Missing Data Replacement in a MultiGaussian Context

Ryan M. Barnett and Clayton V. Deutsch

Unequally sampled data is an important challenge in geostatistics, as multivariate transforms may only be executed with equally sampled observations. To facilitate these transforms, observations not having the full suite of required variables are commonly excluded from the modeling workflow. While leading to the loss of information, this may also introduce a bias since there may be specific reasons for the missing data. A better option is to consider the replacement (or imputation) of missing data values, so that modeling may proceed with all of the sampled data. This missing data replacement must be done in a manner that is unbiased, while also reflecting uncertainty in the imputed values. As methods advocated by data replacement literature are not immediately suited for regionalized random variables, we aim to adapt them to the geostatistical analysis setting. Multivariate geostatistical analysis may be subset based on whether the multiGaussian assumption is reasonable after a normal score transformation. This will dictate which subsequent modeling and/or transformation methods are employed, in addition to what data replacement method is appropriate. The following paper will propose a Bayesian Updating technique for data replacement in the multiGaussian context, while its companion paper[1] considers the complex multivariate setting. Following an overview of the available data replacement methods, the Bayesian Updating Data Imputation (BUDI) method is developed. Using a synthetic case study, very good results are demonstrated in terms of replacement accuracy and resultant gains in geostatistical model accuracy.

Introduction

A large variety of techniques are available for transforming multivariate data to be suitable for geostatistical modeling. A selection of these transforms and their purpose include: (i) logratios [1,11] for removing compositional constraints, (ii) SCT [1,11], MSNT [3], and PPMT [4] for removing non-linearity and heteroscedasticity, and (iii) PCA [1,9] and MAF [1,14] for removing correlation. Unfortunately, all of these transforms may only be executed on homotopically (equally) sampled observations. That is to say, any observations not containing the full suite of variables to be transformed may not be used in these techniques and the subsequent geological modeling workflow.

Consequently, when faced with observations that are missing any number of variables to be modeled, geostasticians must decide between either: (i) eliminating those observations from the modeling framework, or (ii), choosing a method for replacing the missing variable values. The obvious disadvantage to the first approach is the potential for drastically reducing the information that is available for global statistics and local conditioning. As stated by methodologists in the field of missing data replacement [8,16], this also makes strong assumptions of why data is missing in the first place, and could introduce a strong bias to the remaining data. Perhaps even more dangerous, however, is choosing a data replacement method that produces biased results, or does not reflect inherent uncertainty in the replaced values. Modern data replacement theory advocates methods that seek to address these concerns, providing unbiased missing data replacement with associated uncertainty distributions. The following study in conjunction with its companion paper[1] [2] aims to adapt these advocated techniques to the field of geostatistics.

As the complexity of geologic data necessitates the use of differing modeling frameworks, so too will it motivate the use of varying data replacement techniques. Should the data be reasonably multivariate Gaussian (multiGaussian) following the normal score transform [1,6], linear modeling (e.g. co-simulation with the linear model of coregionalization [6]), or decorrelation (e.g. PCA/MAF [1,9,17]) may immediately proceed. This paper will develop a data replacement technique for this multiGaussian setting. On the other hand, if multivariate complexities are present, additional transforms such as those referenced at the top of this introduction may be necessary. The companion paper[1] [2] deals with data replacement in this more complex multivariate setting.

[1] Barnett, R., & Deutsch, C. (2012). Data replacement in a complex multivariate context. *CCG Annual Report 14.* Paper 113.

A brief overview of the available missing data replacement techniques will first be provided. Based on the recommendations of data replacement literature, and its suitability for geostatistical modeling frameworks, Multiple Imputation is selected from these available techniques. Multiple Imputation is then adapted to spatial variables through the use of the familiar Bayesian Updating [13] framework, to form the proposed Bayesian Updating Data Imputation (BUDI) technique. Accuracy of this method and the potential value it will contribute to a multivariate modeling framework is demonstrated on a synthetic case study. This value gained may be calculated analytically in a multiGaussian setting, which is presented in Appendix 1. Parameters for the BUDI software package are presented in Appendix 2.

Data Replacement Methods

As outlined by Enders [7], missing data replacement enjoyed a surge in methodology development during the 1970's with the publication of its two primary "state of the art" methods, maximum likelihood [5] and multiple imputation [15]. A general theoretical framework for missing data replacement [14] was also published during that decade, which continues to see wide spread use. Generally speaking, methodologists in this field seek to replace (or impute[2]) data in a manner that minimizes bias, and provides an accurate distribution of uncertainty. Many methods are practically employed for missing data replacement [7,16], regardless of whether they are theoretically refuted. A selection of these are summarized below in a geostatistical context (a further subset is visually demonstrated in Figure 1):

- *Listwise Deletion:* delete any sample that does not have all variables present.
- *Arithmetic Mean Imputation:* replace missing values with the stationary mean of that variable.
- *Regression Imputation:* determine a regression based model for missing variables based on secondary correlated variables. Then use the collocated secondary variables to impute the missing values.
- *Stochastic Regression Imputation:* the same as regression imputation, but apply random generation methods such as MCS [7] to add realistic variability to the regression model
- *Hot-deck Imputation:* randomly replace missing values with data values of other samples that measure similarly according to colocated secondary variables.
- *Similar Response Pattern Imputation:* also known as nearest neighbour hot-deck imputation. The random selection is more restricted based on additional factors such as spatial considerations.
- *Last Observation Carried Forward:* the name is derived from its application in time series analysis. The spatial equivalent would simply be nearest neighbour imputation.
- *Maximum Likelihood Estimation:* model parameters (e.g. mean and variance) of a distribution are estimated to maximize the log-likelihood of each observation occurring. In missing data analysis, these parameters are estimated through iterative optimization using various subsets of the data.
- *Multiple Imputation:* apply methods such as stochastic regression imputation repeatedly to generate multiple realizations of the data.

