Stoch Environ Res Risk Assess (2017) 31:481–491
DOI 10.1007/s00477-016-1289-4

ORIGINAL PAPER

Resampling of spatially correlated data with preferential sampling for the estimation of frequency distributions and semivariograms

Ricardo A. Olea[1]

Published online: 9 July 2016

Abstract Spatial data are commonly minimal and may have been collected in the process of confirming the profitability of a mining venture or investigating a contaminated site. In such situations, it is common to have measurements preferentially taken in the most critical areas (sweet spots, allegedly contaminated areas), thus conditionally biasing the sample. While preferential sampling makes good practical sense, its direct use leads to distorted sample moments and percentiles. Spatial clusters are a problem that has been identified in the past and solved with approaches ranging from ad hoc solutions to highly elaborate mathematical formulations, covering mostly the effect of clustering on the cumulative frequency distribution. The method proposed here is a form of resample, free of special assumptions, does not use weights to ponder the measurements, does not find solutions by successive approximation and provides variability in the results. The new method is illustrated with a synthetic dataset with an exponential semivariogram and purposely generated to follow a lognormal distribution. The lognormal distribution is both difficult to work with and typical of many attributes of practical interest. Testing of the new solution shows that sample subsets derived from resampled datasets can closely approximate the true probability distribution and the semivariogram, clearly outperforming the original preferentially sampled data.

Keywords Geostatistics · Population · Cluster · Declustering · Bias

✉ Ricardo A. Olea
rolea@usgs.gov

[1] U.S. Geological Survey, 12201 Sunrise Valley Drive, Mail Stop 956, Reston, VA 20192, USA

1 Introduction

Sampling is the process of collecting a limited number of measurements from a population for the purpose of making inferences about such a population. A sampling is said to be preferential when targeting a certain population class at a higher rate than in the underlying frequency. In spatial statistics, such practice results in clusters of sampling locations. Some estimation methods, such as kriging, have a built-in capability to handle preferential sampling, but most other formulations and statistical procedures do not, thus requiring a preprocessing of the data to produce correct results. This is the case of the estimation of the underlying cumulative distribution and the semivariogram.

The problems associated with clustered sampling have been known for some time, with several fixes being formulated. One of the earliest solutions is that of Journel (1983) who proposed to assign a weight to each measurement that is inversely proportional to the number of observations per cell in a regular tessellation of the sampling domain. This method requires calculating several weights sets for different cell sizes. When the preferential sampling favors high values, the solution is the set of weights associated with the cell size that produces the minimum mean cell value. The converse is true for the case favoring low values (Deutsch 1989). The method is heuristic, with minima that are not always clear cut, thus, not guaranteeing an optimal solution. Another early approach uses the data locations to prepare Voronoi polyhedra—polygons in the more common two-dimensional case—and calculate weights now proportional to the volume of the polyhedra (Isaac and Srivastava 1989). One of the main disadvantages of the approach is the large weights assigned to locations near the periphery of the study areas. There are several other more mathematically elaborate but

less frequently applied methods in addition to these two geometrical approaches (Switzer 1977; Omre 1984; Bourgault 1997; Bogaert 1999; Rivoirard 2001; Richmond 2002; Kovitz and Christakos 2004; Pardo-Igúzquiza and Dowd 2004; Emery and Ortiz 2005, 2007; Pyrcz et al. 2006; Reilly and Gelman 2007; Diggle et al. 2010; Marchant et al. 2013; Pyrcz and Deutsch 2014). Most formulations share the disadvantage of having to deal with weights, thus limiting the computer software able to handle a declustered data carrying weights. Some methods are valid only for estimating the frequency distribution, but not the semivariogram, or vice versa. Above all, none of these methods tries to extract some benefit out of the preferential sampling.

In a previous study (Olea 2007), a methodology was proposed that resulted in retaining one observation per cluster. The solution works, but it can be considered suboptimal from the point of view of the usage of the data. Here, in addition to providing a better estimation of the histogram or cumulative distribution and semivariogram, the preferential sampling is used to evaluate uncertainty in the modeling by using multiple versions of the original data according to the procedures below. The objective of this study is to present a method that (a) produces a declustered sample without resorting to weights so that the solution can be handled by a larger number of software applications, (b) generates a declustered sample that can be used to model both the frequency distribution and the semivariogram, and (c) uses the clustered data to provide a measure of uncertainty in the results.

2 Methodology

Preferential sampling of a regionalized variable implies selection of data locations intending to target certain range of values of the underlying attribute, say, high values (Diggle et al. 2010). While the practice may have justifications, such as higher mining venture profit, it has some drawbacks as a sampling practice. The solution to this often forced situation is to preprocess the data to prepare a non-preferential sample adequate for those operations in the modeling that are distorted by preferential sampling. In two-point geostatistics, such operations involve the estimation of the population cumulative distribution and the semivariogram. Further stages in the modeling, such as kriging and stochastic simulation, can properly handle preferential samples. Hence, use of declustered datasets should be limited to the estimation of the cumulative frequency and the semivariogram. Then, the modeler should go back to using the original preferential sampling for running kriging or a simulation.

Given the way a preferential sample is collected, only a few locations actually exhibit a preferential selection because the common practice is to predominantly have a non-preferential sampling where a measurement is not expected to result in a value satisfying the special requirement, say, the observation is high. Hence, spatial data should be split into two classes. Spatially scattered observations, $\mathbf{z}_s(\mathbf{s}_i)$, at location $\mathbf{s}_i \in \Omega$ across the region of interest are the first type of data; they should be unbiased outright, thus supposedly not requiring preprocessing, assumption that should not be taken for granted. Let $\mathbf{z}_c(\mathbf{s}_i)$ be the remainder of the data that were preferentially sampled and are not randomly scattered. Considering that preferential sampling results in clusters of data locations, simultaneous scrutiny by attribute value and distance to the closest neighbor should reveal the data in need of preprocessing.

2.1 Declustering procedure

Let subset $\mathbf{z}_s(\mathbf{s}_i)$ be of size n_s, let c be the number of clusters, and let M be an odd number of resamples. Then:

Step 1 Prepare a cumulative distribution of distance to the closest neighbor for the entire sample.
Step 2 Look for a sudden break in the distribution; this is the critical distance to split the data into two classes: scattered locations and clusters.
Step 3 Prepare a Q–Q plot to confirm that the distributions of the attribute for the two classes are indeed different (e.g., Olea 2008). If not, stop because there is no preferential sampling; mere clustering does not distort the estimation of the cumulative distribution and the semivariogram. Otherwise, continue.
Step 4 Set a counter, k, equal to 1.
Step 5 Prepare resample dataset $\mathbf{z}_k(\mathbf{s}_i)$ by copying all n_s values in subset $\mathbf{z}_s(\mathbf{s}_i)$.
Step 6 From each of the c clusters within $\mathbf{z}_c(\mathbf{s}_i)$, select at random one value per cluster and add it to $\mathbf{z}_k(\mathbf{s}_i)$, thus resulting in a subset of size $n_s + c$.
Step 7 Increase k by 1.
Step 8 If $k < M$, go to Step 5. Otherwise, stop.

