Hierarchical Multivariate Regression for Mineral Recovery and Performance Prediction

J.B. Boisvert, Mario E. Rossi and Clayton V. Deutsch

Mineral recovery and expected plant performance are difficult to predict because they are influenced by a large number of variables such as mineralogy, grade, grain size, plant operation parameters, etc. Often constant recovery factors and plant efficiencies are assumed for a given mine based on past experience and empirical rules. Such methods are acceptable during the feasibility stages of mineral exploration; however, when results of pilot plant trials are available, statistical methods can be utilized to better predict recovery and plant performance. In this paper, 841 bulk samples from flotation and leach tests are used for the calibration of a predictive model. The end result is a model that can be used to predict recovery and plant performance based on available geometallurgical data.

Over 200 geometallurgical variables are available to develop a hierarchical linear regression model for the case study considered. Redundant and unimportant variables are identified and removed from the modeling process, reducing the number of variables to 112. Through a number of hierarchical variable amalgamation steps the variables are condensed into 4 major sub-categories. A linear model based on these 4 amalgamated variables provides a predictive model that is used to estimate potential mineral recovery and plant performance over the mine extents. Minerals of interest in this mine include copper and uranium with secondary production of gold and silver.

Introduction

Plant performance is highly dependent on a large number of variables, such as (1) plant feed (2) operational parameters (3) equipment efficiencies (4) equipment repairs. The purpose of this paper is to relate available geometallurgical data to plant performance. This is done by correlating the available data to pilot plant trials. A total of 841 pilot runs are available with associated plant feed mineralogy, head assays and mineral association data; these data types are described in

Table 1. Plant performance indicators of importance include recovery of Cu and U_3O_8, acid consumption (used in the leaching process), net recovery, drop weight index (DWi) and bond mill work index (BMWi); using the data in

Table 1 as input to a regression model, these six plant performance variables can be predicted at all locations in the deposit.

Table 1: Data available.

Data Type	Description	Notes
Head Assays	This data contains the % content of various important elements, including: Co, As, Mo, Ni, Pb, Zn, Zr, Sr, Bi, Cd, Cs, Ga, In, Sb, Se, Te, Th, Tl	This data is compositional in nature. This will be exploited when modeling these variables in paper 303, also in this report.
Mineralogy	A total of 10 identified minerals make up the bulk of the deposit. These include: Brannerite, Coffinite, Uraninite, Pyrite, Chalcopyrite, Bornite, Chalcocite, Other Sulphides, Acid Soluble Gangue and Acid Insoluble Gangue	This data is also compositional in nature. This will be exploited when modeling these variables in paper 303.
Association Data	A number of thin sections are available. These have been analyzed and the complete matrix of associations between minerals is available. This describes the contact area between two adjacent minerals within a single grain of crushed material.	This data is also compositional in nature. This will be discussed more in paper 303.

Methodology

A linear regression model is used to predict the plant performance variables. One drawback with a linear regression model is that all input variables are required for prediction. Thus, if a single input variable is missing from a sample, the regression model cannot be applied. For this reason, three regression models are generated (Table 2). Each model represents a decreasing number of input parameters. For example, for locations in the deposit where association data is not known the "Full Model" cannot be applied and the "Typical Model" would be appropriate.

Table 2: Description of predictive models generated.

Model	Input variables	Output	Comments
Full Model	-head assays (i.e. %cu, %U …) -10 mineralogy -10x11 matrix of associations -specific gravity	-Cu, U, Au, Ag recoveries -acid consumption -net recovery (U) -BMWi and DWi	This model represents the maximum data available.
Typical Model	- head assays (i.e. %cu, %U …) -10 mineralogy -specific gravity	-Cu, U, Au, Ag recoveries -acid consumption -net recovery (U) -BMWi and DWi	This is the base case model. Field data will most likely contain all these variables
Limited Model	-limited head assays -7 mineralogy variables -specific gravity	-Cu, U, Au, Ag recoveries -acid consumption -net recovery (U) -BMWi and DWi	Only head assays that have many samples in the available database are considered.

The regression models are based on a large set of input variables. The variables are merged into super secondary variables using a likelihood based on the correlations between variables. This is done because there are two few sample data available to accurately determine regression coefficients for the 204 input variables available. The final model is a linear regression on four super secondary variables. The overall methodology consists of six steps:

1) Normal score the input variables.
2) Merge the variables (level 1). This step reduces the 112 input variables to 23 merged variables.
3) Merge the variables (level 2). This step reduces the 23 merged variables to 4.
4) Regression on the 4 variables.
5) Back transform the estimated variables (DWi, BMWi, Cu recovery, U_3O_8 recovery, acid consumption and net recovery).
6) Determine uncertainty in the model

Each step of the model building process will be expanded upon.

