

Measuring Similarity in Nearest Neighbor Imputation: Some New Alternatives

Albert R. Stage, Principal Biometrician (Retired)
Nicholas L. Crookston, Operations Research Analyst
Rocky Mountain Research Station
USDA Forest Service
1221 South Main St.
Moscow, Idaho 83843 USA
astage@moscow.com
ncrookston@fs.fed.us

ABSTRACT

The Most Similar Neighbor (MSN) method of imputation shares many properties of the collection of imputation methods based on measures of similarity between observations. As described by Moeur and Stage (1995), however, MSN had two unique attributes: using a single member of the set of reference observations as the surrogate for the target observation's missing variables; and using canonical correlation analysis to determine attribute weights for measuring the similarity between target and reference observations. In this paper, we report an analysis leading to an alternative weight matrix in the distance function for MSN. We also discuss the statistical properties of the consequent differences between observed and imputed values of the Y-variables. These error properties are relevant to the choice between using a single reference observation for imputation to a target observation rather than means of several reference observations. Newly released MSN software now provides several optional distance-weighting functions, identification of k nearer neighbors, and improved computational algorithms for the canonical correlation analysis.

INTRODUCTION

Most Similar Neighbor (MSN), as originally formulated by Moeur and Stage (1995), differs from related nearest-neighbor methods of imputation in its use of correlations between the variables known for the entire population (X 's) and more important, but costly to obtain, variables known only on a sampled subset (Y 's) of the population. However, experience comparing MSN with imputations based on minimizing a Euclidean distance or a Mahalanobis distance calculated only on the X 's has shown that the addition of the Y vs. X correlations does not always improve the resolution of the imputed values. In this paper, we compare a newer formulation of the weight matrix in MSN to the original version and to alternatives using only the X 's. Our objective is to provide some insight into choosing among alternative methods of imputation.

Effects of the reformulation on estimates of the mean-square differences between the reference observations and their most similar neighbor will be shown for three example data-sets. All three use suites of remotely sensed data and data from digital terrain models to impute data from ground-based observations. For the first example, supplied by Gretchen Moisen, the ground-based data are obtained from routine Forest Inventory and Analysis (FIA) observations for Utah. The other two data-sets use ground data from inventories of stands defined as polygons. One, from the Deschutes National Forest 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, Montana. Table 1 summarizes numbers of variables and sample sizes for the three data-sets.

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

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

REFORMULATION OF THE MSN WEIGHT MATRIX

Moeur and Stage (1995) proposed that members of the vector of Y-variables be selected according to their interest for the intended uses of the completed data-set. A Mahalanobis¹ distance function measuring similarity would ideally consist of the Y-variables, *per se*:

$$d_{ij}^2 = (\mathbf{Y}_i - \mathbf{Y}_j) \mathbf{E}_{\mathbf{Y}\mathbf{Y}}^{-1} (\mathbf{Y}_i - \mathbf{Y}_j)'$$

Unfortunately, they were not observed everywhere, so, in their place, Moeur and Stage (1995) proposed considering their predictions, based on the X-variables that *are* available as a reasonable substitute. Their distance function is:

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

where:

- \mathbf{X}_i is the vector of X-variables for the i^{th} target observation,
- \mathbf{X}_j is the vector of X-variables for the j^{th} reference observation,
- $\mathbf{\Gamma}, \mathbf{\Lambda}$ are the matrices of canonical coefficients for \mathbf{X} and \mathbf{Y} , respectively, and
- $\mathbf{\Lambda}$ is the diagonal matrix of canonical correlations between $\mathbf{X}\mathbf{\Gamma}$ and $\mathbf{Y}\mathbf{\Lambda}$.

We continue that approach, but this derivation leads to a different weight-matrix.

Let the k^{th} canonical pair of variates be $\mathbf{U}_k = \mathbf{Y}\mathbf{V}_k$, and $\mathbf{V}_k = \mathbf{X}\mathbf{b}_k$ with correlation δ_k . Then the least-squares estimate of \mathbf{U}_k is $\hat{\mathbf{U}}_k = \mathbf{b}\mathbf{V}_k = \mathbf{b}\mathbf{X}\mathbf{b}_k$ with $\mathbf{b} = \delta_k$. Because the variance of \mathbf{U}_k is constrained to be unity, its mean-square error of prediction is $1 - \delta_k^2$ (Anderson (1958, p.297-298)). For the full matrices of canonical variates, $\mathbf{U} = \mathbf{Y}\mathbf{\Lambda}$, $\mathbf{V} = \mathbf{X}\mathbf{\Gamma}$, $\hat{\mathbf{U}} = \mathbf{V}\mathbf{\Lambda} = \mathbf{X}\mathbf{\Gamma}\mathbf{\Lambda}$. The corresponding covariance matrix of $\hat{\mathbf{U}}$ is $(\mathbf{I} - \mathbf{\Lambda}^2)$ where $\mathbf{\Lambda}$ is the diagonal matrix of canonical correlations and \mathbf{I} is the identity matrix.