The set of M resamples $\mathbf{z}_k(\mathbf{s}_i)$ is the input data to be used in the estimation of the cumulative distribution and semivariogram. In case the clusters are of significantly different sizes, Step 6 can be generalized by creating a rule to draw values according to cluster size instead of always taking one observation per cluster. For example, the average distance between sampling locations, d_a, can be used to define an area d_a^2. One can retain one measurement per multiple of $\kappa \cdot d_a^2$ in the areas with clusters, where κ is a scaling constant to be set by the modeler.

2.2 Modeling of the cumulative distribution

Each of the M resamples $\mathbf{z}_k(\mathbf{s}_i)$ can be regarded as a partial realization of an unknown random function. Do the following with these data:

Step 1 Sort each of the M resamples $\mathbf{z}_k(\mathbf{s}_i)$ and prepare a table in which each column is one of the resamples.
Step 2 Find the median of the quantile at each of the $n_s + c$ rows and identify the observations matching the value. If the matching is not unique, select one observation randomly.
Step 3 For easier visualization of the results, prepare a joint display of the cumulative frequencies for all M resamples and the median values.

Considering that the number of resamples is odd, the median for any row is always exactly the $((M+1)/2)$th value by magnitude. Because there is no interpolation, each median is one of the values in the dataset. The process of obtaining the media is trivial for the rows away from the values preferentially sampled because all values are the same, but variability increases approaching the values resampled from the clusters.

This set of medians will be collectively denoted by $\mathbf{z}_q(\mathbf{s}_i), \quad i = 1, 2, \ldots, n_s + c$. The median was selected over the mean for two reasons: (a) the median is the minimum absolute error estimate of the true quantile, thus less sensitive large discrepancies, and (b) in general, the mean does not coincide with the value of any observation.

2.3 Modeling of the empirical semivariogram

When it comes to estimating the empirical semivariogram, there are two alternatives: use the sample $\mathbf{z}_q(\mathbf{s}_i)$ or use all the M resamples. Here, estimation of the semivariogram depends on the same assumptions that apply to samples without preferential sampling, such as, minimum size guaranteeing a sufficient numbers pairs of data for a reliable estimations and some form of stationarity (Olea 2006; Chilès and Delfiner 2012).

2.3.1 Semivariogram for the quantiles

Step 1 Chose an estimator to calculate the empirical semivariogram and select all necessary parameters, such as direction and distance increment.
Step 2 Estimate the empirical semivariogram γ_p^* using sample $\mathbf{z}_q(\mathbf{s}_i)$.
Step 3 Display the results.

This solution is straightforward, but it has the inconvenience of not taking advantage of all values in the clusters to model uncertainty in the results.

2.3.2 Semivariogram for the resamples

This approach is more demanding but provides a dispersion of the results.

Step 1 Chose an estimator to calculate the empirical semivariogram and select all necessary parameters, such as direction and distance increment.
Step 2 Estimate the empirical semivariogram for each one of the M resamples $\mathbf{z}_k(\mathbf{s}_i)$.
Step 3 For each distance class, select the median value that collectively provides an estimate γ_k^*.
Step 4 Graphically display each resampled semivariogram and estimate γ_k^*.

3 Case study

3.1 Preparation of the data

This section shares a synthetic example that I prepared for helping to clarify ideas and illustrate the methodology. Real examples have the inconvenience of containing properties not always possible to reproduce mathematically. The main disadvantage is that, unless the sampling is exhaustive, the answer is unknown, thus preventing the ability to compare results to the target population. In real life, sampling to exhaustion is costly, time consuming and impractical. For these reasons, a synthetic dataset is used here.

Figure 1 shows the pixel map and histogram of a synthetic exhaustive sample especially prepared to have both an adequate and challenging dataset to model. There are 45 rows and columns of pixels in the map, thus the sample size is 2025. In particular:

- The attribute is isotropic.
- The attribute is second order stationary.
- The study area is a square.
- The side of the square is more than 2.5 times the size of the semivariogram effective range.
- The attribute follows a positively skewed distribution.
- For the sake of generality, no units are specified for distance and for the attribute.
- A minor requirement is to have the attribute in the range (0, 100) primarily to facilitate display.

Anisotropy and lack of stationarity are not central problems in the estimation of the cumulative distribution or the semivariogram; here they have been avoided to focus on important issues. The condition of a square study area is consistent with the isotropy requirement. The proportion (side of the study area)/range is necessary to properly investigate the semivariogram range. Skewed distributions

