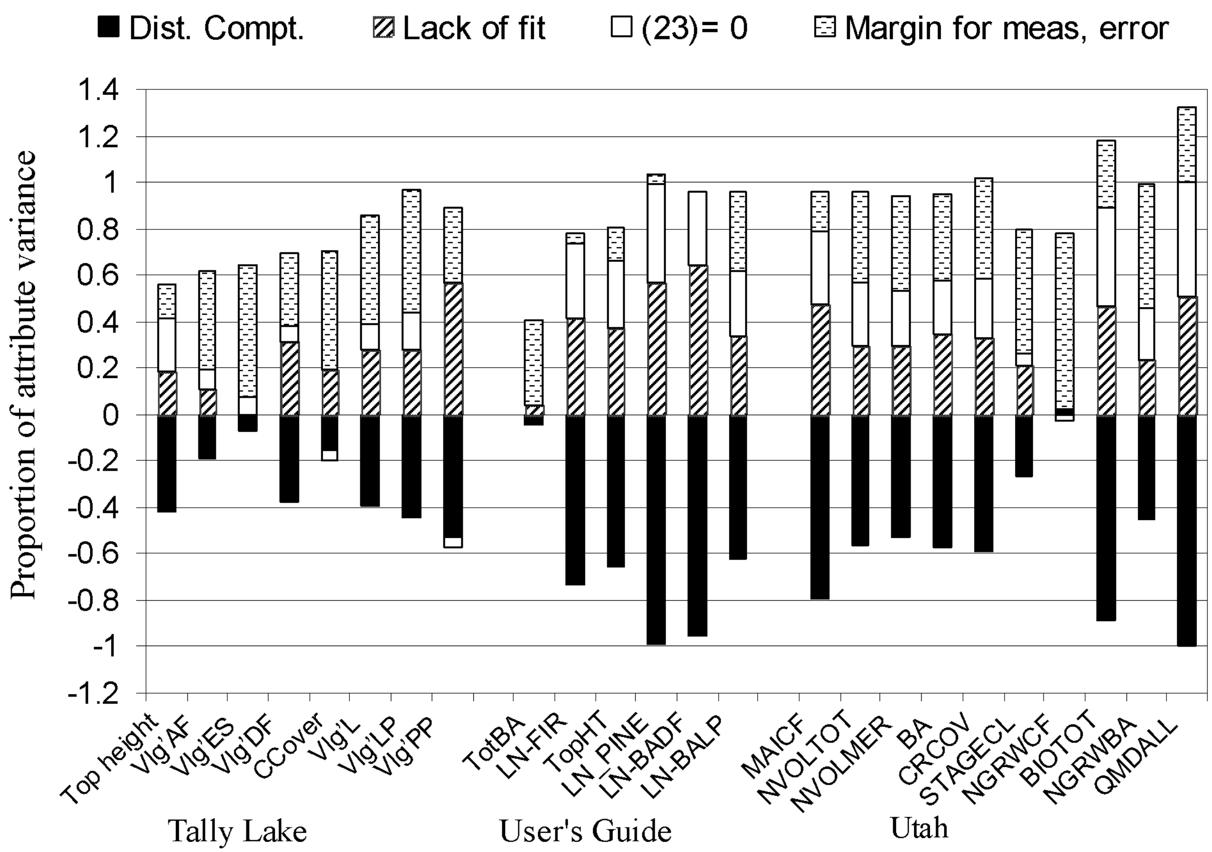
Proportion of zeroes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Top height
Vlg'AF
Vlg'ES
Vlg'DF
CCover
Vlg'L
Vlg'LP
Vlg'PP
TotBA
LN-FIR
TopHT
LN_PINE
LN-BADF
LN-BALP
MAICF
NVOLTOT
NVOLMER
BA
CRCOV
STAGECL
NGRWCF
BIOTOT
NGRWBA
QMDALL
Tally Lake
User's Guide
Utah

