Constraining the Sum of Multivariate Estimates

Behrang Koushavand and Clayton V. Deutsch

Geostatisticians are increasingly being faced with compositional data arising from full geochemical sampling or some other source. Logratios are used at times to ensure the resulting estimates sum to unity; however, the logarithm transform can introduce a bias. We may want to cokrige the variables directly while ensuring that the estimates sum to unity. A linear constraint is added to the Cokriging system. The large cokriging system estimates all variables at the same time such that the sum of all variables to be a specified constant value. The methodology and practical concerns are discussed.

Introduction

Compositional data are represented as vector variables, with individual vector components ranging between zero and a positive maximum value representing a constant sum constraint, usually unity (or 100%). The earth sciences commonly encounter spatial distributions of compositional data, such as concentrations of constituents in natural waters (e.g., mole, mass, or volume fractions), mineral percentages, ore grades, or proportions of mutually exclusive categories (e.g., a water–oil–rock system).

Traditional geostatistical methods, such as kriging or co-kriging, will not respect the compositional nature of data if applied directly because these methods calculate estimates that are independently optimal. A constant sum constraint introduces negative correlations. The constant sum constraint also produces co-kriging systems of equations that are intractable for conventional linear equation solvers (Carle, 2004).

Most of the methods aimed at compositional data are based on data re-expression or transformation. The data must be transformed to new variables to eliminate the constraints. Then geostatistical models of new variables have been created, this model should be back transformed to original unit data, which most of these methods do not have a linear back transformation and introduce serious risk of bias.

The approach presented here is a linear estimation that minimizes estimation variance. In practice, the proposal is the conventional cokriging with an additional constraint.

Methodology

Consider the situation where the composite data $\left\{z_i\left(u_{\alpha_i}\right),\alpha_i=1,..,n_i,i=1,...,N_v\right\}$ at any, possibly different, locations. The linear estimator for each of variables is (Goovaerts, 1997):

$$Z_k^*(u)-m_k(u)=\sum_{i=1}^{N_v}\sum_{\alpha_i=1}^{n_i(u)}\lambda_{\alpha_i}^k(u)\left[Z_i\left(u_{\alpha_i}\right)-m_i\left(u_{\alpha_i}\right)\right] \qquad k=1,...,N_v \tag{1}$$

where $\lambda_{\alpha_i}^k(u)$ is the weight assign to the datum $Z_i\left(u_{\alpha_i}\right)$ to estimate $Z_k^*(u)$. The expected value of the random variable (RV) $Z_k(u)$ is denoted $m_k(u)$. All co-kriging estimator should be unbiased and minimize error variance $\sigma_k^2(u)$, that is (Goovaerts, 1997):

$$\begin{aligned}&\sigma_k^2(u)=\mathrm{Var}\left\{Z_k^*(u)-Z_k(u)\right\}\\&=\sum_{i=1}^{N_v}\sum_{j=1}^{N_v}\sum_{\alpha_i=1}^{n_i(u)}\sum_{\beta_j=1}^{n_j(u)}\lambda_{\alpha_i}^k(u).\lambda_{\alpha_j}^k(u).C_{ij}\left(u_{\alpha_i}-u_{\beta_i}\right) \qquad k=1,...,N_v\\&+C_{kk}(0)-2\sum_{i=1}^{N_v}\sum_{\alpha_i=1}^{n_i(u)}\lambda_{\alpha_i}^k(u).C_{ik}\left(u_{\alpha_i}-u\right)\end{aligned} \tag{2}$$

where C_{ij} is cross covariance function between variables i and j , C_k is covariance function of variable k. The estimation variance is minimized under the constant that the expected error is zero:

$$E\left\{Z_k^*(u)-Z_k(u)\right\}=0 \quad ; \quad k=1,\ldots,N_v \tag{3}$$

It has been shown that simple co-kriging (SCK) is unbiased since stationary mean of each RVs is constant over the entire domain:

$$m_k\left(u_{\alpha_i}\right)=m_k,\forall\alpha_i=1..n_i \quad ; \quad k=1,\ldots,N_v \tag{4}$$

Simple Co-kriging system is written as:

$$\sum_{j=1}^{N_v}\sum_{\alpha_i=1}^{n_i(u)}\sum_{\beta_j=1}^{n_j(u)}\lambda_{\alpha_i}^k(u).\lambda_{\beta_j}^k(u).C_{ij}\left(u_{\alpha_i}-u_{\beta_i}\right)=\sum_{\alpha_i=1}^{n_i(u)}C_{ik}\left(u_{\alpha_i}-u\right) \qquad i=1,\ldots,N_v\ ,\ k=1,\ldots,N_v$$

The matrix notation for full Co-kriging system is shown below:

$$A_{sck}.\Gamma_{sck}=b_{sck} \tag{5}$$

where,

$$A_{sck}=\begin{bmatrix}\left[C_{11}\left(u_{\alpha_1}-u_{\beta_1}\right)\right] & \cdots & \left[C_{1N_v}\left(u_{\alpha_1}-u_{\beta_{N_v}}\right)\right]\\ \vdots & \ddots & \vdots\\ \left[C_{N_v1}\left(u_{\alpha_{N_v}}-u_{\beta_1}\right)\right] & \cdots & \left[C_{N_vN_v}\left(u_{\alpha_{N_v}}-u_{\beta_{N_v}}\right)\right]\end{bmatrix}$$

$$\Gamma_{sck}=\begin{bmatrix}\lambda_{\beta_1}^1(u) & \cdots & \lambda_{\beta_1}^{N_v}(u)\\ \vdots & \ddots & \vdots\\ \lambda_{\beta_{N_v}}^1(u) & \cdots & \lambda_{\beta_{N_v}}^{N_v}(u)\end{bmatrix}$$

$$b_{sck}=\begin{bmatrix}\left[C_{11}\left(u_{\alpha_1}-u\right)\right] & \cdots & \left[C_{1N_v}\left(u_{\alpha_1}-u\right)\right]\\ \vdots & \ddots & \vdots\\ \left[C_{N_v1}\left(u_{\alpha_{N_v}}-u\right)\right] & \cdots & \left[C_{N_vN_v}\left(u_{\alpha_{N_v}}-u\right)\right]\end{bmatrix}$$

