

# Declustering of Clustered Preferential Sampling for Histogram and Semivariogram Inference

Ricardo A. Olea

Received: 15 January 2006 / Accepted: 29 January 2007 / Published online: 22 August 2007  
© International Association for Mathematical Geology 2007

**Abstract** Measurements of attributes obtained more as a consequence of business ventures than sampling design frequently result in samplings that are preferential both in location and value, typically in the form of clusters along the pay. Preferential sampling requires preprocessing for the purpose of properly inferring characteristics of the parent population, such as the cumulative distribution and the semivariogram. Consideration of the distance to the nearest neighbor allows preparation of resampled sets that produce comparable results to those from previously proposed methods. A clustered sampling of size 140, taken from an exhaustive sampling, is employed to illustrate this approach.

**Keywords** Cluster · Preferential sampling · Histogram · Semivariogram

## Introduction

Spatiotemporal samplings are used both for the local characterization of attributes and for the global parameter inference of the parent population. Most efficient patterns for a spatial sampling are those that evenly cover the area or volume of interest, such as regular patterns or stratified samplings. Often, however, samplings are measurements resulting from some earlier activity in which the economics of the enterprise dominated the scene; sampling was an afterthought. Under these circumstances, it is not uncommon to see concentrations—clusters—of measurement in zones proved or

---

R.A. Olea

Department of Petroleum Engineering, Stanford University, Stanford 94305, USA

*Present address:*

R.A. Olea (✉)

U.S. Geological Survey, 12201 Sunrise Valley Dr., MS 956, Reston, VA 20192, USA

e-mail: olea@usgs.gov

expected to have extreme values. Mere clustering is not detrimental to global geostatistical inference, and preferential sampling is not necessarily in the form of clusters. Yet particularly common in oil ventures, preferential sampling is in the form of clusters. This is the type of clustering addressed here, which will be discussed only in the two-dimensional geographical space.

Proper inference of parameters from preferential samplings requires special attention to avoid obtaining unrealistic results. Most existing methods are characterized for being specific to particular aims, with inference of the histogram being the earliest and prevalent effort. Approaches can be grouped into three categories.

(a) In the polygonal method, the sampling locations are employed to tessellate the sampling field into a decomposition called by various authors a Dirichlet tessellation, Thiessen polygons, or Voronoi polygons (Dirichlet 1850; Voronoi 1907; Thiessen 1911). The polygons are used to weight the measurements in the sampling inversely proportional to the area of the polygon containing the sampling location (Isaacs and Srivastava 1989, Chap. 10; Goovaerts 1997, pp. 79–81). The main problems of this method are its sensitivity to boundary conditions and all the measurements are declustered, even at low density areas where it is not needed. Dubois and Saisana (2002) have proposed improvements to this method by incorporating the distance to the nearest neighbor.

(b) In the other equally popular method, the partition of the sampling field is done regularly by resorting to squares, and determining the weight as an inverse proportion of the number of measurement per square cell. As the cell size, origin, and the orientation of the tessellation are decided arbitrarily, the indetermination is decided by trial-and-error by searching the partition that produces the most extreme mean. If the preferential sampling is towards high values, the ad hoc criterion is to accept as best those weights producing the minimum mean, and conversely, for the less predominant sampling biased towards low values (Journel 1983; Deutsch 1989). Schofield (1993) has also proposed entropy as the optimization criteria. Pyrcz and Deutsch (2003) claim that these criteria do not always produce the right inference.

(c) Significantly less popular than the polygonal and the cell methods for histogram inference are some other conceptually more elaborate weighting formulations that assume bivariate or multivariate normality, which accounts for the spatial correlation of the attribute, or both. The authors either assume certain forms of correlation or estimate it from the sampling without addressing the issue of clustering. This is the same disadvantage noted by Deutsch (1989) and Rivoirard (2001) concerning the use of kriging to assign declustering weights to data (e.g. Pardo-Igúzquiza and Dowd 2004). Bourgault (1997) employed the ratio of determinants of correlation matrices. Bogaert (1999) proposed a method based on indicators, which presumes the attribute follows a multivariate Gaussian distribution and depends on a smoothing parameter.

Omre (1984) and Richmond (2002) proposed methods also based on weighting, solely for the purpose of modeling semivariograms. Evidently, the critical step is the proper calculation of the experimental semivariogram. Once such a preliminary step is accomplished, the selection of the best model proceeds as usual and does not require any special approach (Olea 2006). One of Richmond's approaches is the direct generalization of the cell declustering method with the difference the weights

apply to pairs of measurements. In a second method, Richmond does not employ cells and deals directly with clusters defined as groups of sampling locations closer than a critical distance. The weighting applies only to pairs in which at least one of the sampling locations belongs to a cluster.