Step 1: Normal score data

First the number of variables must be reduced. Variables are removed from the analysis because (1) they have a low correlation to the six output variables or (2) they are redundant with one of the other input variables. A variable was considered to have a low correlation if the maximum correlation to any of the output variables was less than 0.13. A variable was considered redundant with another input variable if it had a correlation greater than 0.94. This reduces the number of input variables to 112.

There are a total of 841 samples available for modeling; however, not all samples contain all 112 variables used in the calibration of this model. Due to the nature of a regression model, it is necessary that all 112 variables be present for a sample to be used for calibration. Of the 841 samples, 328 samples

were retained for modeling. Figure 1 shows which variables are lacking in the overall data set, more data are available if the mineral associations are ignored (i.e. the typical model).

All 118 variables (112 input + 6 output) are independently normal score transformed. A visual assessment of the bivariate relationships between the input data indicated very few non-linear relationships; therefore, stepwise conditional transformations are not considered.

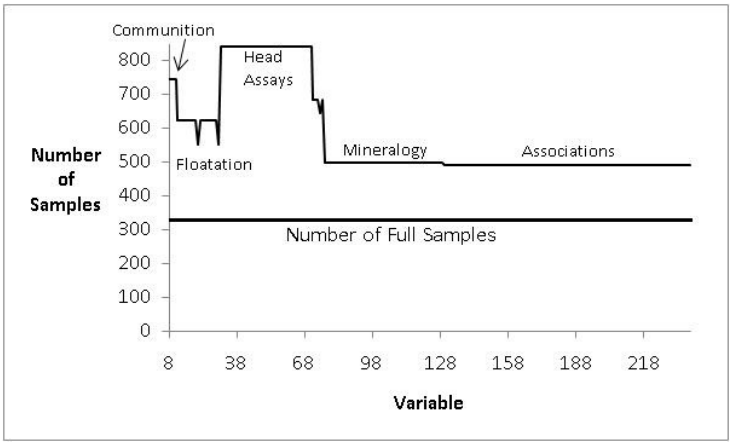


Figure 1: Number of samples available for modeling. Each variable is numbered 8 through 239.

Step 2: Merge variables - reduce 112 input variables to 23 merged super secondary variables.

There is a danger of over fitting the available calibration data if a regression model is constructed on all 112 input variables. Therefore, subsets of the input data were amalgamated to construct super secondary merged variables. These merged variables are linear combinations of a subset of variables and significantly reduces the dimensionality of the problem while also reducing model over fitting; see **Error! Not a valid bookmark self-reference.** and **Error! Not a valid bookmark self-reference.** for the specific subsets of variables used. These subsets were generated using geophysical knowledge to merge related variables. The selection of subsets could be further optimized to improve model performance and is discussed in future work.

Specifically, the merged variables are generated by assigning weights to each variable (Equation 1). These weights are generated by solving Equation 2 for each merged variable and for each of the six output variables. The right hand side of Equation 2 contains the correlation between one of the variables of interest and the *n* input variables to be merged, the left hand side is the correlation between all *n* variables to be merged.

$$M(v) = \sum_{i=1}^{n} \lambda_i v_i \tag{1}$$

where *n* is the number of variables to be merged based on the weights from a likelihood calculation.

$$\begin{bmatrix} \rho_{1,1} & \rho_{2,1} & \cdots & \rho_{n,1} \\ \rho_{1,2} & \rho_{2,2} & \cdots & \rho_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,n} & \rho_{2,n} & \cdots & \rho_{n,n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \rho_{0,1} \\ \rho_{0,2} \\ \vdots \\ \rho_{0,n} \end{bmatrix} \tag{2}$$

These correlation matrices are often poorly conditioned. Poorly conditioned matrices cause extreme weights(λ_i) and introduce unwarranted noise in the predictions. To prevent this, the correlation matrices are *fixed* to improve their stability. This correction is accomplished by decreasing the values of the off diagonal elements of the matrix, which increases the value of the smallest eigenvalue for the matrix and increases stability. The minimum eigenvalue for the correlation matrices was set to 0.05. 24 of the correlation matrices for the full model required a correction, 18 of the correlation matrices for the typical model required a correction and 12 of the correlation matrices for the limited model required a correction.