There are N_v linear systems that should be solved separately. There is no guarantee for such linear systems that sum of estimated variables to be a constant value which is very important in compositional data. A compositional Co-kriging (CCK) system has been defined here, solves a very big linear system such that sum of all estimated variables have to be a constant value. The compositional simple co-kriging (CSCK) system is defined as below:

$$A_{CSCK}.\Gamma_{CSCK}=b_{CSCK} \tag{6}$$

$$A_{CSCK}=\begin{bmatrix}\begin{bmatrix}\left[C_{11}\left(u_{\alpha_1}-u_{\beta_1}\right)\right] & \cdots & \left[C_{1N_v}\left(u_{\alpha_1}-u_{\beta_{N_v}}\right)\right]\\ \vdots & \ddots & \vdots\\ \left[C_{N_v1}\left(u_{\alpha_{N_v}}-u_{\beta_1}\right)\right] & \cdots & \left[C_{N_vN_v}\left(u_{\alpha_{N_v}}-u_{\beta_{N_v}}\right)\right]\end{bmatrix} & [0] & [0] & \begin{bmatrix}\left[Z_1(u_1) & \cdots & Z_1\left(u_{N_1}\right)\right]\\ & \vdots & \\ \left[Z_{N_v}(u_1) & \cdots & Z_{N_v}\left(u_{N_v}\right)\right]\end{bmatrix}^T\\ [0] & \ddots & [0] & \vdots\\ [0] & [0] & \begin{bmatrix}\left[C_{11}\left(u_{\alpha_1}-u_{\beta_1}\right)\right] & \cdots & \left[C_{1N_v}\left(u_{\alpha_1}-u_{\beta_{N_v}}\right)\right]\\ \vdots & \ddots & \vdots\\ \left[C_{N_v1}\left(u_{\alpha_{N_v}}-u_{\beta_1}\right)\right] & \cdots & \left[C_{N_vN_v}\left(u_{\alpha_{N_v}}-u_{\beta_{N_v}}\right)\right]\end{bmatrix} & \begin{bmatrix}\left[Z_1(u_1) & \cdots & Z_1\left(u_{N_1}\right)\right]\\ & \vdots & \\ \left[Z_{N_v}(u_1) & \cdots & Z_{N_v}\left(u_{N_v}\right)\right]\end{bmatrix}^T\\ \begin{bmatrix}\left[Z_1(u_1) & \cdots & Z_1\left(u_{N_1}\right)\right]\\ & \vdots & \\ \left[Z_{N_v}(u_1) & \cdots & Z_{N_v}\left(u_{N_v}\right)\right]\end{bmatrix}^T & \cdots & \begin{bmatrix}\left[Z_1(u_1) & \cdots & Z_1\left(u_{N_1}\right)\right]\\ & \vdots & \\ \left[Z_{N_v}(u_1) & \cdots & Z_{N_v}\left(u_{N_v}\right)\right]\end{bmatrix}^T & [0]\end{bmatrix}$$

$$\Gamma_{CSCK} = \begin{bmatrix} \begin{bmatrix} \lambda^{1}_{\beta_1}(u) \\ \vdots \\ \lambda^{1}_{\beta_{N_v}}(u) \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \lambda^{N_v}_{\beta_1}(u) \\ \vdots \\ \lambda^{N_v}_{\beta_{N_v}}(u) \end{bmatrix} \\ 0 \end{bmatrix} \qquad b_{CSCK} = \begin{bmatrix} \begin{bmatrix} \left[C_{11}\left(u_{\alpha_1} - u\right)\right] \\ \vdots \\ \left[C_{N_v 1}\left(u_{\alpha_{N_v}} - u\right)\right] \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \left[C_{1N_v}\left(u_{\alpha_1} - u\right)\right] \\ \vdots \\ \left[C_{N_v N_v}\left(u_{\alpha_{N_v}} - u\right)\right] \end{bmatrix} \\ \text{Const} \end{bmatrix}$$

The estimation variance for each of variables is larger than SCK estimator. Ordinary and traditional ordinary co-kriging systems also introduce linear constraints for each variable, which should consider at linear system (6). **`Cokt3D_2`** (Deutsch and Journel 1997) program has been modified to solve a full co-kriging which estimated values have a constraint sum to a constant.

Discussion

The constrained cokriging system ensures that the estimates sum to unity, but there is no guarantee to get non-negative estimates. To solve this problem, we need to minimize the estimation variance with other algorithms. This leads to excessive CPU requirements. The second and more reasonable method is to replace negative values with 0 and remove them form system of equations and resolve for the optimal estimates. This method also forces algorithm to solve cokriging two times for each negative estimate, but it is more efficient and faster then solving kriging system with another algorithm. This algorithm is referred to as constrained positive Cokriging.

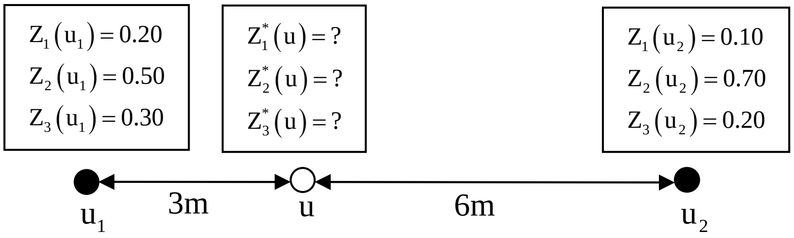
The second issue is bias in the constrained co-kriging system. All kriging estimates are unbiased because the expected value of difference between estimate and true value is 0. This means that these estimates dose not have any systematic error.

$$E\left\{Z^{*}_{k}(u) - Z_{k}(u)\right\} = 0 \text{ So,} \qquad E\left\{\sum_{i=1}^{N} \lambda_i(u) Z(u_i) - Z(u)\right\} = 0 \qquad (7)$$

However, in constrained co-kriging, the system of equations is solved such that estimates are forced to sum to unity. These means that kriging weight $\lambda_i(u)$ are not independent from variables $Z(u)$ and may cause biased estimate value. This issue needs more study.

