Sensitivity Analysis and the Value of Information in Gaussian Multivariate Prediction

John Manchuk and Clayton V. Deutsch

Centre for Computational Geostatistics
Department of Civil & Environmental Engineering
University of Alberta

Integration of multiple variables in a Gaussian context is remarkably useful: the required parameters can be inferred with relatively few data, the approach is simple and robust to implement, and data transformation schemes like the stepwise transformation can handle problematic relationships in a preprocessing step. This note discusses a methodology and small program to gain insight into the importance of different variables. The importance of multiple variables considered simultaneously is investigated.

Introduction

In multivariate geostatistical methods such as Bayesian updating, there can be a large number of secondary variables contributing to the estimation of some variable of interest. Primary variable estimates are made by solving a set of normal equations where the variables are related through correlation coefficients. The formalism is equivalent to simple cokriging. Spatial information may be used in the prediction. It is interesting to know importance of each variable. The importance is not simply measured by the correlation between each secondary variable and the primary variable being estimated; redundancy between multiple correlated secondary variables must be accounted for. The focus in this short note is documenting the importance of secondary data variables.

Redundancy between multiple variables or multiple data at different data is a critical concept in estimation. The normal equations (kriging) resolve redundancy with the correlation between the variables and what we are predicting in an optimal fashion; however, the resulting weights can be unstable and non-intuitive. The techniques developed in this short note are aimed at the larger problem of the stability of the solution to simultaneous linear equations in the context of geospatial prediction.

Normal Equations

The normal equations (simple cokriging) are the formulation of a linear estimate of a primary variable with a corresponding measure of error variance. In a multivariate Gaussian context, the estimate and variance are interpreted as the parameters of the conditional Gaussian distribution. In a non-Gaussian setting, the estimate is best according to a least squares criterion and the estimation variance is a measure of the data configuration. In the multivariate geostatistics case, there are a number of secondary variables used to provide an estimate of a primary variable using correlation coefficients. A linear system of equations is built from the correlation values between the secondary variables and themselves as well as between the secondary variables and primary variable, see Equation 1.

$$A \cdot \lambda = B$$

$$\begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \rho_{1o} \\ \rho_{2o} \\ \vdots \\ \rho_{no} \end{bmatrix} \quad (1)$$

Where ρ_{11} to ρ_{nn} are the correlation values for the secondary variables (matrix $\boldsymbol{A}$), λ_1 to λ_n are the solutions or weights (matrix $\boldsymbol{\lambda}$) and ρ_{1o} to ρ_{no} are the correlation values between secondary variables and the primary variable (matrix $\boldsymbol{B}$). Each of the n data values could be a different variable at the same or different locations; the matrix would be singular if the same variable is measured twice at the same location. The solution to this set of equation is found by inverting $\boldsymbol{A}$ and multiplying it by $\boldsymbol{B}$:

$$A^{-1} \cdot B = \lambda \quad (2)$$

In practice we do rarely invert the matrix. The solution to a set of simultaneous linear equations is quickly established by other techniques. The estimate and variance can then be defined by Equations 3 and 4:

$$estimate = m = \sum_{i=1}^{N} (Y_i \cdot \lambda_i) \quad (3)$$

$$variance = \sigma_K^2 = \sum_{i=1}^{N} (\rho_{io} \cdot \lambda_i) \quad (4)$$

Where Y_i is the value of secondary variable i, and N is the total number of secondary variables. The 0 notation refers to what we are estimating. One goal is to quantify the importance of different variables for a qualitative understanding. Variables with little importance could be removed for the purpose of stabilizing the solution. Future data collection campaigns could use the results of this analysis.

Quantifying Importance

A measure of the importance of variables used in a multivariate problem has been developed. The importance of a variable is measured as the deviation in kriging variance observed when that variable is removed from the system of equations, see Equation 5:

$$Importance = \frac{\sigma_{K,i}^2 - \sigma_K^2}{\sigma_K^2} \quad (5)$$

Where i is the variable removed, $\sigma_{K,i}^2$ is the kriging variance for that configuration, and σ_K^2 is the kriging variance when all variables are considered.

This relation was developed on the premise that the more a variable contributes to reducing the kriging variance, the more important it is. A program was generated to analyze a given matrix by systematically removing each variable and solving for the weights. The inverse matrix is calculated and output for analysis as well. To explore combinations, the option to perform a full combinatorial analysis on the input matrix is available. Variables are removed for every combination starting with 1 and progressing to $N-2$ where N is the total number of variables in the matrix. This analysis should give insight into the redundancy of certain variable combinations.

For the case where 1 variable is removed, two correlation matrices were looked at: one with 3 secondary variables and 4 primary variables and another with 8 secondary variables and one primary. For the later case, Figure 1 displays the input left hand side ($\boldsymbol{A}$) matrix, its inverse ($\boldsymbol{A}^{-1}$) and the solution ($\boldsymbol{\lambda}$).

