Short Note: Merging Secondary Variables for Geophysical Data Integration

Steven Lyster and Clayton V. Deutsch

Department of Civil & Environmental Engineering
University of Alberta

Abstract

Multiple secondary data from geophysical measurements or expert interpretations are often available. Geostatistical methods are capable of dealing with an arbitrarily large number of secondary data; however, building a licit probabilistic model of "redundancy and closeness" to the variables of interest is cumbersome. It is desirable to merge multiple secondary data into a "super secondary" variable with a calculated correlation to the variable of interest. This note describes one procedure to do this under a multivariate Gaussian model.

Merging Secondary Data

The integration of data in geostatistical modeling and simulation is necessary to obtain reasonable models of heterogeneity and estimates of uncertainty. The information represented by secondary variables can help to better express trends in the primary variable. Incorporating secondary data can prove difficult when several variables exist. It would be advantageous to merge these into a single "super" variable, allowing conventional cokriging and cosimulation to be used.

Ideally for this method the secondary variables should be uncorrelated to one another and highly correlated to the primary variable. This situation would minimize the redundancy between the data. For example, consider two secondary variables that each have a correlation ρ=0.6 to the primary variable. The best possible correlation of a super variable to the primary variable would be $\sqrt{(0.6^2+0.6^2)}$=0.686 if the secondary data were completely independent. If the two secondary variables were fully dependent then the correlation coefficient would remain at 0.6 due to the redundancy between the data they contain.

For situations where the secondary data is not fully dependent or independent, the super variable can be found using a linear estimator:

$$y^* = \sum_{i=1}^{n} \lambda_i \cdot y_i \tag{1}$$

where y^* is the value of the super variable and λ_i is the weighting for each of the n variables, y_i. It is useful to work with standardized variables so that the units do not cause confusion. Standardization in this case is achieved through the equation:

$$y_i = \frac{z_i - m_i}{\sigma_i} \tag{2}$$

where z_i is the value of variable i, m_i is the mean, and σ_i is the standard deviation. This results in a variable y_i with a mean of zero and a standard deviation of one. The original shape of the statistical distribution is maintained by equation (2).

Once the variables have been standardized, the weights can be found from an *n*-by-*n* system of normal equations:

$$\sum_{j=1}^{n} \lambda_j \cdot \rho_{i,j} = \rho_{i,0} \qquad i = 1, \ldots, n \tag{3}$$

These weights, once determined, can be used to find the estimate of the super variable from equation (1) and the estimation variance of the super variable:

$$\sigma_L^2 = 1 - \sum_{i=1}^{n} \lambda_i \cdot \rho_{i,0} \tag{4}$$

Correlation with the primary variable will always be higher for the super variable than for any of the individual secondary variables. The correlation coefficient between the super variable and the primary variable to be estimated can be found from the equation:

$$\rho_{y^*,0} = \sqrt{\sum_{i=1}^{n} \lambda_i \cdot \rho_{i,0}} \tag{5}$$

Examples

Suppose a data set has a primary variable and two secondary data sets. The secondary variables have a correlation coefficient of 0.610. Their correlations to the primary variable are 0.852 and 0.778. The system of equations to be solved is

$$\lambda_1 + 0.610\,\lambda_2 = 0.852$$
$$0.610\,\lambda_1 + \lambda_2 = 0.778$$

Solving for the estimator weights yields λ_1=0.601 and λ_2=0.411. The correlation of the resulting super variable and the primary variable is ρ=0.912. This is significantly higher than either of the individual secondary variables in addition to simplifying the estimation of the primary variable.

As another example, suppose there are now three secondary variables to be merged. The correlation coefficients are given in Table 1. Note that variable 2 has a negative correlation to each of the other three.

	1	**2**	**3**	**Primary**
1	1	-0.5	0.6	0.65
2	-0.5	1	-0.65	-0.7
3	0.6	-0.65	1	0.7
Primary	0.65	-0.7	0.7	1

Table 1: Coefficients of correlation between all of the different variables for the second example.

