A variance-stabilizing transformation to mitigate biased variogram estimation in heterogeneous surfaces with clustered samples

Xiaojun Pu and Michael Tiefelsdorf

The University of Texas at Dallas, 800 W. Campbell Rd, Richardson, Texas, USA
Telephone: 972-883-4954
Email: xxp102020@utdallas.edu
Email: tiefelsdorf@utdallas.edu

Abstract

Due to the inherent variance heterogeneity in clustered preferential sampling the underlying variogram cannot be estimated directly. A variance-stabilizing declustering method is proposed here using a modified Box-Cox transformation. In contrast to the traditional Box-Cox transformation that aims at achieving normally distributed data, its modified version has the objective to match the variance in the clustered sample observations to the variance of the remaining more disperse background sample observations. The proposed approach leads to less biased predictions with lower standard errors than alternative proposed methods.

Keywords: clustered preferential sampling, variance-stabilizing transformation, variogram estimation, Kriging prediction.

1. Introduction

Geostatistical techniques, especially Kriging, are widely used to predict natural features with a continuous spatial distribution. The performance of Kriging critically relies on [a] the variogram estimation, which is supposed to capture the spatial autocorrelation structure within the unknown population values (Richmond 2002; Kovitz and Christakos 2004), and [b] the spatial structure of the sample locations among which Kriging interpolation is performed. The sampling locations for the variogram estimation and those for the subsequent surface prediction do not necessarily need to be identical. The most efficient sampling procedure for the variogram estimation requires capturing reliably the semi-variogram at all relevant inter-sample distances, whereas for the purpose of prediction evenly distributed sample locations across the entire study area are preferred. However, for both the variogram estimation and the prediction, clustered preferential sampling may occur due to external factors such as financial limitations and hostile environmental factors (Olea 2007; Menezes et al. 2008). Sample observations coming from a heterogeneous population, like in clustered preferential sampling, lead to compromised variogram estimates. Usually the variability within a cluster will be substantially larger at short distances than that for the remainder of the sample points in the less variable study area. In particular at short distances, this local variance heterogeneity can lead to unrepresentative joint variogram estimation. In clustered preferential sampling the status whether a sample observation belongs to a cluster or the remainder of the study area is well identified, however, when predicting data values in-between the sample locations this status is generally unknown.