There have been a few attempts of declustering the sampling for the simultaneous inference of the histogram and the semivariogram. Switzer (1977) and Rivoirard (2001) have suggested employing directly the same set of weights used for inferring the histogram for the estimation of spatial correlation. Practical attempts have not always been satisfactory (e.g., Richmond 2002). Emery and Ortiz (2005) proposed a comprehensive approach at the core of which there is a weighting method similar to the kriging weight approach for estimating the histogram. In their formulation, the weights come as the solution of a linear system of equations that requires prior knowledge of the semivariogram. The authors suggest starting with an assumed semivariogram or the one resulting from modeling without declustering, then update the system, and solve it again until convergence. The approach involves a fourth order moment that is expressed in terms of second order moments, which requires restricting the solution to multivariate Gaussian models. In addition, there is no assurance of convergence, let alone convergence to the true semivariogram.

The objective of this study is to go beyond existing methods by offering a general and unitary approach to compensate for preferential clustering. This approach is capable of not only providing a realistic representation of histograms, but also of allowing the calculation of experimental points for the modeling of semivariograms and covariances without having to make restrictive assumptions or preparing a different compensated data set.

## Method

The basic idea is to rely on the distance to the nearest neighbor, first for detecting the presence of clusters, and then for censoring the sampling if it is found that the clusters are associated to preferential sampling. The new sampling will be more stratified than the original sampling pattern, thus taking advantage of the fact that the direct use of observations from evenly spaced sampling produces approximately optimal global inferences, independently of the form of the correlation (e.g., Switzer 1977).

The best way to assess the degree of clustering is to analyze the cumulative distribution of the distance to the nearest neighbor. If the distance distribution shows an excess of points towards the lower tail then one must split the sampling into two groups, one comprising all measurements close together (subset 2) and the other one with the remainder of the observations (subset 1). These two numbers are arbitrary and used only to facilitate the exposition. If the attribute histograms show poor to no overlapping, then there is not only clustering, but also a preferential sampling, which is an unacceptable combination for the purpose of global inference of population parameters. Clustering without preferential sampling does not require correction of the sampling.

To remove the preferential taking of measurements from a clustered sampling:

- Calculate the distance to the nearest neighbor within the sampling.

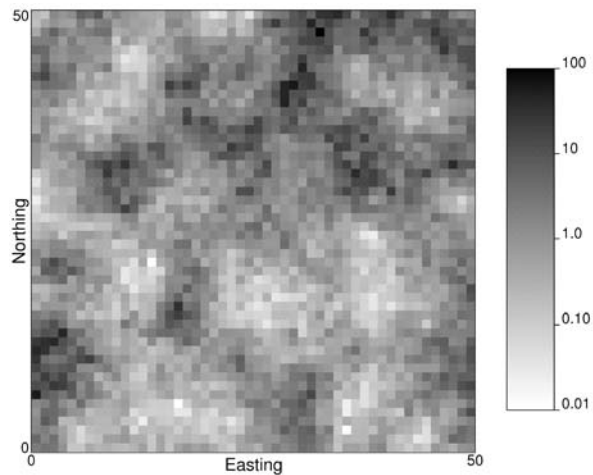
- Prepare a new sampling (subset 1) comprising only of measurements in the group with the largest distances to the nearest neighbor. By construction, this subset will be free of clusters.
- For each measurement in the subset 2 of clusters, calculate the distance to the nearest neighbor in the other subset.
- Move from subset 2 to subset 1 the observation with the largest distance to the nearest neighbor in subset 1.
- Recalculate the distance to the nearest neighbor within the expanded subset 1, and the distances from the remainder of subset 2 to the expanded subset 1.
- Repeat the transfer of points and recalculate the distances. In case of two or more observations in subset 2 are equally distant to measurements in subset 1, decide the tie by random selection. Multiple equally distant observations may require imposition of a maximum number of retained points per distance class. Keep moving points in the reduced subset 2 to the expanded subset 1 until reaching some stopping criteria, such as: (a) the number of transferred points is equal to the original number of clusters or some proportion of them; (b) the number of transferred points is a proportion of the total number of points in the clusters; (c) the distribution of the distance for the subset of retained data originally clustered is sufficiently close to that of subset 1 according to a Q–Q plot; or (d) the Kolmogorov–Smirnov statistic for the cumulative distributions of distances for the retained clustered observations and subset 1 is below a critical value.
- The final expanded subset 1 is the solution to the declustering problem. All the retained observations added to the original subset 1 make subset 3.

Given that the original subset 1 is free of clusters, so is the pattern of the added points in subset 3. The preparation of the expansion also assures that the merging will not result in new clusters.

## Case Study

The new method is applied to a clustered example taken from the literature (Deutsch and Journel 1998, Problem Set One); this example that has been widely used by authors addressing the declustering issue. Figure 1 is a posting of the exhaustive sampling and Fig. 2 displays the global characteristics of the 2500 measurements.

