On the Selection of Secondary Variables for Cokriging and Cosimulation

Miguel Cuba, Olena Babak and Oy Leuangthong

Due to the complexity of fitting several direct and cross-semivariograms for cosimulation, usually only those secondary variables that are highly correlated to the primary variable are considered for geostatistical modeling. Additionally many geomodelers believe that poorly correlated variables will not provide relevant information to improve estimates. This paper shows that even poorly correlated secondary variables can contribute significantly in the prediction of the primary variable. One synthetic example is presented in this document; this consists of one primary variable and two secondary variables, with one highly correlated and the other poorly correlated to the primary variable. The estimation of the primary variable is carried out by simple kriging, simple co-kriging using both secondary variables, and each secondary variable independently. The contribution of the secondary variables are presented considering: 1) correlation coefficient of cross-validation, 2) mean squared error, 3) mean absolute error, and 4) the profile of the simple co-kriging weights of the poorly correlated variable. Our results show that accounting for the poorly correlated secondary variable does improve inference of the primary variable.

Introduction

One of the most important problems in the geosciences is the problem of spatial prediction. Spatial predictions are often required for planning, risk assessment, and decision-making. Typical applications include determining the profitability of mining an orebody, producing a reservoir, management of soil resources, soil properties mapping, pest management, designing a network of environmental monitoring stations, etc. (Weisz et al, 1995; Gotway et al., 1996; Moyeed and Papritz, 2002).

Kriging (and its derivatives) is a well-known and established methodology for spatial prediction. Kriging uses the spatial correlations provided by the variogram to calculate the weights that are applied to the sample values surrounding an unsampled location. The weights obtained from the kriging minimize the estimation variance and account for the spatial correlation between the surrounding samples and the estimation location (that is, closeness to the estimation location) and between samples themselves (that is, data redundancy). Kriging is a statistically optimal interpolator in the sense that it provides the best linear unbiased estimate. In the case of multivariate data, cokriging (CK) is commonly applied for estimation (Vauclin et al, 1983; Wackernagel 1994; Goovaerts, 1997; Wackernagel, 2003). Cokriging allows estimating the variable of interest with data of the same type and auxiliary variables in the neighborhood:

$$Y_1^*(\mathbf{u}) = \sum_{\alpha_1=1}^{n_1(\mathbf{u})} \lambda_{\alpha_1}(\mathbf{u})\left[Y_1\left(\mathbf{u}_{\alpha_1}\right)\right] + \sum_{i=2}^{N_v}\sum_{\alpha_i=1}^{n_i(\mathbf{u})} \lambda_{\alpha_i}(\mathbf{u})\left[Y_i\left(\mathbf{u}_{\alpha_i}\right)\right]$$

As majority of resource characterization problems involve multiple variables, including multiple metals and/or minerals, petrophysical attributes such as porosity, permeability, water saturation, etc, the implementation of cokriging is very time consuming, as well as there might be potential problems of current software limitations to the number of variable being able to considered and in some cases invertability of covariance matrices can be questionable.

The aim of this paper is to investigate whether all of the secondary attributes contribute positively to the estimation and if there can be set a cut-off on the correlation coefficient between primary and multiple secondary data below which the advantage of using a particular secondary data is negligible. The paper is organized as follows. The motivation for using a correlation coefficient as a measure for usefulness of inclusion of particular variable into cokriging estimation equation is explained first by relating regression and (co-)kriging. Then a case study with one primary variable and two secondary variables related with different correlation coefficients to the primary is conducted. The advantage/disadvantage of using poorly correlated variables in estimation is measured in cross validation using a mean squared error and a correlation coefficient between truth and estimate.

Kriging and Regression

Kriging is mathematically very closely related to regression analysis. Regression analysis is a set of methods and techniques for modeling and analyzing several variables with the aim of establishing a the relationship between a dependent (in other words, primary) variable and one or more independent (i.e., secondary) (Draper and Smith, 1998; Glantz and Slinker, 1990). As kriging, regression analysis derives a best linear unbiased estimator, makes assumptions on a covariance model, and is formulated using a very similar formulae (linear regression equation is given below):

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i ,$$

where, for the *i*th observation, Y_i is the response variable, $X_{i,j},\ j = 1,\ldots,p$ are *p* regressors, and ε_i is an zero mean error. The coefficients β's are unknown and are calculated based on least squares method.

To measure the 'goodness of global fit' of the regression model to the data, a coefficient of determination R^2 is used. This coefficient is often interpreted as the proportion of variability in the response variable that is explained by the model. In particular, a coefficient of determination of 1 indicates that the fitted model explains all variability in the variable we are predicting, while indicates that there is no 'linear' relationship between the response variable and regressors.

In the case of a linear regression with only 1 independent (secondary) variable, there exists an interesting connection in-between coefficient of determination R^2 and a coefficient of correlation between response (primary) and independent (secondary) variables. Specifically, R^2 is equal to the square of the correlation coefficient (Nagelkerke, 1991); implying the higher the correlation between primary and secondary variables, the better the fitted model. In particular, if the correlation is above 0.7 (in absolute value), the contribution of the secondary variable is very significant, since more than 50% of variability in primary variable will be explained by a regression model. Due to similarity in-between regression and kriging, it is believed that there exist a similar relationship for correlation coefficient of spatial primary and secondary variables and goodness of estimation (tested in cross validation). Exactly the existance and pattern of this relationship will examined next in a case study.

Case Study

In order to evaluate the impact of secondary variables, the following reference model is considered. One map for the primary variable P1 and two sets of nine maps for secondary variable S1 and S2 respectively, are simulated at a resolution of 100 by 100 pixels, with a relative dimension of each pixel of 100 by 100 uod^2 (uod are units of distance). The primary variable P1 map was simulated using a spherical semivariogram model with no nugget effect and range of 2000; (see Figure 1).