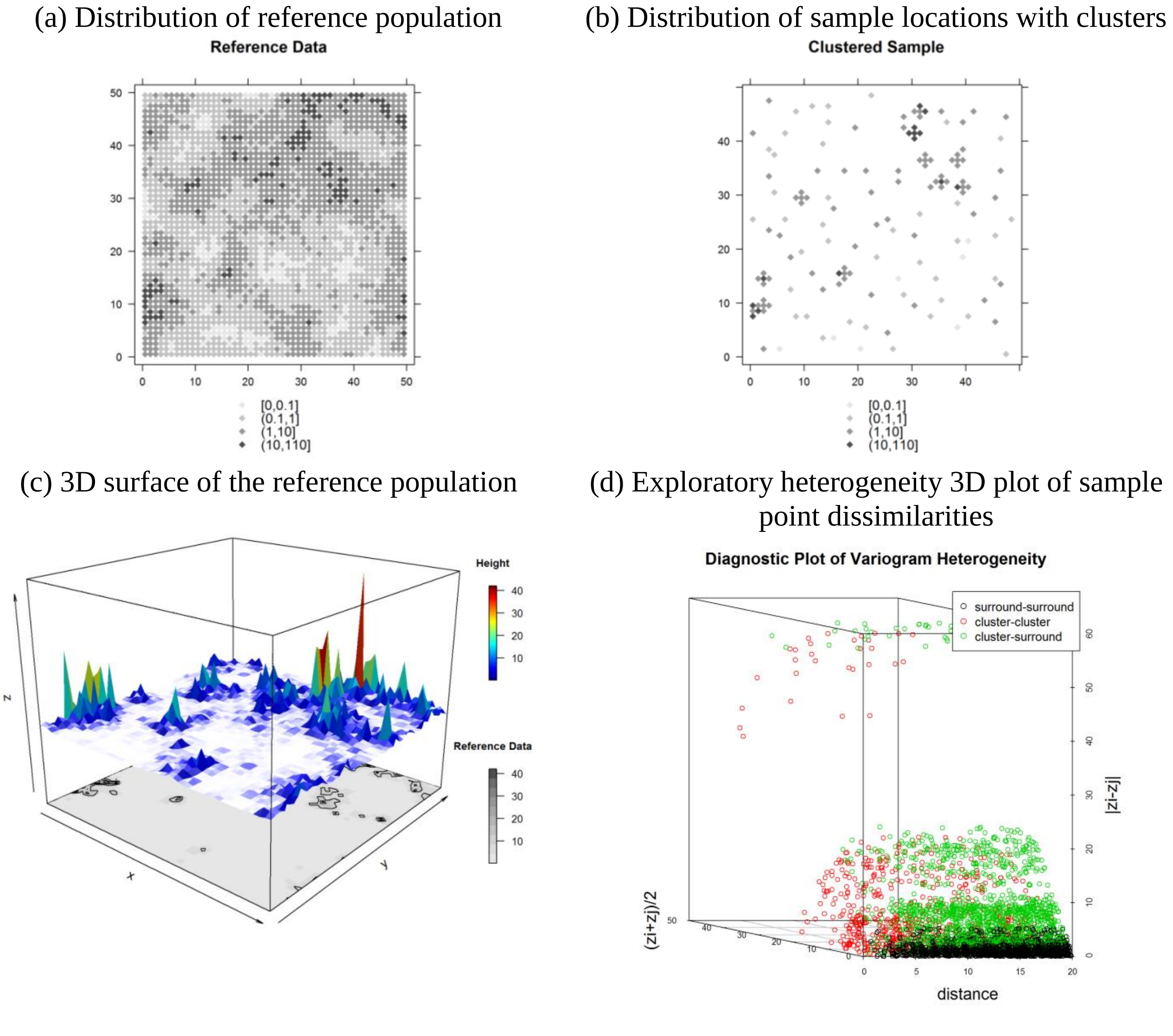
To avoid the problem caused by clustered preferential sampling, a number of declustering methods have been proposed to improve variogram estimation. There are two major branches to control the induced bias in the variogram estimation: the calibration of weighted variograms (Bourgault 1997; Richmond 2002; Kovitz and Christakos 2004; Menezes et al. 2008) and the sub-sampling approach (Olea 2007). Most researchers adopt the weighted variogram approach, such as using the ratio of correlation matrices determinants (Bourgault 1997), two-point declustering method based on cells or clusters (Richmond 2002), declustering weights based on zones of proximity (Kovitzand and Christakos 2004), or the robust kernel variogram estimator (Menezes et al. 2008). The two-point declustering method (Richmond 2002) will be conducted for comparison purposes to evaluate the performance of the proposed method. Regarding grids or clusters as units, Richmond's method (2002) counts the number of point pairs ($\boldsymbol{n}$) for a certain distance ($\boldsymbol{d}$) between two grids ($\boldsymbol{g_i}$ and $\boldsymbol{g_{i'}}$) or two clusters ($\boldsymbol{c_i}$ and $\boldsymbol{c_{i'}}$). Then assign the inverse proportion of the total pair number $\boldsymbol{w_{\alpha\alpha'}} = 1/\boldsymbol{n}$ as the weights to those specified pairs to adjust variogram. Moreover, the sub-sampling approach (Olea 2007), which is to pick up sub-samples free of clusters to build a representative histogram, will be used for evaluation as well. Samples are divided into two subsets: subset 1 (free of clusters) and subset 2 (only including clusters). Based on the maximum nearest neighborhood distance between the two subsets, some points are moved from subset 2 to subset 1 iteratively. The distance distribution of the expanded subset 1 should match that of the original subset 1. Then the expanded subset 1 can be used to model variogram for prediction purposes.

The proposed method in this paper adopts the Box-Cox transformation to improve the variogram estimation for a heterogeneous surface when clustered preferential sampling usually is applied. Instead of the traditional Box-Cox transformation, which treats the sample as coming from a homogeneous population and which aims at achieving a symmetric or Gaussian distribution of the transformed sample observations, the objective of the proposed method is to stabilize the variance within the clustered sample points and those associated with the remainder of the study area.