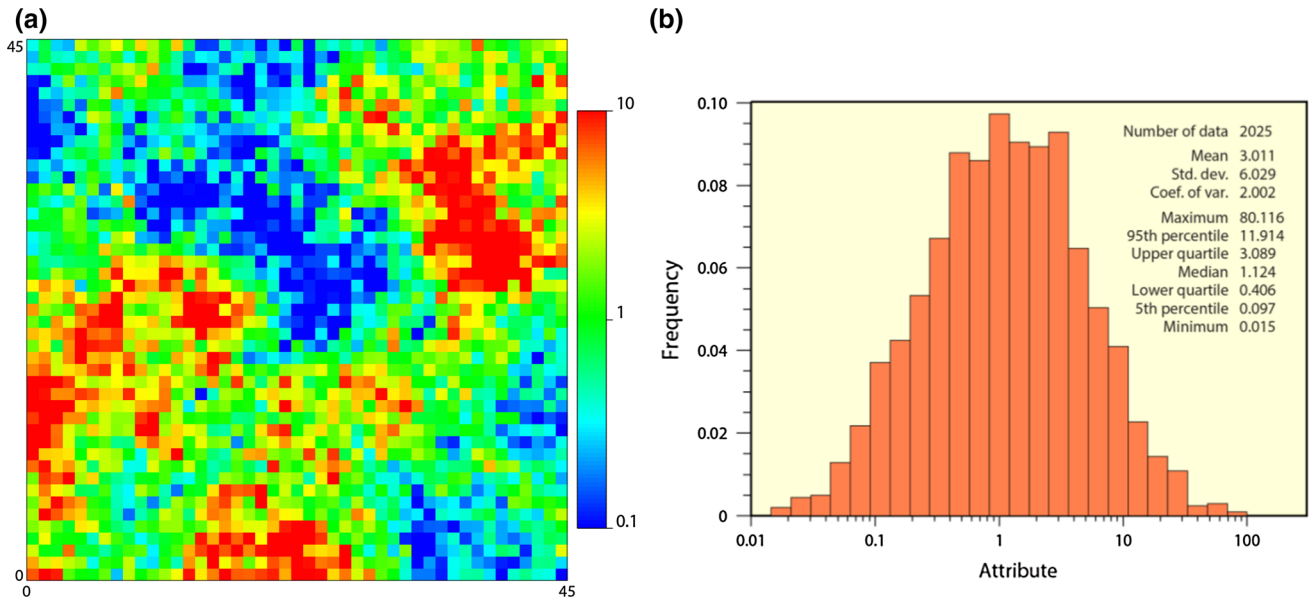

Fig. 1 Synthetic exhaustive example used to illustrate the methodology: **a** pixel map; **b** histogram

are more difficult to model than symmetric ones. Originally the exhaustive sample was generated as a normally distributed realization and then it was skewed through a logarithmic transformation. This exhaustive sample will be used exclusively to evaluate results, not to assist the estimation in any manner.

Important considerations about the dataset to be used in the modeling are:

- The sample size after declustering should be below 100 to have a challenging semivariogram modeling (Webster and Oliver 1992);
- Before starting the preferential sampling, a first set of observations was drawn to conform a stratified sample by taking at random one value within squares of 5 by 5 pixels;
- The preferential drawing was prepared by taking four values immediately North, South, East and West. About 10 clusters were considered a reasonable number for this exhaustive sample. 11 clusters resulted by preferentially sampling all sites in the stratified sample with a value above 6;
- Below the semivariogram range, it should be possible to have at least four distance classes at regular intervals and with enough pairs of data to calculate empirical semivariograms.

I am purposely trying to avoid blaming the preparation of the sample for failures in performance by the methodology. In most cases, a stratified sampling is intermediate in efficiency between a regular and a random sampling (Webster and Oliver 2007; Chilès and Delfiner 2012). Hence, a stratified sample is neither the best configuration nor a subpar option. The minimum number of distance classes in the estimation of the empirical semivariogram is another requirement to make sure that, if the method does not perform well, it is not because of a trivial problem. Figure 2 contains graphical displays for the preferential sample.

3.2 Declustering the data

Figure 3 reveals the two necessary features to have a preferential sample that, in this case, we already know by construction: there is clustering for a distance below 1.4 and the probability distribution for the clusters and the scattered locations are markedly different. In this example, as it is often the case, in the preparation of the dataset, there has been a deliberate attempt to better sample the upper tail of the frequency distribution. Consequently, upon reaching Step 3 in the procedure in Sect. 2.1, it is confirmed that there is indeed a case of preferential sampling.

In this case, $n_s = 70$ and $c = 11$. I decided to prepare 101 resamples, so $M = 101$. The final product upon applying the procedure in Sect. 2.1 is a set of 101 resamples $\mathbf{z}_k(\mathbf{s}_i)$, each of size 81. Table 1 displays the results upon completing the procedure in Sect. 2.1, but because of the limitation of space, there is only a partial display. Row 71 is the first one displaying values taken from the clusters. Note that there is a limit of no more than 5 unique values per row because that is the size of all clusters that are resampled.

3.3 Estimation of the cumulative distribution

Because of space limitations, Table 2 is a partial display of the complete tabulation obtained after completing Steps 1

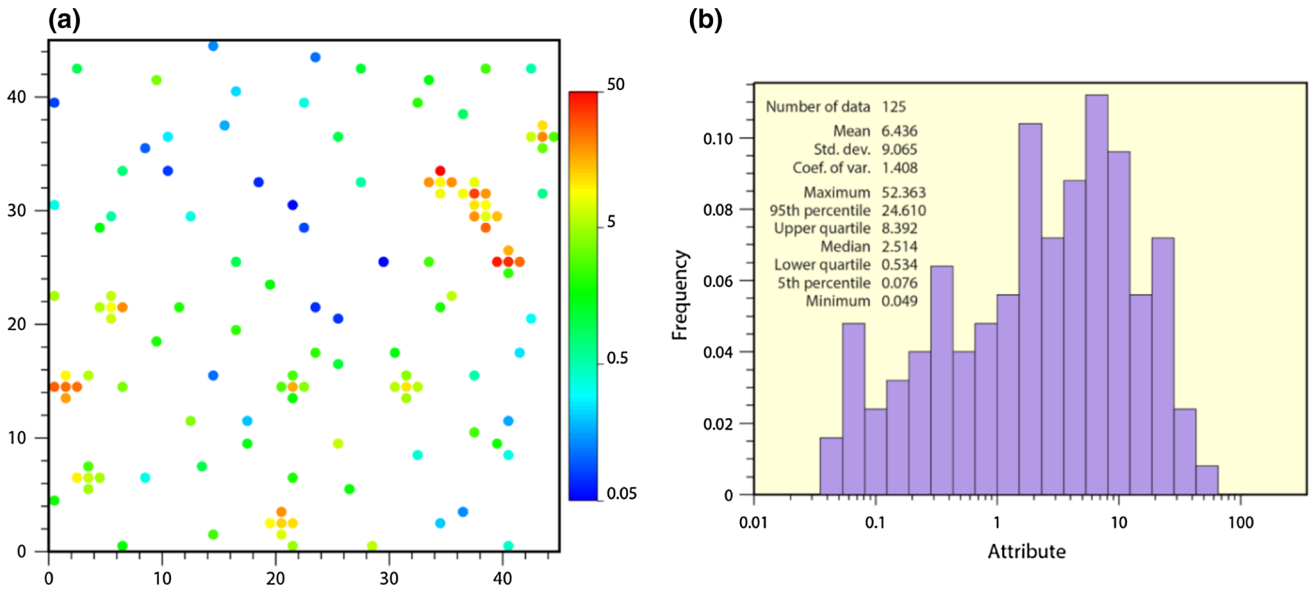

Fig. 2 The preferential sampled subset of the full synthetic dataset: **a** posting of data locations; **b** histogram