Based on the P1 map, the secondary variable S1 and S2 are simulated using full co-simulation with the same semivariogram model of P1. The two sets of nine simulated maps of S1 and S2 are simulated in such a way that their correlation coefficients to the primary are approximately in the range of 0.1 to 0.9 in intervals of 0.1 (see Table 1). The direct semivariogram is scaled by the corresponding correlation coefficient for fitting (see Figures 2-4). For a zero correlation coefficient only the primary variable is considered; this way the range of possible correlation coefficients for the two secondary variables is fairly covered. The maps of the simulated secondary data are shown in Figures 5-6.

Table 1: Correlation coefficients of secondary variables S1 and S2 maps to primary P1 map.

Map #	Corr. Coeff. P1&S1	Corr. Coeff. P1&S2
1	0.109	0.116
2	0.205	0.209
3	0.304	0.299
4	0.414	0.404
5	0.508	0.501
6	0.606	0.602
7	0.707	0.700
8	0.814	0.804
9	0.925	0.902

For evaluation of the impact of the secondary variables S1 and S2 on the estimation of primary variable P1 two sub-datasets of 200 data points each are sampled both from the P1, S1 and S2 maps. The sampling is carried out in two ways: 1) from a regular grid 50% of the values are removed independently for P1, S1 and S2 and 2) 200 samples are drawn from the P1 and S1 maps randomly. The patterns of the secondary datasets are kept the same for all the correlation coefficients of P1-S1 and P1-S2 respectively, see Figures 7-8 for an example pattern of primary and secondary data locations.

The impact of S1 and S2 in the estimation of P1 is calculated using cross-validation. For each correlation coefficient pairing between *P* and *S1* and between *P* and *S2* the following steps are performed: (1) Randomly sample from the reference models of all three variables, (2) Perform cross validation of P using kriging of only primary data – this is the base line for comparison, (3) Perform cross validation of *P* using full cokriging using primary data and *S1* and *S2* data, (4) Calculate and compare the correlation coefficient between estimate and the truth and the mean absolute error (MAE) from cross validation in steps 2 and 3, (5) Repeat steps 1 through 4 many times (say 100) to remove impact of any particular sample set. The same methodology is applied to both gridded and non-gridded samples.

Results and Conclusions

Figures 10-11 show the cross validation correlation coefficient of P1 as a function of correlation coefficient in-between P1 and P2 for two pre-selected values of correlation in between primary and secondary S2 variable. Results in Figures 10-11 correspond to the random grid and random sparse grid. Results for the cross validation correlation coefficient of P1 as a function of both correlation coefficients, that is in-between P1 and S1 and P1 and S2 are shown in Figure 12. Figures 13-14 show the mean absolute error of P1 as a function of correlation coefficient in-between P1 and P2 for several pre-selected values of correlation in between primary and secondary S2 variable.

We conclude that in the presence of only one secondary variable the average cross validation correlation coefficient when using all of the data (primary and secondary) in estimation is always higher (estimation is improved) than when using only primary data. Improvement is insignificant for all the correlations between primary and secondary variables below a cut-off of about 0.2-0.3. Probability to improve cross validation correlation coefficient between truth and estimate is close to 100% for all the correlations between primary and secondary variables above a cut-off of about 0.3-0.4.

In the case of estimation using 2 variables we come to the same conclusion. The average cross validation correlation coefficient increases when using primary and both secondary data, see Figure 15. The probability to improve cross validation correlation coefficient between truth and estimate is almost always close to 100% except for the case when both variables are very poorly correlated with the primary, see Figure 17. Furthermore, it is also worth noting that there is a probability that cross validation correlation coefficient and mean absolute error when less correlated secondary variables are used could be better than slightly more correlated secondary variables are used due to the change in the sampling pattern. Overall, the sampling pattern plays a significant role in the implementation of cokriging. Semi-regular grid of samples yields higher cross validation correlation coefficient.

References

Draper, N.R. and Smith, H. (1998). Applied Regression Analysis. Wiley-Interscience.

Glantz, S.A. and Slinker, B.K., (1990). Primer of Applied Regression and Analysis of Variance. McGraw-Hill.

Goovaerts, P. (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York.

Gotway, C.A., Ferguson, R.B., Hergert, G.W., and Peterson, T.A. (1996) Comparison of kriging and inverse-distance methods for mapping soil parameters. Soil Science Society of America Journal, 60:1237–1247.

Moyeed, R.A., and Papritz, A. (2002) An empirical comparison of kriging methods for nonlinear spatial point prediction. Mathematical Geology, 34**:**365-386.

Nagelkerke, N. (1991). A Note on a General Definition of the Coefficient of Determination, Biometrika 78: 691-692.

Vauclin, M., Vieira, S.R., Vachaud, G., Nielsen, D.R. (1983) The use of cokriging with limited field soil observations. Soil Science Society of America Journal, 47:175-184.

Wackernagel, H. (2003): Multivariate geostatistics. Springer, Berlin New York.

Wackernagel, H. (1994) Cokriging versus kriging in regionalized multivariate data analysis. Geoderma, 62:83–92.

Weisz, R., Fleischer, S., and Smilowitz, Z. (1995) Map generation in high-value horticultural integrated pest management: appropriate interpolation methods for site-specific pest management of Colorado Potato Beetle (Coleoptera: Chrysomelidae). Journal of Economic Entomology, 88:1650–1657.