The weights for each variable are λ_1=0.2962, λ_2=-0.3678, and λ_3=0.2832. λ_2 is negative, which is to be expected because of the negative correlation to the other variables. The correlation between the merged super variable and the primary variable is ρ=0.805. The linear estimation variance is 0.352. Correlation to the primary variable is once again significantly higher for the merged super variable than any of the secondary variables alone.

Real data was used to determine the weights for merging three secondary variables. The publicly-available data set has four variables, using different units; for this example one was arbitrarily chosen as the primary variable. There are 67 data points contained in the file. To determine the correlation between the different data types, declustering weights were first found using the program `declus`. Next, all of the variables were plotted against one another using `scatplt`. The results from this are shown in Figure 1. Table 2 shows a summary of the coefficients of correlation.

	1	**2**	**3**	**Primary**
1	1	.526	.451	.463
2	.526	1	.387	.680
3	.451	.387	1	.422
Primary	.463	.680	.422	1

Table 2: Coefficients of correlation between all of the different variables for the third example.

Solving this problem gives weights of λ_1=0.0907, λ_2=0.5702, and λ_3=0.1604. The coefficient of correlation between primary variable data and the merged super secondary variable is ρ=0.705. This is slightly better than the correlation between the primary variable and secondary variable 2. A scatterplot of the primary variable vs. the super variable is shown in Figure 2. Note that the primary and super variables have been back-transformed, which does not affect the correlation.

Conclusions

The use of a single merged secondary variable greatly simplifies estimation accounting for secondary data. This could prove useful in a variety of cases, such as when many different metal types are present in ore or when numerous seismic data are available. This secondary data must be accounted for somehow to properly estimate the resource in question. The relative simplicity of simulating with only one secondary variable instead of many makes this technique attractive.

References

1. C.V. Deutsch and A.G. Journel. *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press, New York, 2nd Edition, 1998.
2. O. Leuangthong. Multivariate Geostatistical Simulation at Red Dog Mine, Alaska, USA. In *Centre for Computational Geostatistics*, Report 5, Edmonton, AB, 2003.