The clustered sampling is shown in Fig. 3. It comprises only 140 observations, or 5.6% of the exhaustive sampling, with 10 clusters. 97 of the measurements were taken on a pseudo-regular grid, with the remaining data coming from observations added to create 5-point clusters centered on values above the median. As a result of the clustering expansion, two clusters merged and one point was not possible to consider because of boundary conditions. Figure 3(b) clearly shows an abnormal concentration of values only 1 unit apart from each other. The 54 points with a distance of 1 to the nearest neighbor define a subset of locations clearly clustered (corresponding to subset 2 in the previous section), and the remaining 86 points make a perfectly declustered subset 1 (Fig. 4). Note that the subset means are 1.29 and 9.23 with a minimum overlapping in values (Fig. 5). The maximum deviation from the main

**Fig. 1** Posting of values for exhaustive sampling in case study**Table 1** Comparison of distributional parameters for exhaustive sampling and three declustering methods

|                                         | Mean | Standard deviation | Lower quartile | Median | Upper quartile |
|-----------------------------------------|------|--------------------|----------------|--------|----------------|
| Exhaustive sampling                     | 2.58 | 5.15               | 0.34           | 0.98   | 2.58           |
| Cell (Deutsch and Journel 1998, p. 245) | 2.52 | 4.57               | 0.34           | 1.20   | 2.75           |
| Correlation (Bourgault 1997)            | 2.57 | 4.67               | 0.34           | 1.21   | 2.76           |
| This paper                              | 2.60 | 4.34               | 0.34           | 1.02   | 2.54           |

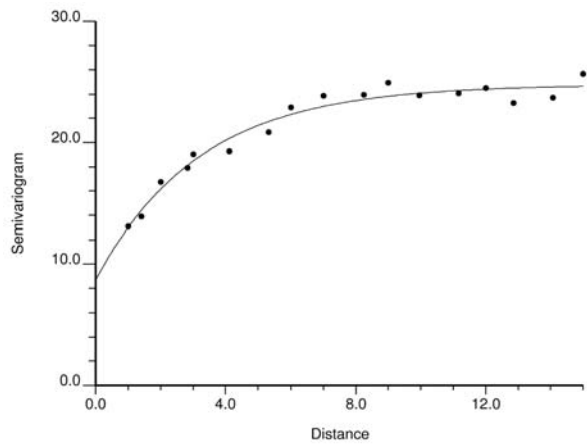
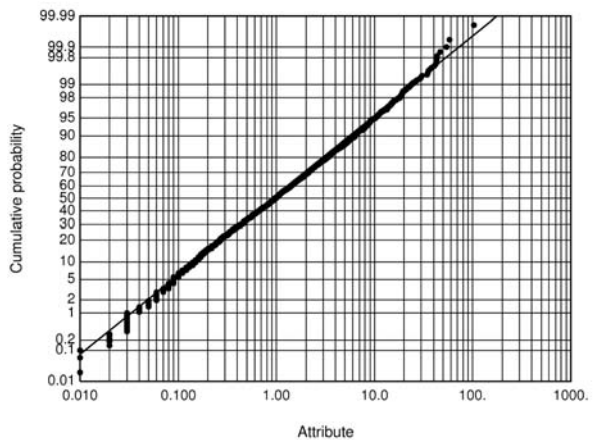
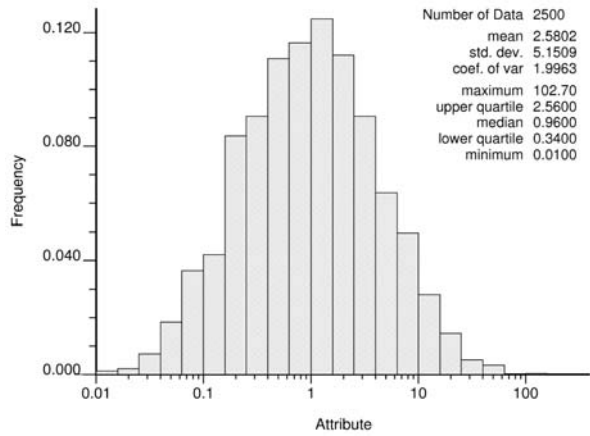
diagonal in the Q–Q plot in Fig. 5(c) is 53.2. Clearly, we are in the presence of two different distributions.

Figures 5(a) and 6(a) should be sufficient to dispel any notion that a naïve elimination of all clusters is sufficient to produce a satisfactory declustered sampling for the purpose of inferring the global characteristics of the parent population.