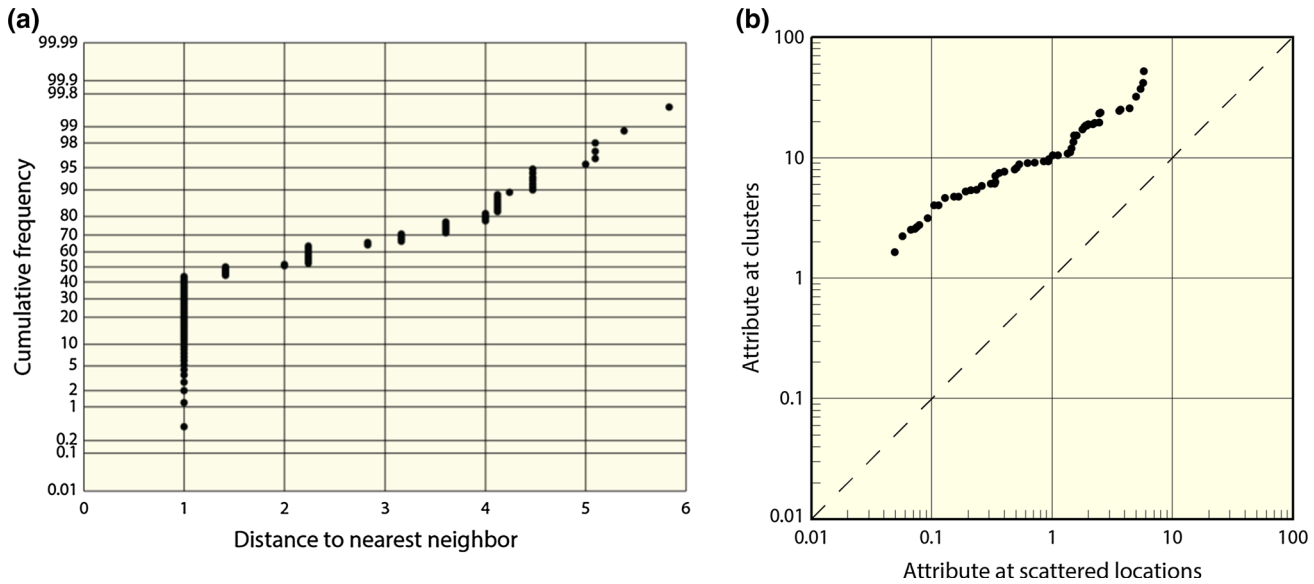

Fig. 3 Confirmation of preferential sampling: **a** cumulative distribution of distance to closest neighbor; **b** Q–Q plot

and 2 of the procedure in Sect. 2.2. Now all observations in a resample are sorted by increasing value. The lowest value in the clusters is 1.632, which is smaller than the highest value of 5.813 among the scattered locations. Hence, lateral change in values starts earlier than in Table 1. The values under "Row median" refer to $\mathbf{z}_q(\mathbf{s}_i)$.

Figure 4 is the graphical summary of Step 3, Sect. 2.2. As seen in Table 2, below row 48 all values in a row are equal, so are the resamples and the median. Figure 4 shows all those values coded as scattered observations. Dispersion in values is not noticeable until the 77th percentile $(\approx 100 \cdot 62/81)$ and is only important above the 85th percentile $(\approx 100 \cdot 69/81)$. Expanding the dispersion to incorporate uncertainty in the range of values covered by the scattered data would require bootstrapping them.

Figure 5 is a posting of all observations in the last column of Table 2 making the solution $\mathbf{z}_q(\mathbf{s}_i)$ to the estimation of the cumulative distribution. Note that three of the clusters retained two observations, while another three clusters are not represented. This is a result of overlapping in the intervals of values for scattered and clustered locations (Fig. 3b), always a realistic possibility. Hence, clustering is not completely precluded in what is called here the "declustered" solution.

Table 1 The 101 resamples, each one of size 81

Seq. no.	$\mathbf{z}_1(\mathbf{s}_i)$	$\mathbf{z}_2(\mathbf{s}_i)$	$\mathbf{z}_3(\mathbf{s}_i)$	…	$\mathbf{z}_{67}(\mathbf{s}_i)$	$\mathbf{z}_{68}(\mathbf{s}_i)$	$\mathbf{z}_{69}(\mathbf{s}_i)$	…	$\mathbf{z}_{99}(\mathbf{s}_i)$	$\mathbf{z}_{100}(\mathbf{s}_i)$	$\mathbf{z}_{101}(\mathbf{s}_i)$
1	0.049	0.049	0.049	…	0.049	0.049	0.049	…	0.049	0.049	0.049
2	0.051	0.051	0.051	…	0.051	0.051	0.051	…	0.051	0.051	0.051
3	0.066	0.066	0.066	…	0.066	0.066	0.066	…	0.066	0.066	0.066
…	…	…	…	…	…	…	…	…	…	…	…
69	5.744	5.744	5.744	…	5.744	5.744	5.744	…	5.744	5.744	5.744
70	5.813	5.813	5.813	…	5.813	5.813	5.813	…	5.813	5.813	5.813
71	9.692	2.514	9.692	…	2.514	4.640	4.640	…	6.043	4.722	6.043
72	18.537	25.256	18.537	…	7.496	9.003	25.256	…	13.603	13.603	7.496
73	6.287	6.287	19.406	…	19.406	6.287	5.262	…	8.755	6.287	19.406
74	18.949	18.949	18.949	…	9.358	52.363	18.390	…	7.960	9.358	18.949
75	5.378	5.430	10.523	…	5.430	4.759	5.378	…	4.047	4.047	5.378
76	18.887	18.887	11.910	…	18.887	7.048	18.887	…	11.910	18.887	9.133
77	1.632	1.632	2.556	…	2.671	4.026	4.026	…	4.026	4.026	1.632
78	3.150	6.110	19.641	…	2.747	19.641	10.889	…	3.150	19.641	6.110
79	23.356	17.121	23.676	…	23.356	23.356	9.344	…	23.676	17.121	23.676
80	8.271	10.458	7.639	…	7.639	10.458	7.639	…	8.271	18.098	18.098
81	2.229	37.262	2.229	…	2.229	15.337	37.262	…	2.229	37.262	15.337

The scattered data are sorted by increasing value, but the values drawn from the clusters, which start at row 71, are in the order of visitation of the 11 clusters

Table 2 A subset of 10 of the 101 resamples sorted by increasing value plus the median for every row