The merged variables are a linear combination of N(0,1) variables. Thus, the mean of the merged variables will be 0 but the variance will not be 1. The merged variables are standardized by the standard deviation determined from the following classical relationship:

$$\sigma^2(M(v)) = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Cov(v_i, v_j) \tag{3}$$

Thus, the final merged variable becomes:

$$M(v) = \frac{\sum_{i=1}^{n}\lambda_i v_i}{\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Cov(v_i, v_j)} \tag{4}$$

Step 3: Merge variables - reduce 23 input variables to 4 merged variables for regression.

There are two levels of variable amalgamation. The first level grouped related variables into 16 merged variables and retains 7 additional variables for a total of 23 variables (**Error! Not a valid bookmark self-reference.**). The second level groups the variables into the final 4 super-secondary variables used for regression (**Error! Not a valid bookmark self-reference.** and **Error! Not a valid bookmark self-reference.**):

A. Retained variables
B. Head assays
C. Mineralogy
D. Associations

Step 4: Regression.

The typical and limited models are generated by regression on variables A, B and C while the full model considers variables A through D. Regression is preformed with both linear and quadratic terms but through a cross validation exercise it was found that increasing the number of terms beyond the linear coefficients resulted in little consistent gain and the linear model is sufficient. Thus, the final model becomes:

$$Prediction = av_1 + bv_2 + cv_3 + dv_4 \tag{5}$$

Step 5: Back Transformation.

Once the predictions are made in normal space for each of the six output variables, they must be transformed back into original units using the original transformation tables from Step 1.

Figure 2: Variables used in the limited model. A total of 28 input variables are considered.

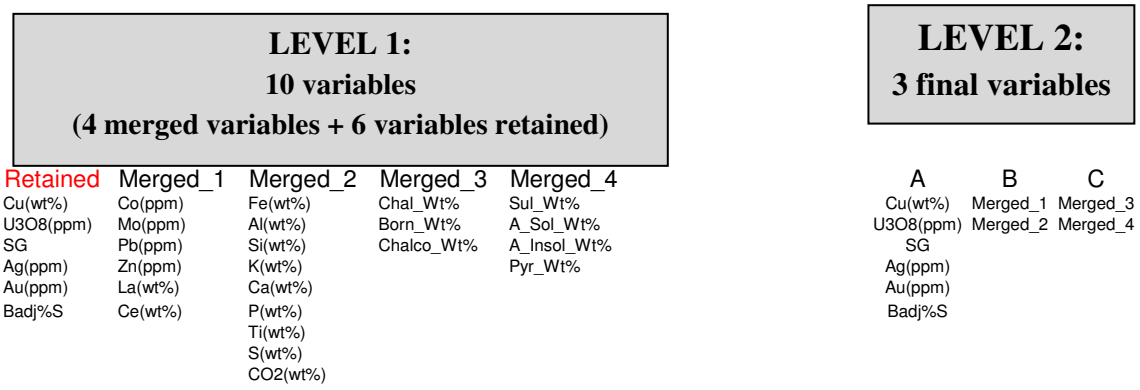


Figure 3: Variables used in the regression models.

Retained	Merged_1	Merged_2	Merged_3	Merged_4	Merged_5
Cu(wt%)	Ba(wt%)	La(wt%)	Uran_Wt%	Chal_Wt%	Sul_Wt%
U3O8(ppm)	Fe(wt%)	Mg(wt%)	Cof_Wt%	Born_Wt%	A_Sol_Wt%
SG	Al(wt%)	Mn(wt%)	Bran_Wt%	Chal_Wt%	A_Insol_Wt%
K:Al	Si(wt%)	Na(wt%)			Pyr_Wt%
	K(wt%)	P(wt%)			
	Ca(wt%)	Ti(wt%)			
	S(wt%)				
	CO2(wt%)				
	F(wt%)				

Merged_6	Merged_7	Merged_8	Merged_9
Bran_Pyr_assoc	Cof_Bran_assoc	Uran_Cof_assoc	Pyr_Cof_assoc
Bran_Chalcopy_assoc	Cof_Uran_assoc	Uran_Chalcopy_assoc	Pyr_Uran_assoc
Bran_Bornite_assoc	Cof_Pyr_assoc	Uran_Bornite_assoc	Pyr_Chalcopy_assoc
Bran_Chalcocite_assoc	Cof_Chalcopy_assoc	Uran_A_Sol_assoc	Pyr_Sulphides_assoc
Bran_A_Sol_assoc	Cof_Chalcocite_assoc	Uran_A_Insol_assoc	Pyr_A_Sol_assoc
Bran_A_Insol_assoc	Cof_Sulphides_assoc		Pyr_Free_Surf_assoc
Bran_Free_Surf_assoc	Cof_A_Sol_assoc		
	Cof_A_Insol_assoc		
	Cof_Free_Surf_assoc		