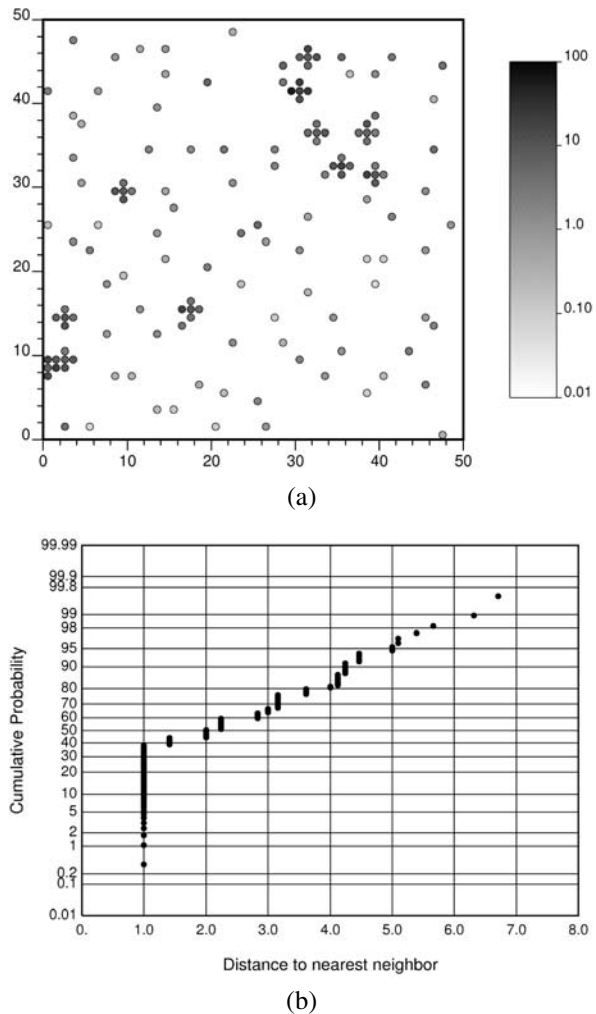
Figure 7 shows that, after transferring 11 points from subset 2 to subset 1 following the method outlined in the previous section, one reaches a distribution for subset 3 that closely approximates the distance distribution for the original subset 1. There are at most three observations in subset 3 with the same distance to the closest observation in subset 1. The maximum deviation from the diagonal in the Q–Q plot is 0.88 (Fig. 7(a)), and the Kolmogorov–Smirnov statistic  $D$  is 0.13 (Fig. 7(b)). This new sampling of size 97 posted in Fig. 7(c) is, by construction, free of clusters and preferential sampling. The following graphs evaluate the quality of the declustered sampling for the purpose of characterizing the histogram and semivariogram of the exhaustive sampling.

Figure 8 displays the distribution of values for the final declustered sampling. Table 1 compares some parameters with both the exhaustive system and with solutions by Bourgault (1997) and Deutsch and Journel (1998) using weights. The standard deviation for the new method is the only parameter with a significantly worse approximation to the exhaustive sampling than the cell and the correlation methods.

**Fig. 2** Global inference for exhaustive sampling: **(a)** histogram; **(b)** cumulative distribution, with straight line denoting a normal distribution with the same mean and variance as the exhaustive sampling; **(c)** semivariogram



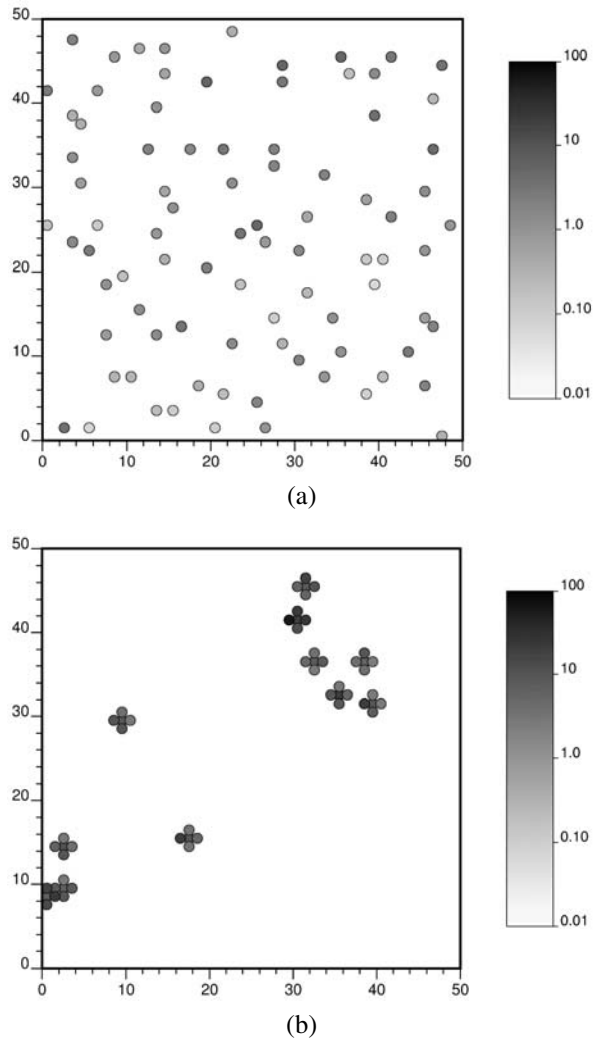
**Fig. 3** The clustered sampling:  
**(a)** posting of values;  
**(b)** distance to nearest neighbor



The main reason for developing the new declustering method was to provide samplings that one can also use for modeling semivariograms, hence covariances, without again having to preprocess the original clustered sampling. Figure 9 displays the experimental semivariogram and the best model for the exhaustive and clustered sampling of size 140. It is clear that the latter semivariogram is completely distorted.