Seq. no.	$\mathbf{z}_1(\mathbf{s}_i)$	$\mathbf{z}_2(\mathbf{s}_i)$	$\mathbf{z}_3(\mathbf{s}_i)$	…	$\mathbf{z}_{67}(\mathbf{s}_i)$	$\mathbf{z}_{68}(\mathbf{s}_i)$	$\mathbf{z}_{69}(\mathbf{s}_i)$	…	$\mathbf{z}_{99}(\mathbf{s}_i)$	$\mathbf{z}_{100}(\mathbf{s}_i)$	$\mathbf{z}_{101}(\mathbf{s}_i)$	Row median
1	0.049	0.049	0.049	…	0.049	0.049	0.049	…	0.049	0.049	0.049	0.049
2	0.051	0.051	0.051	…	0.051	0.051	0.051	…	0.051	0.051	0.051	0.051
3	0.066	0.066	0.066	…	0.066	0.066	0.066	…	0.066	0.066	0.066	0.066
…	…	…	…	…	…	…	…	…	…	…	…	…
48	1.561	1.561	1.561	…	1.561	1.561	1.561	…	1.561	1.561	1.561	1.561
49	1.632	1.632	1.766	…	1.766	1.766	1.766	…	1.766	1.766	1.632	1.766
50	1.766	1.766	1.794	…	1.794	1.794	1.794	…	1.794	1.794	1.766	1.794
…	…	…	…	…	…	…	…	…	…	…	…	…
69	4.964	5.359	5.359	…	4.550	4.964	5.262	…	4.550	4.964	5.378	5.359
70	5.359	5.430	5.651	…	4.964	5.359	5.359	…	4.964	5.339	5.631	5.378
71	5.378	5.651	5.744	…	5.359	5.651	5.378	…	5.359	5.651	5.744	5.651
72	5.651	5.744	5.813	…	5.430	5.744	5.651	…	5.651	5.744	5.813	5.744
73	5.744	5.813	7.539	…	5.651	5.813	5.744	…	5.744	5.813	6.043	5.813
74	5.813	6.110	9.692	…	5.744	6.287	5.813	…	5.813	6.287	6.110	6.287
75	6.287	6.287	10.523	…	5.813	7.048	7.639	…	6.043	9.358	7.496	8.755
76	8.271	10.458	11.910	…	7.496	9.003	9.344	…	7.960	13.603	9.133	9.692
77	9.692	17.121	18.537	…	7.639	10.458	10.869	…	8.271	17.121	15.337	11.910
78	18.537	18.887	18.949	…	9.358	15.337	18.390	…	8.755	18.098	18.098	18.098
79	18.887	18.949	19.406	…	18.887	19.641	18.887	…	11.910	18.887	18.949	18.949
80	18.949	25.256	19.641	…	19.406	23.356	25.256	…	13.603	19.641	19.406	23.676
81	23.356	37.262	23.676	…	23.356	52.363	37.262	…	23.676	37.262	23.676	37.262

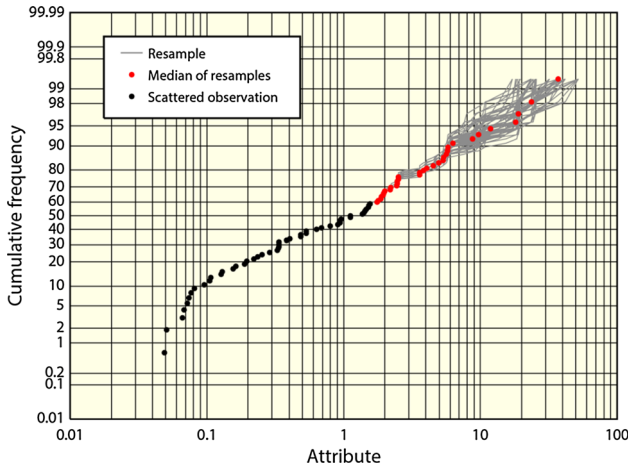

Fig. 4 Simultaneous display of all 101 cumulative distributions. Although not clear because of unavoidable overlappings, there are 101 resamples, all starting at the lowest value of 0.049. Up to 1.561 all resamples are the same and have been coded as "scattered observation" because all values come from subset $\mathbf{z}_s(\mathbf{s}_i)$

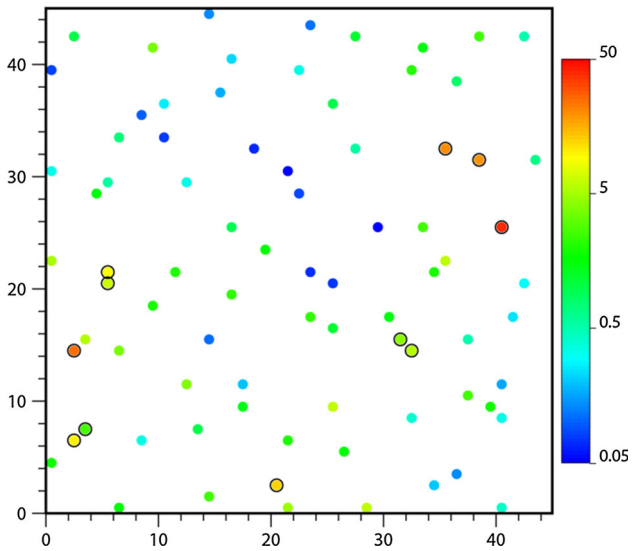

Fig. 5 Posting of the declustered dataset, $\mathbf{z}_q(\mathbf{s}_i)$, partially displayed in the last column of Table 2. *Circles* around dots indicate observations retained from the clusters

3.4 Estimation of the semivariogram

I used the traditional estimator

$$\gamma^*(\mathbf{h}) = \frac{1}{2 \cdot N(\mathbf{h})} \cdot \sum_{i=1}^{N(\mathbf{h})} [z(\mathbf{s}_i) - z(\mathbf{s}_i + \mathbf{h})]^2 \quad (1)$$

where $z(\mathbf{s}_i)$ is an observation at location $\mathbf{s}_i$, and $N(\mathbf{h})$ is the number of pairs of observations within a distance class on average $\mathbf{h}$ units apart (e.g., Chilès and Delfiner 2012). Omnidirectional modeling will suffice because the attribute is isotropic and second order stationary (Olea 2006).

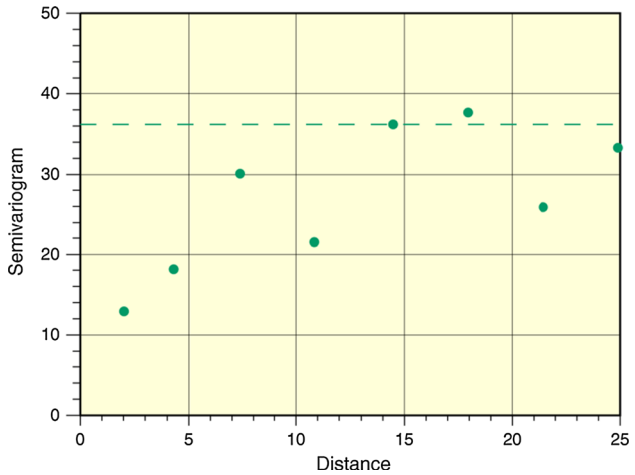

Fig. 6 The empirical semivariogram of the median resample $\mathbf{z}_q(\mathbf{s}_i)$. The *segmented line* shows the asymptotic value for the underlying sill

Figure 6 shows the results when using as data the median resample $\mathbf{z}_q(\mathbf{s}_i)$ displayed in Fig. 4 and Table 2.