Resulting importance measures are displayed using tornado diagrams, see Figure 2. The importance measures have been centered at zero and initial kriging weights and correlation coefficients are shown for comparison purposes. If we recall the two methods of removing variables when unacceptable weights result and apply them to the results for system 2 in Figure 2, we would not be removing variables of least importance. Method 1 would initially eliminate variable 1, which is deemed most important. With method 2 we would remove variable 5 first, which has been given a high importance measure as well.

Importance measures of various variable-sets produced interesting results for the matrix of Figure 1. Variables were removed by sets of 1 up to sets of 6 and half-tornado charts were created in a MS-Excel spreadsheet, see Figure 3. The output file from the program which will be explained later was imported into Excel as space delimited text and then conditionally formatted to shade in removed variables. Half-tornadoes were calculated using the importance measure centered at zero, see Equation 6. This equation normalizes the importance measure by the kriging variance of the system with no variables removed.

$$S_N = 2 \cdot \left[\sqrt{S+1} - 1 \right] \tag{6}$$

Where S_N is the normalized importance measure and S is the importance as calculated by Equation 5.

In Figure 3, the grid on the left indicates the variable set combination where darkened blocks are omitted variables. Two trends are immediately visible: (1) as more variables are removed from the matrix, more patterns exist with higher importance, and (2) variable sets that include variable 1 tend to be more important than others. To reinforce trend (1), with six variables removed, the importance measure remains fairly consistent for all 28 combinations, see Figure 4. Another aspect to note refers to the variables that remain in the matrix while others are removed: If the importance measure is lower, the remaining variables offer less redundancy. This statement is developed from the notion that if the kriging variance increases dramatically (higher importance) when a variable is removed, the remaining variables may contain more redundancy, which tends to cause anomalous kriging weights and higher kriging variances.

Importance of Spatially Correlated Data

An additional study was carried out using covariance matrices for three spatial data configurations. Configurations were developed in MS-Excel and covariance matrices were calculated using a spherical variogram with zero nugget effect and a range of 32 units. Positioning of conditioning data were selected to give one case with no apparent spatial issues, one with screening, and a third with the string-effect [1]. Figures 5 to 7 show the spatial configurations, covariance and inverse matrices and resulting importance tornado diagrams.

In a spatial context, it seems that for most cases data receiving higher weights have a higher importance. Results for the screening configuration are interesting in that the data receiving the highest weight was rated second-most important. Another observation is that conditioning data 1 and 6, which are furthest from the estimate, are deemed more important than data 2 and 5, which are somewhat closer. When looking at the string effect configuration (Figure 7) data 1 and 6 at the ends of the string are deemed most important.

Inverse matrices for each configuration become more unstable when screening and the string-effect are observed. For the first configuration, magnitudes of inverse values rarely exceed 2, whereas with screening and the string-effect, we see magnitudes exceeding 4 and 10 respectively and ranging very close to zero.

Numerical Stability

The main purpose of this short note was to show how the importance of different variables could be quantified. A secondary goal was to dig deeper into certain numerical instability problems that occur from time to time. As with certain spatial data configurations with significant screening, this matrix equation ($\boldsymbol{A\lambda=B}$) could be somewhat unstable, which leads to unacceptably large positive and/or negative weights and erratic estimates. In extreme cases, the estimation variance could be negative. A standard approach is to iteratively remove highly redundant data or data that are poorly correlated to what we are predicting. Data are removed and the solution recalculated until the results appear stable. The following two cases are considered:

1. If too high a weight results, the variable linked to that weight is removed and the system is resolved. This is performed until all weights are acceptable. Unacceptable weights are those beyond some absolute maximum.
2. This case is similar to 1 in that an unacceptable solution is based on the production of high weights; however, instead of removing variables linked to the high weight, the least correlated variable is removed. This process is carried out until all weights are below some absolute maximum.

Both of these methods eventually result in a solution with acceptable weights; however, data is removed from the system with no real measure of its importance. Variables that are potentially very important to an estimate could be removed while data of little importance may remain. When considering the second case, variables that correlate poorly with the primary variable are removed, but the problematic variable or combination there of may involve a high correlation coefficient.

Program

A program, `krig_tests`, was written to calculate the required estimation variances as well as the importance measure with various data removed. The program parameters are as follows:

Line	START OF PARAMETERS:	
1	matrix.dat	-File containing the matrices
2	1	- matrix type: 1=spatial, 2=multivariate
3	2	- number of variables to ignore
4	4 5	- columns
5	7	- column for the RHS matrix
6	0	-Single variable (0) or full combinatorial analysis (1)
7	0	-Number of variables, 0 for automatic
8	sensative.out	-Name of output file for sensativity values
9	matrix_check.out	-Name of output file for matrix checks

Input for the program (Line 1) is a file containing the left and right-hand-side matrices together. If the matrix is for spatial data, the matrix input will typically have one more column for the right-hand-side; however, if the matrix is that for multivariate analysis, the right-hand-side is typically located in a row as well as a column (the matrix is still square upon input). An option to ignore variables is available as well (Lines 3 and 4). For the analysis type, if only single variable analysis is selected (value of 0 for Line 6) then the matrix check output will show every inverse

matrix for each removed variable. If a full combinatorial analysis is done, only the initial input and inverse matrices are output. If the number of variables is not indicated in the input file then it can be specified in Line 7.

