Advances in Multivariate Analysis of Spatial Data

J.B. Boisvert and C.V. Deutsch

The analysis of data sets with large numbers of variables is becoming more common. With the expense of drilling and the relatively inexpensive analysis of sampling, often many measurements are made on each sample collected. Moreover, a large number of variables are available from diverse sources such as geophysical measurements, remote sensing, production data, geological interpretations, etc. Many of the variables available are of little interest by themselves; they are only used to better predict a small number of important primary variables such as grade, porosity, saturations, etc. It is reasonable to believe that the inclusion of these secondary variables should improve modeling of the primary variables. Unfortunately, few modeling methodologies can handle a large number of secondary variables. The proposed methodology can be used to (1) identify and merge redundant secondary variables (2) identify and remove unimportant secondary variables and (3) merge the remaining secondary variables into an arbitrary number of clusters for further multivariate spatial modeling methodologies such as kernel methods, Bayesian updating or stepwise.

1. Introduction

Many projects begin with an initial data analysis stage, typically deemed Exploratory Data Analysis (EDA). One aspect of the EDA is the assessment of which variables will be used in the analysis. This article discusses methodologies to select relevant secondary variables to carry forward in further analysis (such as the generation of geostatistical models of the primary variables of interest). The nature of earth sciences problems is that there are often a fixed number of primary variables, such as mineral grade, petro-physical variables, contaminate concentrations, etc. The variables are often fixed by the project objectives. The difficulty arises when there are a large number of secondary variables that are more readily available due to cost effectiveness or physical availability. We would like to use the secondary variables to make better predictions of the primary variables. This is a typical framework arising in many industries, such as in mining when geometallurgical samples typically contain numerous samples of many elements/compounds (204 variable example in Boisvert et al., 2009) or in the petroleum case when the cost to drill a well is high, as such, for each well there are a large number of measurements taken which may help predict variables of interest such as permeability. In addition, there may be many exhaustively sampled secondary variables arising from a number of sources such as geophysical measurements, remote sensing, production data, geological interpretations, etc.

The combination of all of these secondary data sources often results in an large number of variables to consider as secondary. A naïve approach is to deem all secondary variables relevant and attempt to use them all in the study; however, few techniques exist that can consider such a large number of variables. There are more sophisticated techniques that can consider the high dimensional multivariate relationship between secondary and primary data (Table 1). Such techniques often require large amounts of data and/or extensive CPU requirements; as such, it is important to assess the importance of a secondary variable and its potential impact on the modeling of the primary variable. The proposed methodology to assess the importance of each secondary variable has three possible outcomes:

a) The secondary variable is retained as is for use in modeling the primary.
b) The secondary variable is deemed unrelated to the primary and is removed from the analysis.
c) The secondary variable is deemed similar to another secondary variable and these variables are merged into a single variable.

An assumption is made in the proposed methodology that there is only one primary variable. The methodology for multiple primary variables is to repeat the analysis independently for each primary.

Table 1: Techniques that can consider the multivariate relationship between secondary variables.

Technique	Reference	Drawbacks
Kernel Methods	Hong (2010)	CPU intensive for >5 variables
Stepwise	Leuangthong (2010)	Requires many sample data as the number of variables increases
Bayesian Updating	Deutsch and Zanon (2007)	Equivalent to cokriging, only considers linear multivariate relationship between data

2. Methodology

There are three steps to the proposed methodology: 1) initial variable assessment 2) removal of unimportant variables; 3) linear combination of important variables. In each step of the methodology the number of variables is reduced with the goal being to reduce the number of variables to an arbitrary user supplied number that depends on the modeling methodology to be applied (Table 1) while maximizing the predictive capabilities of the secondary variables.

1) **Initial variable assessment**: secondary variables that are highly correlated and known to measure identical features are merged.
2) **Removal of unimportant variables:** Secondary variables that are deemed unimportant to the analysis are removed. Importance is measured by cross validation results and a linear proxy model.
3) **Linear combination of important variables:** If there are still too many variables to consider after the initial merging (Step 1) and the removal of unimportant variables (Step 2), the secondary variables are clustered and merged into a user supplied number of clusters based on k-means clustering.

It should be noted, that this methodology is similar to the methodology suggested in Boisvert et al. (2009), the novel contribution here is the set of tools available to help perform each step and the use of k-means clustering to determine the final groups. These tools are discussed and expanded upon in each section.

Step 1: Initial Variable Assessment

The purpose of this step is to identify which secondary variables are redundant and to reduce the total number of variables by considering only one merged variable to be representative of the group. Identification of redundant secondary variables can be done with a correlation matrix (Figure 1). Consider the two groups of redundant variables [2,3,4,5,6,8,9,10] and [11,12]. Each of these variables is known to be related (measures of topography) and are therefore highly redundant. Typically a single variable would be selected, perhaps the one with the highest correlation to the primary (secondary variable 2 and 11) and the remainder discarded. This has the undesirable effect of discarding additional information that may help predict the primary.