Figure 7 displays the results for the more demanding modeling in Sect. 2.3.2. The 9 dots are part of the empirical semivariograms of 8 different resamples, with a maximum of 2 from the same resample, #69. Substantial fluctuations in results despite that at least 70 out of the 81 values (86 %) used in the calculations are the same should not be a complete surprise when comparing to sensitivity analyses reported in the literature (e.g., Webster and Oliver 1992).

4 Discussion

Figure 8 and Table 3 allow an evaluation of the results in terms of the cumulative distributions. The declustered sample $\mathbf{z}_q(\mathbf{s}_i)$ is a significant improvement over the clustered sample. The maximum discrepancy between the clustered and the exhaustive sample is 23.9 percentage units at 4.550, which, according to Fig. 3b, is in the range of values common to clusters and scattered values. The maximum discrepancy between the declustered and the exhaustive sample is only 10.8 percentage units at 0.337. Curiously, the most persistent large deviations are for attribute values below 1, which is in the range of values of exclusive occurrence among scattered locations and happens to be the best interval for the clustered sample. The source of such a discrepancy in the declustered sample seems to be an excess of observations in the intervals 0.065–0.080 and 0.32–0.35.

Resample #69 ($\mathbf{z}_{69}(\mathbf{s}_i)$) could be considered another candidate to be a solution to the estimation problem given its close approximation to the median values of the experimental semivariograms in Fig. 7. Indeed, $\mathbf{z}_{69}(\mathbf{s}_i)$,

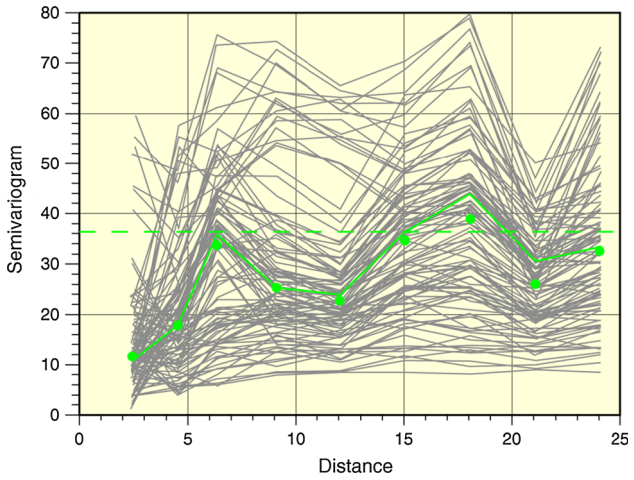

Fig. 7 Collection of empirical semivariograms resulting from using all 101 resamples. The *segmented line* is the asymptotic value for the underlying sill. The *green dots* denote the median value for each distance class. The *green line* indicates the empirical semivariogram for the 69th resample, $\mathbf{z}_{69}(\mathbf{s}_i)$, which is the one with the minimum discrepancy to the median points in an absolute value sense

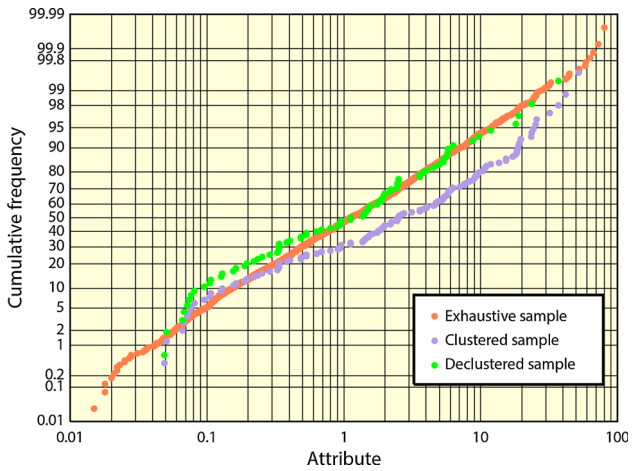

Fig. 8 Cumulative frequency distributions. The *green dots* are the median values partly displayed to the right of Table 2

found solely from considerations about estimation of the semivariogram, slightly outperforms solution $\mathbf{z}_q(\mathbf{s}_i)$.

An issue not specifically related to the declustering methodology is the reduction in the range of observations, a typical problem associated with any sampling. As it can be seen in Table 3, the extreme values for the exhaustive system are 0.015 and 80.116. The minimum value for the clustered data is 0.049, which is the same for the declustered sample because no corrective action was taken at the low range of values in this example in which the preferential sampling is determined by the extreme high values. The largest value in the clustered data is 52.363. This value as well as an observation of 41.948 did not appear enough

Table 3 Statistics of selected samples. D is the maximum absolute discrepancy of a cumulative frequency distribution to that of the exhaustive sample and D_m is the mean of those absolute discrepancies

	Exhaustive	Clustered	$\mathbf{z}_p(\mathbf{s}_i)$	$\mathbf{z}_{69}(\mathbf{s}_i)$
Size	2025	125	81	81
Mean	3.011	6.436	2.999	3.004
Standard deviation	6.029	9.065	5.675	5.716
Maximum	80.116	52.363	37.262	37.262
95th percentile	11.914	24.610	14.695	14.214
Upper quartile	3.089	8.392	2.516	2.789
Median	1.124	2.514	1.124	1.124
Lower quartile	0.406	0.534	0.280	0.280
5th percentile	0.097	0.076	0.070	0.070
Minimum	0.015	0.049	0.049	0.049
D	–	23.862	−10.765	−10.765
D_m	–	12.334	3.563	3.607

times in the resamples, consequently they vanish in the calculation of the median for the maximum value of the resamples—the last line in Table 2—where, by chance, the value 41.948 does not even show among the only 9 resamples displayed, despite being in about 20 other resamples. However, as it can be observed in Fig. 8, the loss of values at the tails did not have an important impact in approximating the underlying cumulative distribution and none in terms of estimating the most important percentiles. Selecting the maximum value instead of median for the bottom row in Table 2 is always a possibility to expand the range of values in the declustered solution, but changes are marginal, without assurance to reproduce the always non-robust prediction of the true maximum value (Beirlant et al. 2004).