In narrowing down the available choices, it is worth noting that data replacement methodologists [7,16] advocate the use of either maximum likelihood and multiple imputation due to their unbiased estimates and associated distributions for uncertainty. These are very attractive properties that will be necessary of any geostatistical data replacement method. Multiple imputation is likely the more immediately suitable to geostatistics of the two techniques, as it will form multiple realizations of the data. These data realizations may be used for generating multiple geostatistical model realizations, allowing for seamless integration into popular simulation frameworks. Log-likelihood methods will estimate model parameters rather than the missing data itself, making it comparatively difficult to integrate with subsequent geostatistical modeling.

[2] Rather than the word replace, missing data methodologists prefer the word impute. Replace and impute will be used interchangeably throughout this paper, in order to maintain consistency with the literature where required. From the Oxford English Dictionary: "Impute - to assign a value to something by inference based on the value of the products or processes to which it contributes".

Bayesian Updating Data Imputation (BUDI)

Multiple Imputation can be conceptualized based on the stochastic imputation panel of Figure 1, where regression is performed, before sampling stochastically from the conditional distribution. Adapting this to a geostatistical setting, a method will be required for inferring this conditional distribution. This method should take advantage of the spatial correlation of a missing variable that is to be imputed, in addition to any correlated and collocated secondary variable(s).

As this study works in a multiGaussian setting, Bayesian Updating may be appropriately applied to construct the conditional distributions at every location and variable requiring imputation. Bayesian Updating is a powerful method for integrating primary and secondary information into the construction of a regionalized random variable's local conditional distribution. Readers are referred to Ren's PhD thesis [13] for a full development of Bayesian Updating methodology, which the following outline of essential theory will follow. Working with normal scores of the original data [1,6], an arbitrary variable being imputed is termed the primary, while the collocated and correlated sampled variables are considered the secondary. The Bayesian Updating workflow is then given by the following steps and visually presented in Figure 2.

Prior Distribution

At location $\mathbf{u}$ requiring imputation, estimate the primary distribution mean, $\overline{y}_P(\mathbf{u})$, and variance, $\sigma_P^2(\mathbf{u})$, according to Equations 1 and 2 respectively. Here, weights, λ_i, are calculated based on the covariance between the location $\mathbf{u}$ and the location $\mathbf{u_i}$ of the i^{th} sample (Equation 3). This amounts to using the normal equations to perform simple kriging based on correlated surrounding samples of the primary variable (consider `kt3d` [6] in cross-validation mode).

$$\overline{y}_P(\mathbf{u}) = \sum_{i=1}^{n} \lambda_i \cdot y_P(\mathbf{u}_i) \tag{1}$$

$$\sigma_P^2(\mathbf{u}) = 1 - \sum_{i=1}^{n} \lambda_i C(\mathbf{u}, \mathbf{u}_i) \tag{2}$$

$$\sum_{j=1}^{n} \lambda_j C(\mathbf{u}_i, \mathbf{u}_j) = C(\mathbf{u}, \mathbf{u}_i) \quad i = 1, ..., n \tag{3}$$

Likelihood Distribution

Next, estimate the likelihood distribution mean, $\overline{y}_L(\mathbf{u})$, and variance, $\sigma_L^2(\mathbf{u})$, according to Equations 4 and 5 respectively. Here, weights, λ_i, are calculated based on correlation between the variable y and the colocated i^{th} variable x_i (Equation 6). This amounts to using the normal equations to perform linear least squares regression based on correlated and colocated secondary variables (consider `likelihood_di` - Appendix 2).

$$\overline{y}_L(\mathbf{u}) = \sum_{i=1}^{m} \lambda_i \cdot x_i(\mathbf{u}) \tag{4}$$

$$\sigma_L^2(\mathbf{u}) = 1 - \sum_{i=1}^{m} \lambda_i \rho_{i,0} \tag{5}$$

$$\sum_{j=1}^{m} \lambda_j \rho_{i,j} = \rho_{i,0}, \quad i = 1, ..., m \tag{6}$$

Updated Distribution

Finally, merge the prior and likelihood distributions to form the posterior or Updated mean, $\overline{y}_U(\mathbf{u})$, and variance, $\sigma_U^2(\mathbf{u})$, according to Equations 7 and 8 respectively. As a Gaussian distribution is fully defined by its mean and variance [10], the missing value now has its full distribution of uncertainty defined based on the spatially correlated primary variables (prior), and the correlated/colocated secondary variables (likelihood) (consider `update_di` - Appendix 2).

$$\overline{y}_U(\mathbf{u}) = \frac{\overline{y}_L(\mathbf{u})\sigma_P^2(\mathbf{u}) + \overline{y}_P(\mathbf{u})\sigma_L^2(\mathbf{u})}{\sigma_P^2(\mathbf{u}) - \sigma_P^2(\mathbf{u})\sigma_L^2(\mathbf{u}) + \sigma_L^2(\mathbf{u})} \tag{7}$$

$$\sigma_U^2(\mathbf{u}) = \frac{\sigma_L^2(\mathbf{u})\sigma_P^2(\mathbf{u})}{\sigma_P^2(\mathbf{u}) - \sigma_P^2(\mathbf{u})\sigma_L^2(\mathbf{u}) + \sigma_L^2(\mathbf{u})} \tag{8}$$

Stochastically Simulate

Once the above steps are executed for every missing value in a dataset, stochastically simulate the data realizations by randomly sampling from the empirical Gaussian CDF's defined by each imputed value's $\overline{y}_U(\mathbf{u})$ and $\sigma_U^2(\mathbf{u})$. Keep in mind that for every realization of the data, actual sampled values remain constant (0 uncertainty), while imputed values will vary based on the degree of uncertainty in its Updated distribution, $\sigma_U^2(\mathbf{u})$ (consider `busim_di` - Appendix 2).