Small Example

Consider two data with three variables:



where the variogram of this data set is:

$$\gamma_{Z_1}(h)=0.5.\mathrm{Sph}\left(\frac{h}{5}\right)+0.5.\mathrm{Sph}\left(\frac{h}{20}\right)$$

$$\gamma_{Z_2}(h)=0.2.\mathrm{Sph}\left(\frac{h}{5}\right)+0.8.\mathrm{Sph}\left(\frac{h}{20}\right)$$

$$\gamma_{Z_3}(h)=0.8.\mathrm{Sph}\left(\frac{h}{5}\right)+0.2.\mathrm{Sph}\left(\frac{h}{20}\right)$$

$$\gamma_{Z_1Z_2}(h)=-0.2.\mathrm{Sph}\left(\frac{h}{5}\right)-0.3.\mathrm{Sph}\left(\frac{h}{20}\right)$$

$$\gamma_{Z_1Z_3}(h)=-0.3.\mathrm{Sph}\left(\frac{h}{5}\right)-0.1.\mathrm{Sph}\left(\frac{h}{20}\right)$$

$$\gamma_{Z_2Z3}(h)=0.3.\mathrm{Sph}\left(\frac{h}{5}\right)+0.3.\mathrm{Sph}\left(\frac{h}{20}\right)$$

Tables below show the Left hand side matrix A_{CSCK} and right hand side vector b_{CSCK}.

1	-0.5	-0.4	0.19	-0.11	-0.04	0	0	0	0	0	0	0	0	0	0	0	0	0.20
-0.5	1	0.6	-0.11	0.30	0.11	0	0	0	0	0	0	0	0	0	0	0	0	0.50
-0.4	0.6	1	-0.04	0.11	0.07	0	0	0	0	0	0	0	0	0	0	0	0	0.30
0.19	-0.11	-0.04	1	-0.5	-0.4	0	0	0	0	0	0	0	0	0	0	0	0	0.10
-0.11	0.30	0.11	-0.5	1	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0.70
-0.04	0.11	0.07	-0.4	0.6	1	0	0	0	0	0	0	0	0	0	0	0	0	0.20
0	0	0	0	0	0	1	-0.5	-0.4	0.19	-0.11	-0.04	0	0	0	0	0	0	0.20
0	0	0	0	0	0	-0.5	1	0.6	-0.11	0.30	0.11	0	0	0	0	0	0	0.50
0	0	0	0	0	0	-0.4	0.6	1	-0.04	0.11	0.07	0	0	0	0	0	0	0.30
0	0	0	0	0	0	0.19	-0.11	-0.04	1	-0.5	-0.4	0	0	0	0	0	0	0.10
0	0	0	0	0	0	-0.11	0.30	0.11	-0.5	1	0.6	0	0	0	0	0	0	0.70
0	0	0	0	0	0	-0.04	0.11	0.07	-0.4	0.6	1	0	0	0	0	0	0	0.20
0	0	0	0	0	0	0	0	0	0	0	0	1	-0.5	-0.4	0.19	-0.11	-0.04	0.20
0	0	0	0	0	0	0	0	0	0	0	0	-0.5	1	0.6	-0.11	0.30	0.11	0.50
0	0	0	0	0	0	0	0	0	0	0	0	-0.4	0.6	1	-0.04	0.11	0.07	0.30
0	0	0	0	0	0	0	0	0	0	0	0	0.19	-0.11	-0.04	1	-0.5	-0.4	0.10
0	0	0	0	0	0	0	0	0	0	0	0	-0.11	0.30	0.11	-0.5	1	0.6	0.70
0	0	0	0	0	0	0	0	0	0	0	0	-0.04	0.11	0.07	-0.4	0.6	1	0.20
0.20	0.50	0.30	0.10	0.70	0.20	0.20	0.50	0.30	0.10	0.70	0.20	0.20	0.50	0.30	0.10	0.70	0.20	0

0.49
-0.27
-0.14
0.28
-0.17
-0.06
-0.27
0.66
0.30
-0.17
0.45
0.17
-0.14
0.30
0.32
-0.06
0.17
0.11
1.00

The table below shows the solution of linear system of CSCK, estimate and estimation variance values.

Z	W1	W2	W3	Z*W1	Z*W2	Z*W3	
0.20	0.47988	0.06788	0.05942	0.095976	0.013576	0.011884	
0.50	-0.05182	0.69148	0.16156	-0.02591	0.34574	0.08078	
0.30	0.10286	-0.11011	0.24838	0.030858	-0.033033	0.074514	
0.10	0.21997	0.05108	0.04785	0.021997	0.005108	0.004785	
0.70	0.00141	0.38811	0.15109	0.000987	0.271677	0.105763	
0.20	0.05748	-0.09881	0.01779	0.011496	-0.019762	0.003558	
				$Z_1^*=0.135404$	$Z_2^*=0.583306$	$Z_3^*=0.281284$	$\sum_{i=1}^{3} Z_i^*=0.999994$
				$\sigma^2_{\mathrm{csck}}=0.705$	$\sigma^2_{\mathrm{csck}}=0.443$	$\sigma^2_{\mathrm{csck}}=0.586$	

Large Data Set

There is three variable in this data set that should be sum to 1 at each sample or estimated location. Figure 1 shows the location map and figure 2 shows histograms and Scatter plot of variables. **Cokt3D** and **Cokt3D_2** with and without forced to positive estimate were used to an ordinary full co-kriging estimation to compare the results. All three methods were used with the same LMC model used in small data set above. Pixel plot of estimate and estimation variance values are presented at figure 3 and 4 respectively. The blank blocks on right column maps, which appear on the second variable, are zero estimates replaced with negative estimate values on the middle column. Histograms of estimated values and scatter plot of estimate variance values are shown at figure 5, 6 and 7 respectively. There is no such a big difference at estimate variance values between three methods, but it is clear that adding sum to 1 constraint and forcing to get positive estimates increase estimation variance.

There are 1316 negative estimates with Cokriging and 235 estimates with constrained Cokriging and the sum of all negative values of second variable for first and second method respectively are -83.7 and -3.2. Adding sum to 1 constraint to Cokriging system decrease number of negative estimates for second variable but we still have negative estimates, which are replaced with zero by constrained positive Cokriging;

`Sgsim` (Deutsch and Journel 1997) was used to generate 100 realizations of unconditional simulation for each of variables in Gaussian space and then all realizations were back transferred to original units and standardized to get sum of variables in every location to 1.