Merged_10	Merged_11	Merged_12
Chalcopy_Bran_assoc	Bornite_Cof_assoc	Chalcocite_Chalcopy_assoc
Chalcopy_Cof_assoc	Bornite_Pyr_assoc	Chalcocite_Bornite_assoc
Chalcopy_Uran_assoc	Bornite_Chalcopy_assoc	Chalcocite_Sulphides_assoc
Chalcopy_Pyr_assoc	Bornite_Chalcocite_assoc	Chalcocite_A_Sol_assoc
Chalcopy_Bornite_assoc	Bornite_Sulphides_assoc	Chalcocite_A_Insol_assoc
Chalcopy_Chalcocite_assoc	Bornite_A_Sol_assoc	Chalcocite_Free_Surf_assoc
Chalcopy_Sulphides_assoc	Bornite_A_Insol_assoc	
Chalcopy_A_Sol_assoc	Bornite_Free_Surf_assoc	
Chalcopy_A_Insol_assoc		
Chalcopy_Free_Surf_assoc		

Merged_13	Merged_14	Merged_15
Sulphides_Uran_assoc	A_Sol_Bran_assoc	A_Insol_Bran_assoc
Sulphides_Pyr_assoc	A_Sol_Cof_assoc	A_Insol_Cof_assoc
Sulphides_Chalcopy_assoc	A_Sol_Uran_assoc	A_Insol_Uran_assoc
Sulphides_Bornite_assoc	A_Sol_Pyr_assoc	A_Insol_Chalcopy_assoc
Sulphides_A_Sol_assoc	A_Sol_Chalcopy_assoc	A_Insol_Bornite_assoc
Sulphides_A_Insol_assoc	A_Sol_Bornite_assoc	A_Insol_Sulphides_assoc
	A_Sol_Chalcocite_assoc	A_Insol_A_Sol_assoc
	A_Sol_Sulphides_assoc	A_Insol_Free_Surf_assoc
	A_Sol_A_Insol_assoc	
	A_Sol_Free_Surf_assoc	

LEVEL 2: 4 Final Variables

Variable A contains individual variables retained.
Variable B contains the remainder of the head assays.
Variable C contains all mineralogy variables.
Variable D contains all association variables.

Step 6: Determine uncertainty in the model

When a prediction is made the uncertainty in that prediction is also determined. The uncertainty is obtained by examining the distribution of true values for a given estimate. Consider the difference in making an acid consumption prediction of 60 vs. 220 (Figure 4). There is more uncertainty in the estimate of 220. The measure of uncertainty used is the spread of the true values around the estimate. A p_{10} and p_{90} range is given for each estimate based on the true values near the estimate.

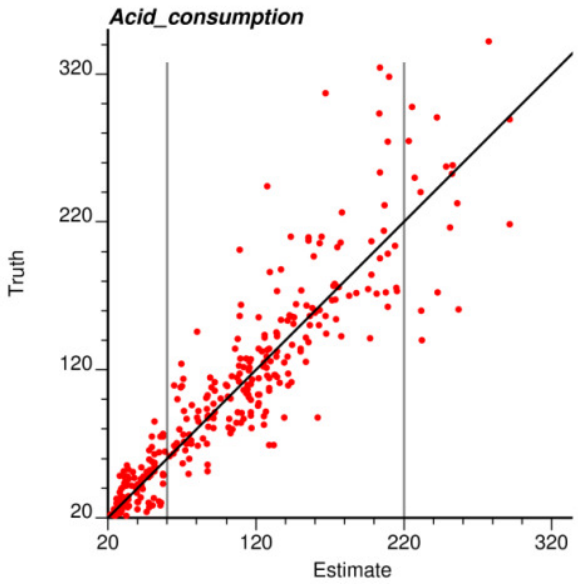


Figure 4: Determining the uncertainty in an estimate of 60 vs. 220. There is more uncertainty at 220.

Analysis

This section will explore the effectiveness of the different models. All samples were used to generate the regression models with the above methodology. High correlation between the estimate and the truth is desirable. Figure 5 shows the models built on all possible data points available for the different models.

Rather than show the 768 coefficients for variable merging and the 24 regression coefficients, a tornado chart (**Error! Not a valid bookmark self-reference.**) will be used to illustrate the influence of each of the 112 variables on the overall model. The lower limit is determined by selecting the p_{10} value for the input variable of interest and setting all remaining 111 variables to their p_{50} value. An estimate is made for each of the six output variables, giving the lower limit on the tornado chart. Similarly, the p_{90} value is selected for the variable of interest to generate the upper limit on the tornado chart. A short horizontal line to the left of the variable indicates that the variable is negatively correlated with the output variable (i.e. the p_{10} response is higher than the p_{90}). Bars are shaded based on the origin of the variable: *White* – Head Assays; *Gray* – Associations; *Red* – Mineralogy; *Black* – Specific gravity.