Case Study

The BUDI methodology will be demonstrated on a synthetic case study. Exhaustive True synthetic models are first generated, creating five correlated multiGaussian variables of varying spatial continuity, from which 283 homotopic observations are sampled (Figure 3). From these 283 samples, 30 observations of each variable are independently and randomly selected for removal (Figure 4). This results in a dataset of 174 complete observations, with 109 that are incomplete to varying degrees. The scatterplots of all five variables following this data removal are displayed for these sampled observations in Figure 5, before and after normal score transformation. While not perfectly multivariate Gaussian, this normal score data is illustrative of what may be considered reasonably multivariate Gaussian, allowing for BUDI to be appropriately applied. Note that perfect multivariate Gaussianity would enhance BUDI results.

The BUDI workflow is applied next to form 100 realizations of complete data. As outlined, the quality of these imputed values will largely be a function of the variable's spatial continuity (Figure 3), and its degree of correlation with the secondary variables (Figure 5). Following the generation of these data realizations, the imputed values may be compared with the removed True values for cross-validation (Figure 6). Observe in this figure that the mean and variability of these imputed realizations is unbiased and accurate when compared to the True removed samples.

The question arises of whether this effort of data replacement is justified by a measurable gain in the quality of modeling results. While this may be studied analytically since the replacement is taking place within the multiGaussian domain (Appendix 1), a more intuitive demonstration will be presented using the above data realizations. Identical geostatistical modeling workflows will be executed with and without the use of data replacement. That is to say, one workflow will use the data replacement realizations attained above, while the other will eliminate the incomplete samples (necessary so that a decorrelation transform may be applied). Comparing the resultant geostatistical models with and without data replacement to the True model from which the samples were originally drawn (Figure 3) will provide an indication of value gained from the BUDI replacement.

Dealing first with details of the modeling workflow, the 100 data realizations are individually MAF [1,17] and normal score [1,6] transformed to form independent Gaussian variables. These 100 data realizations are used to condition an SGSIM [6] based simulation of 100 models, which are then back-transformed. An identical modeling workflow is then executed using a single dataset, where incomplete observations have been eliminated so that MAF may be applied.

Next, to ascertain the value gained in terms of local accuracy, E-Type estimates are formed from the 100 realizations of the two modeling workflows (Figure 7) and compared with the True model (Figure 3). This comparison displayed uniformly better results for the workflow involving data replacement, which is summarized by Table 1 according to the MSE and Covariance improvement (as compared to the True model). In addition to maps of the E-Type estimates, Figure 8 also displays (i) E-types of the imputed realizations at the removed data locations, (ii) True values at the removed data locations, and (iii) a comparison between the E-Type models based on data replacement and data elimination. Note from this figure that while data replacement has improved quality of the estimates overall, poorer local estimates will inevitably occur in regions where there is poor data imputation. Cross-referencing poorly imputed

locations with the True map in Figure 3, it is unsurprising that they generally occur in regions where the missing value was fairly unrelated to its spatially surrounding samples. This would lead to an inaccurate prior distribution.

Table 1: Improvement in the MSE and covariance of E-Type estimates vs. the True model (using data replacement realizations rather than data elimination).

Variable	% Improvement	
	Mean Squared Error	Covariance
1	27.73	20.19
2	13.36	23.56
3	12.17	13.94
4	16.09	3.70
5	9.54	36.42

Conclusion

The replacement of missing data is very important for multivariate geostatistical modeling with unequally sampled data. It allows for the application of multivariate transforms to a potentially far greater number of observations, providing the subsequent modeling workflow with additional information for local conditioning and global statistics. Following the outline of a number of available data replacement techniques, multiple imputation was selected based on its theoretical properties and suitability within geostatistical frameworks. After being adapted to geostatistics through Bayesian Updating, multiple imputation was demonstrated on a multiGaussian synthetic case study. A high degree of accuracy was seen in the uncertainty distributions of the imputed data using this technique. This imputed data was demonstrated to greatly improve the geostatistical modeling accuracy, as compared to a parallel workflow that used data elimination. A note on the analytical approximation of value gain in a multiGaussian setting is provided in Appendix 1, while the BUDI software package is given in Appendix 2.

References

1 Barnett, R. (2011). *Guidebook on Multivariate Geostatistical Tools.* Edmonton, Alberta: Centre for Computational Geostatistics.

2 Barnett, R., & Deutsch, C. (2012). Missing Data Replacement in a Complex Multivariate Context. *CCG Annual Report 14*, Paper 113.

3 Barnett, R., & Deutsch, C. (2012). MSNT Advances and Case Studies. *CCG Annual Report 14*, Paper 101.

4 Barnett, R., Manchuk, J., & Deutsch, C. (2012). Projection Pursuit Multivariate Transform. *CCG Annual Report 14*, Paper 103.

5 Beale, E., & Little, R. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, series B, vol.37, pp.129-145.

6 Deutsch, C., & Journel, A. (1998). *GSLIB: A geostatistical software library and user's guide, second edition.* Oxford University Press.

7 Deutsch, J., & Deutsch, C. (2012). Latin hypercube sampling with multidimensional uniformity. *Journal of Statistical Planning, vol.142*, pp.763-773.

8 Enders, C. (2010). *Applied Missing Data Analysis.* New York: Guilford Pess.

9 Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417-441.