Rather than select a single secondary to represent a group of redundant variables, the group of variables is linearly merged to extract the maximum amount of information inherent in each variable; thus, the number of variables is reduced with minimal loss of information. Note that the merging of variables is a linear combination, thus if the relationship between the secondary variables is non-linear, some information is lost. This merging is discussed in Boisvert et al. (2009). To summarize, linear regression is performed on the variables of interested to

determine the weights (λ) to assign to each variable (*v*) in Equation 1. The resulting merged variable has a larger correlation to the primary variable than any single variable. The merged variable (*M*) is carried forward and replaces the group of redundant secondary:

$M(v) = \sum_{i=1}^{n} \lambda_i v_i$ where the weights are determined from: $\begin{bmatrix} \rho_{1,1} & \rho_{2,1} & \cdots & \rho_{n,1} \\ \rho_{1,2} & \rho_{2,2} & \cdots & \rho_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,n} & \rho_{2,n} & \cdots & \rho_{n,n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \rho_{0,1} \\ \rho_{0,2} \\ \vdots \\ \rho_{0,n} \end{bmatrix}$ (1)

There are a number of useful modifications to the standard correlation matrix. First, the uncertainly in the correlation between each pair of variables is indicated by a bar above the correlation value in the upper triangle (Figure 1). This bar indicates the p_{10}, p_{50} and p_{90} correlation as calculated from Equations 2 and 3 (Johnson, Kotz and Balakrishnan, 1995). We prefer variables that have a high correlation that also have a low uncertainty. The uncertainty in the correlation depends only on the correlation coefficient (r) and the number of samples (n).

$$p_R(r) = \frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\sqrt{\pi}\Gamma(\frac{1}{2}(n-1))\Gamma(\frac{1}{2}n-1)} \times \sum_{j=0}^{\infty} \frac{[\Gamma(\frac{1}{2}(n-1+j))]^2}{j!}(2\rho r)^j \quad (2)$$

where:

$$N_{Independent} = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_{ij}} \quad (3)$$

The second modification to the correlation matrix is the addition of the bivariate scatterplots between pairs of variables in the lower triangle. Typically, the correlations are repeated in this area which is redundant. In the remainder of this paper, variables will be linearly combined. These scatter plots can be used to indentify variables that have a nonlinear relationship.

In addition to manually assessing the nonlinearity of the bivariate relationships through the scatterplots, the Gaussianity measure proposed by Deutsch and Deutsch (2009) is also used to shade the scatterplot if the relationship violates their measure of non-Gaussianity. This is used as an indication the two variables may be non-Gaussian and further visual examination of the scatterplots should be conducted.

Note that if variables are considered to have a highly non-Gaussian relationship (such as Secondary Variable 13 and Merged variable 1 in Figure 2) these variables should not be merged as the linear merging suggested in this paper will not capture this relationship. These variables should be carried forward unmerged if the relationship is deemed important to the study (i.e. if the non-Gaussian relationship is between a secondary and primary variable).

To summarize, the following issues should be identified at this stage:

1) Non-linear and non-Gaussian relationships that will not be captured by the correlation should be determined by examining each scatter plot.
2) Variables that have a low correlation with the primary variable should be identified for removal from the modeling process.
3) Highly correlated secondary variables should be identified for possible merging.
4) Sparsely sampled secondary data should be removed.

Step 2: Removal of Unimportant Variables

If step 1 reduced the total number of variable to a sufficiently low number that modeling can be accomplished with the desired multivariate modeling technique (Table 1) then the process ends. Otherwise, unimportant variables are identified for removal from the modeling process to further reduce the number of secondary variables.

The idea is to use a proxy model for multivariate modeling, predict the primary variable using the secondary and assess how important each secondary is when predicting the primary. A linear model (i.e. cokriging or linear regression) is used to incorporate all the secondary variables. The main input to this technique is the correlation matrix which is used to calculate the weights (*b*) to assign to each secondary variable:

$$\sum_{j=1}^{n_{sec}} b_j \bullet \rho_{s_j,i} = \rho_{i,0} \qquad i = 1,\ldots,n \tag{4}$$

The weight given to each secondary variable (b_j) is used to assess how important each secondary is. Variables with a high weight are more influential on the primary variable. This importance measure is shown as a tornado chart in (Figure 3). To account for the uncertainty in the correlation matrix, a resampling (bootstrap) technique is applied to generate multiple (*nreal*) correlation matrices. This provides a distribution of the importance of each variable. The range (p_{10}-p_{90}) of the importance is shown as a box for each secondary with the p_{50} indicated by a vertical line. The chart in Figure 3 is ordered by the correlation to the primary. This allows for a quick assessment of inconsistencies (i.e. variables that have a high importance but low correlation to the primary should be further examined or merged with another variable).