2. Methodology

2.1 Data

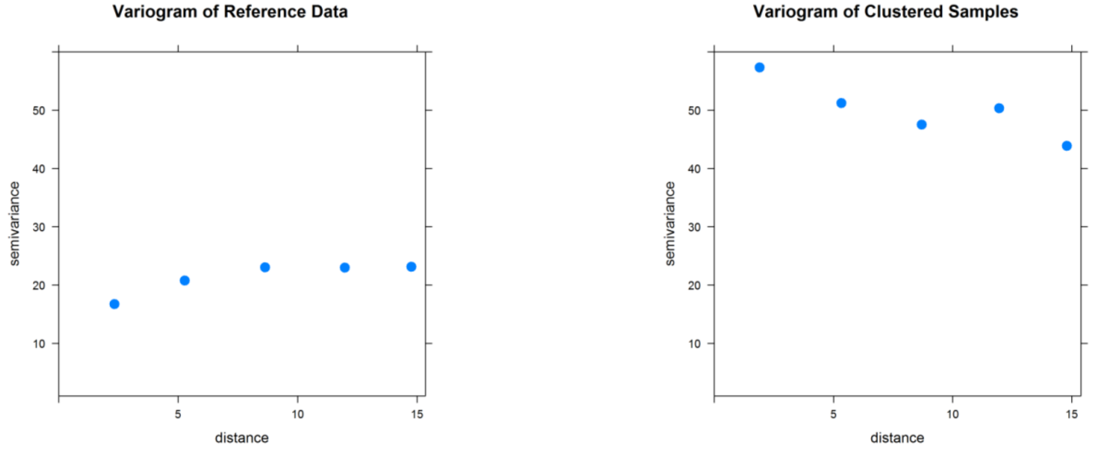
Clustered data from the GSLIB (Deutsch and Journel 1997) have been used in the two-point declustering method (Richmond 2002) and Olea's sub-sampling approach (2007). There were in total 140 sample observations (Figure 1b) selected from a 50×50 regular grid image (Figure 1a), including 86 single points (subsample 1) in the disperse study area and 10 clusters (subsample 2). Most clusters consist of 5 sample points with distance of 1 around a central sample point with the exception one 9-point cluster, which is a combination of two 5-point clusters with one edge point being cut-off (Olea 2007). Figure 1c displays the 3D map of the reference population data with sharp high peaks representing clusters of large values. Figure 1d displays the abnormal variogram cloud of sample point pairs in 3 dimensions: the spatial distance between two sample points (z_i and z_j), the average attribute value $(z_i + z_j)/2$ identifying the different baseline levels in the cluster and the surrounding background data values, and the absolute difference between two sample points $|z_i - z_j|$ measuring their dissimilarity. There are three groups of point pairs identified in the variogram cloud: [a] the relationship between clustered-surrounding sample point pairs in green, [b] clustered-clustered point pairs in red, and [c] surrounding-surrounding point pairs in black. Due to the larger variance in the clustered-surrounding group and clustered-clustered group, the variogram cloud does not display the usually increasing attribute dissimilarity trend with increasing inter-point distance. Compared to the variogram of the full referenced population (Figure 1e), the calibrated variogram of the clustered sample (Figure 1f) displays an irregular pattern with the high dissimilarities at the small distance intervals.