Some interesting relationships were discovered in the cross plots and the tornado charts:

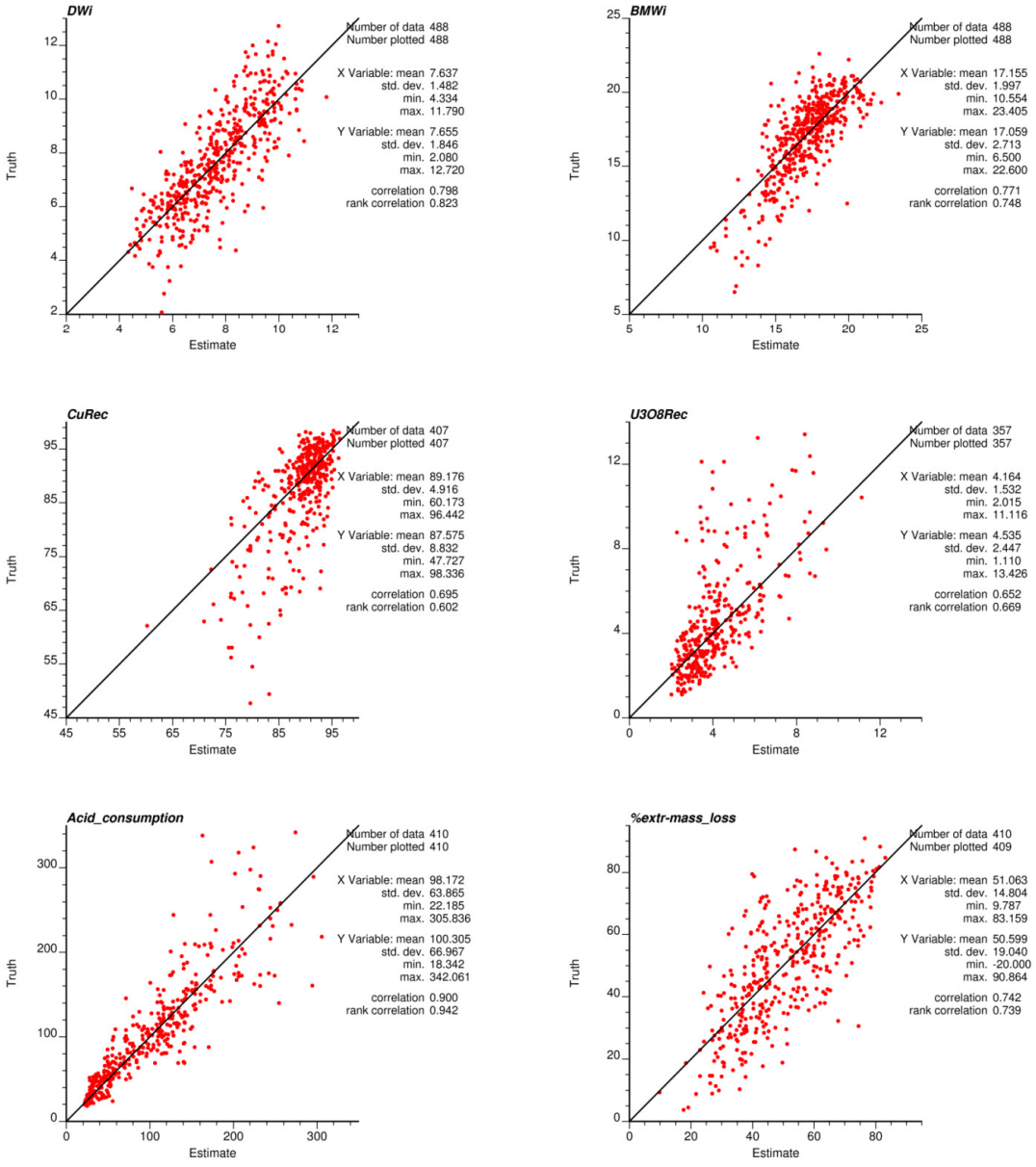
- Na is a significant contributor for DWi/BMWi – indicates different mineralogy
- SG is important for DWi but not BMWi – this is expected as it matters whether the rock is brittle or not, and this is related to the ratio of iron/silica content in the rock matrix
- BMWi is heavily influenced by the head assays (top 6 variables contributing to BMWi are from head assays)
- Individual mineralogy variables have little significance (Cu recovery is the exception).
- Presence of Chalcopyrite and acid (in)soluble gangue are critical to Cu recovery.
- Cu wt% has a large effect on U_3O_8 recovery but little effect on Cu recovery.
- Based on the tornado charts, associations are important for DWI, Cu recovery, acid consumption and net recovery. This is also seen in the comparison of the typical and full models (Figure 5) as the BMWi and U_3O_8 recovery predictions are not significantly altered by removing the association data.
- Recoveries are the most difficult variables to predict (lowest correlation on Figure 5). This is expected, as recovery is dependent on a large number of complex interactions.