It should be noted that in order to apply this proxy model based on the correlation matrix, the correlation matrix must be positive definite. Because of the heterotopic sampling of typical data available, positive definiteness is not ensured. To obtain a correlation matrix that is positive definite, the correction presented in Kumar and Deutsch (2009) is employed. Moreover, a typical correction that is used to improve the conditioning of matrices is to apply a Tikhonov Regularization. Essentially the diagonals of the correlation matrix are increased to improve stability. This regularization parameter is an input option to the program.

In addition to using the weight to determine the importance of a secondary variable, a loss/gain measure is provided for each secondary variable. This measure uses cross validation to determine if each variable is *contributing to* or *detracting from* the prediction of the primary variable. The loss/gain is calculated as the difference in cross validation correlation with and without the secondary. If including a variable *decreases* the cross validation this is considered a loss (i.e. prediction of the primary would be improved by removing the secondary):

$$loss\,or\,gain = \rho_n^{Cross\,Validation} - \rho_{n-1}^{Cross\,Validation} \tag{4}$$

If the inclusion of the secondary variable increases correlation, the variable contributes to the prediction of the primary (gain). Due to the limited data available, there will be some uncertainty in the loss/gain for each variable. The bootstrap procedure is again used to provide multiple correlation matrices for each realization which are used to generate a distribution of loss/gain. The bar is colored red if the p_{10}-p_{90} range of the loss/gain measure is negative (surely the variable detracts from predicting the primary), the bar is colored green if the range is positive (surely the variable contributes positively to predicting the primary) and if there is uncertainty to the variable contributing the bar is colored yellow (Figure 3). Variables that are red should be removed from the analysis. Variables that are colored yellow should be further examined. Variables that contribute little to the estimation should also be removed.

Step 3: Linear Combination of Important Variables

Through steps 1 and 2 redundant variables have been merged and unimportant variables have been removed. All remaining variables have been deemed relevant to the study; if the number of variables remaining requires further

reduction, the variables can be clustered into groups of similar variables. A standard statistical plot showing how variables are related is the dendrogram (Figure 4) and groups variables based only on the correlation coefficient.

The dendrogram represents a standard hierarchical clustering of variables where variable pairs with the largest correlation coefficient are iteratively grouped into clusters. Once two variables have been clustered, it is necessary to calculate the correlation between the cluster and the remaining variables. This can be done by (1) single linkage: using the maximum correlation between variables in the cluster and the remaining variables (2) complete linkage: using the minimum correlation between variables in the cluster and the remaining variables or (3) average linkage: using the average correlation between variables in the cluster and the remaining variables. Each linkage option results in a different dendrogram (Figure 4).

The dendrogram provides a useful analysis of the relationship between variables. The user can select the number of clusters desired and group variables based on their correlations. However, there are often more considerations involved with variable amalgamation than is captured by the correlation coefficient alone. As such, k-means clustering is implemented to merge variables while accounting for a number of important features such as: (1) correlation (2) a heterotopic sampling penalty and (3) the spatial variability of each variable. The dissimilarity between two variables will be described by a pseudo-distance metric where the distance between variables *i* and *j* is:

$$D(v_i, v_j) = c \cdot C(v_i, v_j) + p \cdot P(v_i, v_j) + s \cdot S(v_i, v_j) \tag{5}$$

where *c*, *p* and *s* are weights given to each dissimilarity measure based on the perceived importance of each measure. This dissimilarity (or distance) measure is formulated to be flexible. It would be possible to incorporate additional terms if other attributes of the variables are relevant to the analysis. These three measures will be assumed to capture the (dis)similarity between variables, the linear nature of Equation 5 makes the incorporation of additional measures of dissimilarity between variables easy.

Correlation $C(v_i,v_j)$

The correlation coefficient is a similarity measure; however a dissimilarity measure is required:

$$C(v_i, v_j) = 1 - abs[\rho(v_i, v_j)] \tag{6}$$

The range of possible values for $C(v_i,v_j)$ are [0,1]. The two remaining dissimilarity measure have a different range and will be standardized to [0,1] so that the weights, *c*, *p* and *s* are consistent.

Heterotopic sampling penalty $P(v_i,v_j)$

This penalty is due to missing samples in variables that are to be merged. Merging variables linearly requires all variables to be present in the sample, if one variable is not present, all samples used in the merging at this location must be discarded. This penalty encourages groupings of variables that are homotopically sampled:

$$P(v_i, v_j) = \sum_{k=1}^{n} \begin{cases} 1, & \text{if variable i exists in sample k and variable j does not} \\ 1, & \text{if variable j exists in sample k and variable i does not} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