Fig. 2



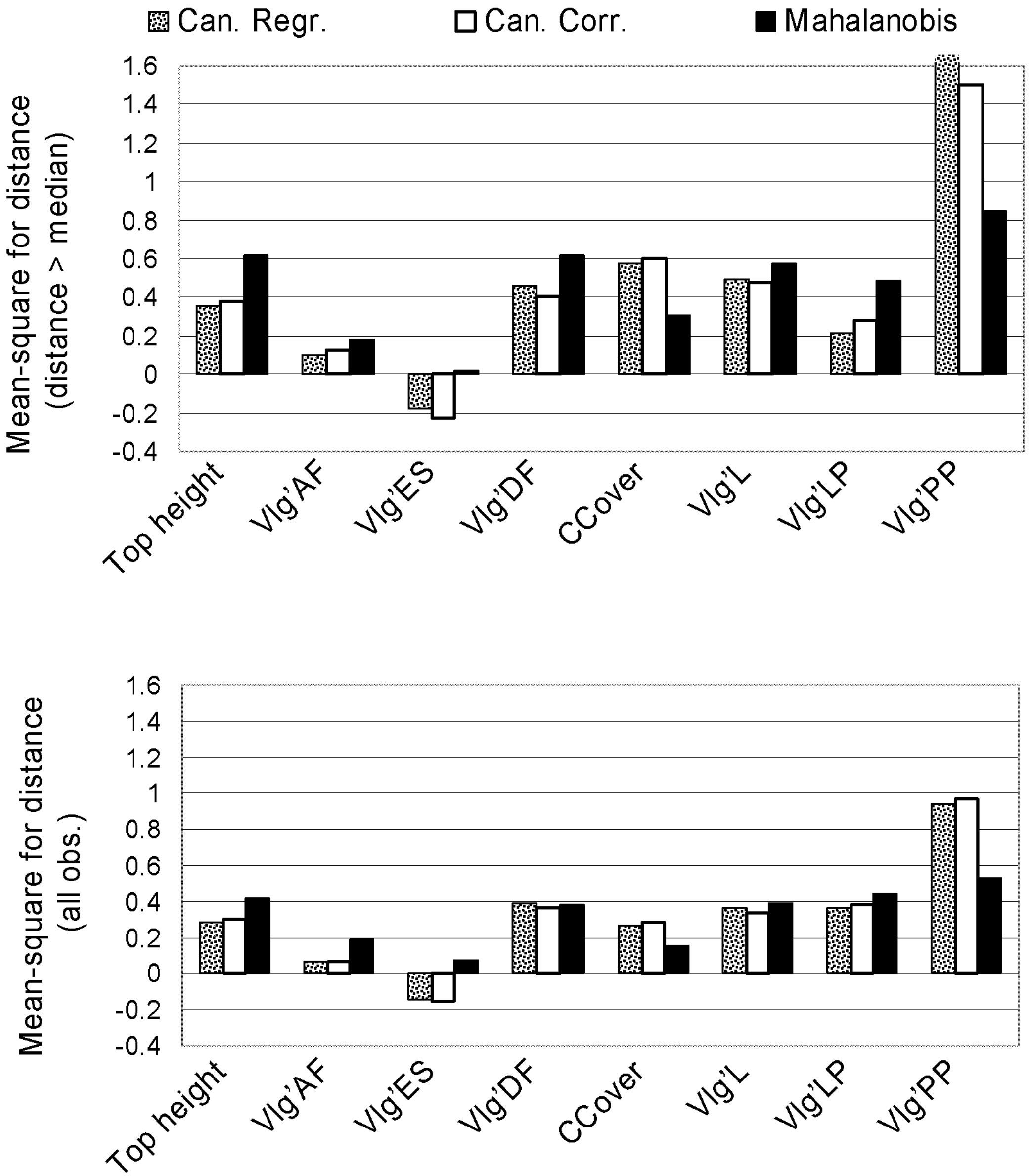
Distance component
Pure error + Meas. Error
Proportion of attribute variance
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Top height
Vlg'AF
Vlg'ES
Vlg'DF
CCover
Vlg'L
Vlg'LP
Vlg'PP
TotBA
LN-FIR
TopHT
LN_PINE
LN-BADF
LN-BALP
MAICF
NVOLTOT
NVOLMER
BA
CRCOV
STAGECL
NGRWCF
BIOTOT
NGRWBA
QMDALL
Tally Lake
User's Guide
Utah