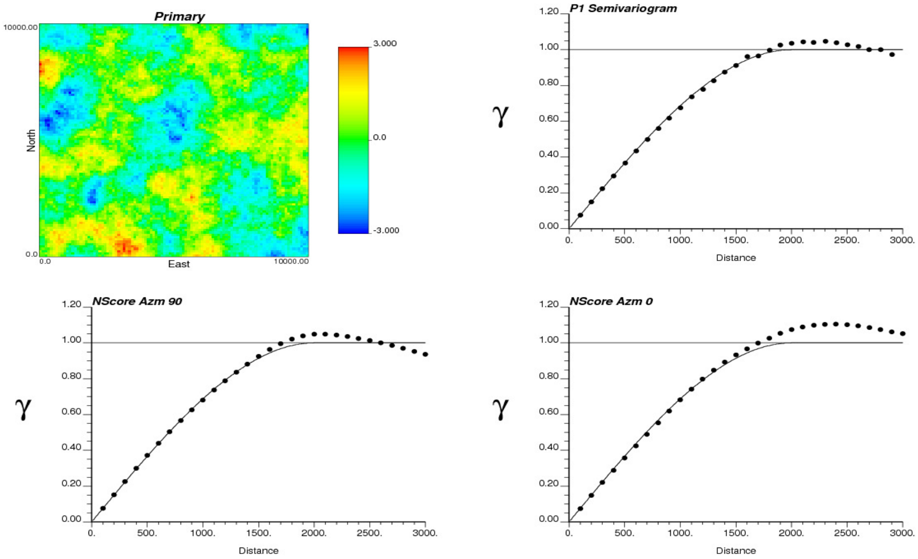


Figure 1: Simulated map for primary variable P1 (top left), omni-directional semivariogram (top right), 90º azimuth direction semivariogram (bottom left) and 0º azimuth semivariogram (bottom right); in the semivariogram plots, experimental semivariogram (black dots) and semivariogram model (solid line).

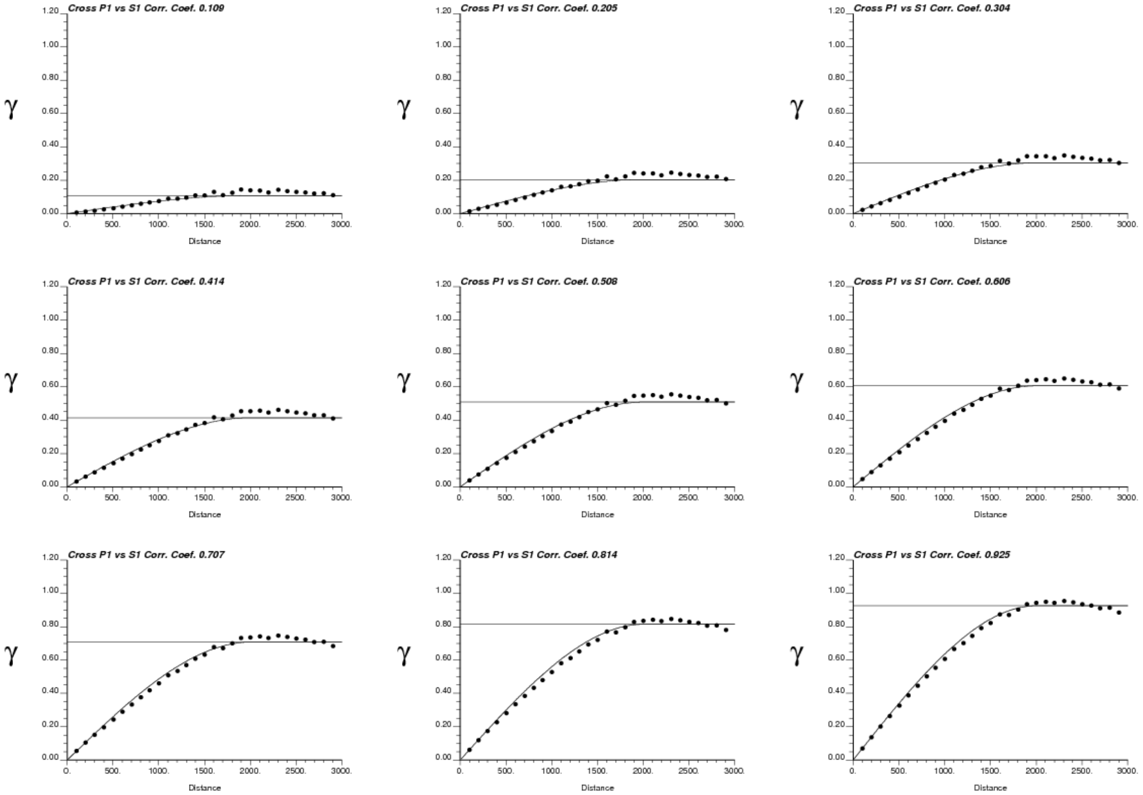


Figure 2: Cross semivariograms of P1 and S1.