This is standardized by the largest penalty so that the range is [0,1] to be consistent with the other dissimilarity measures

$$P(v_i, v_j) = \frac{P(v_i, v_j)}{\max P} \tag{8}$$

Spatial variability $S(v_i,v_j)$

It would be inappropriate to merge variables that have very different spatial continuity. Ideally the variograms of the variables would be compared; however, with sparse sampling the variograms are rarely meaningful and are likely to be noisy. A summary measure of spatial variability is the ratio of the point support variance to the block support variance. The user inputs *m* different block sizes to upscale the variance and the dissimilarity between two variables is:

$$S(v_i,v_j)=\sum_{k=1}^{m}\left(\frac{block\,\sigma_k^i}{block\,\sigma^i}-\frac{block\,\sigma_k^j}{block\,\sigma^j}\right) \tag{9}$$

This is standardized by the largest penalty so that the range is [0,1] to be consistent with the other dissimilarity measures

$$S(v_i,v_j)=\frac{S(v_i,v_j)}{\max S} \tag{10}$$

Once these measures have been calculated for the available variables, the distance between each variable is used to perform k-means clustering. Initially, all variables are assigned to a random cluster. K-means then iteratively assigns variables to the centroid of the nearest cluster until all variables are nearest to the centroid of their respective cluster. The initial configuration of variables into random clusters affects the final clustering. The clustering is performed a large number of times and the clustering configuration that scores the lowest SE is retained. SE is the sum of the distance between each variable and the centroid of its cluster:

$$SE=\sum_{i=1}^{\text{n vars}}D(v_i,centroid_i) \tag{11}$$

Table 2: Analysis of clustering into three clusters.

#	Name	n samples rejected	Spatial Var score [average]	Average Correlation to all Clusters: 1	2	3
3	Secondary 13	0	2.74 [2.69]	0.17	0.18	0.47
3	Secondary 15	0	2.65	0.21	0.07	0.47
2	merged 1	0	3.34 [3.30]	0.16	0.24	0.10
2	Secondary 7	0	3.26	0.16	0.24	0.15
1	merged 2	13	3.11 [2.89]	0.43	0.15	0.18
1	Secondary 14	13	2.68	0.30	0.24	0.23
1	Secondary 1	13(0)	2.89	0.21	0.10	0.15

Table 2 is a summary of the clustering of the 7 variables from Figure 2 into three clusters. These three variables would then be used as secondary variables in further modeling (i.e. Table 1). Average correlation between a variable and the other variables in its cluster are indicated in gray. Variables that have a higher correlation to variables in a cluster other than its own are indicated in red.

When analyzing the number of samples rejected it is reasonable to question which variables are missing samples. If a number appears in () adjacent to the number rejected, this indicates the number of samples that would be rejected if this variable was removed from the cluster. Consider Secondary 1, if this variable is removed from cluster 1, 0 samples would be rejected. Due to the nature of linear merging (Equation 1), if a variable is missing (such as Secondary 1) these 13 samples must be discarded when merging cluster 1.

3. Conclusions

The proposed methodology can be used to (1) identify and merge redundant secondary variables (2) identify and remove unimportant or uncertain secondary variables and (3) merge variables in a flexible fashion with k-means clustering such that the desired final number of variables is not exceeded. Four GSLIB programs are presented to generate the plots and analysis as presented in this paper and are discussed below.

4. Description of Programs

There are four programs. The input parameters will be discussed for each program.

Program Name	Description	Output(s)
`Var_Sel`	Used to generate the correlation matrix between all variables, also generates the tornado chart	Correlation matrix, tornado chart, data file with merged variables
`Dendrogram`	Generate a dendrogram plot	Dendrogram plot
`Merge_Vars`	Perform K-means clustering to determine variable groups	Chart with clusters, data files containing (1) correlation, penalty and spatial variation scores and (2) information on each cluster
`Linear_merge`	Linearly merge variables based on correlation to primary	Data file with merged variables

Var_Sel program

```
START OF PARAMETERS:
Line 1: data.dat           -file with data
Line 2: data_merged.out    -file with data to output after merging/elimination
Line 3: 5                  -number of secondary variables
Line 4: 1 2 3 4 5          -columns
Line 5: 0 1 2 2 1 1        -ncategories (0=skip, also add titles on next lines), category number for each variable
Line 6: 6                  -col for the primary
Line 7: 1.0                -regularization parameter (tikhonov regularization)
Line 8: 63154              -rand number seed
Line 9: 30                 -number of realizations to use for tornado chart uncert
Line 10: 1                 -plot cormat?? 1=yes
Line 11: cor.ps            -file for Postscript output
Line 12: tor.ps            -file for Postscript output
Line 13: 0.5               -bullet size in the scatter plots on the correlation matrix: 0.1sml-1reg-10big
Line 14: -998 1e21         -trimming limits
Line 15: 1                 -include correlation uncert in the correlation matrix plot  1=yes 0=no
Line 16: 1                 -sort the correlation mat based on the primary variable correlation to match up with the tornado chart (1=yes 0=no)
Line 17: 1                 -number of mergings to do, if you merge variables, they will be assigned the category of the first merged variable
Line 18: 3 1 3 4           -number of variables, col numbers, repeat for more mergins
Line 19: 0                 - number of variables to remove
Line 20: 98 99             -variables to remove, list all variable number on this row
Line 21: Topography        -variable names
Line 22: Z cmfs20
Line 23: Z BK
Line 24: Z cmfs10
Line 25: Z Rollins
Line 26: cor matrix         -title
```