10 Johnson, R., & Wichern, D. (1988). Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall.

11 Leuangthong, O., & Deutsch, C. (2003). Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology, vol.35, no.2*, pp.155-173.

12 Pawlowsky-Glawh V, E. J. (2006). Compositional data and their analysis: an introduction. In A. M.-F.-G. Buccianti, *Compositional data analysis in the geosciences: from theory to practice.* (pp. vol.264, pp.1-10). London Geological Society Special Publication.

13 Ren, W. (2007). *Bayesian Updating for Geostatistical Analysis, PhD. Thesis.* Edmonton: University of Alberta.

14 Rubin, D. (1976). Inference and missing data. *Biometrika*, vol.63, pp.581-592.

15 Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* Hoboken, NJ: Wiley.

16 Rubin, D., & Little, R. (2002). *Statistical analysis with missing data.* Hoboken, N.J.: Wiley.

17 Switzer, P., & Green, A. (1984). Min/Max autocorrelation factors for multivariate spatial imaging. *Stanford University: Department of Statistics Technical Report No.6*, pp.14.

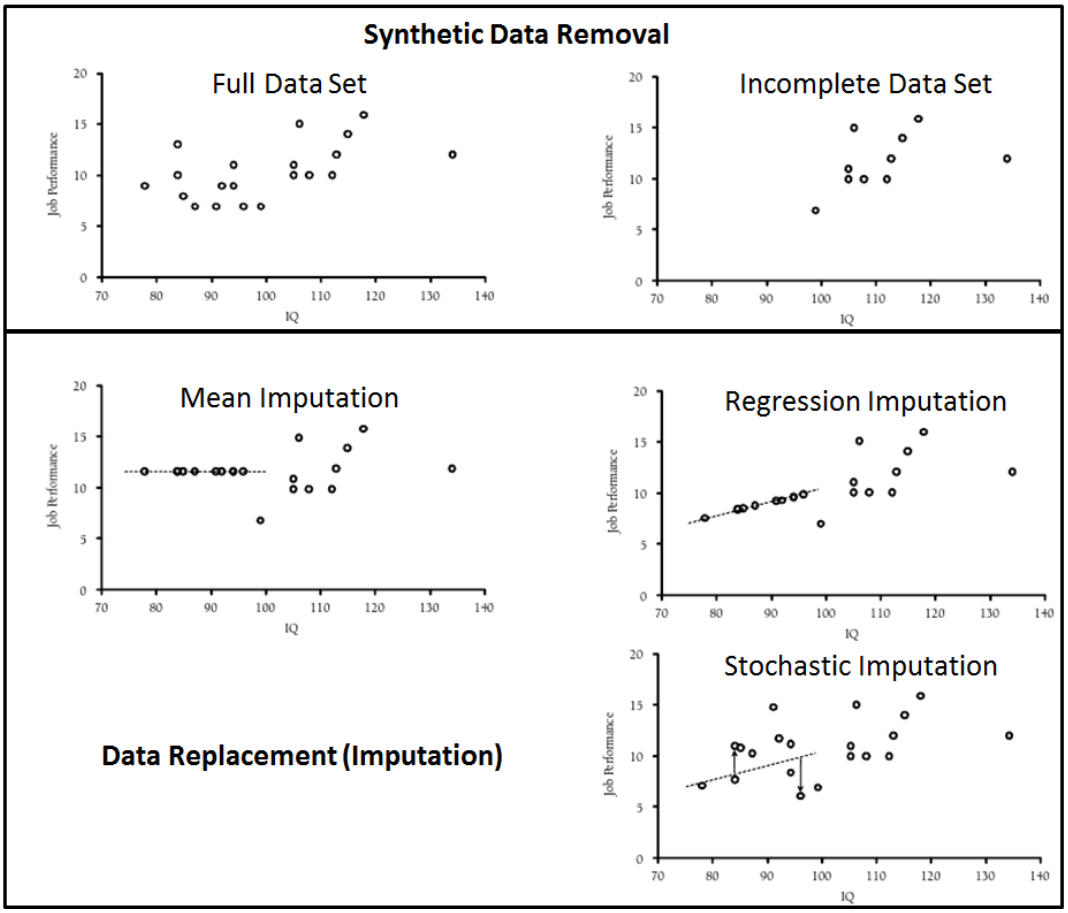


Figure 1: Simplified example of missing data replacement with multiple techniques [8].

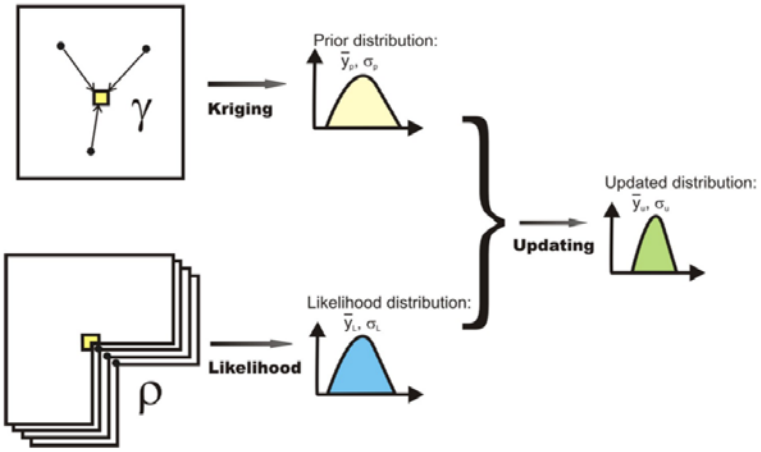


Figure 2: Visual schematic of the Bayesian Updating process [13].