Figure 10 exhibits the results of modeling five different semivariograms of common interest in geostatistical characterization of attributes, all of them highly satisfactory except for the sill of the model in the original space. It is worth noting that this subpar performance is not restricted to the method proposed here. Richmond's (2002) methods applied to the same data do not work in the original space as well as they do after employing a logarithmic transformation (Richmond 2007, personal communication). In addition, in situations like this in which the attribute is highly skewed, the semivariogram in Fig. 10(a) is the one having the least practical inter-

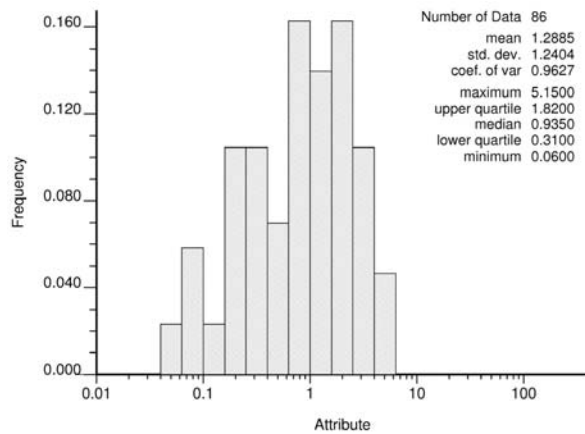
**Fig. 4** The decomposed sampling: (a) subset 1 devoid of clusters; (b) subset 2 of clusters



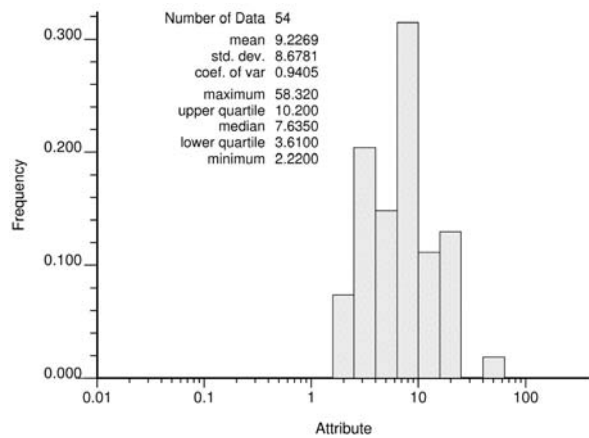
**Table 2** Comparison of semivariogram modeling for natural logarithm of exhaustive sampling and three declustering methods

|                         | Model type | Nugget | Sill-nugget | Range |
|-------------------------|------------|--------|-------------|-------|
| Exhaustive sampling     | Spherical  | 0.13   | 1.80        | 9.51  |
| Cell (Richmond 2002)    | Spherical  | 0.10   | 1.90        | 9.75  |
| Cluster (Richmond 2002) | Spherical  | 0.10   | 1.90        | 10.25 |
| This paper              | Spherical  | 0.06   | 1.77        | 9.52  |

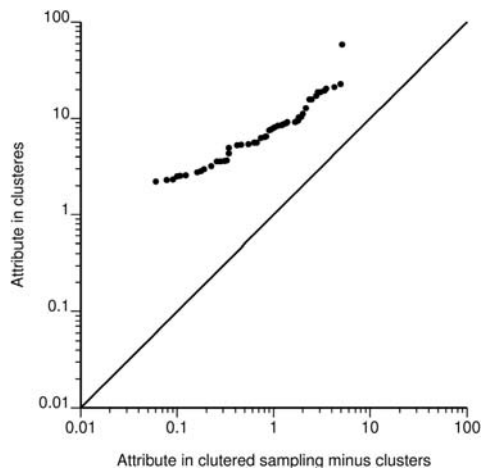
**Fig. 5** Attribute distributions for the decomposed sampling: (a) declustered subset 1; (b) clustered subset 2; (c) Q–Q plot of subset 1 versus subset 2



(a)