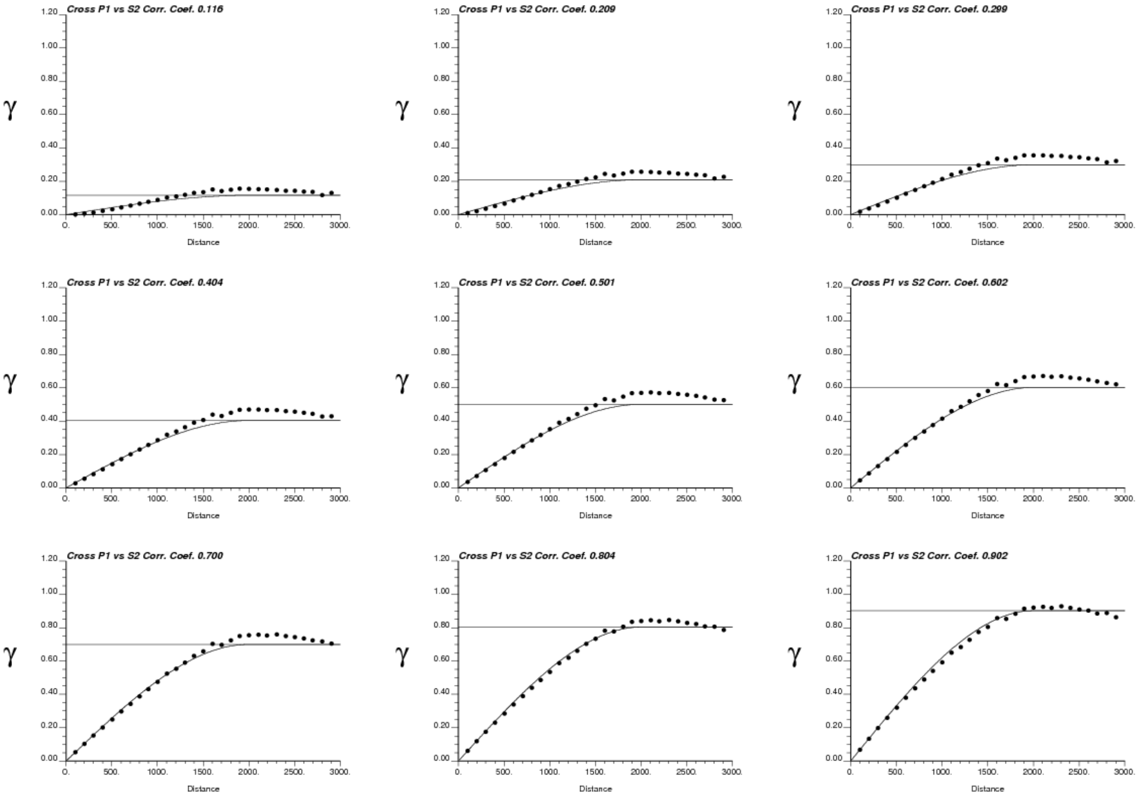


Figure 3: Cross semivariograms of P1 and S2.

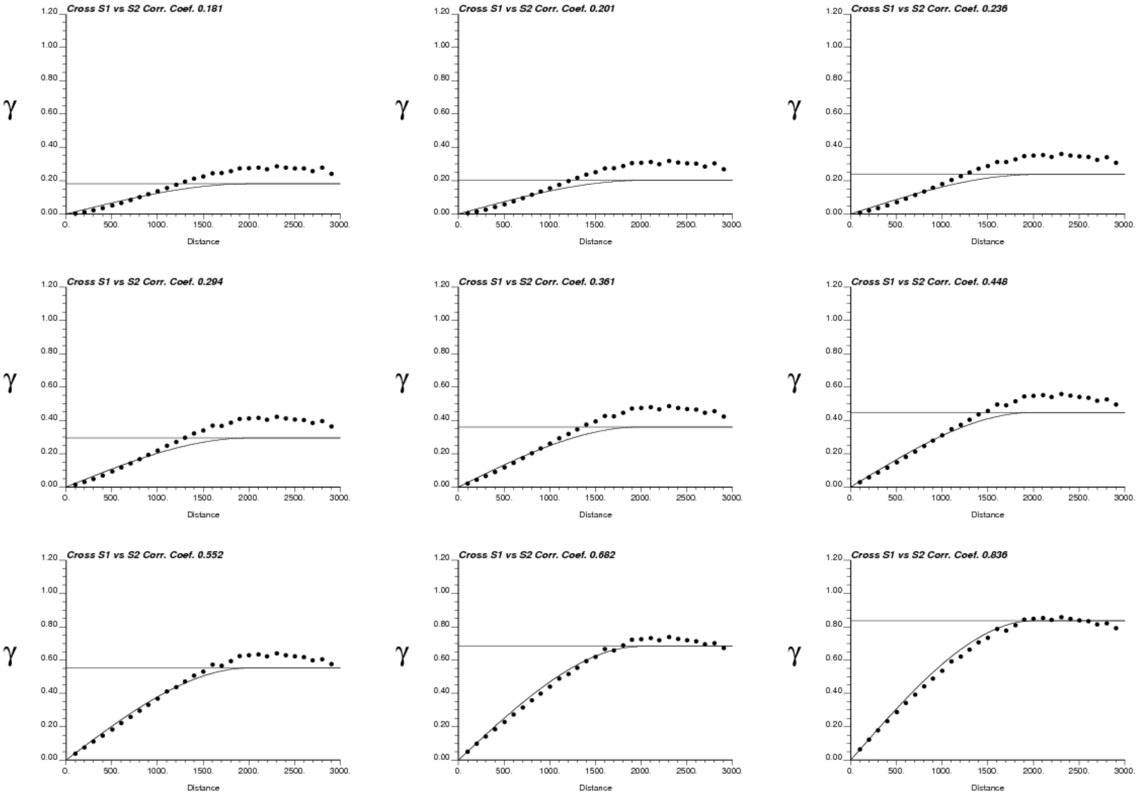


Figure 4: Cross semivariograms of S1 and S2.