Output of the program is a file containing the number of removed variables, the lexicographic combination of removed variables and the importance measure for that combination:

Line	Output of variable orders and importance measures
1	8
2	Number Removed
3	Variable 1
4	Variable 2
5	Variable 3
6	Variable 4
7	Variable 5
8	Variable 6
9	Importance
10	1 1 0 0 0 0 0 0.033307
11	1 0 1 0 0 0 0 0.183239
12	1 0 0 1 0 0 0 0.033477
13	1 0 0 0 1 0 0 0.058244
14	1 0 0 0 0 1 0 0.027477
15	1 0 0 0 0 0 1 0.031448

In the case of a full combinatorial analysis, the output is the same; however, the number of variables removed increases from 1 to *N*-2 where *N* is the number of variables. The lexicographic combination indicators contain the same number of 1's as there are variables removed and their position is associated to the variable they describe. Looking at Line 13, there is one variable removed, that being variable 4 and the resulting importance is 0.058244.

Conclusions

The multiGaussian framework for data integration has proven remarkably useful for many years. We present a small program to calculate the importance of each variable by direct means – calculate the reduction in the estimation variance when the variable is dropped. This was implemented for one variable at a time and for sets of multiple variables at a time. In most cases, the single-variable sensitivity is sufficient; however, there are times when groups of secondary data are important. In the presence of many secondary data (>10) or when the secondary data are highly redundant (ρ>0.95) the solution to the multiGaussian set of equations may become unstable.

References

Deutsch, C.V., *Kriging in a Finite Domain*, Mathematical Geology, Vol. 25, No. 1, 1993