Future Work

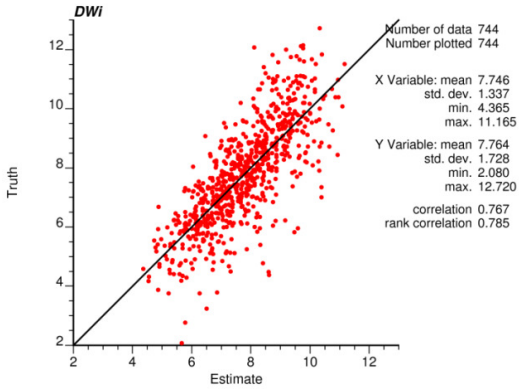
There are a number of opportunities for potential improvement on the modeling methodology presented in this paper: (1) Regression Model: Optimize the merging of the variables at the two different levels. The merging of the variables was done using logical groupings of the 112 variables. An optimization procedure could be developed to select ideal subsets of variables to increase the predictive power of the regression model. (2) Regression Model: There may be an ideal set of variables to use for each variable predicted. In this work, all 112 variables were used for all 6 output variables. Eliminating some of the less significant variables may reduce noise and increase model accuracy. This could be considered during the optimization of the merging of the variables (see point #1 above).

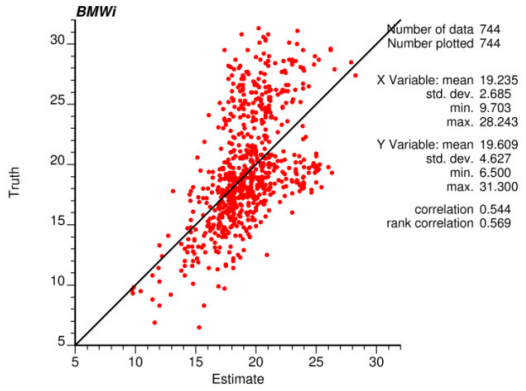
Figure 5: Cross plots of the truth/estimated values based on the full model (this page) and the typical model (next page) and the limited model (next page).