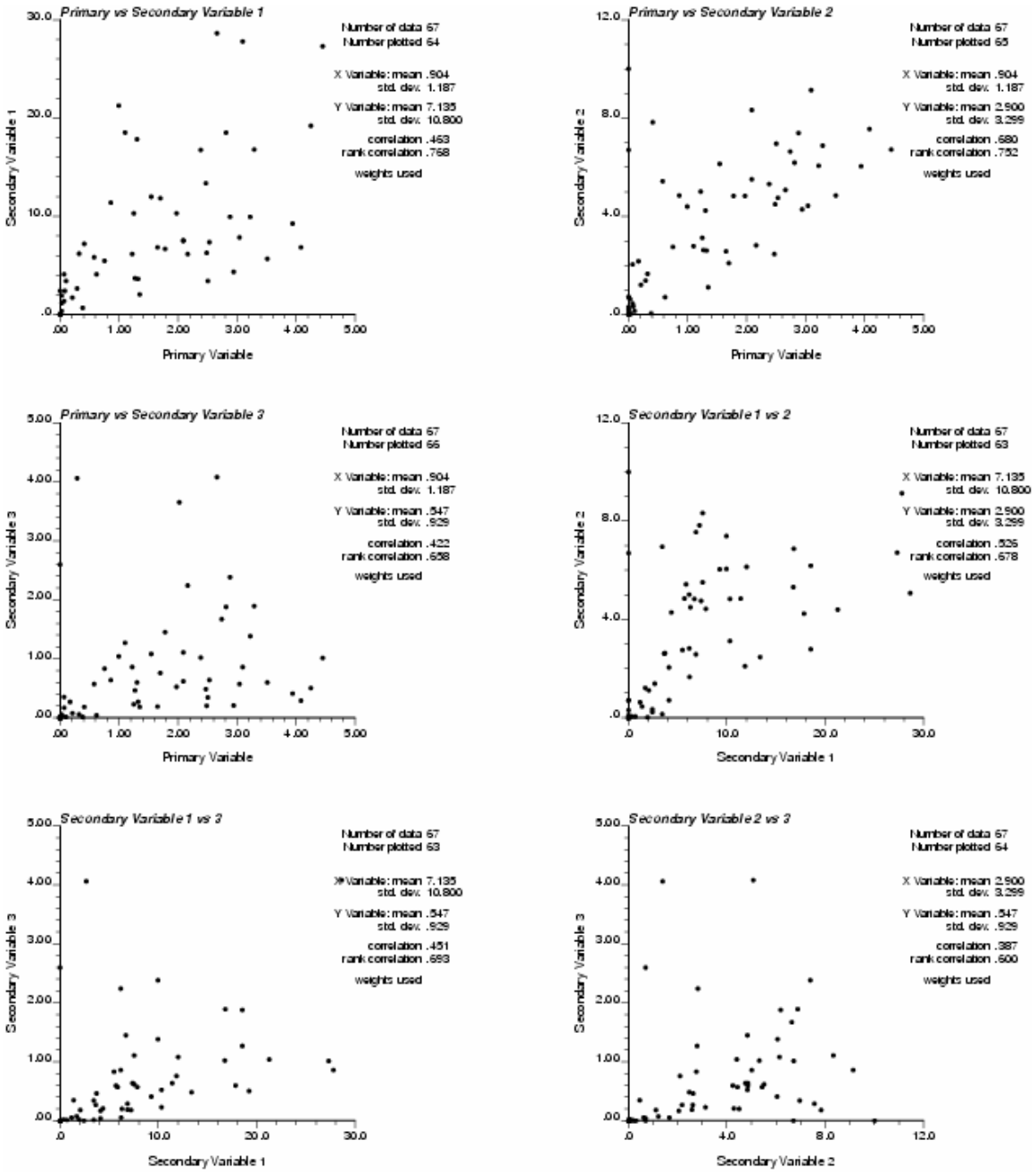

Figure 1: Scatterplots used to find the correlation between the different variables in the third example.

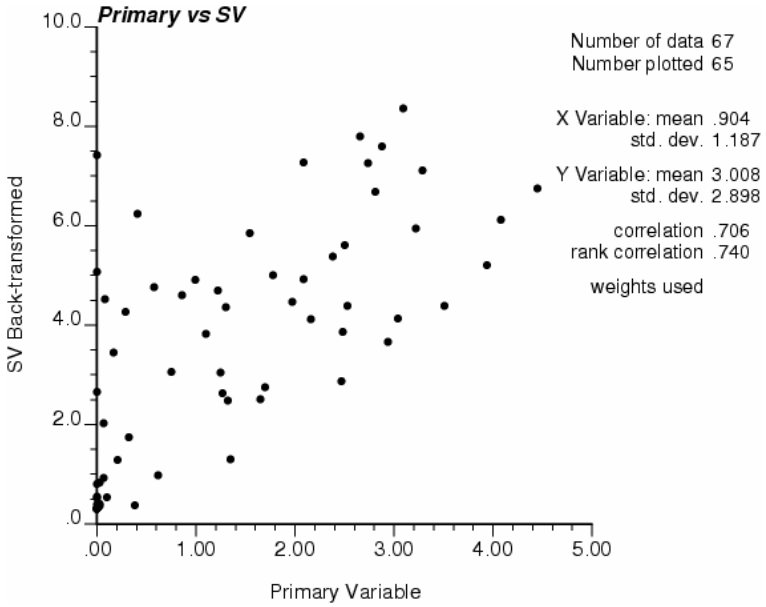

Figure 2: A scatterplot of the primary variable vs. the merged super secondary variable.