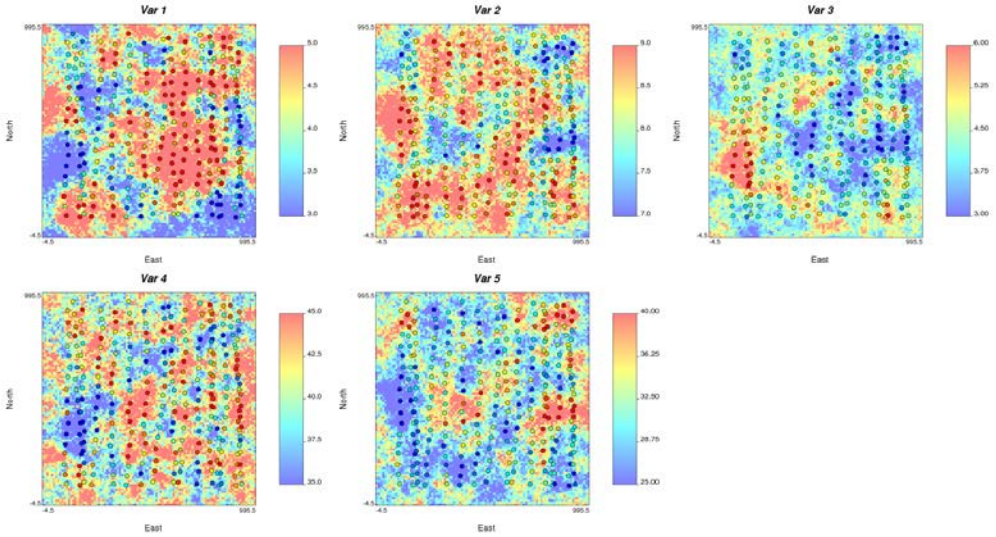


Figure 3: Synthetic true models, with circles indicating the locations of 283 homotopic samples.

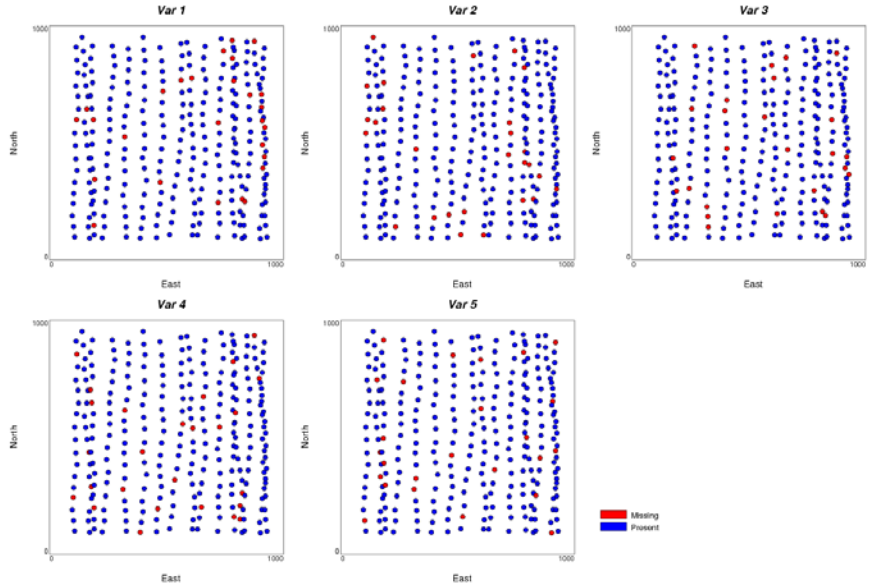


Figure 4: Locations of randomly removed observations for each variable.