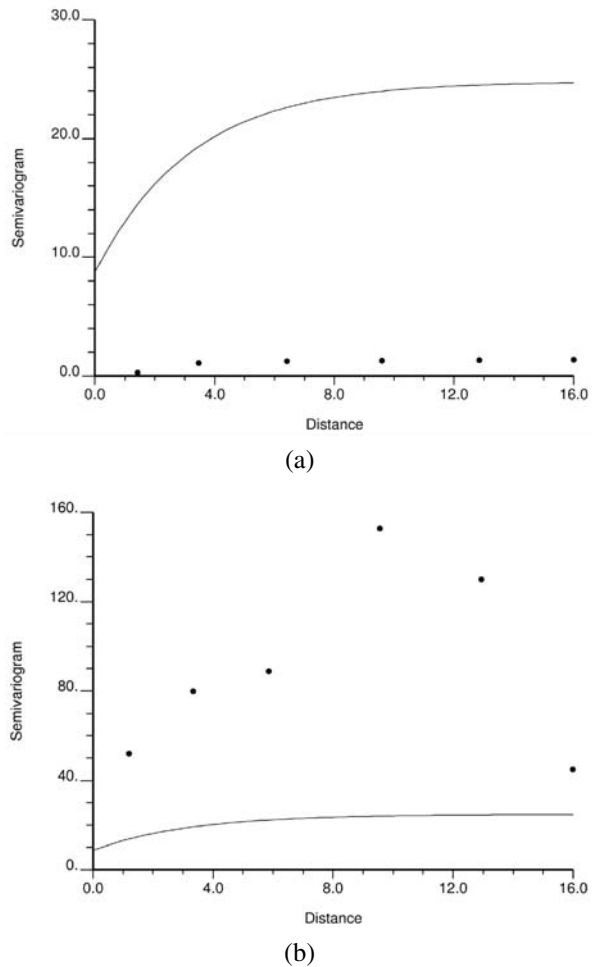


(b)



(c)

**Fig. 6** Semivariograms for the decomposed sampling, with *dots* denoting experimental values and the *continuous line* the model for the exhaustive sampling: **(a)** declustered subset 1; **(b)** clustered subset 2



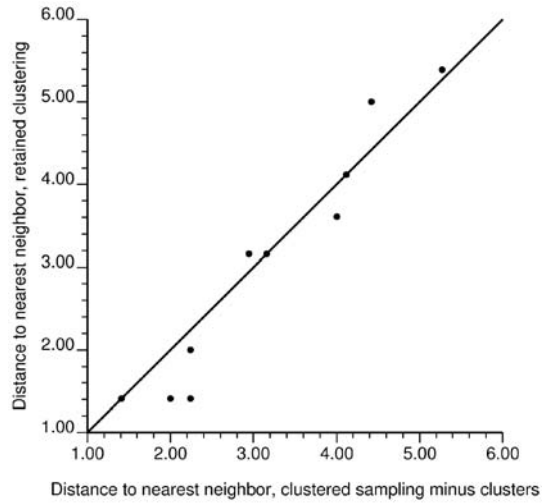
est. Modelers will use any of the other semivariograms before paying any attention to the one in the attribute space. Table 2 compares results for the case of the natural logarithm semivariogram.

For the purpose of estimation, the sampling to use is still the original clustered sampling, as geostatistical methods are capable of automatically dealing with clustering (e.g., Olea 1999, pp. 28–30).

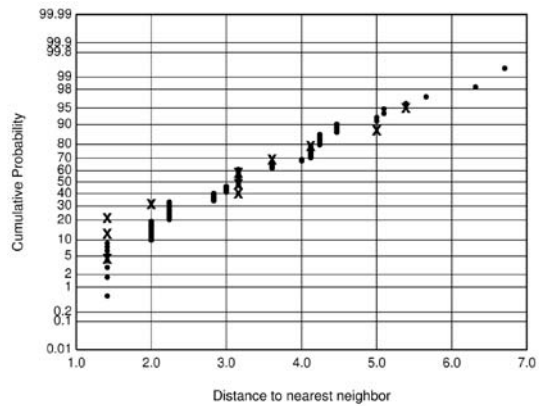
## Conclusions

Partial elimination of clustered measurements allows for the preparation of a new sampling for global geostatistical inference. The method is simple, general, and directly corrects the preferential sampling. Above all, this method offers a unique declustered data subset that, different from most other existing methods, is valid for any global inference independent of the parameter or function of interest.

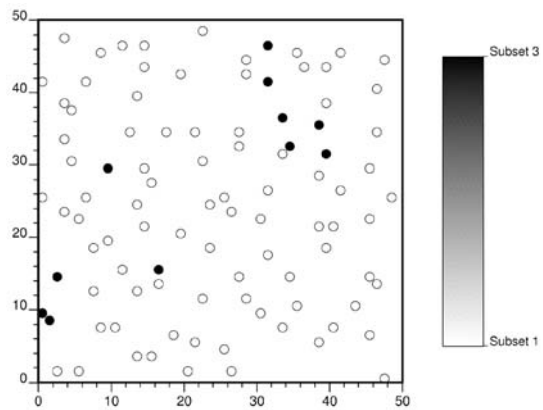
**Fig. 7** Declustered sampling components: **(a)** Q–Q plot comparing the distributions of subset 1 and subset 3; **(b)** cumulative distribution of distance to nearest neighbor, with *dots* representing subset 1 and *crosses* subset 3; **(c)** posting of the final solution to the declustering problem



(a)