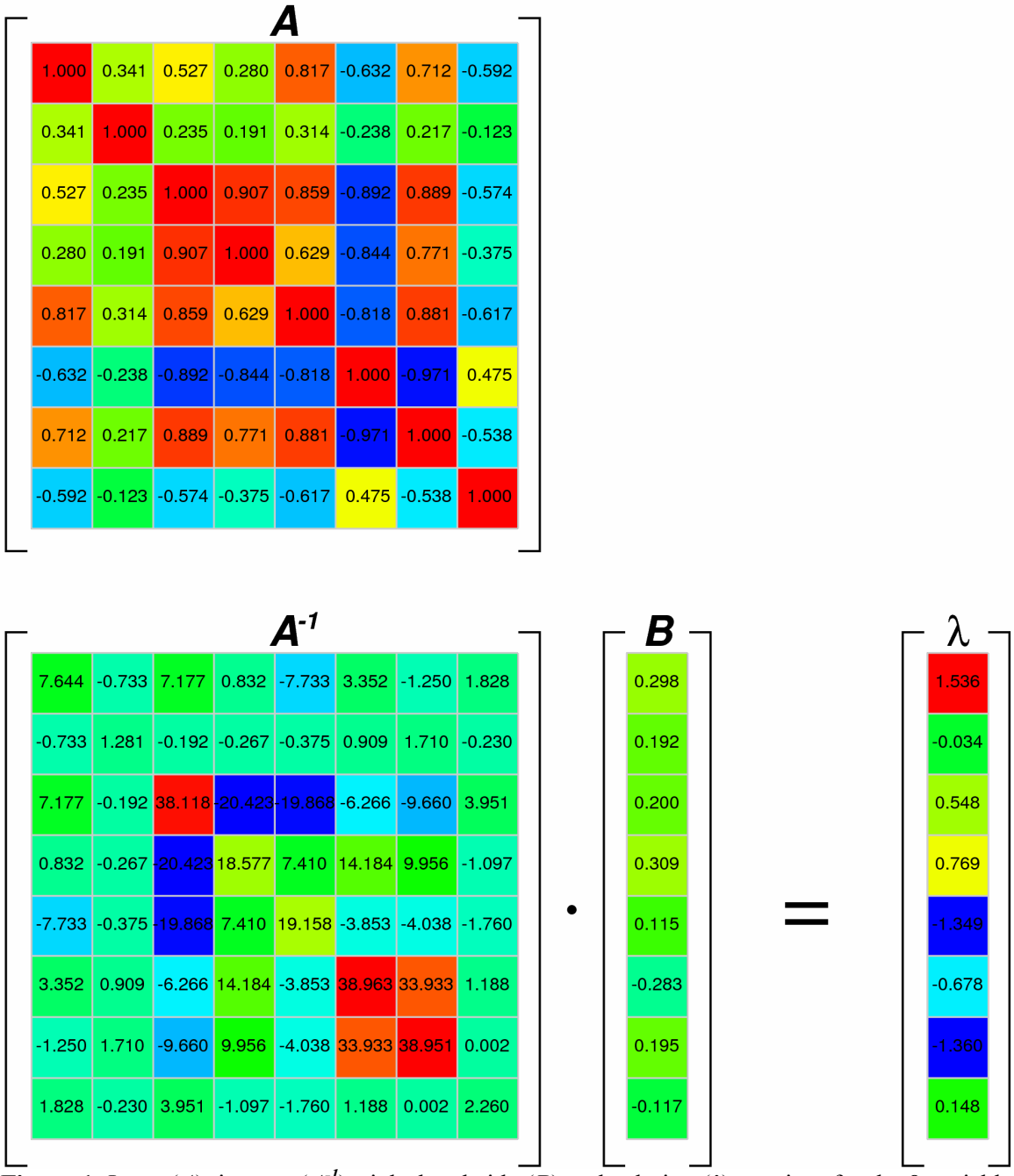


Figure 1: Input (***A***), inverse (A^{-1}), right-hand-side (***B***) and solution (λ) matrices for the 8-variable problem.

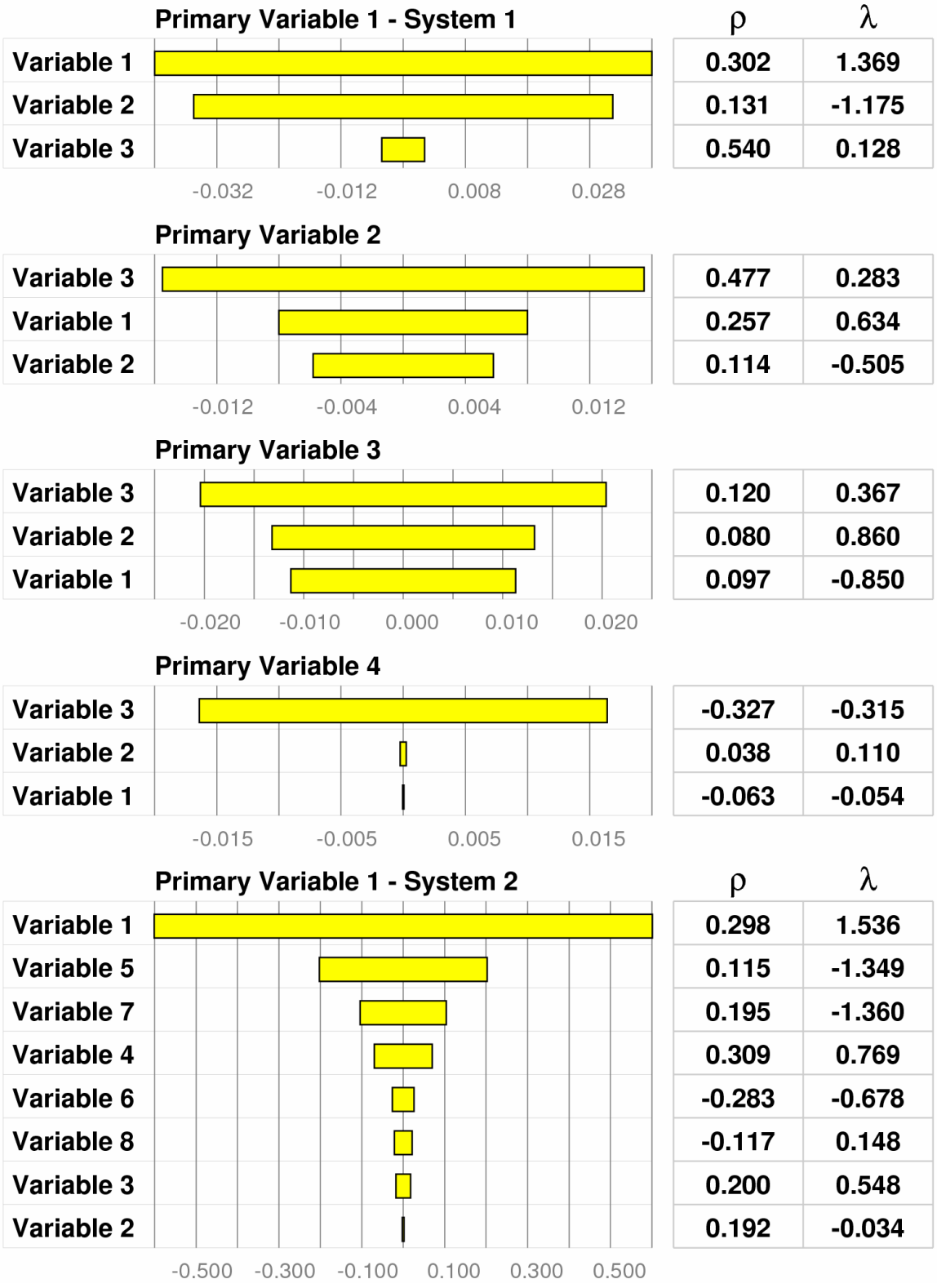


Figure 2: Importance tornado diagrams and associated right-hand-side correlation coefficients and solution weights.