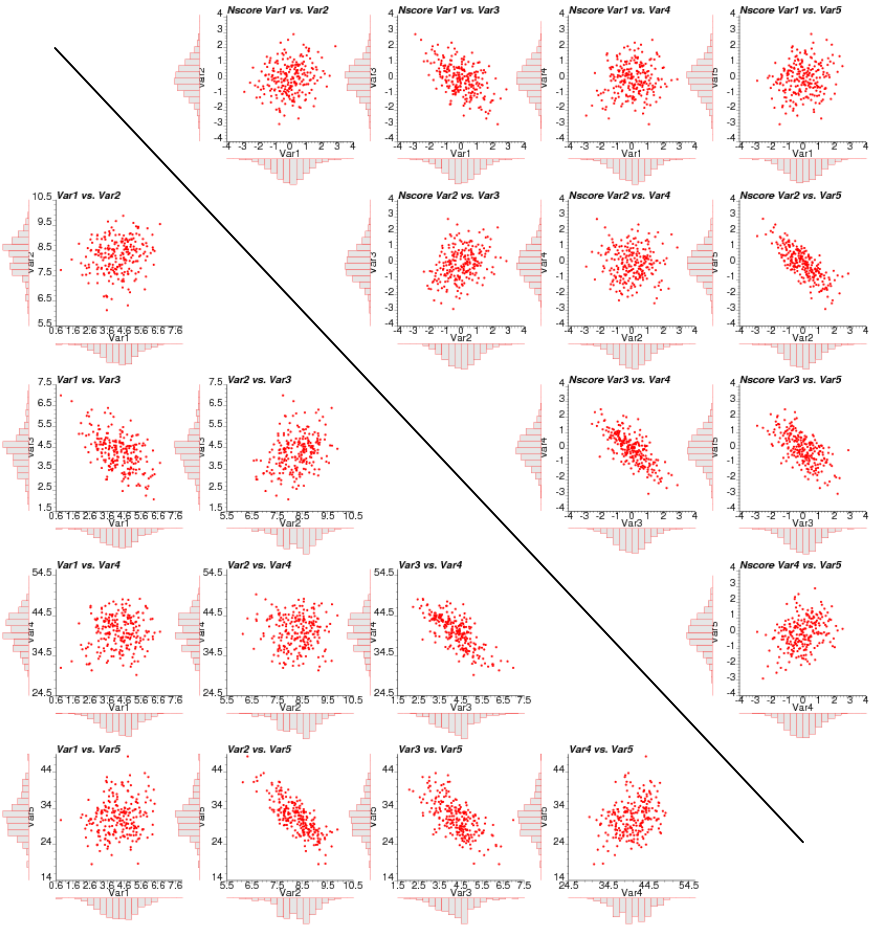
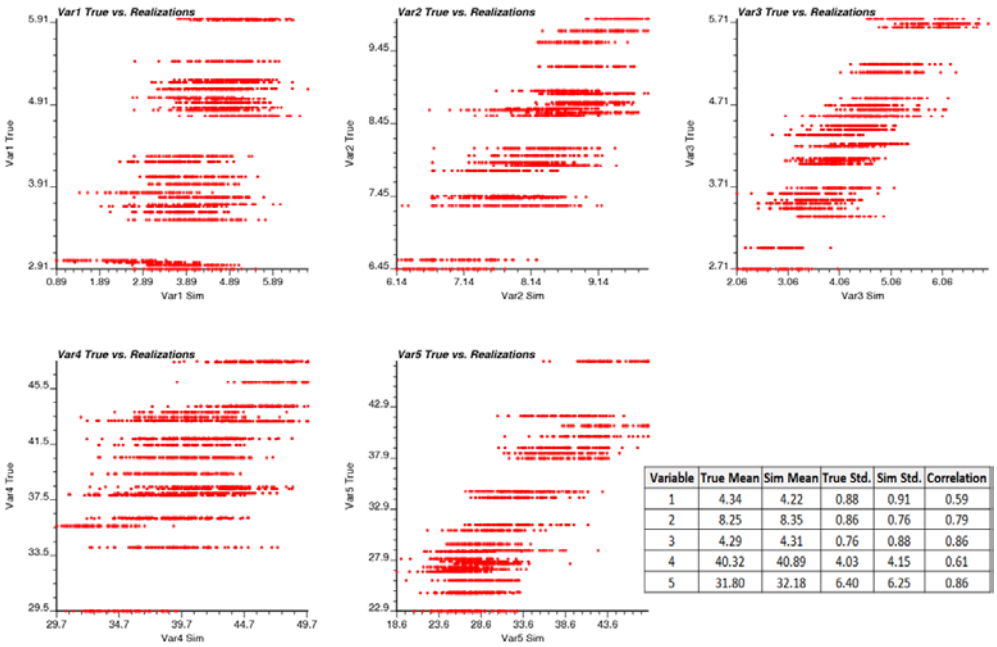


Figure 5: Scatterplots of the sampled observations before (bottom covariance triangle) and after (upper covariance triangle) normal score transformation.



Variable	True Mean	Sim Mean	True Std.	Sim Std.	Correlation
1	4.34	4.22	0.88	0.91	0.59
2	8.25	8.35	0.86	0.76	0.79
3	4.29	4.31	0.76	0.88	0.86
4	40.32	40.89	4.03	4.15	0.61
5	31.80	32.18	6.40	6.25	0.86

Figure 6: Comparison between removed True values (y-axis) and the associated 100 realizations of imputed values (x-axis).