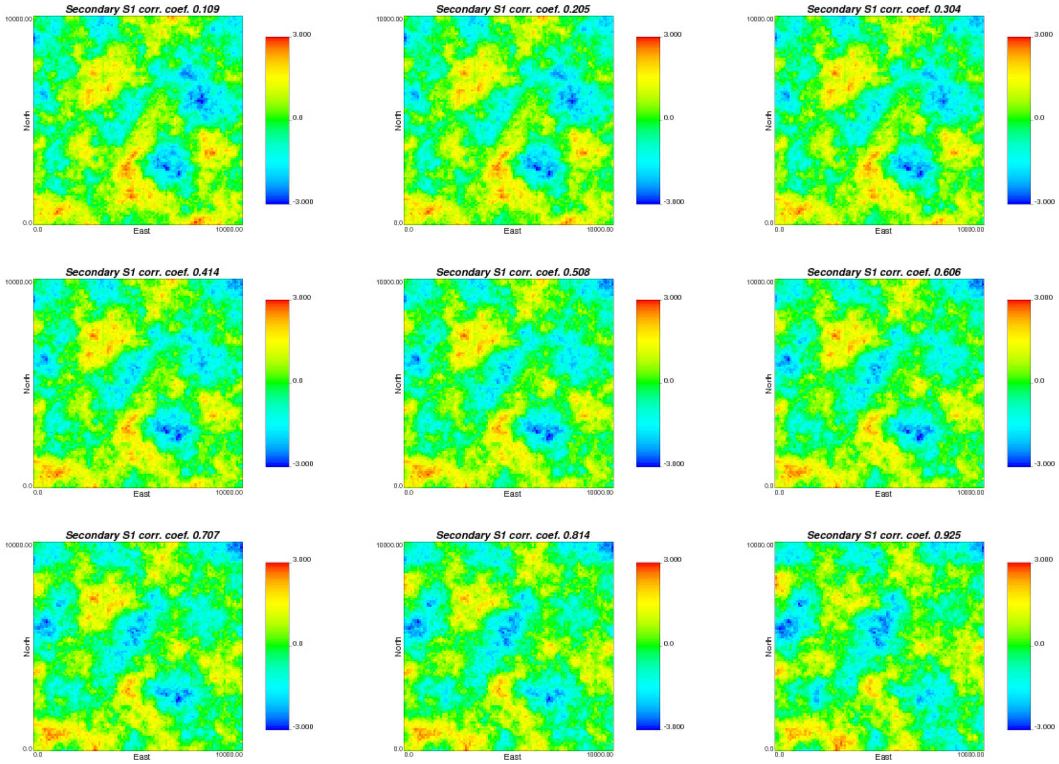

Secondary S1 corr. coef. 0.109
Secondary S1 corr. coef. 0.205
Secondary S1 corr. coef. 0.304
Secondary S1 corr. coef. 0.414
Secondary S1 corr. coef. 0.508
Secondary S1 corr. coef. 0.606
Secondary S1 corr. coef. 0.707
Secondary S1 corr. coef. 0.814
Secondary S1 corr. coef. 0.925
10000.00
North
0.0
East
10000.00
3.000
0.0
-3.000

Figure 5: Simulated maps for secondary variable S1. Different plots correspond to different values for the correlation coefficient (0.1 to 0.9 with step 0.1) in-between primary P1 and secondary S1 data.

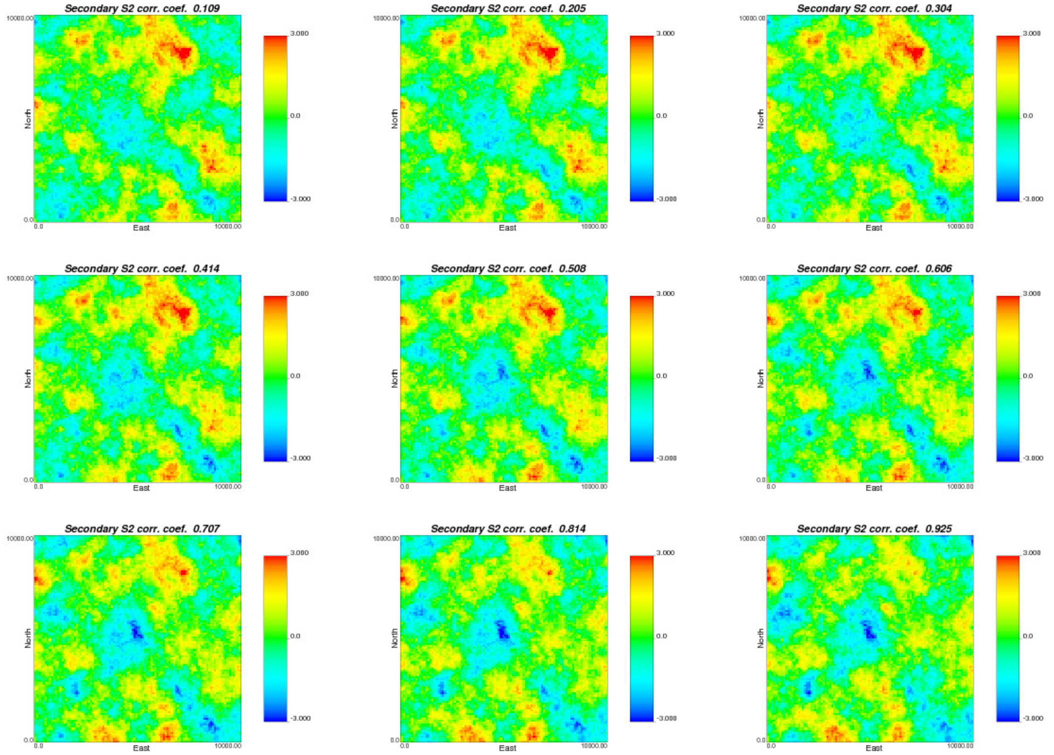

Secondary S2 corr. coef. 0.109
Secondary S2 corr. coef. 0.205
Secondary S2 corr. coef. 0.304
Secondary S2 corr. coef. 0.414
Secondary S2 corr. coef. 0.508
Secondary S2 corr. coef. 0.606
Secondary S2 corr. coef. 0.707
Secondary S2 corr. coef. 0.814
Secondary S2 corr. coef. 0.925
10000.00
North
0.0
East
10000.00
3.000
0.0
-3.000

Figure 6: Simulated maps for secondary variable S2. Different plots correspond to different values for the correlation coefficient (0.1 to 0.9 with step 0.1) in-between primary P1 and secondary S2 data.