Considering that each of the 101 resamples $\mathbf{z}_k(\mathbf{s}_i)$ is a sample that could have been collected when planning a sampling without preference, Fig. 7 shows the empirical semivariograms that could have been obtained under those circumstances. The results are a reminder of the risk of modeling semivariograms with a minimum number of points, which is in many circumstances a realistic situation in need of better estimation methods. Inspection of Figs. 7 and 8 show a positive side of preferential sampling, which has always been regarded as a detrimental sampling practice: for small size samples, adequate processing of preferentially sampled data can produce more accurate estimates of frequency distributions and semivariograms than those derived from samples of comparable size devoid of clusters.

Modeling of the semivariogram is always more challenging than approximating the cumulative distribution because the semivariogram is a second order moment. Figure 9 confirms the well-known fact that clustering can

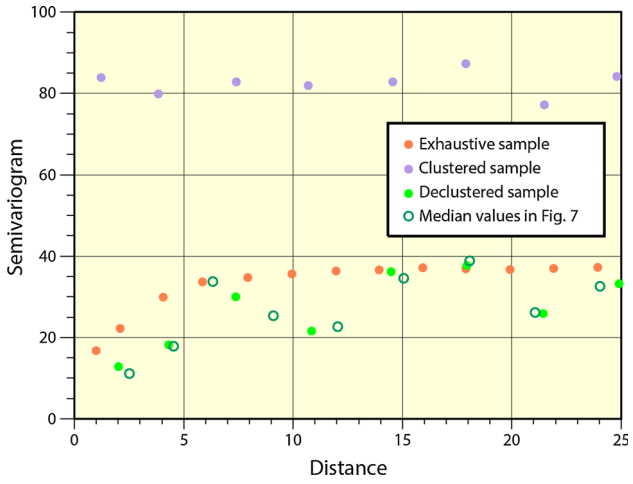

Fig. 9 Comparison of four empirical semivariograms

completely mask spatial correlation when it comes to modeling semivariograms (e.g., Bourgault 1997); the semivariogram for the clustered data may be pure nugget effect. By comparison, in general terms, declustering improvements in the estimation of the semivariogram are even more remarkable than those obtained for the cumulative frequency. Paradoxically, dataset $\mathbf{z}_q(\mathbf{s}_i)$, found without considering estimation of the semivariogram, provides as good a result as $\mathbf{z}_{69}(\mathbf{s}_i)$ or the set of median points in Fig. 7, indicating an overall conformity between the two alternatives in Sect. 2.3. Under closer scrutiny, when comparing the results to the semivariogram of the exhaustive sample, the stand-alone result in Fig. 6 does not look as good anymore. The low semivariogram values for short distances would be consistent with an excess of small values in the lower data interval of the scattered data.

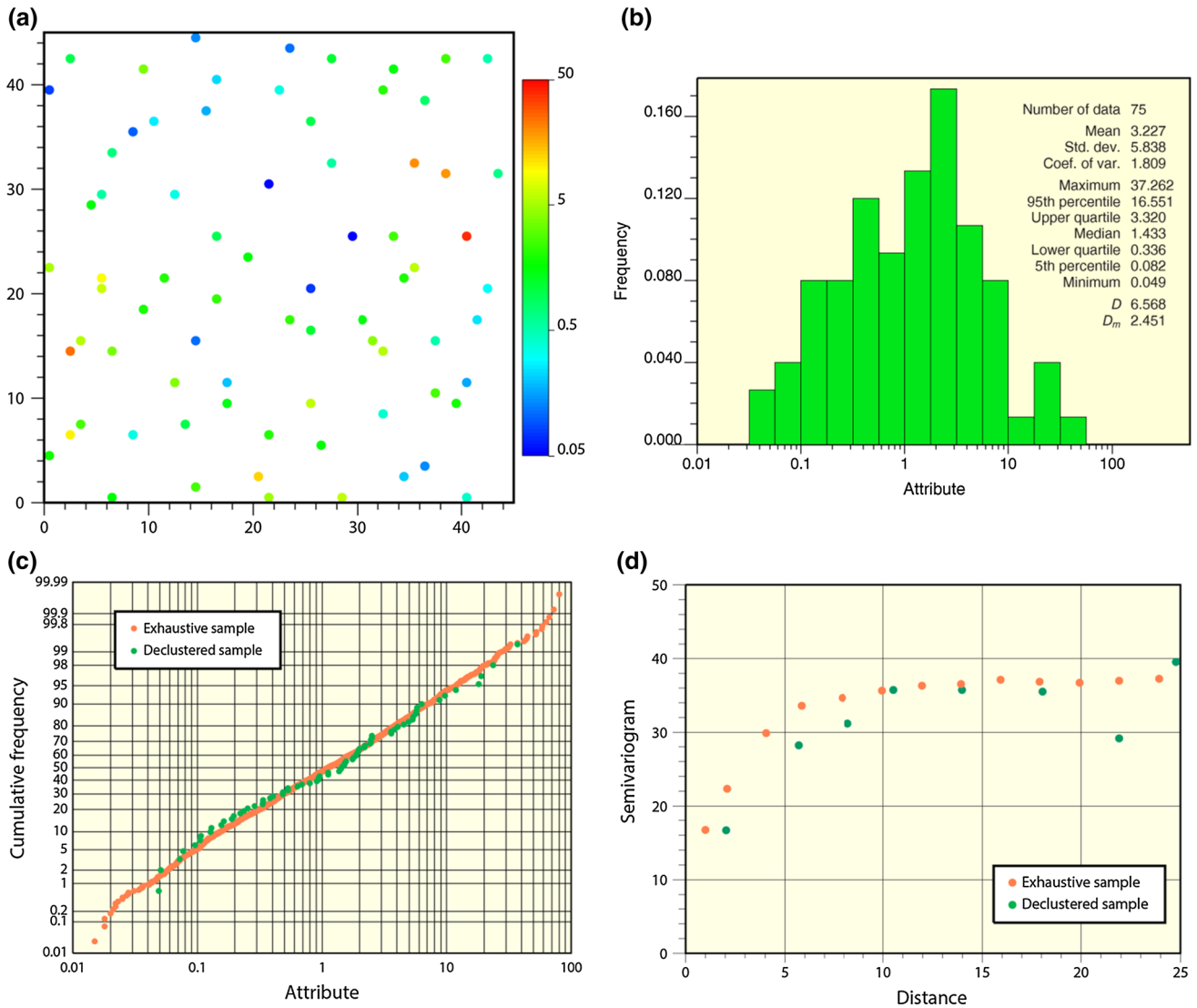

Fig. 10 Final results after postprocessing the declustered sample: **a** posting of data; **b** histogram and tabulation of statistics; **c** cumulative frequency, and **d** semivariogram