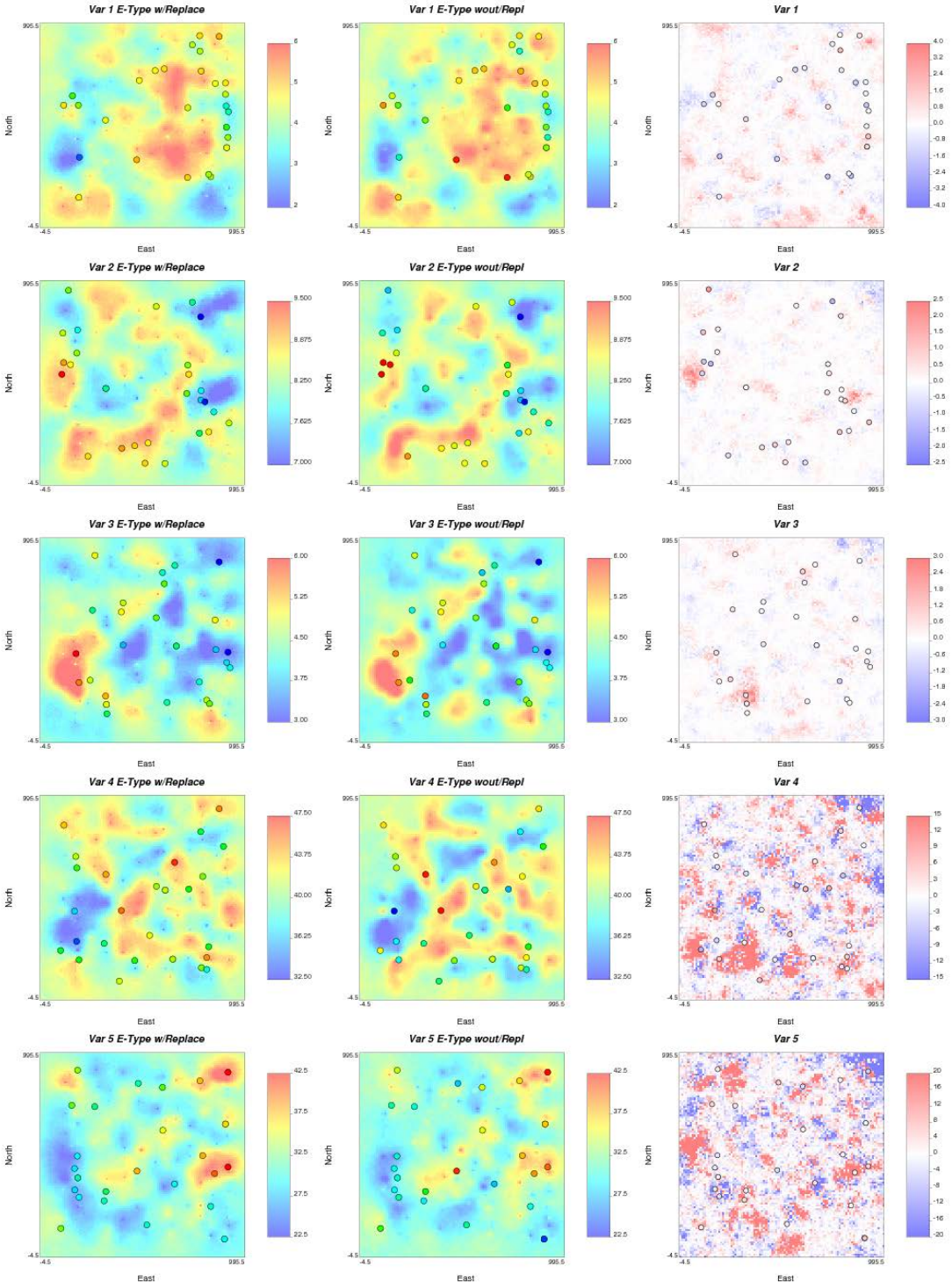


Figure 7: E-type estimate maps for the five variables, as constructed from the data with (left) and without (middle) data replacement. These two maps are compared to the True model (Figure 3) in the far right column, where accuracy improvements in the E-type estimate from data replacement is indicated by red, while estimates which are better without data replacement are indicated by blue. Circle values in these maps display the mean of imputed value realizations (left), True removed values (middle), and mean minus True (right).

Appendix 1: Analytical Value Gained from Data Replacement

While illustrative, the value gained from data replacement need not be determined numerically in the multiGaussian setting. Consider that the accuracy of a multiGaussian simple co-kriging estimate may be approximated by its kriging variance σ^2_{SCK}. Using a simple bivariate setting consisting of two variables *Z1* and *Z2*, suppose that *Z1* is fully sampled at N number of observations, while *Z2* is only sampled in a smaller subset, n number of observations. The value gained from data replacement, as measured by the reduction in MSE may then approximated according to Equation 9.

$$\text{Gain with Data Replacement} = \frac{\text{MSE}_{\text{No Replace}} - \text{MSE}_{\text{Replace}}}{\text{MSE}_{\text{No Replace}}} = \frac{\sigma^2_{SCK^n} - \sigma^2_{SCK^N}}{\sigma^2_{SCK^n}} \tag{9}$$

The value gained from data replacement for the Z1 variable estimation (based on greater secondary conditioning information from the replaced Z2) is given by equation 10, which reduces to Equation 11.

$$\text{Z}_1\text{ Gain} = \frac{\left(\sigma^2_{Z_1} - \sum_{i=1}^{N}\lambda_i\rho_{z_{1i}z_{1\square}} - \sum_{i=1}^{n}\upsilon_i\rho_{z_{2i}z_{1\square}}\right) - \left(\sigma^2_{Z_1} - \sum_{i=1}^{N}\eta_i\rho_{z_{1i}z_{1\square}} - \sum_{i=1}^{N}\psi_i\rho_{z_{2i}z_{1\square}}\right)}{\sigma^2_{Z_1} - \sum_{i=1}^{N}\lambda_i\rho_{z_{1i}z_{1\square}} - \sum_{i=1}^{n}\upsilon_i\rho_{z_{2i}z_{1\square}}} \tag{10}$$

$$\text{Z}_1\text{ Gain} = \frac{\sum_{i=1}^{n}\left(\eta_i - \lambda_i\right)\rho_{z_{1i}z_{1\square}} + \sum_{i=1}^{N}\psi_i\rho_{z_{2i}z_{1\square}} - \sum_{i=1}^{n}\upsilon_i\rho_{z_{2i}z_{1\square}}}{\sigma^2_{Z_1} - \sum_{i=1}^{N}\lambda_i\rho_{z_{1i}z_{1\square}} - \sum_{i=1}^{n}\upsilon_i\rho_{z_{2i}z_{1\square}}} \tag{11}$$

The value gained from data replacement for the Z2 variable estimation (based on greater condition primary information from the replaced Z2) is given by equation 12, which reduces to Equation 13.

$$\text{Z}_2\text{ Gain} = \frac{\left(\sigma^2_{Z_2} - \sum_{i=1}^{n}\lambda_i\rho_{z_{2i}z_{2\square}} - \sum_{i=1}^{N}\upsilon_i\rho_{z_{1i}z_{1\square}}\right) - \left(\sigma^2_{Z_2} - \sum_{i=1}^{N}\eta_i\rho_{z_{2i}z_{2\square}} - \sum_{i=1}^{N}\psi_i\rho_{z_{1i}z_{1\square}}\right)}{\sigma^2_{Z_2} - \sum_{i=1}^{n}\lambda_i\rho_{z_{2i}z_{2\square}} - \sum_{i=1}^{N}\upsilon_i\rho_{z_{1i}z_{1\square}}} \tag{12}$$

$$\text{Z}_2\text{ Gain} = \frac{\sum_{i=1}^{N}\left(\psi_i - \upsilon_i\right)\rho_{z_{1i}z_{2\square}} + \sum_{i=1}^{N}\eta_i\rho_{z_{2i}z_{2\square}} - \sum_{i=1}^{n}\lambda_i\rho_{z_{2i}z_{2\square}}}{\sigma^2_{Z_2} - \sum_{i=1}^{n}\lambda_i\rho_{z_{2i}z_{2\square}} - \sum_{i=1}^{N}\upsilon_i\rho_{z_{1i}z_{1\square}}} \tag{13}$$