The same sample locations were extracted from each realization. Cokriging and constrained positive Cokriging were used to estimate un-sampled locations that we already have the true values. Figure 8 shows the histograms of Error (Estimate-True) with these two methods. As we expected there is a very small biased on estimation of V1 and V2, but for this case study constrained positive Cokriging dose not inject significant biased to the estimation.

Conclusion

Geostatisticans commonly encounter compositional data. Non linear transformation of these variables with techniques like logratios is problematic because of significant bias. Kriging in original units provides a best linear unbiased estimator. Traditional cokriging does not constrain the sum of estimates. The approach presented here leads to a large cokriging system that has a constraint to force estimated values to sum to a constant. Negative estimated variables was set to 0 and removed form system of equations and a new system of equations is solved. This method requires more CPU time than traditional Cokriging because algorithm solves a big linear system rather than N (number of variables) smaller linear systems in Cokriging.

References

Carle, S. F., 2004, Geostatistical Analysis of Compositional Data : Oxford University Press, New York

Deutsch, C. V., and Journel, A. G., 1998, GSLIB: Geostatistical software library and users guide, 2nd ed.: Oxford University Press, New York, 369 p.

Goovaerts, P., 1997, Geostatistics for Natural Resources Evaluation: Oxford University Press, New York

Woronow, A. 1995, A Pseudo-Genetic Algorithm Suited to Compositional Data: Math Geology, v. 27, no. 2, p. 625–645