Fig 3



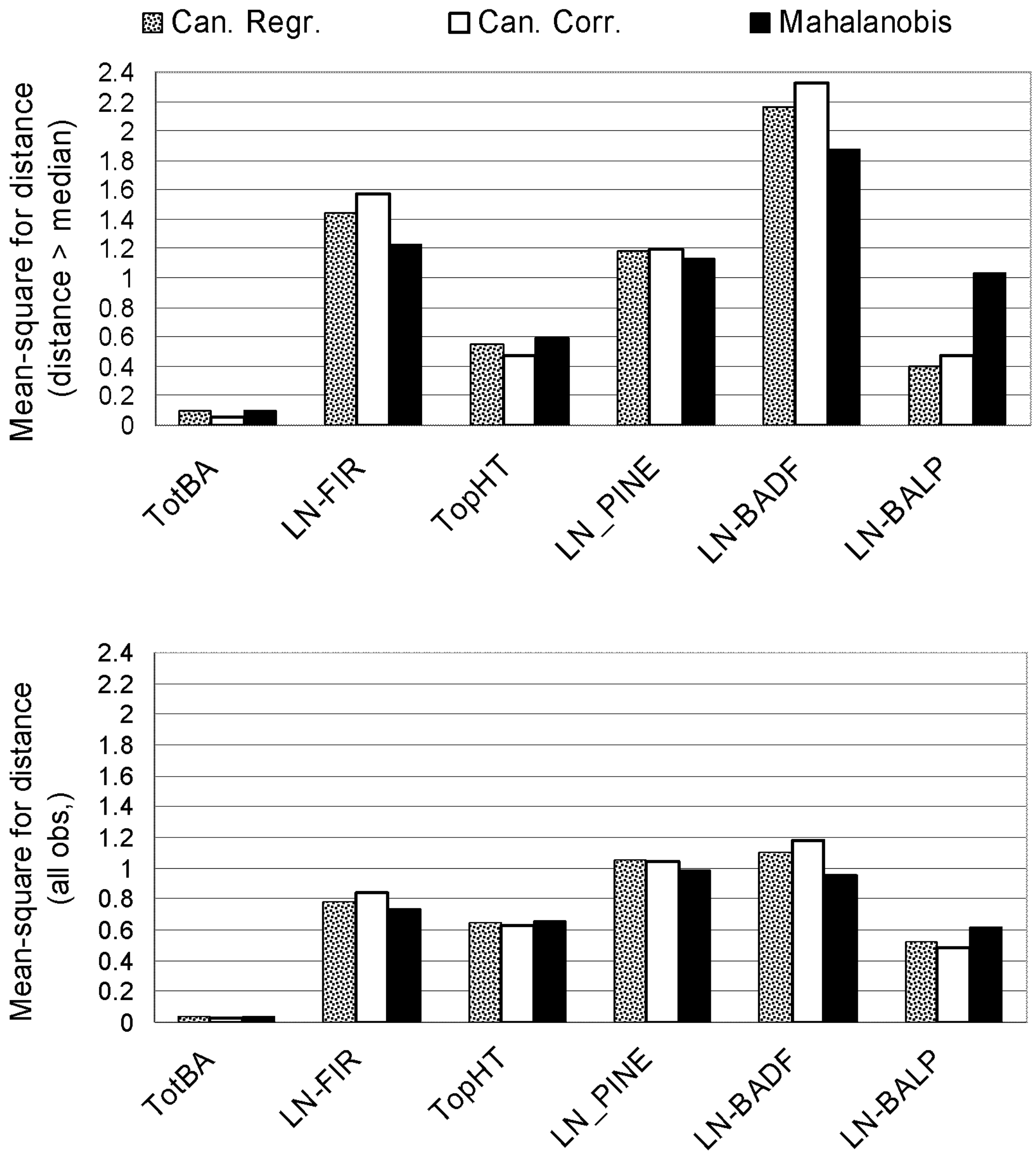
Dist. Compt.
Lack of fit
(23)= 0
Margin for meas, error
Proportion of attribute variance
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
Top height
Vlg'AF
Vlg'ES
Vlg'DF
CCover
Vlg'L
Vlg'LP
Vlg'PP
TotBA
LN-FIR
TopHT
LN_PINE
LN-BADF
LN-BALP
MAICF
NVOLTOT
NVOLMER
BA
CRCOV
STAGECL
NGRWCF
BIOTOT
NGRWBA
QMDALL
Tally Lake
User's Guide
Utah

Fig. 4



Can. Regr.
Can. Corr.
Mahalanobis
Mean-square for distance
(distance > median)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Top height
Vlg'AF
Vlg'ES
Vlg'DF
CCover
Vlg'L
Vlg'LP
Vlg'PP
Mean-square for distance
(all obs.)