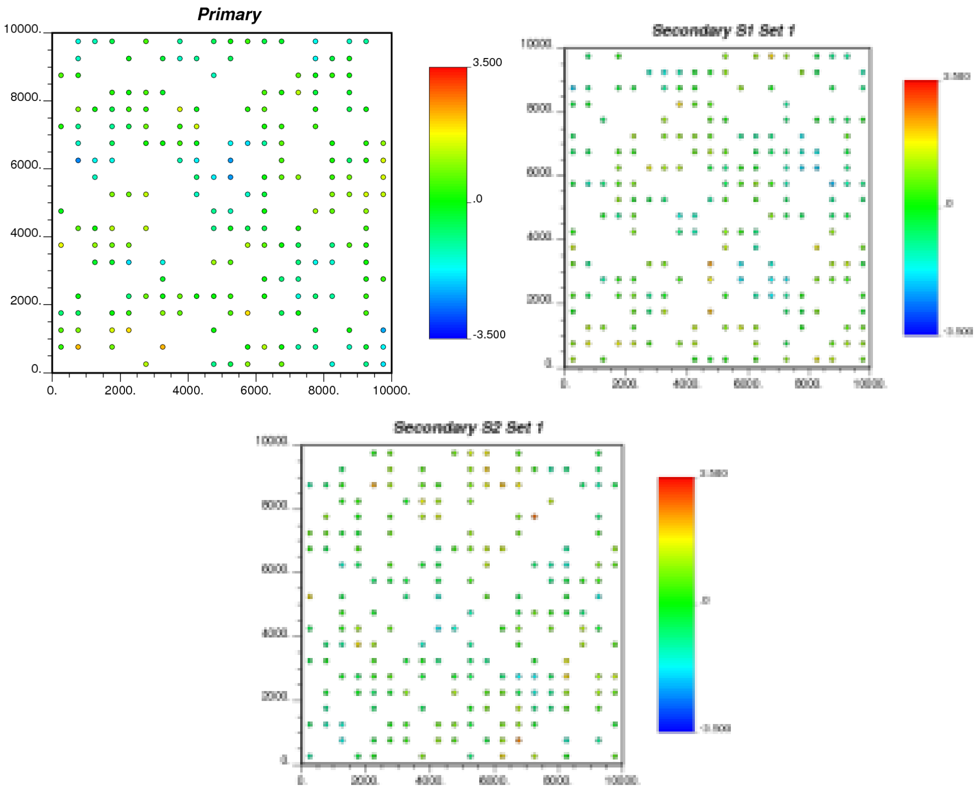


Figure 7: Primary P1 (top left), secondary P1 (top right; the same for all correlation values) secondary P2 (bottom right; the same for all correlation values) data locations selected randomly from a regular grid.

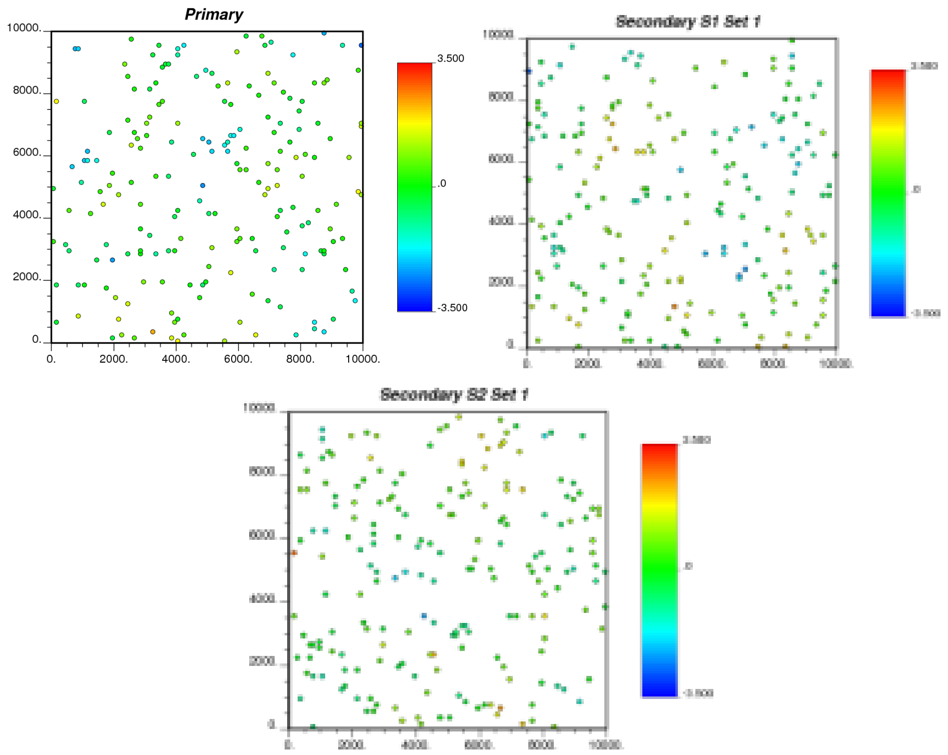


Figure 8: Primary P1 (top left), secondary P1 (top right; the same for all correlation values) secondary P2 (bottom right; the same for all correlation values) data locations selected randomly from the simulated maps.

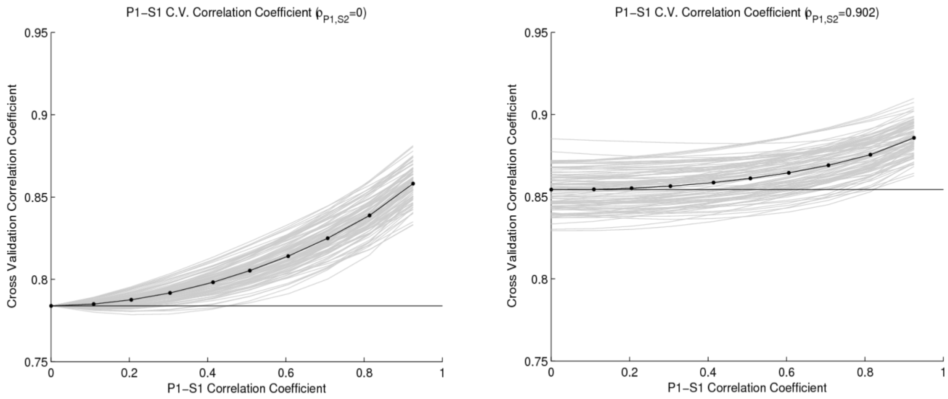


Figure 10: Cross validation correlation coefficient of P1 as a function of correlation coefficient in-between P1 and P2 with $\rho_{P1,S2}$=0 (left) and $\rho_{P1,S2}$=0.902 (right) for the random grid. Average contribution is shown in dotted line and contribution of realizations in gray lines.