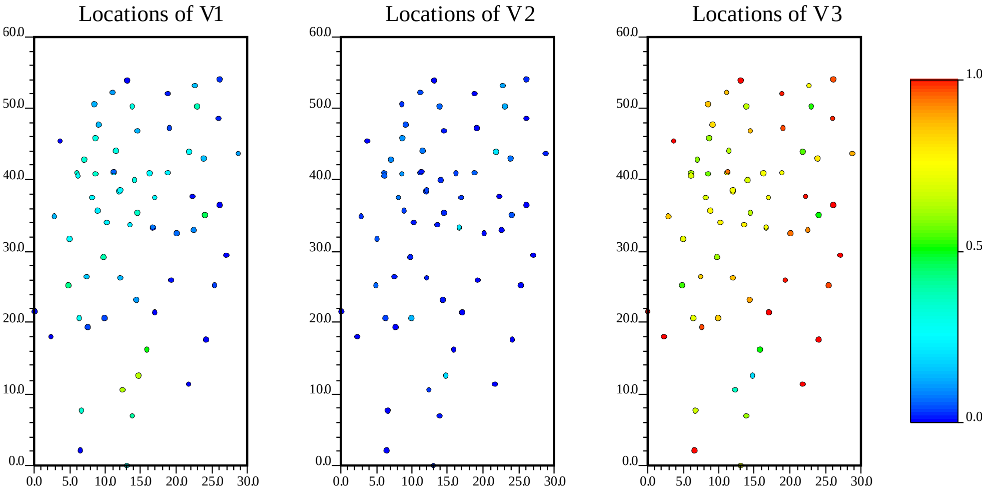


Figure 1. Location map of samples and three variables content

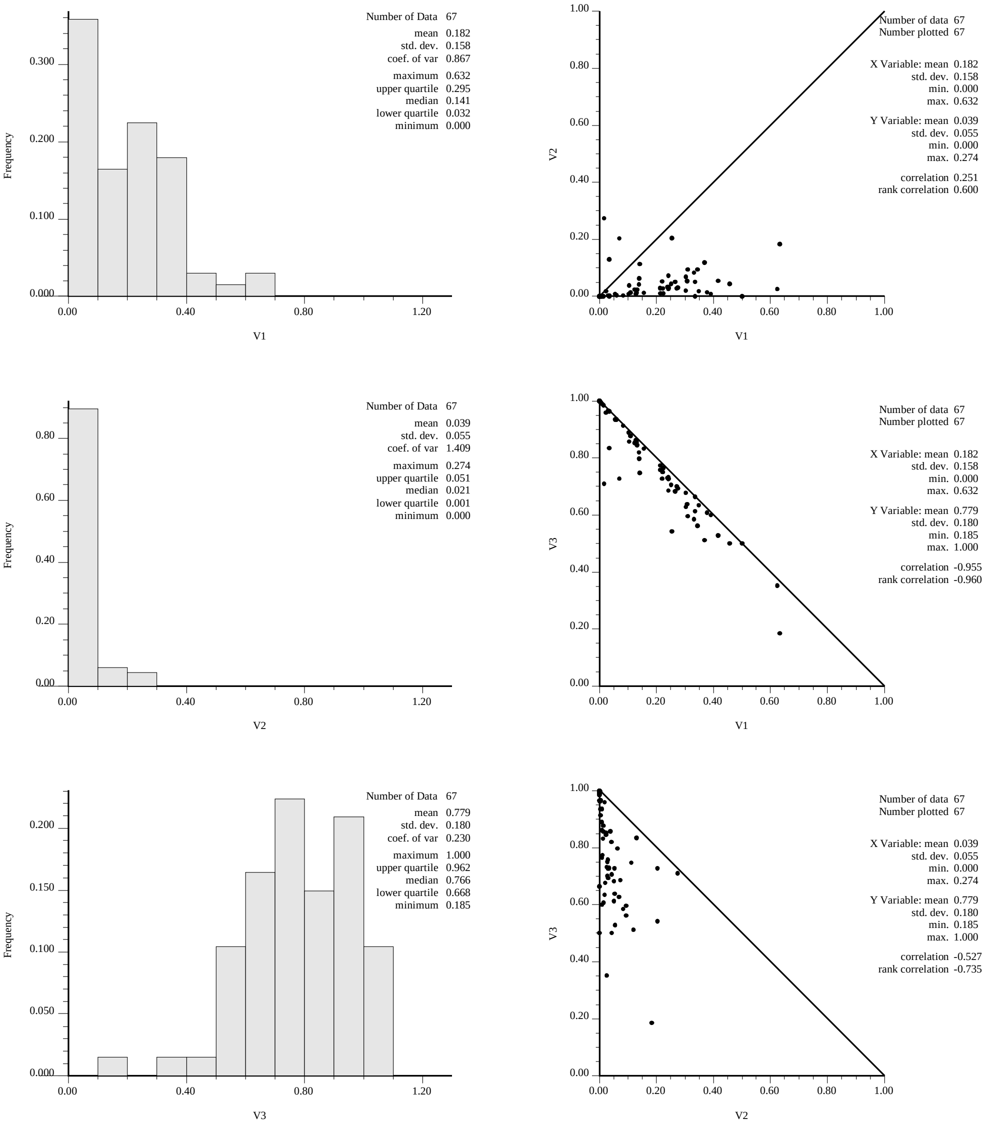


Figure 2. Histograms and Scatter plot of variables

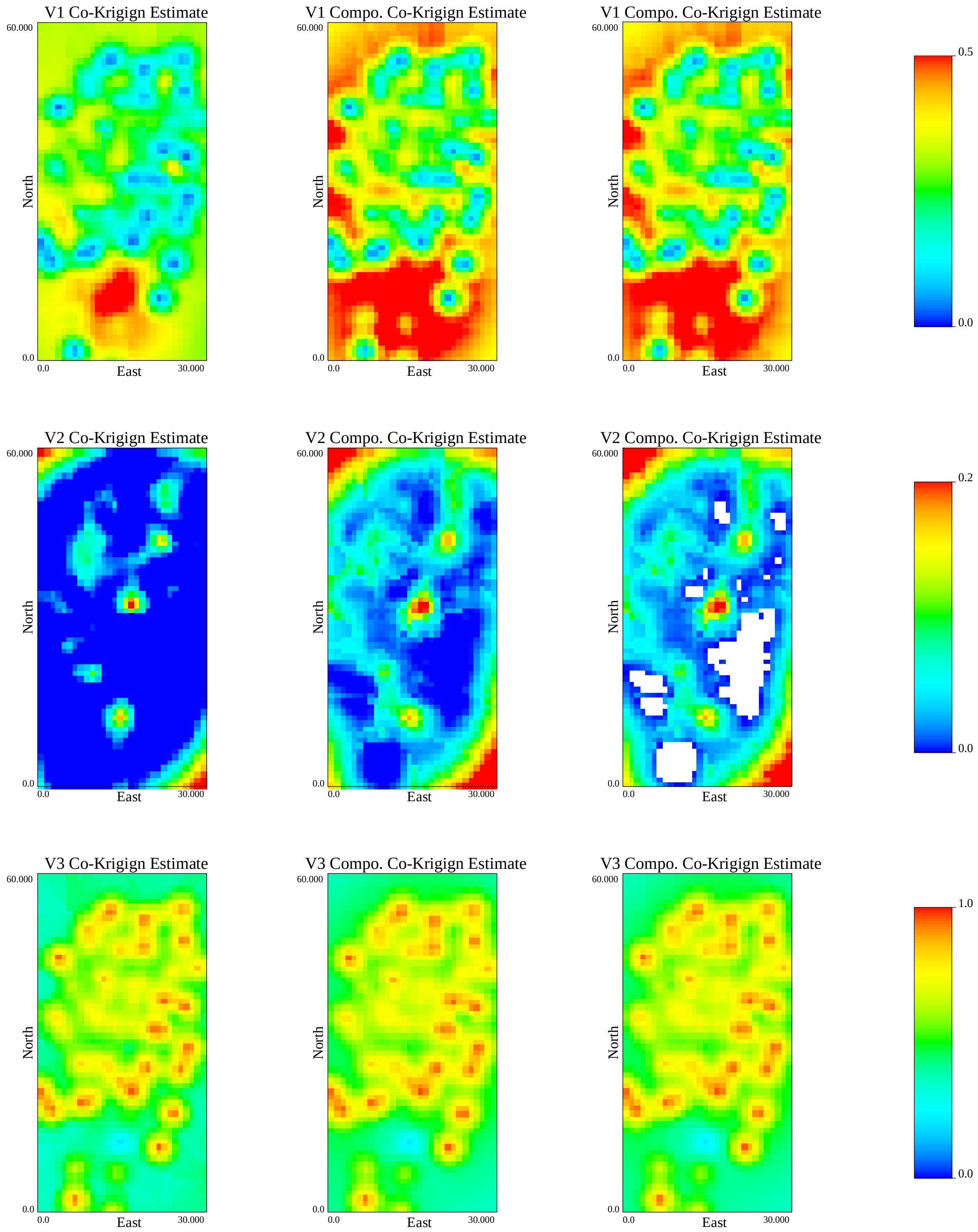

V1 Co-Krigign Estimate
V1 Compo. Co-Krigign Estimate
V1 Compo. Co-Krigign Estimate
0.5
0.0
V2 Co-Krigign Estimate
V2 Compo. Co-Krigign Estimate
V2 Compo. Co-Krigign Estimate
0.2
0.0
V3 Co-Krigign Estimate
V3 Compo. Co-Krigign Estimate
V3 Compo. Co-Krigign Estimate
1.0
0.0
60.000
North
0.0
0.0
East
30.000

Figure 3. Co kriging Estimate map: without constraints(left) , constraint sum to 1 (middle) and constraint sum to 1 which forced to positive values (right)