Then a Mahalanobis distance in the space of $\mathbf{U} = \mathbf{Y}\mathbf{\Lambda}$ is:

$$d_{ij}^2 = (\mathbf{U}_i - \mathbf{U}_j) \mathbf{E}_{\mathbf{U}\mathbf{U}}^{-1} (\mathbf{U}_i - \mathbf{U}_j)'$$

$$\text{dhat}_{ij}^2 = (\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j) \mathbf{E}_{\hat{\mathbf{U}}\hat{\mathbf{U}}}^{-1} (\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j)'$$

$$= (\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j) (\mathbf{I} - \mathbf{\Lambda}^2)^{-1} (\hat{\mathbf{U}}_i - \hat{\mathbf{U}}_j)'$$

Replacing $\hat{\mathbf{U}}$ by $\mathbf{V}\mathbf{\Lambda}$:

$$\text{dhat}_{ij}^2 = (\mathbf{V}_i \mathbf{\Lambda} - \mathbf{V}_j \mathbf{\Lambda}) (\mathbf{I} - \mathbf{\Lambda}^2)^{-1} (\mathbf{V}_i \mathbf{\Lambda} - \mathbf{V}_j \mathbf{\Lambda})'$$

¹ A Mahalanobis distance differs from an Euclidean distance by the insertion of the inverse of the covariance function into the middle of the quadratic form. By transforming elliptical contours of the distribution of data points into circular (spherical or hyper-spherical) contours, the weights given to members of pairs of highly correlated variates are reduced.

$$\begin{aligned}
 &= (\mathbf{X}_i \mathbf{\Gamma} \mathbf{\Lambda} - \mathbf{X}_j \mathbf{\Gamma} \mathbf{\Lambda}) (\mathbf{I} - \mathbf{\Lambda}^2)^{-1} (\mathbf{X}_i \mathbf{\Gamma} \mathbf{\Lambda} - \mathbf{X}_j \mathbf{\Gamma} \mathbf{\Lambda})' \\
 &= (\mathbf{X}_i - \mathbf{X}_j) \mathbf{\Gamma} \mathbf{\Lambda} (\mathbf{I} - \mathbf{\Lambda}^2)^{-1} \mathbf{\Lambda} \mathbf{\Gamma}' (\mathbf{X}_i - \mathbf{X}_j)' \\
 &= (\mathbf{X}_i - \mathbf{X}_j) \mathbf{\Gamma} \mathbf{\Lambda}^2 (\mathbf{I} - \mathbf{\Lambda}^2)^{-1} \mathbf{\Gamma}' (\mathbf{X}_i - \mathbf{X}_j)'
 \end{aligned}$$

where $\mathbf{\Gamma}$ is the matrix of X-variable canonical vectors and $\mathbf{\Lambda}$ is the diagonal matrix of canonical correlations. The diagonal elements of $\mathbf{\Lambda}^2 (\mathbf{I} - \mathbf{\Lambda}^2)^{-1}$ are $8_k^2 / (1 - 8_k^2)$ and the off-diagonal elements are zeroes.

The same result can be derived from the starting point of canonical regression estimation conditional on arbitrary, fixed \mathbf{x} 's (Anderson 1958). In this formulation, the canonical procedure should share many of the properties of reduced rank regression.

This solution is qualitatively the same as that in Moeur and Stage (1995). The coefficients in $\mathbf{\Gamma}$ are unchanged. However, $\mathbf{\Lambda}^2 (\mathbf{I} - \mathbf{\Lambda}^2)^{-1}$ replaces their $\mathbf{\Lambda}^2$. If only the first canonical variate is used, then distances will only change proportionately, leading to the same choice of most similar neighbors. However, if more than one canonical variate is included in the weight matrix, then relatively more weight will be given to the more highly correlated canonical variates. The significance of this alternative depends on the covariance structure of particular data sets. Table 2 displays relative changes in weights of the canonical variates for the three data-sets.