Line 1 – input data file, data have been standardized (mean=0 variance=1) or normal score transformed

Line 2 – output file containing the variables merged on line 18

Line 3 – number of secondary variables to consider

Line 4 – columns in the data file

Line 5 – category for each variable. If the first category is 0, the categories are ignored

ADD LINES – add one line for each category title, the category title is written out next to the variables in the plots

Line 6 – column in the data file corresponding to the primary variable of interest

Line 7 – increases the diagonal of the covariance matrix, 1.0=no change, 1.5 would be a large increase (50%)

Line 8 – random number seed for the bootstrapping of the correlations

Line 9 – number of realizations to determine the distributions of uncertainty in the tornado chart

Line 10 – output a correlation matrix plot (1=yes 0=no)

Line 11 – name for the correlation matrix plot file

Line 12 – name for the tornado chart file

Line 13 – bullet size on the scatter plots in the correlation matrix

Line 14 – trimming limits for the variables (all variables use the same limits)

Line 15 – include the uncertainty bar on each element of the correlation matrix. This may require significant computation time

Line 16 – sort the output based on the correlation to the primary

Line 17 – number of amalgamations to consider

Line 18 – add a new line for each merging. First entry indicates how many variables to merge, followed by variable number

Line 19 – number of variables to remove from the analysis

Line 20 – list all variables to remove on this line (if number to remove=0, keep line but it will be skipped)

Lines 21-25 – one variable name per line. Number of lines equals Line 3

Line 26 – title for the plot

Dendrogram

```
START OF PARAMETERS:
Line 1: data_merged.dat          -file with input data
Line 2: -998 1e21                -trimming limits
Line 3: data_merged.ps           -file with dendrogram
Line 4: 5                        -number of secondary variables
Line 5: 1 2 3 4 5                -columns
Line 6: 1                        -type of linkage (1=min, 2=max)
Line 7: 1.4 1.5 1.5              -scale in x,y,font (increase/decrease plot size, plot may expand larger than the page)
Line 8: Topography               -variable names
Line 9: Z cmfs20
Line 10:Z BK
Line 11:Z cmfs10
Line 12:Z Rollins
```

Line 1 – input data file

Line 2 – trimming limits

Line 3 – output file with plot

Line 4 – number of secondary variables to consider

Line 5 – columns in the data file

Line 6 – type of linkage to consider when merging, 1=use the minimum correlation between cluster and variables, 2=maximum

Line 7 – increase plot size in x or y (1.0 = standard), font size can also be uniformly increased.

Lines 8-12 – one variable name per line. Number of lines equals Line 4

Merge_vars

```
START OF PARAMETERS:
Line 1: data_merged.dat          -file with data
Line 2: 1 2                      -col for xy (for spatial variability)
Line 3: 3 100  500 2000          -nblock sizes to use with the spatial variability measure, block sizes
Line 4: 2                        -nclusters
Line 5: 65413 2000               -random number seed,nreal: kmeans requires initial config of vars in clusters, try nreal initial configs
Line 6: 1.0 0.5 1.0              -relative weight to give to the correlation, heterotopic sampling, spatial variability
Line 7: clusters_info.dat        -file with clusters data
Line 8: data_merged.ps           -file with chart output
Line 9: 5                        -number of secondary variables
Line 10:3 4 5 6 7                -columns
Line 11:-998 1e21                -trimming limits
Line 12:Topography               -variable names
Line 13: Z cmfs20
Line 14: Z BK
Line 15: Z cmfs10
Line 16: Z Rollins
```

Line 1 – input data file

Line 2 – xy coordinates required to calculate the spatial variability of each variable

Line 3 – number of block sizes to compare point scale distribution to. Will consider block averaging to 100/500/2000.

Line 4 – number of clusters to merge the data into. There is no control on the number of variables in each cluster.

Line 5 – number seed and number of realizations for initial k-means configurations, can be large (50,000) as program is fast

Line 6 – weight to give to each of the three dissimilarity measures between variables

Line 7 – text output file with clusters information

Line 8 – chart output file

Line 9 – number of secondary variables to consider

Line 10 – columns in the data file

Line 11 – trimming limits

Lines 12-16 – one variable name per line. Number of lines equals Line 9

Linear_merge

```
START OF PARAMETERS:
Line 1: data_merged.dat                  -file with data
Line 2: clusters.out                     -file with clustered variables
Line 3: 42                               -col for primary data
Line 4: -998 1e21                        -trimming limits
Line 5: 7                                -number of mergings to do
Line 6: 8  4 5 6 7 8 9 14 21             -number of variables in this cluster, cols in input file
Line 7: 2 29 30                          -number of variables in this cluster, cols in input file
Line 8: 3 17 19 20                       -number of variables in this cluster, cols in input file
Line 9: 6 22 23 24 25 26 27              -number of variables in this cluster, cols in input file
Line 10:2 32 40                          -number of variables in this cluster, cols in input file
Line 11:2 15 16                          -number of variables in this cluster, cols in input file
Line 12:3 31 37 39                       -number of variables in this cluster, cols in input file
```

Line 1 – input data file

Line 2 – output file with the clustered/merged variables

Line 3 – primary data is required to weight each secondary variable. Merging will be different for each primary variable.