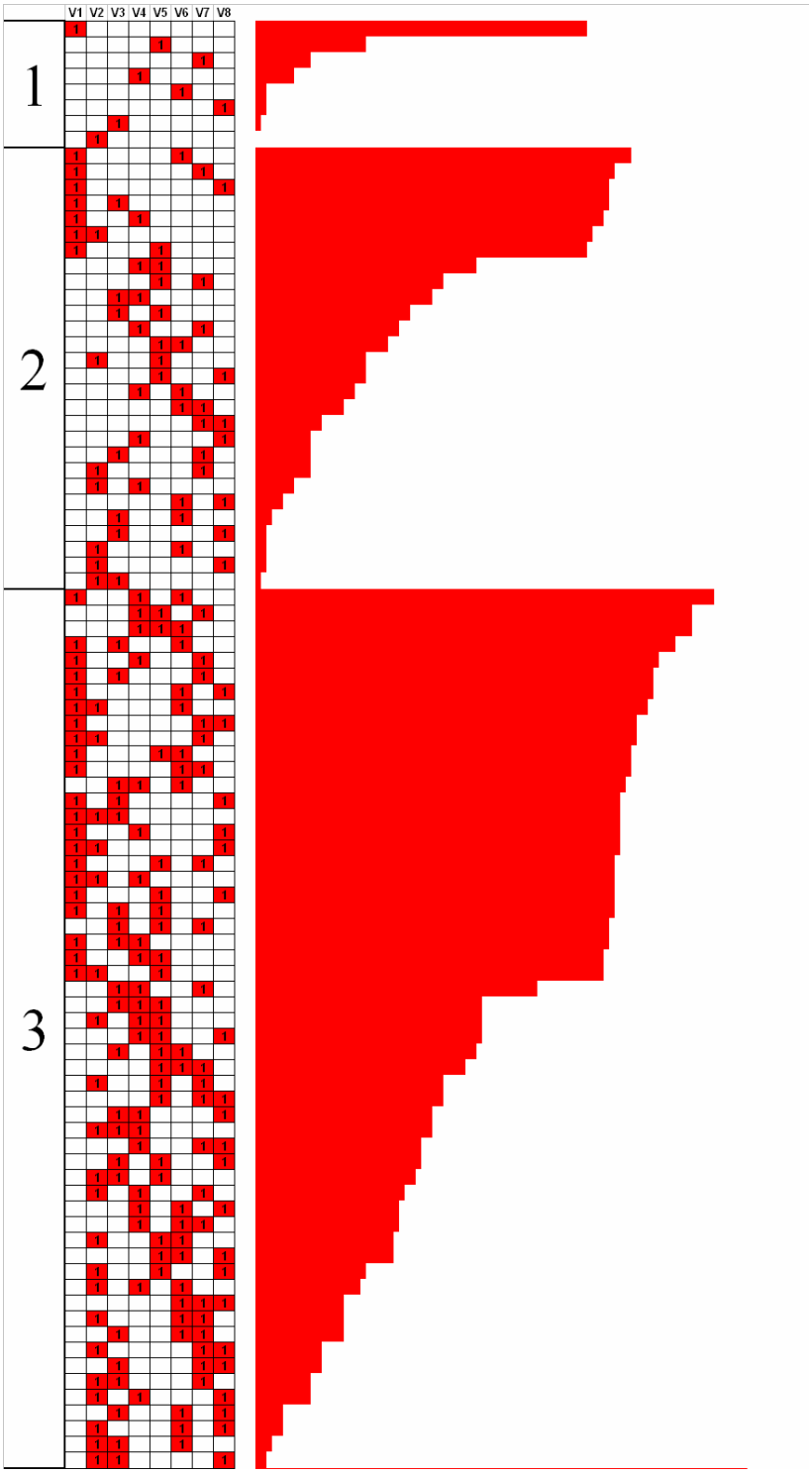


Figure 3: Half-Tornado diagrams (right) of importance for removed variable sets (left). Large numbers on the left indicate number of variables removed.