Table 2. Relative weights of first four canonical variates under new formulation compared to original formulation. Canonical correlation (λ^2)

| Canonical pair | Utah | | Tally Lake | | User's Manual | |
|----------------|-------------|-------------------------|-------------|-------------------------|---------------|-------------------------|
| | λ^2 | Relative Weight New/old | λ^2 | Relative Weight New/old | λ^2 | Relative Weight New/old |
| 1 | 0.465 | 1.00 | 0.626 | 1.00 | 0.691 | 1.00 |
| 2 | 0.159 | 0.64 | 0.348 | 0.57 | 0.454 | 0.57 |
| 3 | 0.125 | 0.61 | 0.327 | 0.56 | 0.247 | 0.41 |
| 4 | 0.042 | 0.56 | 0.227 | 0.49 | 0.219 | 0.40 |
| Total | 0.863 | | 1.861 | | 1.823 | |

VARIABILITY OF IMPUTED VALUES

MSN has assumed that the observed data comprising the reference observations are true representations of the state of nature. In contrast, regression analysis assumes that the “true” value (i.e the expectation) is unobservable. In the regression context, it is the observed value of the dependent variable that includes the error. Although not usually separable, this residual error consists of two components: observation error, and model error. The latter component includes all the causes of variation that would be explained by a perfect model. Regression estimates remove both components from the predictions.

The same two error components may be present in the observation data used in MSN. However, the imputation process retains both sources. The most similar neighbor can be considered as an observation on the same “super-stand” as the reference observation if their measure of dissimilarity (distance) is zero. If e_i is the measurement error of the i^{th} observation of Y_i and Y_j is the most similar neighbor, then we have been calculating a residual as $Y_i - Y_j$. But: $Y_i - Y_j = (\Theta_i + e_i) - (\Theta_j + e_j) + f(X_i - X_j)$ where Θ is the observation-error-free, unobservable value of Y . The last term measures the contribution to error attributable to the difference in the vectors of X-variables between the reference observation and its most similar neighbor. Heretofore, the contribution of this last term has been considered the major component of MSN error because of our assumption of Y 's observed without error. However, note that as a minimum the expected value of the squared difference includes **twice** the expected value of the observation error variance. Thus, the real imputation variance is half the reported mean square “residual”. This conclusion is analogous to the semi-variance of time-series and the usual formula for estimating variance from mean-square successive differences.

EFFECT OF REFORMULATION ON RESIDUALS

The new formulation of the weight matrix reduced the residuals for some, but not all variables and data-sets (Figures 1 a, b, c). When data are sparsely distributed in the space in which distance is calculated, small differences in the distance function will not change the identity of the nearest neighbor. When that density is higher, as indicated by the average number of observations per parameter to be estimated (last line of table 1) for the Utah data, the reformulation appears to improve the selection of neighbors. Unfortunately, the data-sets also differ in the relative change in weights because of their differences in variance structure (Table 2). The structure of the Utah data also produced the least change in weighting due to the new formulation.

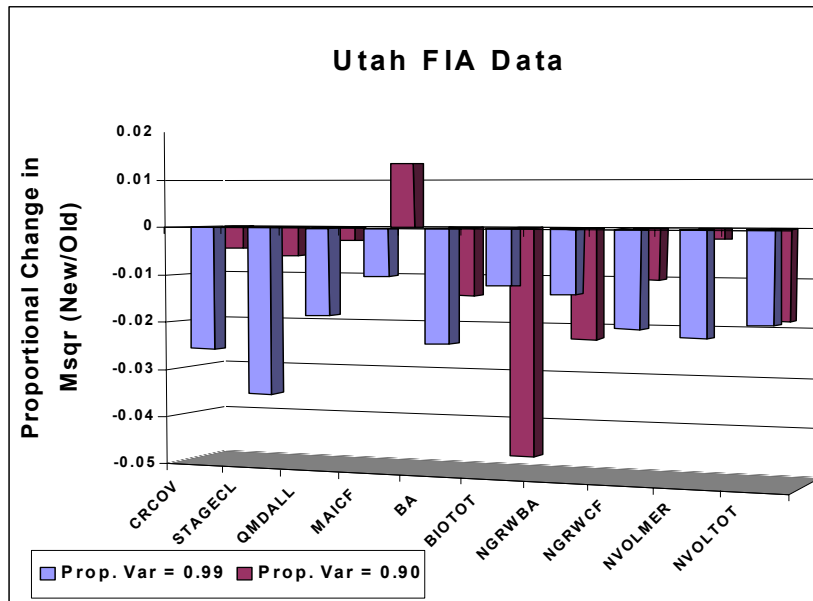


Figure 1a. Comparisons of changes in root-mean-square errors for Utah FIA data. Ratios of new to old RMSE's based on canonical variates for two levels of cumulative correlation. Negative bars indicate variables for which the new formulation is an improvement.