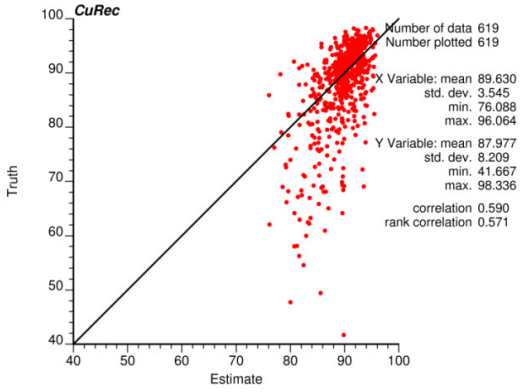
FULL MODEL

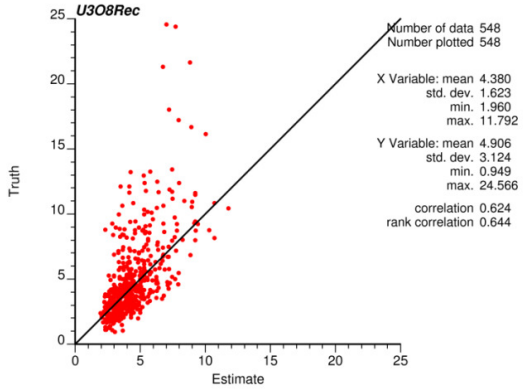


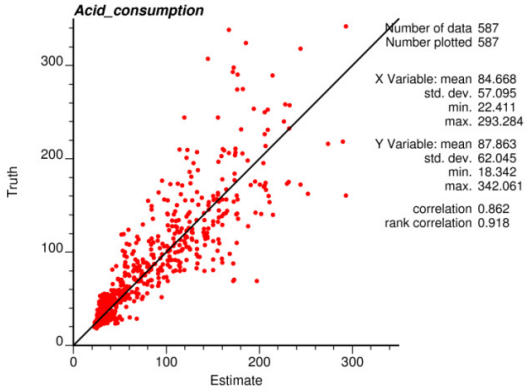
TYPICAL MODEL

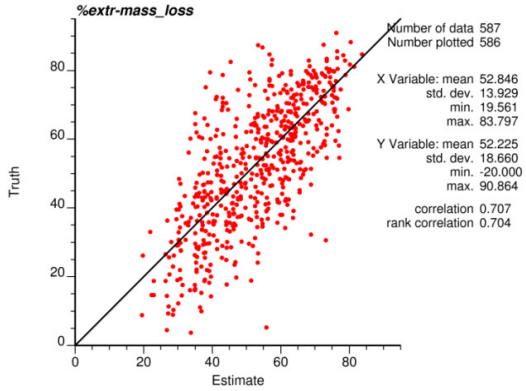




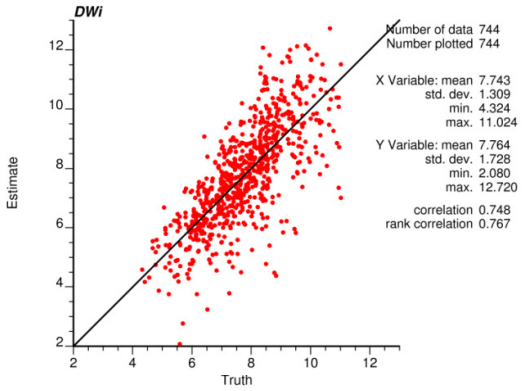


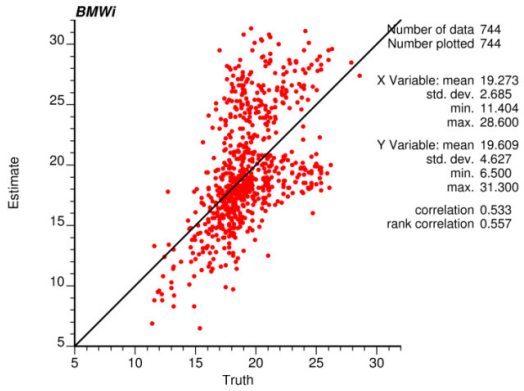


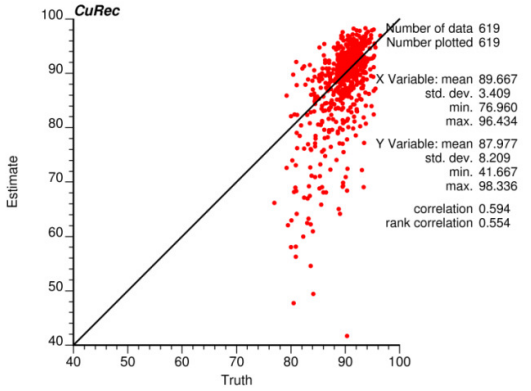


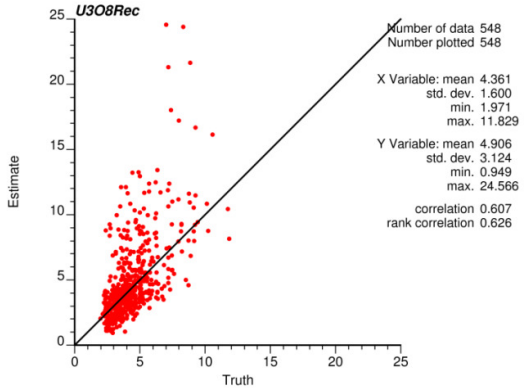