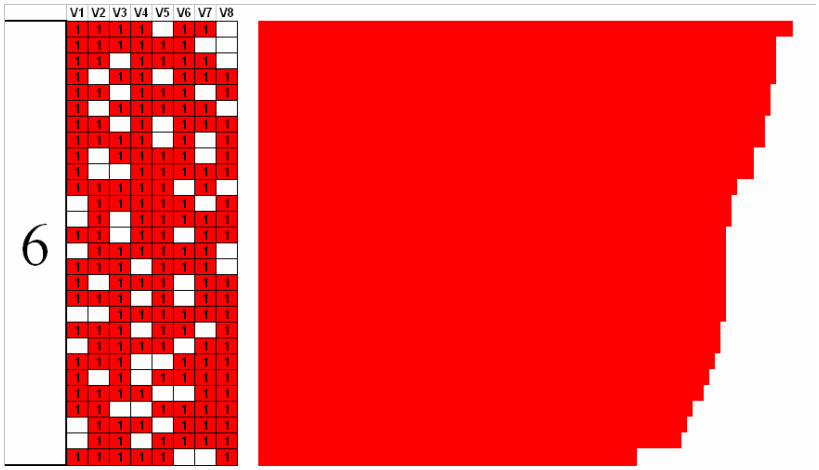
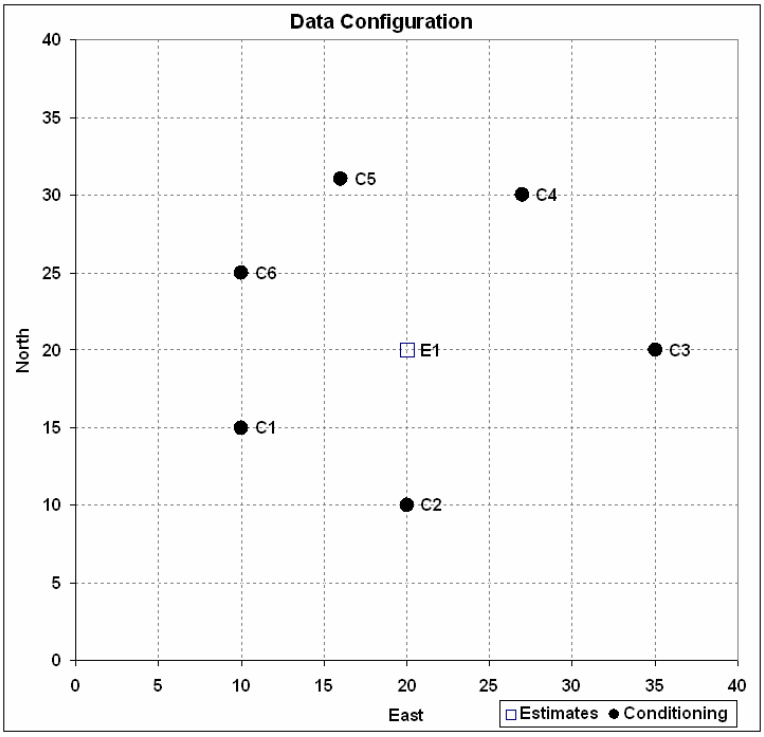
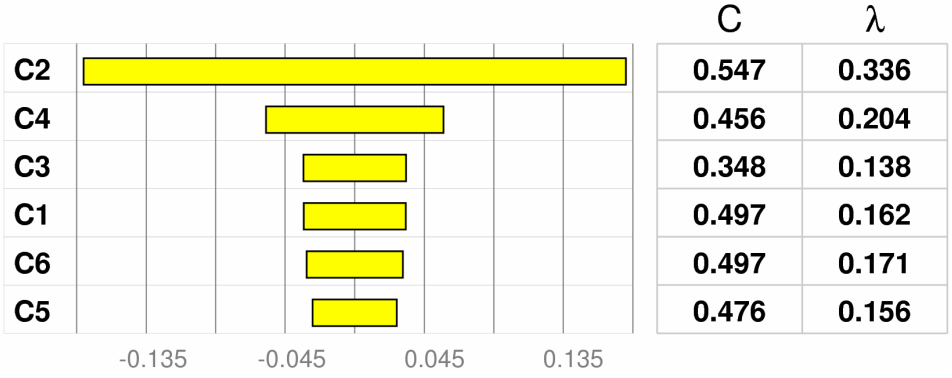


Figure 4: Combinatorial analysis results for 6 variables removed.