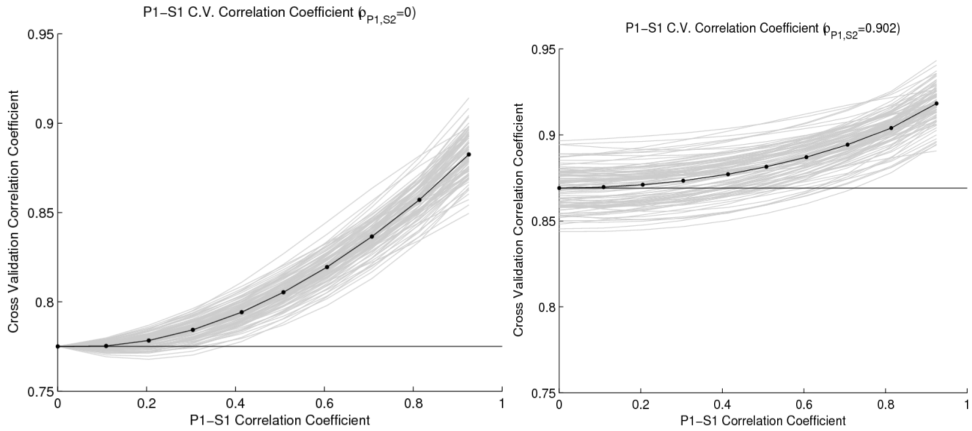


Figure 11: Cross validation correlation coefficient of P1 as a function of correlation coefficient in-between P1 and P2 with $\rho_{P1,S2}$=0 (left) and $\rho_{P1,S2}$=0.902 (right) for the random sparse grid. Average contribution is shown in dotted line and contribution of realizations in gray lines.

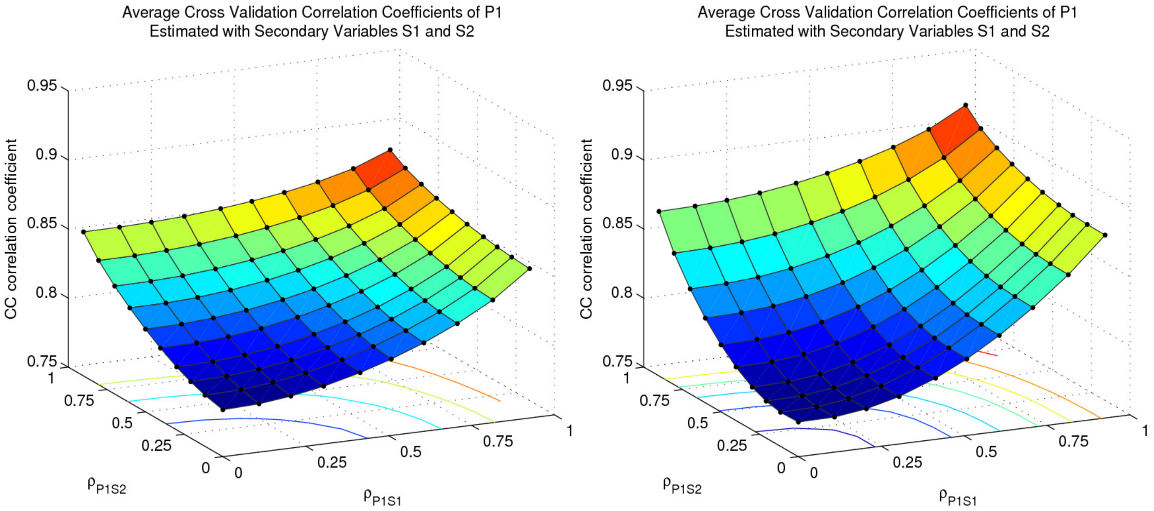


Figure 12: Average cross validation correlation coefficient for P1 in a sparse random sampling patterns (left) and gridded random sampling (right).

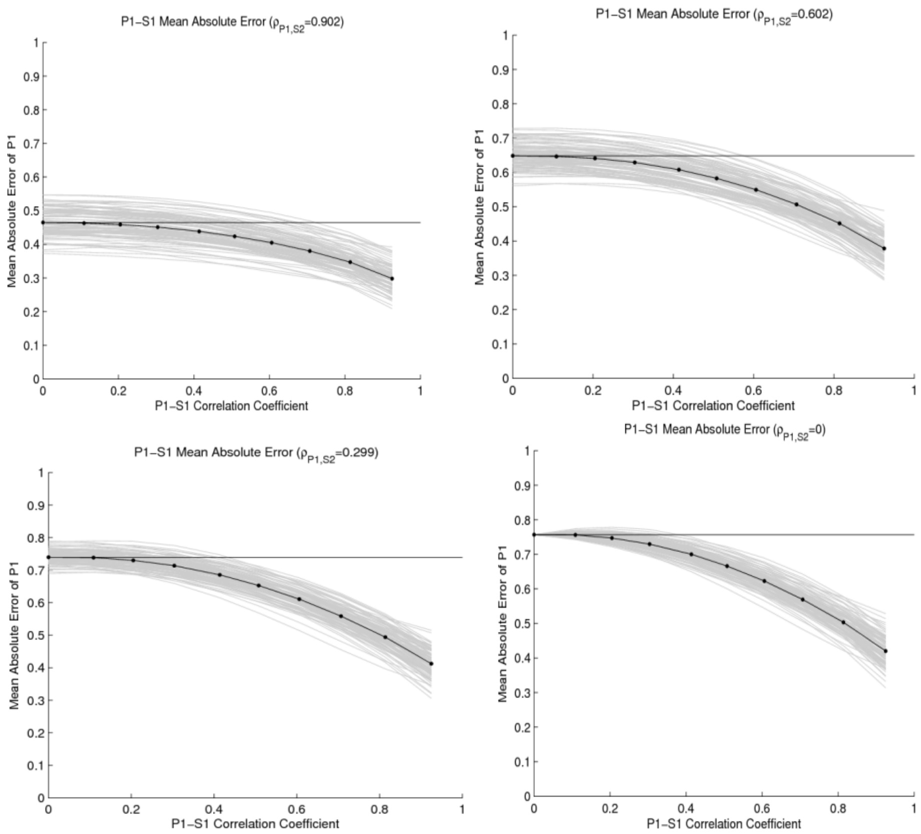
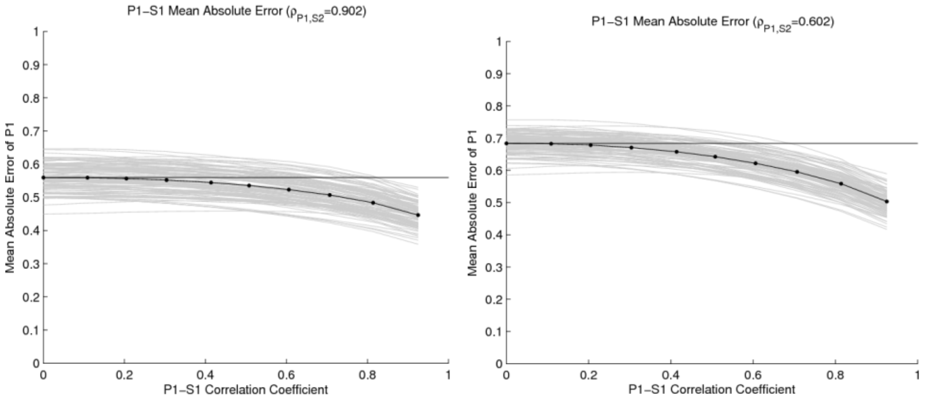