Fig 5a



Can. Regr.
Can. Corr.
Mahalanobis
Mean-square for distance (distance > median)
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
TotBA
LN-FIR
TopHT
LN_PINE
LN-BADF
LN-BALP
Mean-square for distance (all obs,)
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
TotBA
LN-FIR
TopHT
LN_PINE
LN-BADF
LN-BALP

Fig 5b

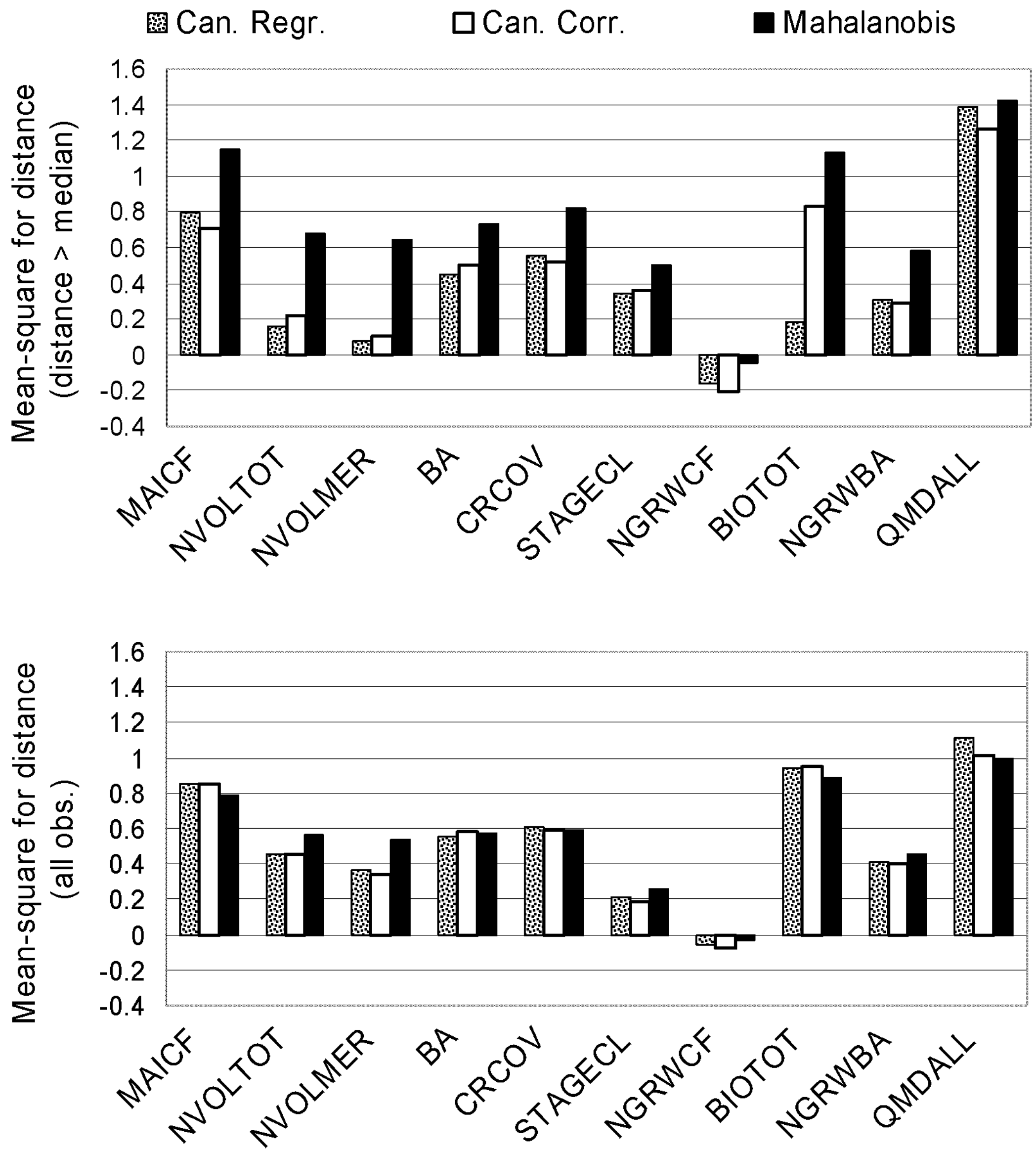

Can. Regr.
Can. Corr.
Mahalanobis
Mean-square for distance
(distance > median)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
MAICF
NVOLTOT
NVOLMER
BA
CRCOV
STAGECL
NGRWCF
BIOTOT
NGRWBA
QMDALL
Mean-square for distance
(all obs.)
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
MAICF
NVOLTOT
NVOLMER
BA
CRCOV
STAGECL
NGRWCF
BIOTOT
NGRWBA
QMDALL

Fig 5c