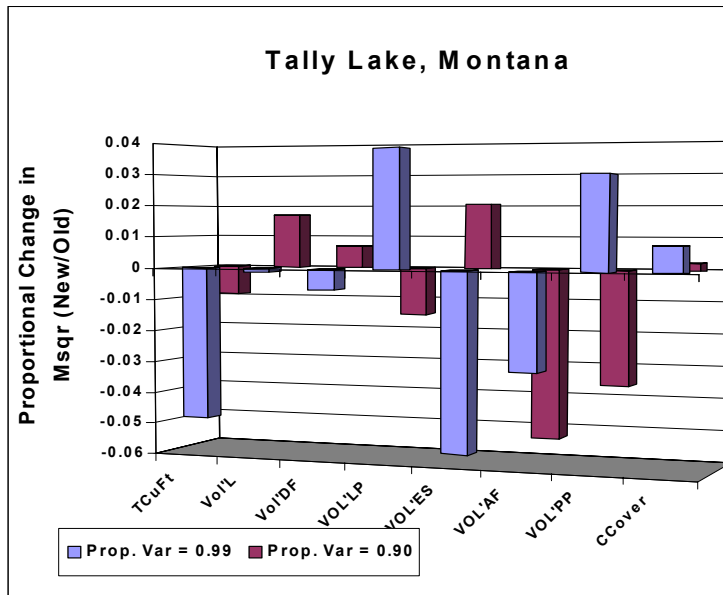


Figure 1b. Comparisons of changes in root-mean-square errors for Tally Lake data. Ratios of new to old RMSE's based on canonical variates for two levels of cumulative correlation. Negative bars indicate variables for which the new formulation is an improvement.