(a) Distribution of reference population

(b) Distribution of sample locations with clusters

(c) 3D surface of the reference population

(d) Exploratory heterogeneity 3D plot of sample point dissimilarities

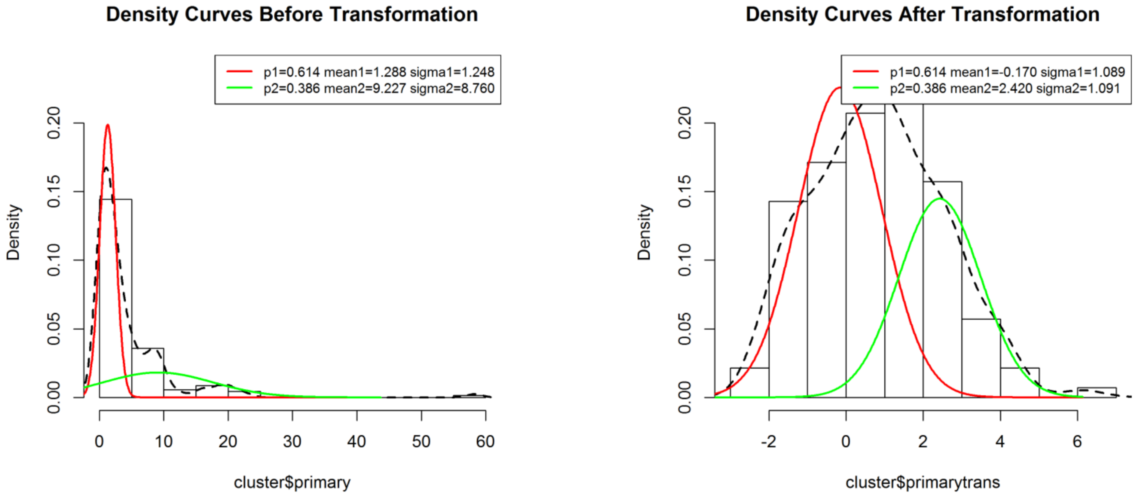

(e) Variogram of reference population

(f) Variogram of sampling locations with clusters

Figure 1 Exploratory data analysis including variogram estimation of the untransformed data.

2.2 Box-Cox transformation and Kriging prediction

The observed sample values range from 0.06 to 58.32. However, nearly half of the observations are smaller than 1, which implies two underlying different distributions are contributing to this heterogeneous surface. Most of single points are from the sub-population with small mean and small variance, whereas the clusters are from the sub-population with large mean and large variance (Figure 2a). Therefore, the dissimilarities captured at short distances cannot reflect the true differences of the entire study area. To address this problem, we can conduct a Box-Cox transformation (Equation 1) with the objective of making the variances of both sub-populations as similar as possible (Figure 2b).


(a) Density curves before transformation

(b) Density curves after transformation

Figure 2 Distributions of two populations

The equation of Box-Cox transformation is:

$$y_i^{(\lambda)} = \begin{cases} (y_i^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \ln y_i & \lambda = 0 \end{cases} \tag{1}$$

where y_i denotes the sample data i on the original scale which needs to be transformed and λ is the optimal transformed power. Note that we were not seeking a transformation parameter λ here to make the joint distribution symmetric but the optimal λ-value was identified iteratively so that both sub-samples exhibit comparable variances. In this case the optimal transformation parameter is $\lambda = 0.19$ (see Figure 2b). We regard sub-samples 1 and 2 to be representative for the underlying heterogeneous population. After the optimal λ has been identified, we calibrate the variogram model parameters on the transformed scale and conduct ordinary kriging to predict the interpolated values for the whole study area. In addition to these predicted values on the transformed scale, also prediction standard errors and the 95% confidence interval bounds ($CI_{0.025}$ and $CI_{0.975}$) are calculated for each prediction location.

The back-transformation (see Equations 2 and 3) into the original data scale yields the expected predicted values $E(Y)$, confidence interval bounds $E(CI_{0.025})$ and $E(CI_{0.975})$ as well as prediction standard errors $\sqrt{Var(Y)}$. In Equations 2 and 3 μ_λ is the predicted value in the transformed scale and $\sigma^2_{(\lambda)}$ is the local prediction variance at each predicted location (Tiefelsdorf 2013).

$$E(Y) \approx (\lambda \cdot \mu_\lambda + 1)^{1/\lambda} \cdot \left(1 + \frac{1}{2} \cdot \sigma^2_{(\lambda)} \cdot \frac{(1-\lambda)}{\left(\lambda \cdot \mu_{(\lambda)} + 1\right)^2}\right) \tag{2}$$

$$Var(Y) \approx \sigma^2_{(\lambda)} \cdot \left(\lambda \cdot \mu_{(\lambda)} + 1\right)^{\frac{2}{\lambda} - 2} \tag{3}$$

Note that both expressions are derived from a truncated Taylor-series expansion. In particular the variance expression may exert a noticeable truncation error (Tiefelsdorf 2013). Therefore, the back-transformed confidence interval is also reported. In order to remain consistent with the geo-statistical practice and software implementations, the expectation rather than the median were used.

The variogram modeling parameters (Richmond 2002; Olea 2007) of three earlier methods and those of the proposed variance-stabilizing approach are displayed in Table 1. Note that (Richmond 2002; Olea 2007) by default applied a log-transformation on the original data, which is equivalent to choosing $\lambda = 0$, before calibrating their variograms. In contrast, the optimal variance stabilizing transformation parameter for these sample observation is $\lambda = 0.19$.