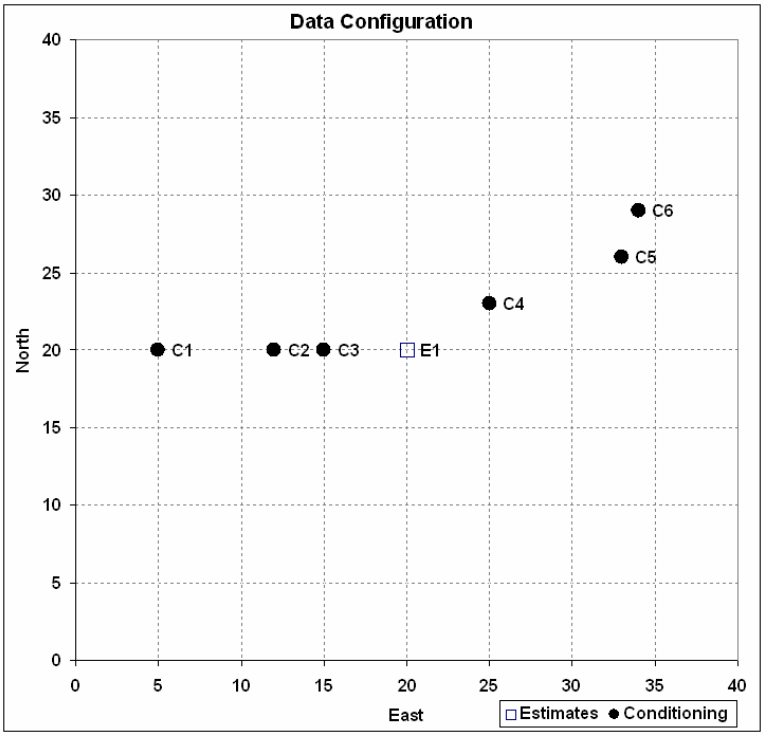


$$A = \begin{bmatrix} 1.000 & 0.497 & 0.058 & 0.115 & 0.275 & 0.547 \\ 0.497 & 1.000 & 0.244 & 0.152 & 0.147 & 0.244 \\ 0.058 & 0.244 & 1.000 & 0.432 & 0.132 & 0.058 \\ 0.115 & 0.152 & 0.432 & 1.000 & 0.503 & 0.254 \\ 0.275 & 0.147 & 0.132 & 0.503 & 1.000 & 0.612 \\ 0.547 & 0.244 & 0.058 & 0.254 & 0.612 & 1.000 \end{bmatrix}, A^{-1} = \begin{bmatrix} 1.811 & -0.726 & 0.098 & 0.019 & 0.142 & -0.910 \\ -0.726 & 1.420 & -0.301 & -0.019 & -0.006 & 0.076 \\ 0.098 & -0.301 & 1.310 & -0.605 & 0.141 & 0.012 \\ 0.019 & -0.019 & -0.605 & 1.644 & -0.817 & 0.110 \\ 0.142 & -0.006 & 0.141 & -0.817 & 2.042 & -1.126 \\ -0.910 & 0.076 & 0.012 & 0.110 & -1.126 & 2.139 \end{bmatrix}, B = \begin{bmatrix} 0.497 \\ 0.547 \\ 0.348 \\ 0.456 \\ 0.476 \\ 0.497 \end{bmatrix}$$

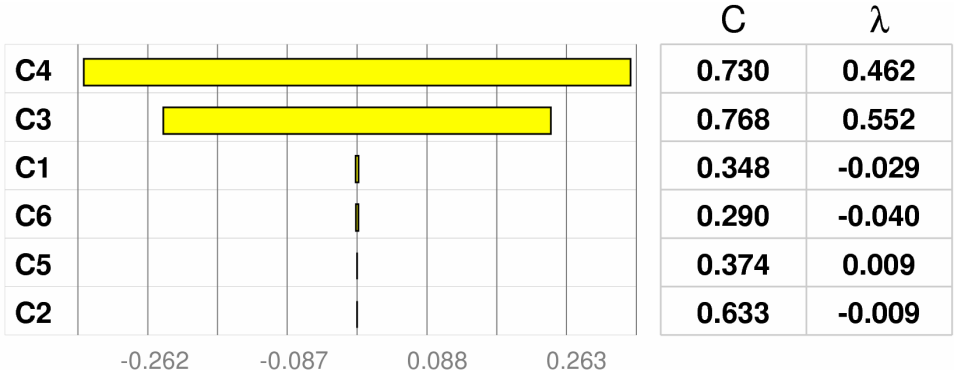


	C	λ
C2	0.547	0.336
C4	0.456	0.204
C3	0.348	0.138
C1	0.497	0.162
C6	0.497	0.171
C5	0.476	0.156

Figure 5: Estimate and conditioning data locations along with covariance (***A***), inverse ($\boldsymbol{A^{-1}}$) and right-hand-side (***B***) matrices for a problem-free configuration.