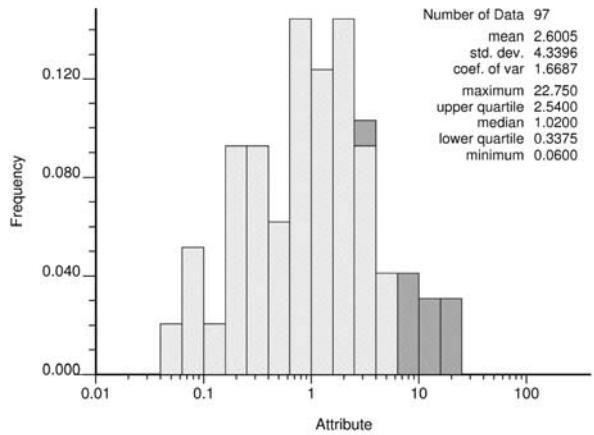


(b)

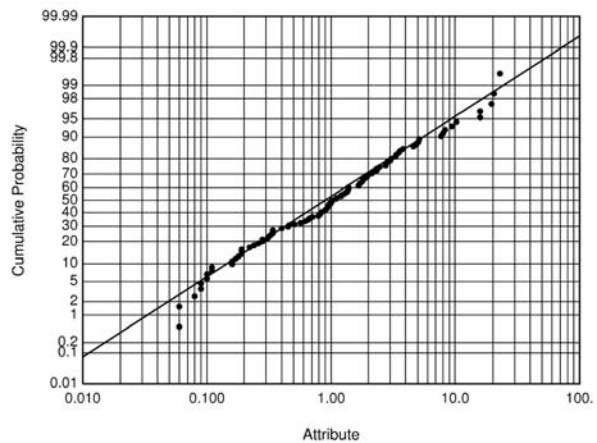


(c)

**Fig. 8** The declustered sampling of size 97: (a) histogram, with *light shading* denoting contribution of subset 1 and *dark shading* that of subset 3; (b) cumulative distribution, with *straight line* denoting a normal distribution with the same mean and variance as the exhaustive sampling

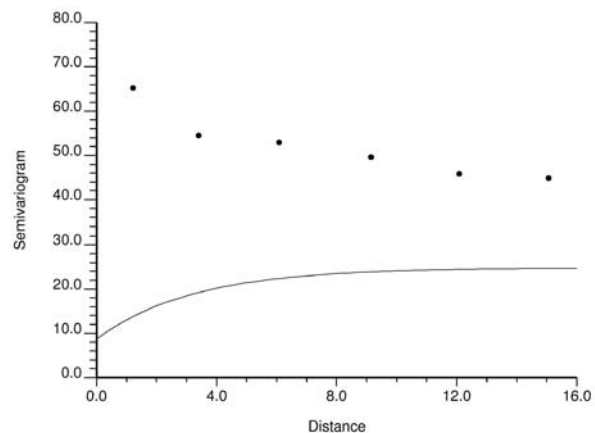


(a)

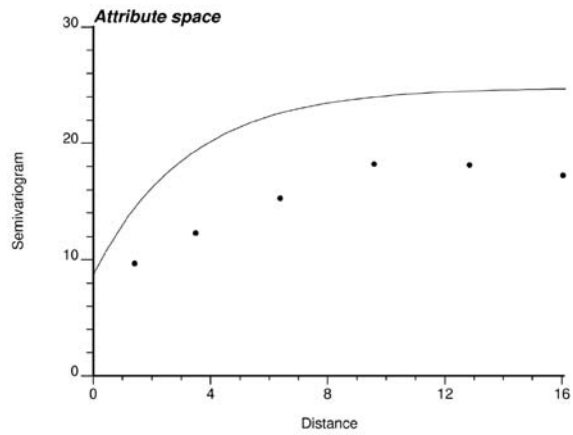


(b)

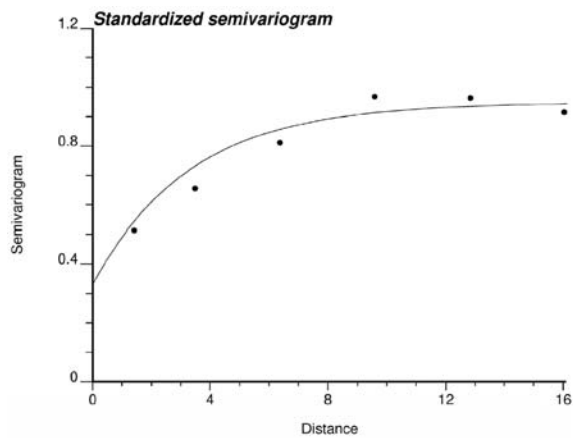
**Fig. 9** Semivariograms. *Dots* denote experimental semivariogram for the clustered sampling of size 140, and *solid line* is the best model for the exhaustive sampling