	Model type	Nugget	Partial Sill	Range
Cell (Richmond 2002)	Spherical	0.10	1.90	9.75
Cluster (Richmond 2002)	Spherical	0.10	1.90	10.25
Sub-sampling (Olea 2007)	Spherical	0.06	1.77	9.52
Variance-stabilization	*Spherical*	*0.56*	*2.07*	*11.41*

Table 1: Variogram parameter estimates of the different declustering methods

3. Preliminary results

The distributional characteristics of the predicted values in the original measurement scale of the four declustering methods are different from the true characteristics of the underlying reference population data (Table 2). The means and medians of the predicted back-transformed values for all four methods are biased but the mean and skewness of the proposed variance-stabilizing method are closer to those of the reference population.

	Mean	1st Quartile	Median	3rd Quartile	Standard Deviation	Skewness
Reference population	***2.58***	***0.34***	***0.96***	***2.56***	***5.15***	***6.83***
Cell (Richmond 2002)	2.43	0.89	1.70	3.12	2.65	5.85
Cluster (Richmond 2002)	2.40	0.86	1.65	3.07	2.66	5.80
Sub-sampling (Olea 2007)	2.33	0.84	1.63	2.97	2.60	6.16
Variance-stabilization	*2.42*	*1.11*	*1.80*	*2.89*	*2.45*	*6.99*

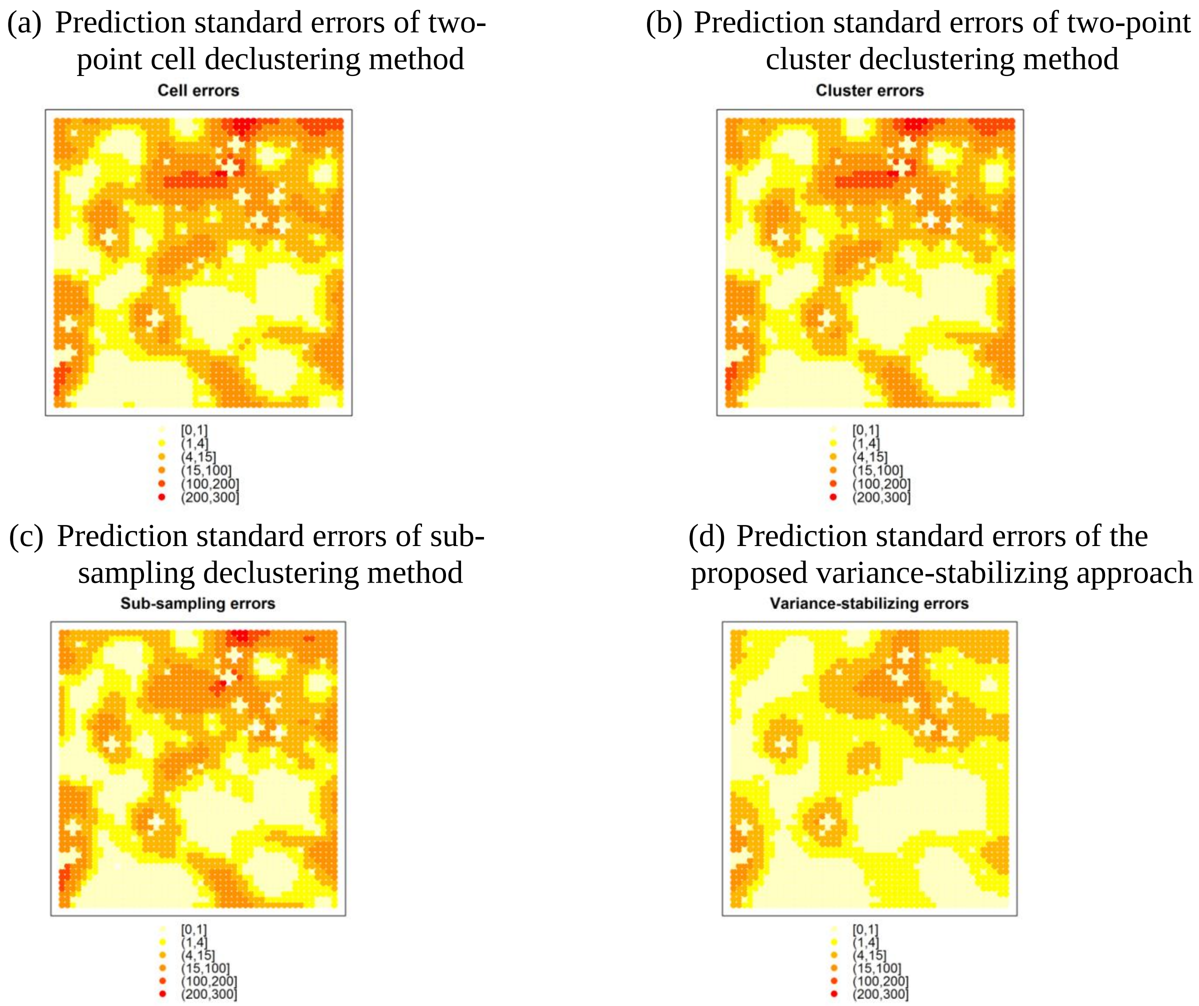
Table 2: Comparison of distributional characteristics for the four declustering methods against those of the true reference distribution