Line 4 – trimming limits

Line 5 – number of amalgamations to consider

Lines 6-12 – add a new line for each merging. First entry indicates how many variables to merge, followed by column number

References

Boisvert J, Rossi ME and Deutsch CV. 2009, Hierarchical Multivariate Regression for Mineral Recovery and Performance Prediction. CCG report 11, 10 pages.

Deutsch, CV and Zanon S. 2007, Direct Prediction of Reservoir Performance with Bayesian Updating, JCPT, 10 p.

Deutsch, J and Deutsch CV, 2009, Multiple bivariate Gaussian plotting and checking. CCG report 11, 10 pages.

Hong, S. 2010. Multivariate Analysis of Diverse Data for Improved Geostatistical Reservoir Modeling. University of Alberta, PhD Thesis. 188 p.

Johnson, Richard A., and Dean W. Wichern. 2007. Applied multivariate statistical analysis. 6th ed. Upper Saddle River, NJ: Pearson prentice Hall.

Kumar, A and C. V. Deutsch, 2009. Optimal Correction of Indefinite Correlation Matrices. CCG report 11, 8 pages.

Leuangthong, O. 2003. Stepwise Conditional Transformation for Multivariate Geostatistical Simulation. University of Alberta, PhD Thesis. 187 p.

Figures

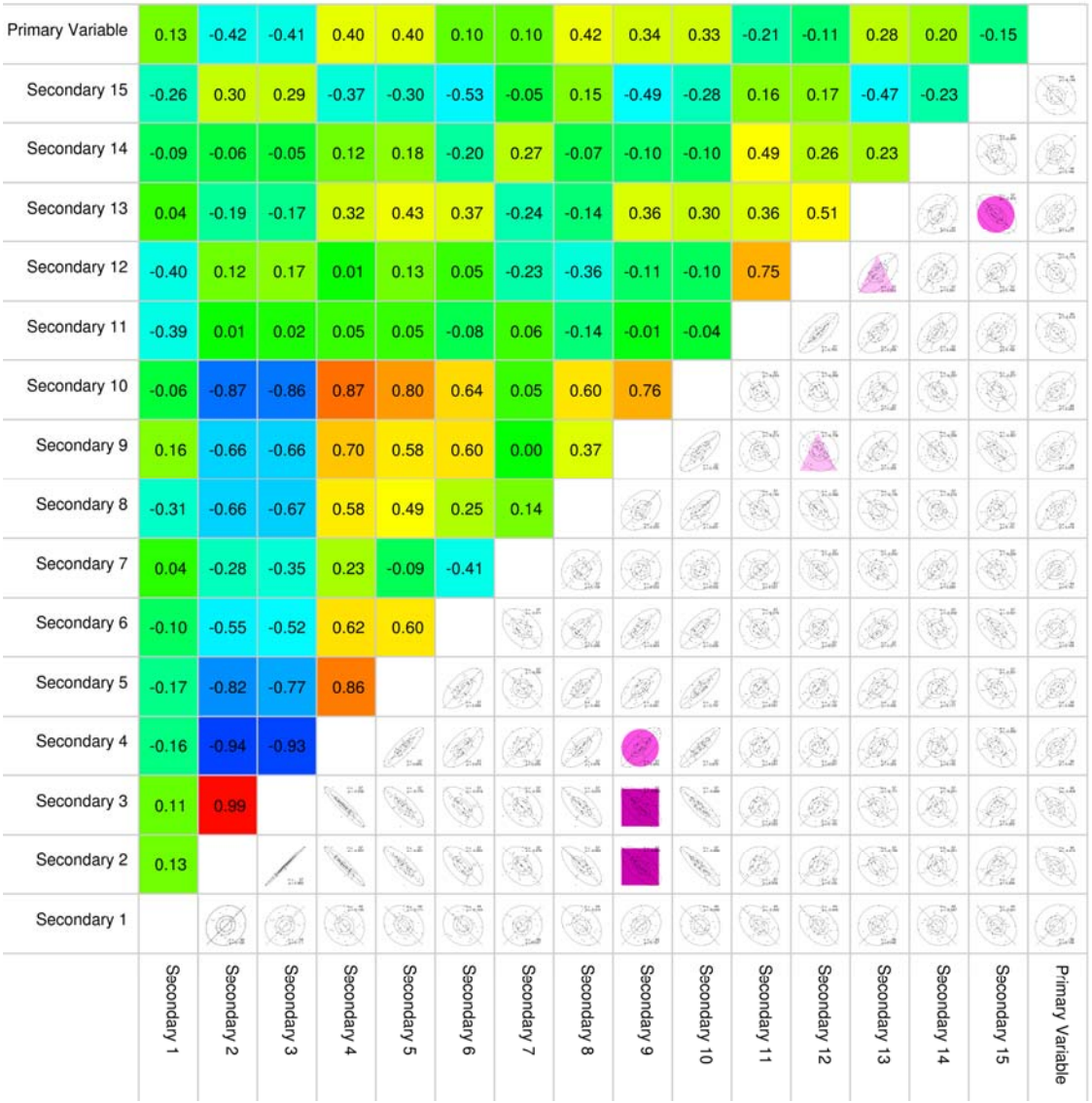

Figure 1: Correlation matrix. Bar above correlation values indicates uncertainty in the correlation. Dark squares indicates variables that are likely nonGaussian, circles indicates moderately nonGaussain and light triangles indicates potentially nonGaussain.

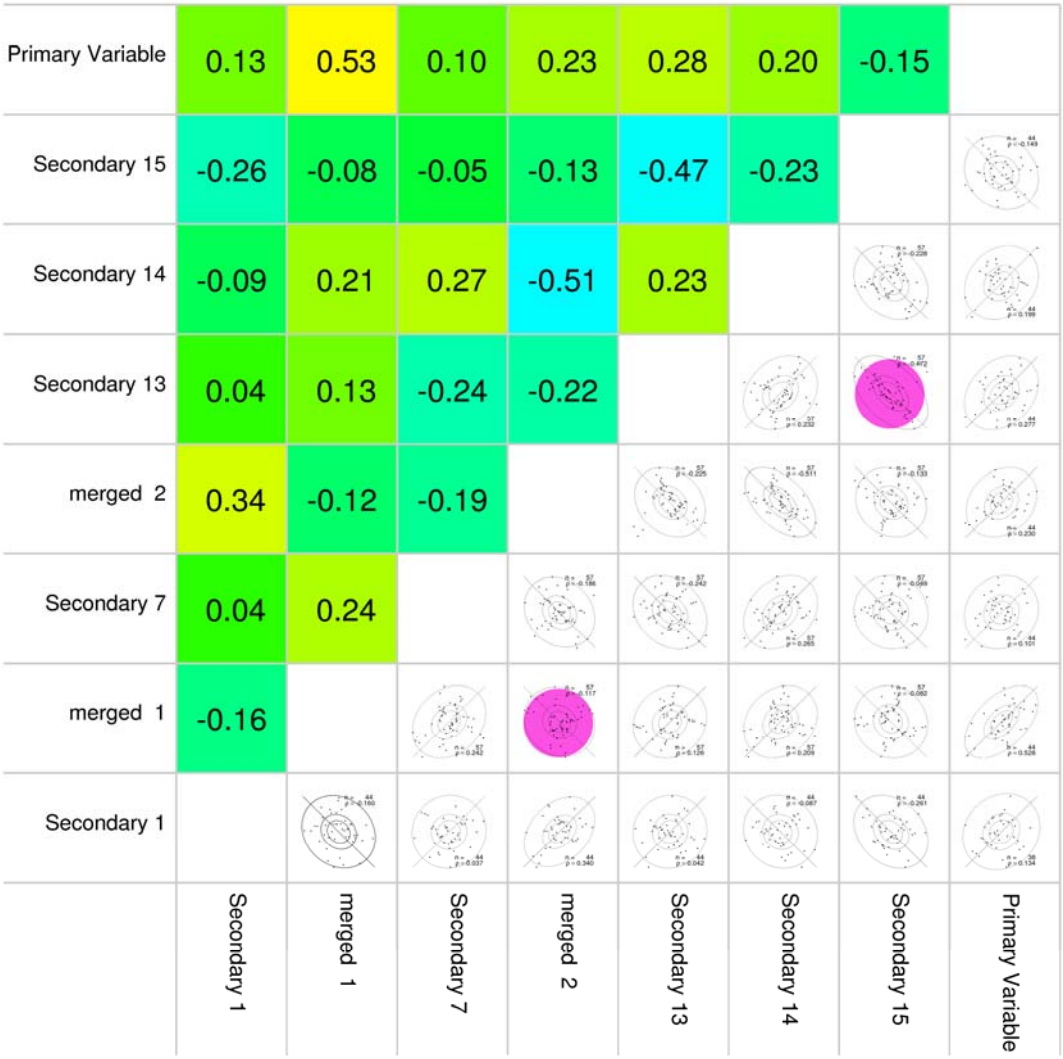

Figure 2: Correlation matrix with redundant secondary variables linearly merged.