Appendix 2: BUDI Software Package

The Bayesian Updating Data Imputation (BUDI) software package is very closely adapted from the original Bayesian Updating FORTRAN software developed by Clayton Deutsch and Weishan Ren for geostatistical modeling. Primary differences from the original code include input files, formats, and parameters, as well as the output files and formats. The first BUDI program is `likelihood_di`, which is used for estimating the likelihood distribution. Parameters are displayed in Figure 9 and given below:

- **outfl:** output file containing the likelihood mean and variance for every location requiring imputation of the primary variable (s)
- **nvar:** total number of variables (primary + secondary)
- **corrfl:** input file containing the correlation between every variable at every location requiring imputation. Correlation is defined at every location because the program adopts locally varying correlation [13] methodology
- **tmin,tmax:** trimming limits that are applied to all variables
- **npred:** number of primary variables to be predicted
- **pvar(i), i=1,…,npred:** location(s) of the primary variable(s) to be predicted in the **corrfl**
- **gridfl**: input file containing data values of the secondary variables
- **ndata:** number of secondary variables

Repeat the following line for **i=1,…,ndata**

- **dvar(i),dcol(i):** location of the **i**th secondary variable in the **corrfl (dvar(i))** and **gridfl (dcol(i))** respectively

```
1                   Parameters for Likelihood_DI
2                   ****************************
3
4
5   START OF PARAMETERS:
6   likelihood.out                  -file for output
7   5                               -number of variables
8   locorrel_data.out               -file with correlations
9   -6.0     6.0                    -trimming limits
10  1                               -number of prior variables
11  1                               -variable number (corr) for prior 1
12  ../2-Nscore/nscore.out          -file with likelihood data
13  4                               -number of likelihood variables
14  2 4                             -variable number (corr),column(data) for likelihood 1
15  3 5                             -variable number (corr),column(data) for likelihood 2
16  4 6                             -variable number (corr),column(data) for likelihood 3
17  5 7                             -variable number (corr),column(data) for likelihood 4
```

Figure 8: Parameter file for the `likelihood_di` program.

The second BUDI program is `update_di`, which is used for estimating the updated distribution. Parameters are displayed in Figure 10 and given below:

- **priorfl:** input file containing the prior mean and variance for every location requiring imputation of the primary variable. This file may be formed using cross-validation mode in `kt3d` [6]
- **icprm,icprv:** columns in **priorfl** for the prior mean and variance
- **likefl:** input file containing the likelihood mean and variance for every location requiring imputation of the primary variable. This file is formed using `likelihood_di`
- **iclim,icliv:** columns in **likefl** for the likelihood mean and variance
- **outfl**: output file containing the updated mean and variance
- **tmin,tmax:** trimming limits that are applied to all variables

```
1                Parameters for UPDATE_DI
2                ************************
3
4  START OF PARAMETERS:
5  ../4-Prior/kt3d1.out            - file with prior distribution
6  5 6                             - columns for mean and variance
7  ../6-Likelihood/likelihood1.out - file with likelihood  distribution
8  1 2                             - columns for mean and variance
9  ../7-Update/Updated1.out        - file for output
10 -9.0    100.0                   - trimming limits
```

Figure 9: Parameter file for the `update_di` program.

The third BUDI program is `busim_di`, which is used for simulating datasets based on sampled data and updated distributions. Parameters are displayed in Figure 11 and given below:

- **datafl(1):** input file containing the original normal score values and any other data that may be of interest (e.g. coordinates). It is this file that will be made into realizations, where missing values are imputed
- **nvar:** number of variables requiring imputation
- **icol(i),i=1,…,nvar:** column locations within **datafl(1)** for the variables to be imputed
- **tmin,tmax:** trimming limits that will determine which values are missing and requiring imputation

Repeat the following line for **i=1,…,nvar**

- **datafl(i):** input file containing the updated distribution for the **i**th variable to be imputed. It is assumed the mean and variance reside in the first and second column respectively
- **nreal:** number of realizations of the data that should be generated
- **ixv(1):** random number seed
- **outname**: prefix name (may include a directory) for the output data realizations. This prefix will have the realization number and '.out' appended to form the final name of each file
- **indfl:** output file containing a binary indicator classification of whether a data value was imputed (1) or not (0)

```
1
2                  Parameters for BUSIM_DI
3                  ***********************
4
5  START OF PARAMETERS:
6  ../2-Nscore/nscore.out          -file with original nscore values
7  5                               -number of variables to simulate
8  3 4 5 6 7                       -columns of variables to simulate
9  -1.0e21   1.0e21                -trimming limits
10 ../7-Update/Updated1.out        -file with updated mean and var. for variable 1
11 ../7-Update/Updated2.out        -file with updated mean and var. for variable 2
12 ../7-Update/Updated3.out        -file with updated mean and var. for variable 3
13 ../7-Update/Updated4.out        -file with updated mean and var. for variable 4
14 ../7-Update/Updated5.out        -file with updated mean and var. for variable 5
15 100                             -number of realizations
16 696969                          -random number seed
17 ./Data_Real/nscore_di           -output files prefix (will append realization number and .out)
18 di_indicator.out                -output file for imputation indicator
```

Figure 10: Parameter file for the `busim_di` program.