After the back-transformation of the predicted interpolation values into the original measurement scale the aggregated width of the confidence intervals $\sum(CI_{0.975} - CI_{0.025})$, root mean square errors (RMSE) and sum of prediction standard errors are calculated as the evaluation criterion. Although the predictions based on the proposed variance stabilizing transformation have the largest RMSE, the aggregated widths of the confidence intervals and the sum of local predication standard errors are much lower than those of the other three methods (Table 3).

	RMSE	$\sum(Q_{95\%} - Q_{5\%})$	Sum of uncertainty
Cell (Richmond 2002)	3.98	36936.81	37762.89
Cluster (Richmond 2002)	3.98	34962.81	35410.85
Sub-sampling (Olea 2007)	4.00	31798.75	29281.93
Variance-stabilization	*4.16*	*26602.85*	*9504.08*

Table 3: Comparison of predicted accuracy and uncertainty of four declustering methods

Figure 3 shows the prediction standard errors of the four declustering methods. Figure 3d displays that the proposed variance-stabilizing approach also decreases the prediction uncertainty in non-clustered sub-regions. In other words, the method proposed in this paper achieves the highest prediction certainty which improves the prediction accuracy.



(a) Prediction standard errors of two-point cell declustering method

(b) Prediction standard errors of two-point cluster declustering method

(c) Prediction standard errors of sub-sampling declustering method

(d) Prediction standard errors of the proposed variance-stabilizing approach

Figure 3: Predicted uncertainty of four declustering methods

4. Conclusions

Highly heterogeneous spatial surfaces are not only observed in the natural sciences, they emerge also in other disciplines. For instance, the population density will be mainly flat in rural regions and then will peak sharply once one enters urbanized areas; alternatively, air pollution measurement stations will focus on areas with a high emission potential. Methods to representatively sample these heterogeneous surfaces and subsequently to perform accurate interpolations are highly relevant. The proposed variance stabilizing transformation is one approach to handle these diverse scenarios, which at least as long as one evaluates dissimilarities within each sub-sample are resistant to the detrimental effects of incongruity. However, the authors feel that so far none of the proposed methods have satisfactorily addressed this interpolation challenge without additional exogenous information to capture the dichotomy of clusters and their surrounding background surface. In order to gain insights into how to handle and model these kind interpolation scenarios most appropriately, well-designed simulation studies are a required starting point.

5. References

Bourgault, G., 1997, “Spatial declustering weights”, *Mathematical Geology*, 29(2): 277-290.

Deutsch, C.V., Journel, A.G., 1997, *GSLIB: Geostatistical Software Library and User’s Guide*, 2nd edition. Oxford University Press, New York, 369.

Richmond, A., 2002, “Two- point declustering for weighting data pairs in experimental variogram calculations”, *Computers and Geosciences*, 28(2): 231-241.

Kovitz, J.L. & Christakos, G., 2004, “Spatial statistics of clustered data”, *Stochastic Environmental Research and Risk Assessment,* 18(3): 147-166.

Olea, R.A., 2007, “Declustering of Clustered Preferential Sampling for Histogram and Semivariogram Inference”, *Mathematical Geology*, 39: 453-467.

Menezes, R., Garcia-Soidán, P. & Febrero-Bande, M., 2008, “A kernel variogram estimator for clustered data”, *Scandinavian Journal of Statistics,* 35(1): 18-37.

Tiefelsdorf, M., 2013, “A note on the reverse Box-Cox transformation”, Unpublished lecture notes.