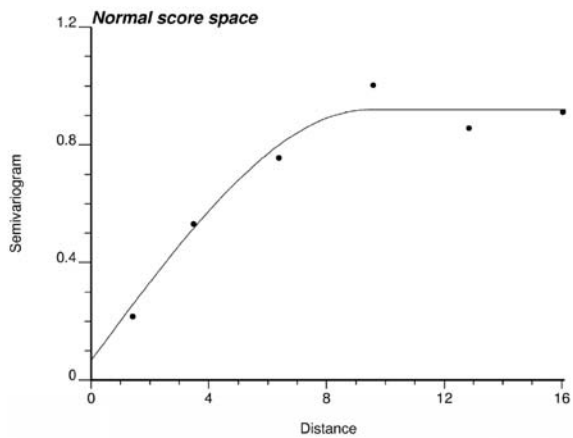
**Fig. 10** Experimental semivariograms for the new declustered sampling and corresponding model for the exhaustive system



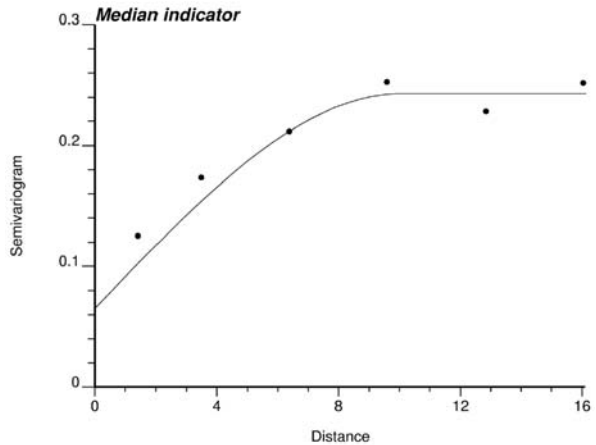
(a)



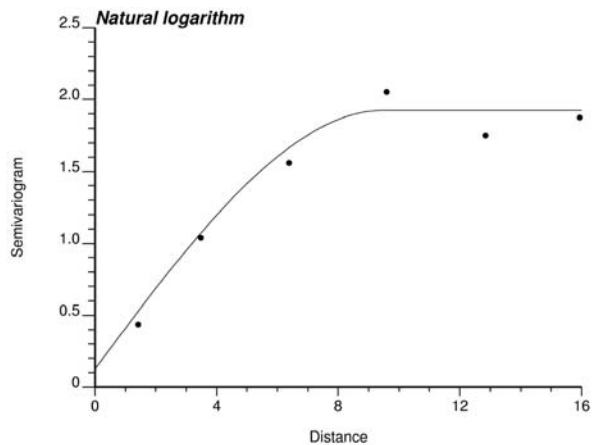
(b)



(c)

**Fig. 10** (Continued)

(d)



(e)

The new approach has performed equally well or better than alternative methods, offering flexibility to adapt to different practical situations.

**Acknowledgements** The author is grateful to the Blaustein Visiting Scholars Program, Stanford University, for providing support for activities that partly included research on clustered samplings, and to André Journel for extending the invitation and hosting discussions leading to the preparation of this publication. I am indebted to two reviewers of my original submission who asked to remain anonymous.

## References

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

- Deutsch CV, Journel AG (1998) GSLIB—Geostatistical software library and user's guide, 2nd edn. Oxford University Press, New York (384 p)
- Dirichlet GL (1850) Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. *J Reine Angew Math* 40:209–227
- Dubois G, Saisana M (2002) Optimizing spatial declustering weights—comparison of methods. In: Proceedings of the annual conference of the international association for mathematical geology, Berlin, pp 479–484
- Emery X, Ortiz JM (2005) Histogram and variogram inference in the multigaussian model. *Stoch Environ Res Risk Assess* 19(1):48–58
- Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York (483 p)
- Isaacs EH, Srivastava RM (1989) Introduction to applied geostatistics. Oxford University Press, New York (561 p)
- Journel AG (1983) Nonparametric estimation of spatial distributions. *Math Geol* 15(3):445–468
- Olea RA (1999) Geostatistics for engineers and earth scientists. Kluwer Academic, Norwell (303 p)
- Olea RA (2006) A six-step practical approach to semivariogram modeling. *Stoch Environ Res Risk Assess* 20(5):307–318
- 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 CV (2004) Normality test for spatially correlated data. *Math Geol* 36(6):659–681
- Pyrz MJ, Deutsch CV (2003) Declustering and debiasing. Downloaded in January 2007 from <http://www.gaa.org.au/pdf/DeclusterDebias-CCG.pdf>
- 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
- Schofield N (1993) Using the entropy statistic to infer population parameters from spatially clustered sampling. In: Soares A (ed) Geostatistics Tróia'92, vol 1. Kluwer Academic, Dordrecht, pp 109–119
- Switzer P (1977) Estimation of spatial distributions from point sources with applications to air pollution measurement. In: Proceedings of the 41st ISI session, New Delhi. *Bull Int Stat Inst* 47(2): 123–137
- Thiessen AH (1911) Precipitation average for large areas. *Monthly Weather Rev* 39:1082–1084
- Voronoi G (1907) Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J Reine Angew Math* 133:97–178