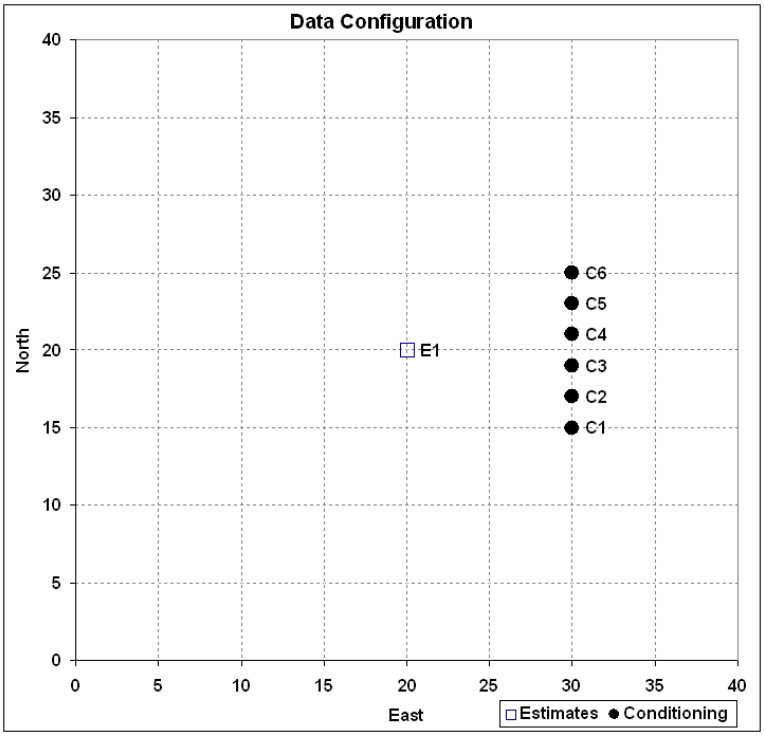


$$A = \begin{bmatrix} 1.000 & 0.677 & 0.547 & 0.178 & 0.016 & 0.004 \\ 0.677 & 1.000 & 0.860 & 0.411 & 0.135 & 0.091 \\ 0.547 & 0.860 & 1.000 & 0.528 & 0.215 & 0.156 \\ 0.178 & 0.411 & 0.528 & 1.000 & 0.609 & 0.512 \\ 0.016 & 0.135 & 0.215 & 0.156 & 1.000 & 0.852 \\ 0.004 & 0.091 & 0.156 & 0.512 & 0.852 & 1.000 \end{bmatrix}, A^{-1} = \begin{bmatrix} 1.893 & -1.469 & 0.125 & 0.181 & 0.072 & -0.047 \\ -1.469 & 5.023 & -3.539 & 0.005 & 0.058 & 0.051 \\ 0.125 & -3.539 & 4.485 & -1.016 & 0.045 & 0.102 \\ 0.181 & 0.005 & -1.016 & 2.179 & -1.080 & -0.039 \\ 0.072 & 0.058 & 0.045 & -1.080 & 4.305 & -3.129 \\ -0.047 & 0.051 & 0.102 & -0.039 & -3.129 & 3.666 \end{bmatrix}, B = \begin{bmatrix} 0.348 \\ 0.633 \\ 0.768 \\ 0.730 \\ 0.374 \\ 0.290 \end{bmatrix}$$

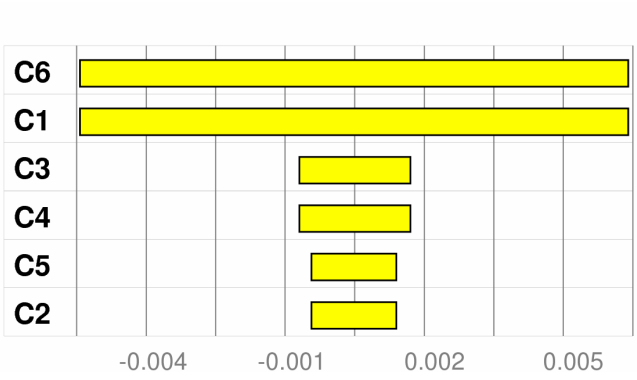


	C	λ
C4	0.730	0.462
C3	0.768	0.552
C1	0.348	-0.029
C6	0.290	-0.040
C5	0.374	0.009
C2	0.633	-0.009

Figure 6: Estimate and conditioning data locations along with covariance (***A***), inverse (***A*** $^{-1}$) and right-hand-side (***B***) matrices for a configuration displaying screening effects.



$$A = \begin{bmatrix} 1.000 & 0.906 & 0.813 & 0.722 & 0.633 & 0.547 \\ 0.906 & 1.000 & 0.906 & 0.813 & 0.722 & 0.633 \\ 0.813 & 0.906 & 1.000 & 0.906 & 0.813 & 0.722 \\ 0.722 & 0.813 & 0.906 & 1.000 & 0.906 & 0.813 \\ 0.633 & 0.722 & 0.813 & 0.906 & 1.000 & 0.906 \\ 0.547 & 0.633 & 0.722 & 0.813 & 0.906 & 1.000 \end{bmatrix}, A^{-1} = \begin{bmatrix} 5.642 & -5.314 & 0.026 & 0.027 & 0.027 & 0.215 \\ -5.314 & 10.639 & -5.340 & 0.000 & 0.000 & 0.027 \\ 0.026 & -5.340 & 10639 & -5.340 & 0.000 & 0.027 \\ 0.027 & 0.000 & -5.340 & 10.639 & -5.340 & 0.026 \\ 0.027 & 0.000 & 0.000 & -5.340 & 10.639 & -5.314 \\ 0.215 & 0.027 & 0.027 & 0.026 & -5.314 & 5.642 \end{bmatrix}, B = \begin{bmatrix} 0.497 \\ 0.528 \\ 0.544 \\ 0.544 \\ 0.528 \\ 0.497 \end{bmatrix}$$



	C	λ
C6	0.497	0.149
C1	0.497	0.149
C3	0.544	0.092
C4	0.544	0.092
C5	0.528	0.081
C2	0.528	0.081

Figure 7: Estimate and conditioning data locations along with covariance (***A***), inverse (***A*** $^{-1}$) and right-hand-side (***B***) matrices for a configuration showing the string effect.