The subset of scattered data, $\mathbf{z}_s(\mathbf{s}_i)$, is not completely devoid of bias, a bias in the sampling space must be corrected in order not to compromise the quality of the declustering results. As mentioned at the beginning of this Sect. 4, there are two unique concentrations of values that were detected in this case by analyzing increments in the ranked data, one of 6 points between 0.068 and 0.081 and another group of 4 points between 0.033 and 0.035 with increments below 0.005, which is two orders of magnitude below the average increment of 0.34 in variable space. Consequently, the decision was made to eliminate at random 4 points in the first group and 2 in the other. Figure 10 displays the results showing significant improvements. Inspection of Figs. 8 and 10c indicates a reduction not only in maximum deviation, but also in terms of the average discrepancy. In the case of the semivariogram, there was a significant change for the better, particularly below the semivariogram range, which is the most important interval. Given the importance of correctly estimating the probability distribution and semivariogram of any attribute for further adequate modeling, say, stochastic simulation, analysts should not fall short in their attempts to obtained the most accurate approximations for the underlying histogram and semivariogram.

The final point is that sometimes good declustering requires paying attention to additional details beyond spatial declustering, which, in the case presented here, has been to address the clustering in variable space among the spatially scattered locations reveled as sudden steps in the cumulative distribution (Fig. 8).

5 Conclusions

Adequate preprocessing of preferential sampling for the purpose of estimating the cumulative frequency and the semivariogram can turn a liability into an asset. By resampling the clusters of a preferential sample of size 125, without introducing special restrictive assumptions in the methodology, it is observed in this particular case that:

- It is possible to generate a large number of different resamples of smaller size than the original sample.
- For any quantile, it is also possible to find the median of all resamples. The set of median values is a minimum absolute error approximation to the underlying cumulative frequency distribution.
- The resamples can be used to generate an equal and corresponding number of empirical semivariograms. For the distances considering in the modeling, the median is now a minimum absolute error estimate of the empirical semivariogram.
- The resample whose empirical semivariogram more closely approximates the set of median points was another reasonable approximation to the cumulative frequency distribution.
- The two modeled semivariograms more closely fit the exhaustive sample semivariogram for large distances than near the origin.
- The set of all resamples provides measures of uncertainty in the results associated with the preferential sampling.

Further improvements were obtained by addressing bias in attribute space at the subset of scattered data by eliminating 6 observations in two concentrations of values.

Acknowledgments The author wishes to thank Mark Engle (U.S. Geological Survey), Michael Pyrcz (Chevron Energy Technology Company), John Schuenemeyer (Southwest Statistical Consulting) and an anonymous reviewer appointed by the journal for reviewing an earlier version of the manuscript and making suggestion to improve its contents.

References

Beirlant J, Goegebeur Y, Segers J, Teugels J (2004) Statistics of extremes: theory and applications. Wiley, Chichester, 490 p

Bogaert P (1999) On the optimal estimation of the cumulative distribution function in presence of spatial dependence. Math Geol 31(2):213–239

Bourgault G (1997) Spatial declustering weights. Math Geol 29(2):277–290

Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, 2nd edn. Wiley, Hoboken, 734 p

Deutsch CV (1989) DECLUS: a FORTRAN 77 program for determining optimum spatial declustering weights. Comput Geosci 15(3):325–332

Diggle J, Menezes R, Su TL (2010) Geostatistical Inference under preferential sampling. J R Stat Soc Ser C 59(2):191–232

Emery X, Ortiz JM (2005) Histogram and variogram inference in the multigaussian model. Stoch Environ Res Risk Assess 19(1):48–58

Emery X, Ortiz JM (2007) Weighted sample variograms as a tool to better assess the spatial variability of soil properties. Geoderma 140(1–2):81–89

Isaac EH, Srivastava RM (1989) Introduction to applied geostatistics. Oxford University Press, New York, 561 p

Journel AG (1983) Nonparametric estimation of spatial distributions. J Int Assoc Math Geol 15(3):445–468

Kovitz JL, Christakos G (2004) Spatial statistics of clustered data. Stoch Environ Res Risk Assess 18(3):147–166

Marchant BP, Viscarra Rossel RA, Webster R (2013) Fluctuations in method-of-moments variograms caused by clustered sampling and their elimination by declustering and residual maximum likelihood estimation. Eur J Soil Sci 64(4):401–409

Olea RA (2006) A six-step practical approach to semivariogram modeling. Stoch Environ Res Risk Assess 20(5):307–318

Olea RA (2007) Declustering of clustered preferential sampling for histogram and semivariogram inference. Math Geol 39(6): 453–467

Olea RA (2008) Basic statistical concepts and methods for earth scientists. U.S. Geological Survey Open-File Report 2008–1017, 191 p

Omre H (1984) The variogram and its estimation. In: Verly G, David M, Journel AG, Meréchal A (eds) Geostatistics for natural resources characterization, part 1. Reidel, Dordrecht, pp 107–125

Pardo-Igúzquiza E, Dowd PA (2004) Normality test for spatially correlated data. Math Geol 36(6):659–681

Pyrcz MJ, Deutsch CV (2014) Geostatistical reservoir modeling, 2nd edn. Oxford University Press, New York, 433 p

Pyrcz MJ, Gringarten E, Frykman P, Deutsch CV (2006) Representative input parameters for geostatistical simulation. In: Coburn TC, Yarus JM, Chambers RL (eds) Stochastic modeling and meostatistics: principles, methods and case studies, vol II. AAPG Computer Applications in Geology 5, pp 123–137

Reilly C, Gelman A (2007) Weighted classical variogram estimation for data with clustering. Technometrics 49(2):184–194

Richmond A (2002) Two-point declustering for weighting data pairs in experimental variogram calculations. Comput Geosci 28(2): 231–241

Rivoirard J (2001) Weighted semivariograms. In: Kleingeld WJ, Krige DG (eds) Proceedings of the 6th International Geostatistics Congress, Cape Town, pp 145–155

Switzer P (1977) Estimation of spatial distributions from point sources with applications to air pollution measurement. Proceedings of the 41st ISI Session, New Delhi. Bulletin of the International Statistical Institute 47(2):123–137. Also available as Technical Report No. 9, Department of Statistics, Stanford University 20 p. https://statistics.stanford.edu/sites/default/files/SIMS%2009.pdf

Webster R, Oliver MA (1992) Sample adequately to estimate variograms of soil properties. J Soil Sci 43(1):177–192

Webster R, Oliver MA (2007) Geostatistics for environmental scientists. Wiley, Chichester, 315 p