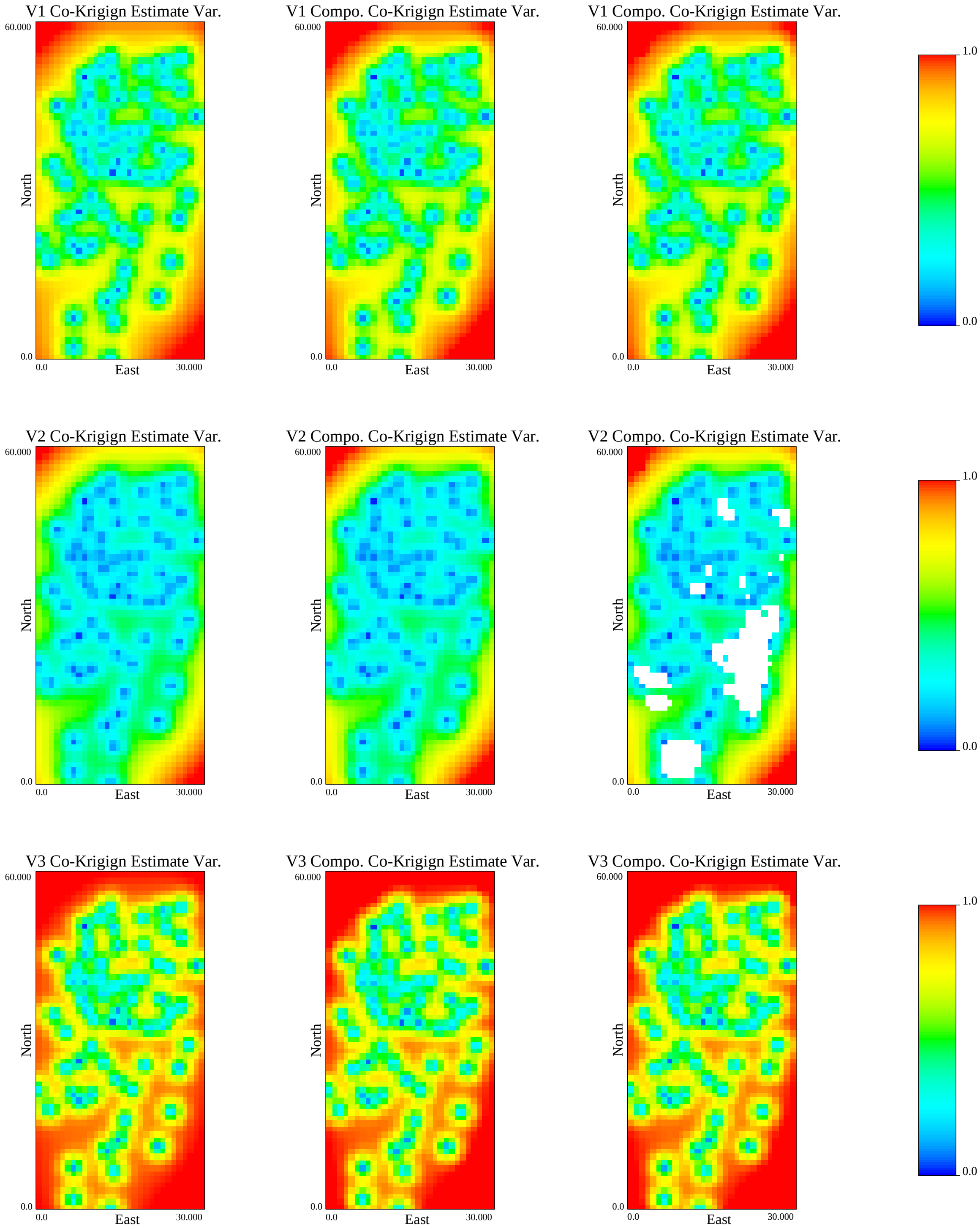

V1 Co-Krigign Estimate Var.
V1 Compo. Co-Krigign Estimate Var.
V1 Compo. Co-Krigign Estimate Var.
V2 Co-Krigign Estimate Var.
V2 Compo. Co-Krigign Estimate Var.
V2 Compo. Co-Krigign Estimate Var.
V3 Co-Krigign Estimate Var.
V3 Compo. Co-Krigign Estimate Var.
V3 Compo. Co-Krigign Estimate Var.
60.000
0.0
North
East
30.000
1.0
0.0

Figure 4. Co kriging Estimation variance map: without constraints(left) , constraint sum to 1 (middle) and constraint sum to 1 which forced to positive values (right)

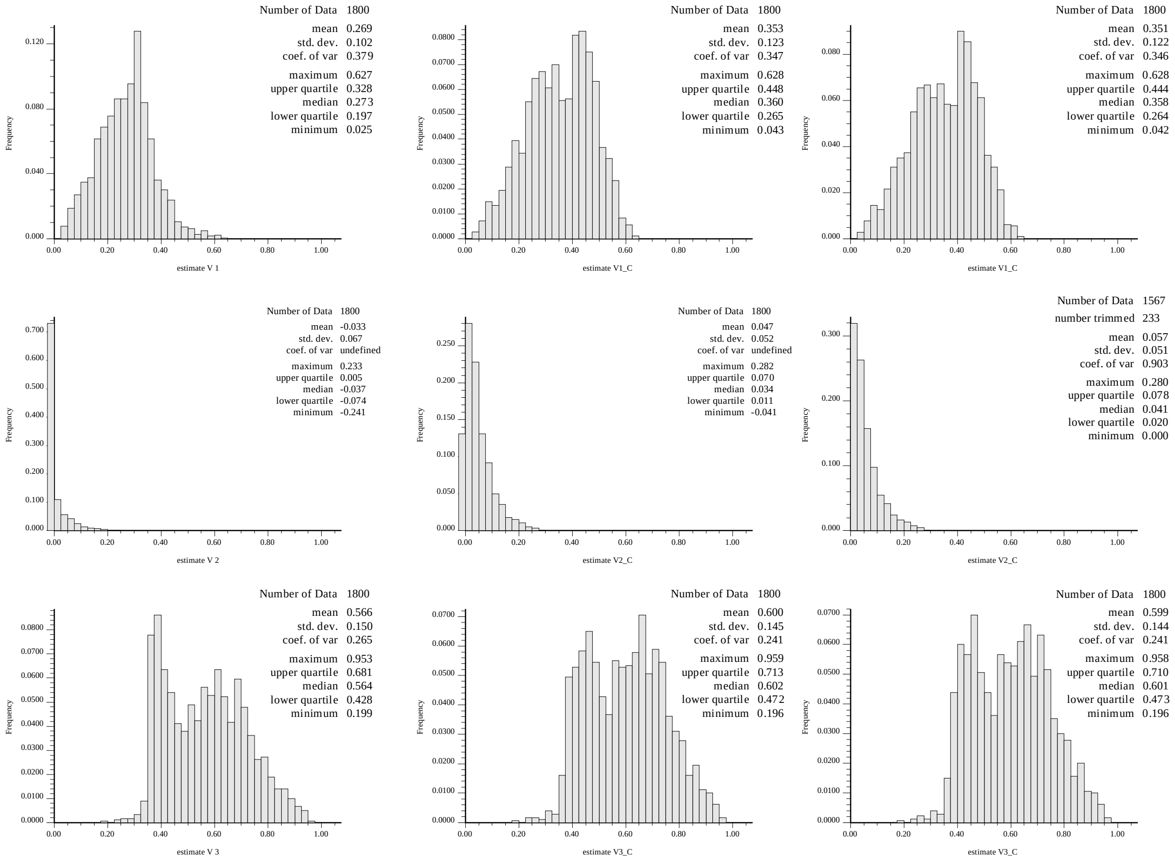


Figure 5. histogram of Co kriging Estimate: without constraints(left), constraint sum to 1 (middle) and constraint sum to 1 which forced to positive values (right)