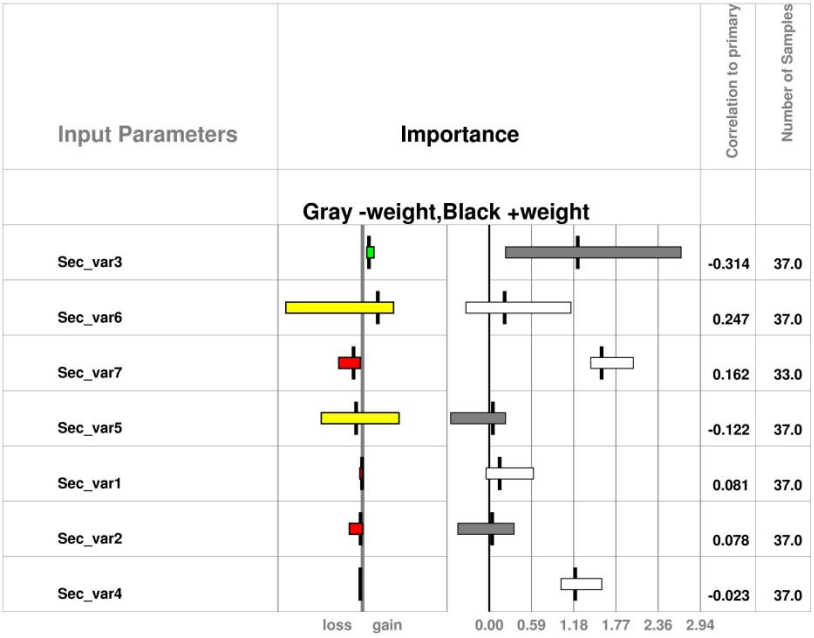

Figure 3: Variables are ordered by their correlation to the primary variable. Bars indicate weight given to each variable. Loss/gain as determined by cross validation.

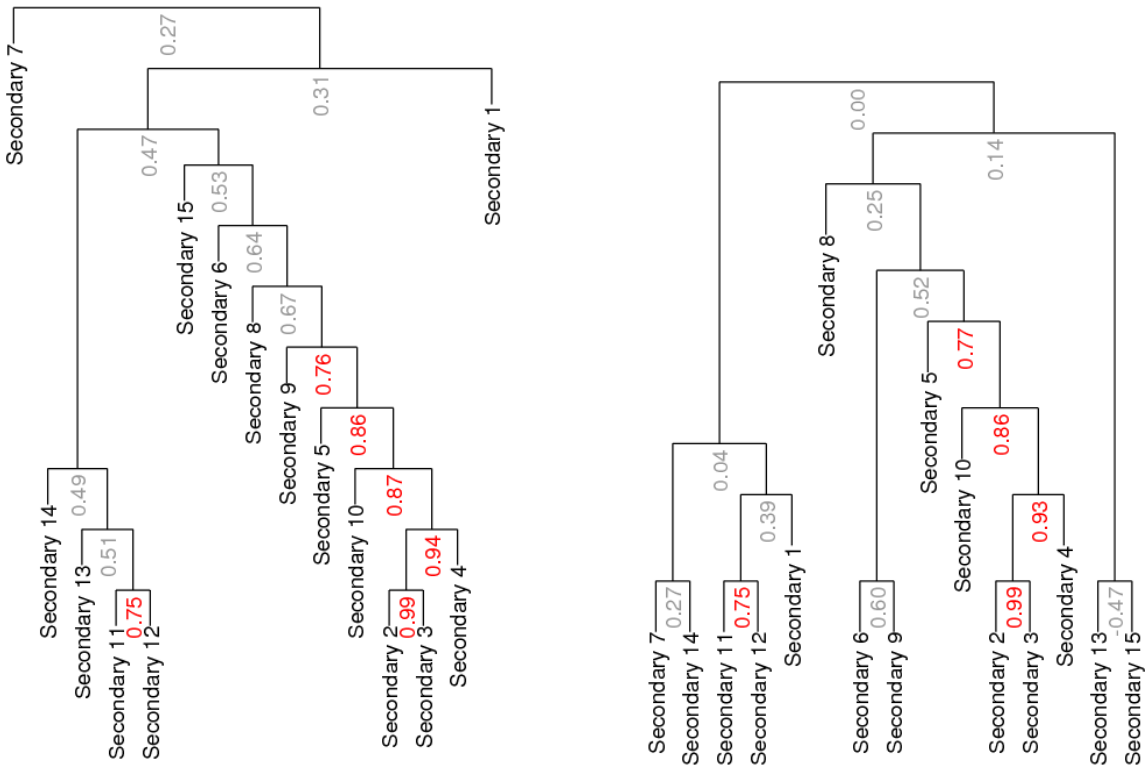

Figure 4: Dendrogram for 15 secondary variables. Left: Complete linkage. Right: Single linkage.