Figure 13: Mean absolute error of P1 as a function of correlation coefficient in-between P1 and P2 with $\rho_{P1,S2}$=0.902 (top left), $\rho_{P1,S2}$=0.602 (top right), $\rho_{P1,S2}$=0.299 (bottom left), and $\rho_{P1,S2}$=0 (bottom right) for the random grid. Average contribution is shown in dotted line and contribution of realizations in gray lines.



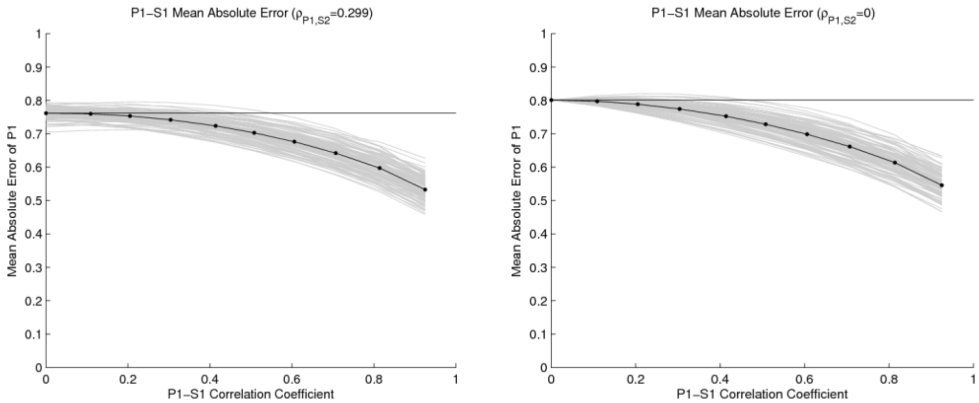


Figure 14: Mean absolute error of P1 as a function of correlation coefficient in-between P1 and P2 with $\rho_{P1,S2}$=0.902 (top left), $\rho_{P1,S2}$=0.602 (top right), $\rho_{P1,S2}$=0.299 (bottom left), and $\rho_{P1,S2}$=0 (bottom right) for the random sparse grid. Average contribution is shown in dotted line and contribution of realizations in gray lines.

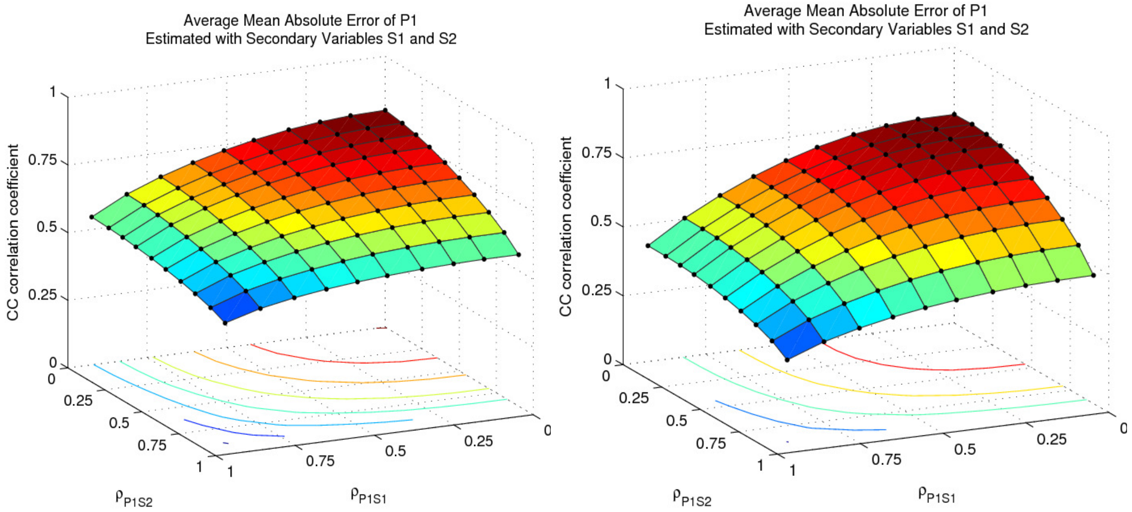


Figure 15: Average mean absolute error of P1 in a sparse random sampling patterns (left) and gridded random sampling (right).

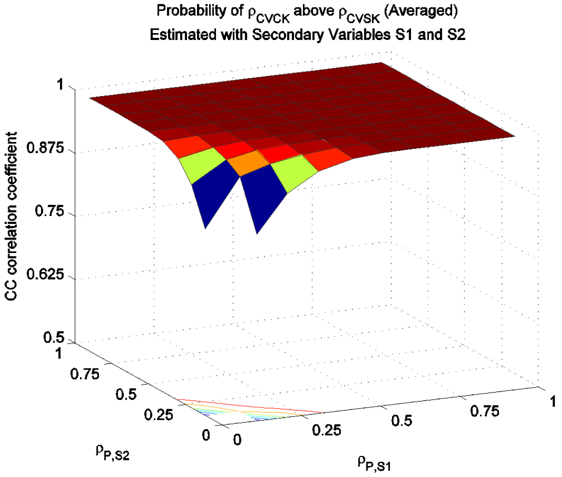


Figure 16: Probability of cross validation correlation coefficient in co-kriging to be higher than in kriging (averaged).