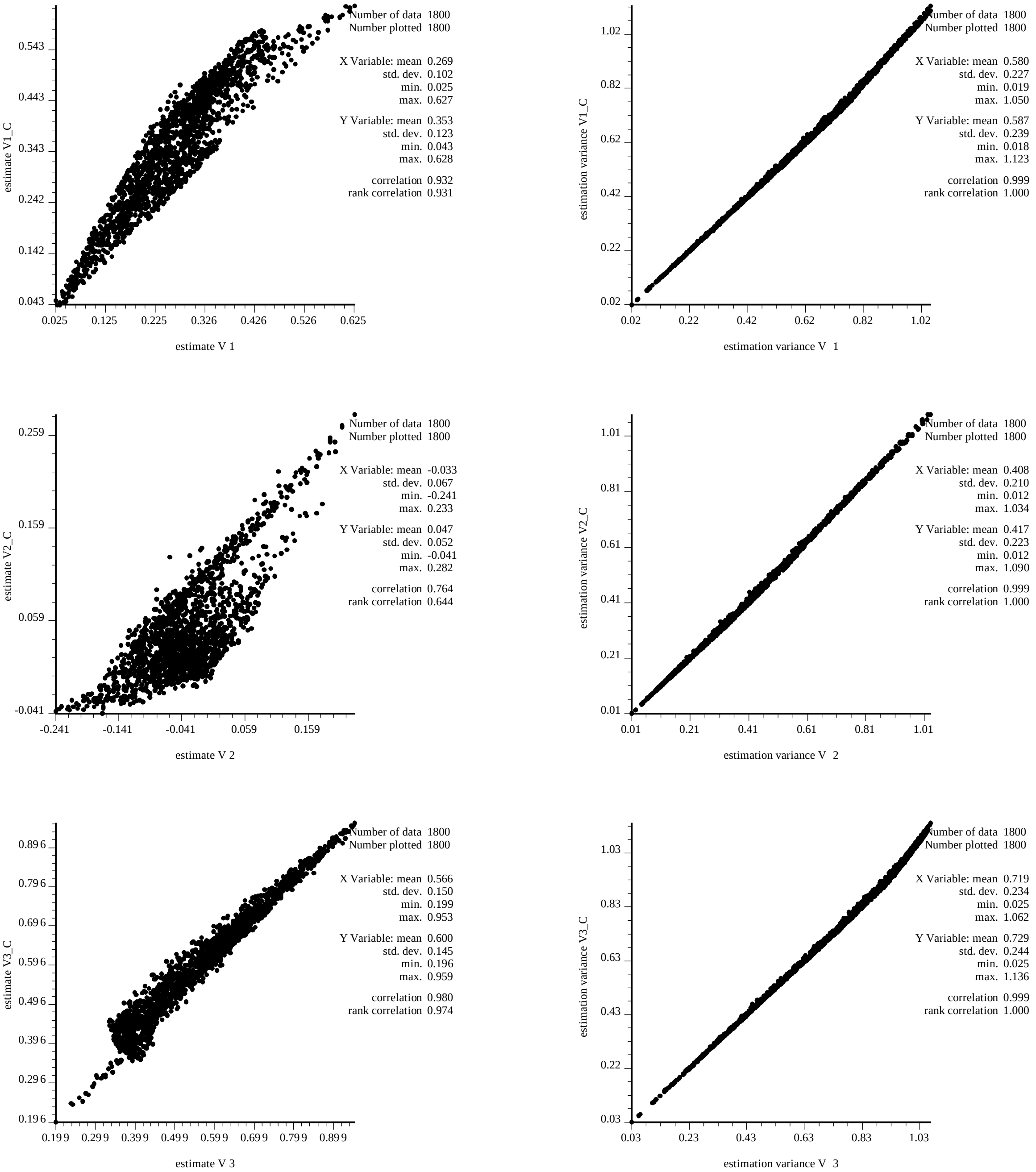


Figure 6. Scatterplots of estimated values and estimation variance values: Cokriging vs. Constrained Cokriging