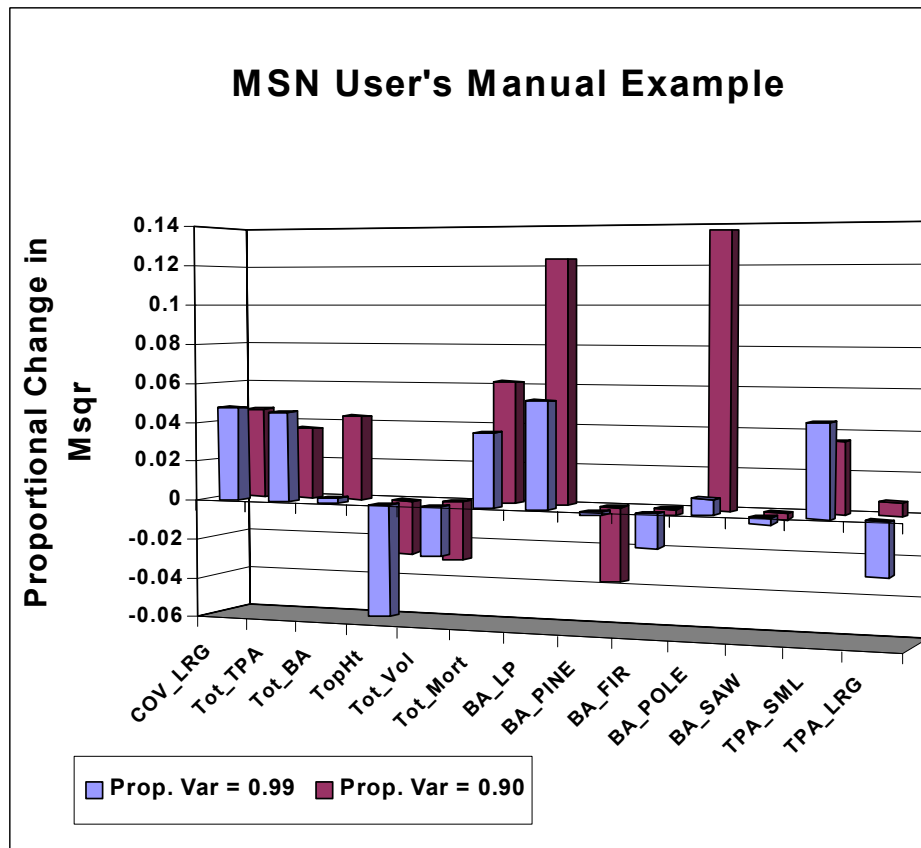


Figure 1c. Comparisons of changes in root-mean-square errors for User's Manual data. Ratios of new to old RMSE's based on canonical variates for two levels of cumulative correlation. Negative bars indicate variables for which the new formulation is an improvement.

Comparison with Mahalanobis distance on X variables

Figure 2 compares residuals using the new formulation of the canonical-based distance function with residuals using a Mahalanobis distance computed directly from the X variables. In this comparison, selection of neighbors in the Deschutes data from the User's Manual was not improved by using the relationships of the Y's to the X's even though that data set had the highest canonical correlations. In contrast, Tally Lake showed that MSN improved the selection of neighbors although its correlations between Y's and X's were weaker than in the User's Manual data.

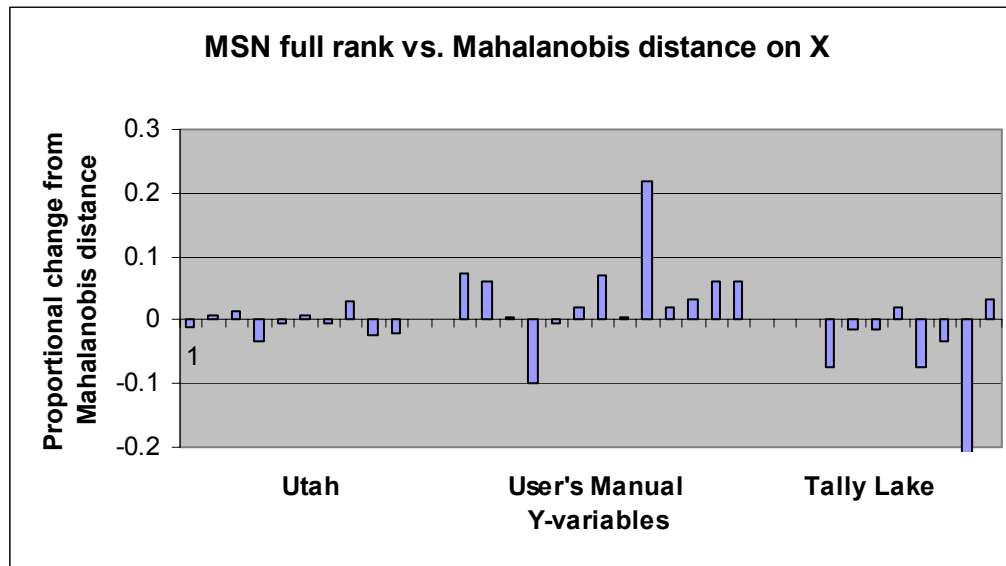


Figure 2. Comparisons of MSN selection of neighbors with a distance function ignoring correlations between Y variables and X-variables. Negative bars indicate variables for which MSN is better than the Mahalanobis distance on only the X-variables. Sequence of variables within a data set is the same as in previous figures.

CONCLUSIONS

The analyses reported here demonstrate that when selecting surrogates for imputation, no one measure of similarity is superior in all cases. Therefore, new software for selecting surrogates has provided a suite of distance measures as options (Crookston, *et al.* 2002). Exploration of alternatives is made easier by providing consistent measures of the quality of imputation. In addition, the same software can provide identification of k nearer neighbors for each distance function.

REFERENCES

- Anderson, T.W. 1958 *An introduction to multivariate analysis*. New York. John Wiley & Sons, Inc. 374 + xii p.
- Crookston, N.L., Moer, 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.
- Moer, M. 2000. Extending stand exam data with most similar neighbor inference. P. 99-107. In: Proceedings of the Society of American Foresters National Convention:1999 September 11-15, Portland, OR. SAF Publication 00-1; Bethesda, MD. Society of American Foresters:
- Moer, M. and Stage, A.R. 1995. Most Similar Neighbor: An improved sampling inference procedure for natural resource planning. *Forest Science* 41:337359.