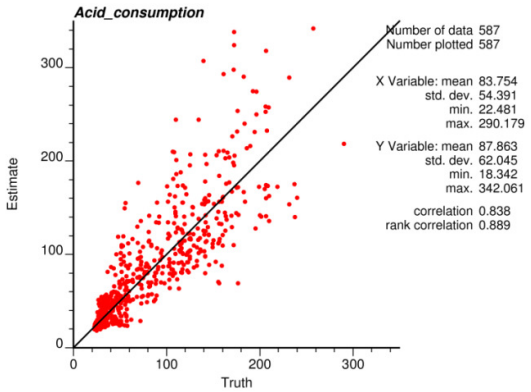
LIMITED MODEL

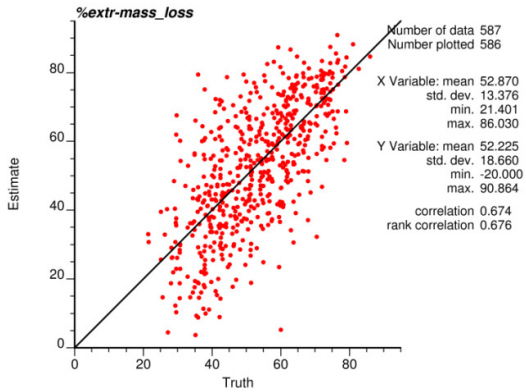












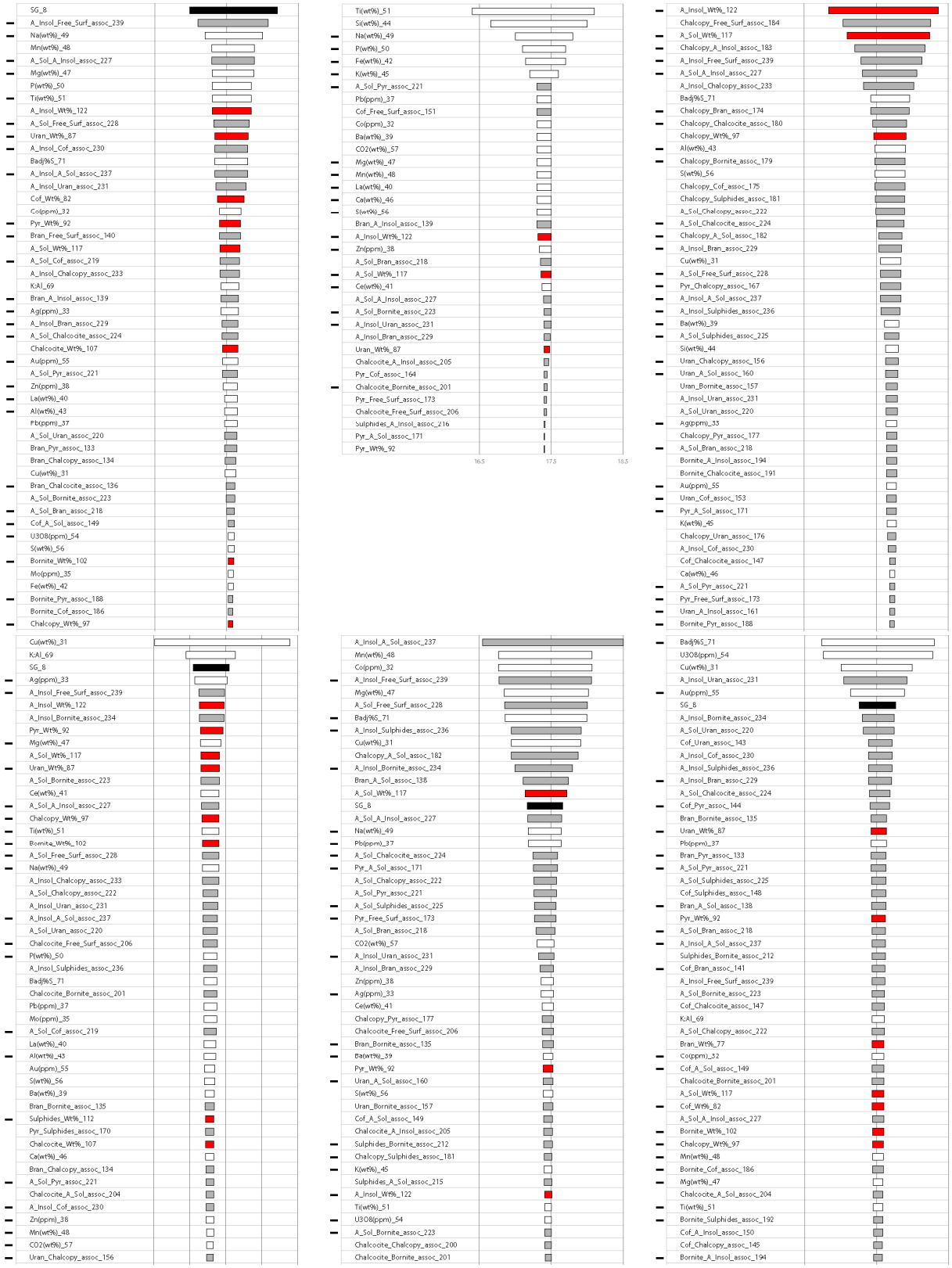


Figure 6: FULL model Tornado charts for DWi, BMWi and Cu recovery (this page) as well as U_3O_8 recovery, acid consumption and net recovery (bottom). White – Head Assays; Gray – Associations; Red – Mineralogy; Black – SG.