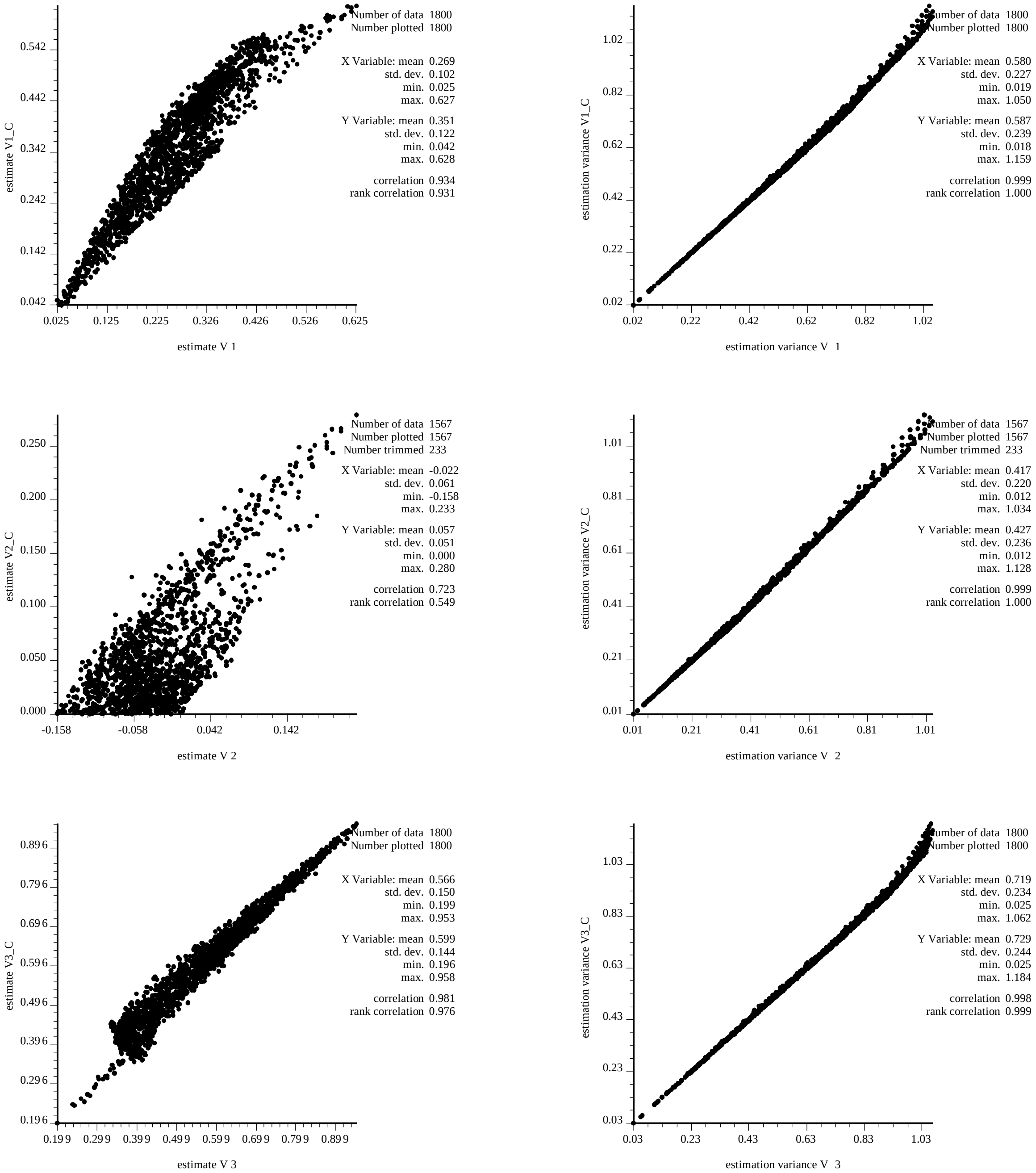


Figure 7. Scatterplots of estimated values and estimation variance values: Cokriging vs. Constrained Positive Cokriging

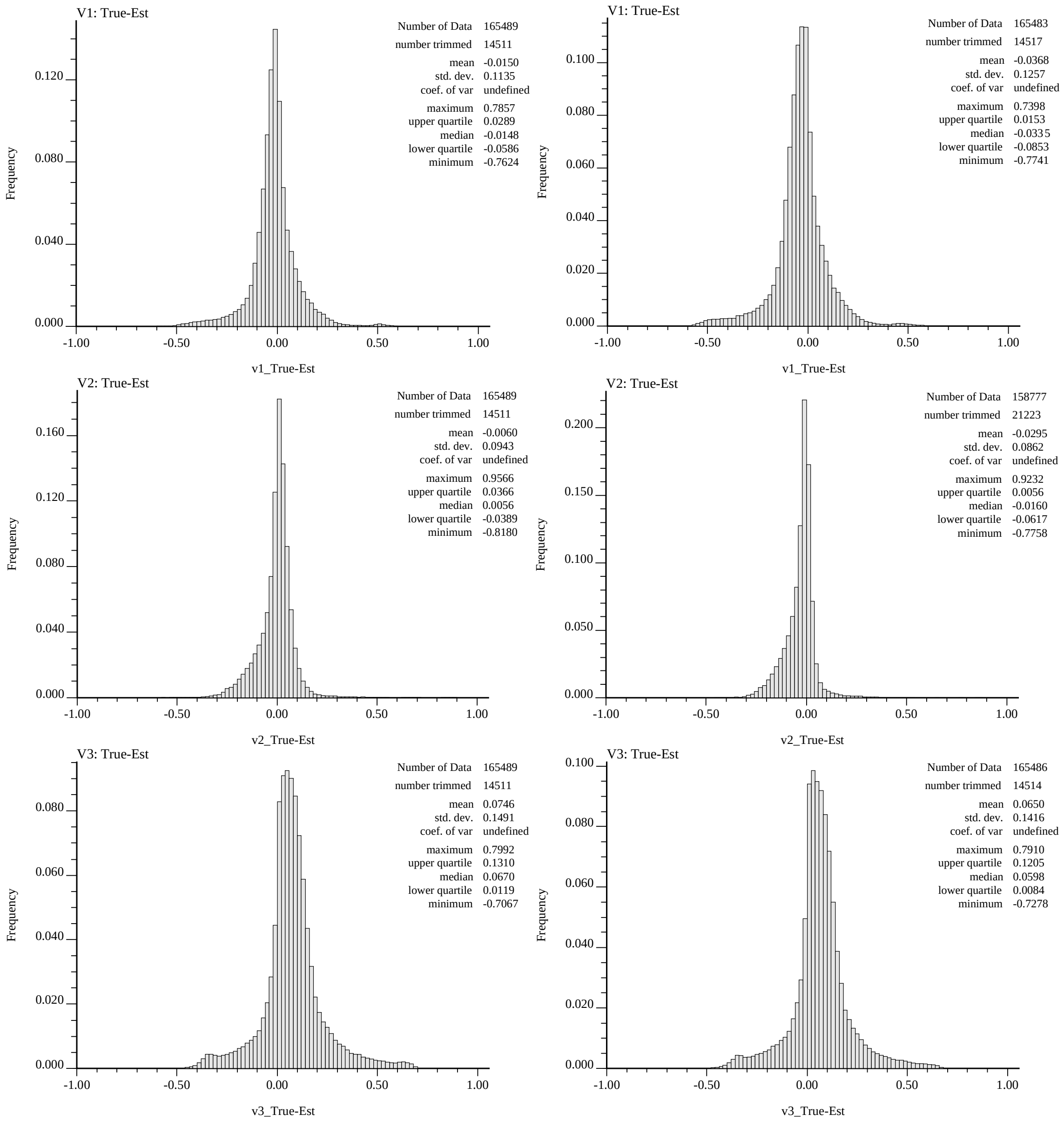


Figure 8. Histograms of Error: Cokriging Estimate – True Value ; Left: Cokriging (left) and Right: Constrained positive Cokriging