Assessing the effect of clustered and biased multi-stage sampling

Raquel Menezes* and Jonathan Tawn†

*Department of Mathematics for Science and Technology, University of Minho,
4800-058 Guimarães, Portugal
rmenezes(at)mct.uminho.pt

†Department of Mathematics and Statistics, Lancaster University
Lancaster LA1 4YF, England

Abstract

We propose a method for detecting biased multi-stage sampling of spatial data and a method to adjust for biased clustering of samples. We assess the effect of these methods for the analysis of radioactivity contamination data from Rongelap island, with the scientific problem being the estimation of the maximum level of radioactivity over the island. These data were collected over a two-stage process of uniform and clustered samples, which may have an impact on conclusions from a standard analysis that does not account for either of these features.

KEY WORDS: variogram estimation; multi-stage sampling; biased sampling; clustered data; radioactivity data.

1 Introduction

The traditional geostatistical methods rely on the expected assumption that the sampling design for locations $\mathbf{x_i}$, $i = 1, ..., n$ is deterministic or it is stochastic but independent of the data process, and all analyses are carried out conditionally on $\mathbf{x_i}$ [1]. It is then assumed that the sampling points have been chosen independently of the values of the spatial variable. However, dependencies can occur due to the adopted sampling method, such as the favored selection of specific areas that are believed critical (e.g. maximum values search).

Schlather et al. in [2] propose methods to detect the dependence between marks and locations of marked point processes. As described in [3], the random field and the marked point process are two type of spatial processes such that:

- The former is defined in every point of the observed region, and the sample positions can be determined by the scientist himself (example of deterministic sampling design);
- For the latter, the locations are always given by a stochastic point process, and interactions among the locations and the marks are normally expected. Otherwise, one has

the so called *random field model* (marked point process becomes a special class of a random field).

If the data are consistent with a random field model, the point pattern and the marks can be analysed separately using standard techniques for point processes (e.g. [4] and [5]) and for geostatistical data (e.g. [6]). Therefore, this analysis is greatly simplified. Furthermore, the examination of second-order characteristics, like the variogram, of a spatial process should consider if data come from a random field or a genuine marked point process. Example of references concerned with this subject are [7] and [8].

Schlather et al. in [2] indicate next two likely situations for point and data processes being dependent, and subsequent failure of this important geostatistics assumption. Firstly, if the dependency is an intrinsic property of the data themselves, for example the relative positions of trees impact on their size due to their competition for light and nutrient. This is the case of genuine marked point processes. Alternatively, this dependency can be justified by a prior scientific knowledge of the spatial variable of interest, for example of the expected local level of contamination in air pollution. This can lead to the gathering of samples in areas with atypical values. Our work concerns the problems resulting from the second situation, that we think of major importance in geostatistics because of its high likelihood of occurrence on actual field measurements, and often either ignored or addressed by generic techniques like declustering ones applied to the first-order characteristics (e.g. [9] and [10]).

In this paper, we are motivated by the application example of the radioactivity data from Rongelap island, where a two-stage data collection was used, leading to the presence of clustered data. So that we restrict our attention to multi-stage samples, aiming to assess the presence of multi-stage dependence, or also referred to as *sequential dependence*, where the choice of sampling points is driven by previous measurements.

We propose a data exploratory method which is intended to detect biased multi-stage collection of spatial data. We then investigate corrector models that aim to minimize the impact on variogram estimation due to the adoption of the type of non-standard sampling designs just described.

2 Motivating example

Our example is related to data collected on Rongelap island. This island is located in the Pacific Ocean approximately 4000 kilometres south-west of Hawaii. The data were collected for the analysis of current levels of radioactivity contamination that resulted from a nuclear weapons testing programme during the 1950s. The scientific problem has been the estimation of the maximum level of radioactivity over the island, as part of a wider investigation to decide whether Rongelap can safely be resettled. See [6] or [11] for more detail on these data.

The sampling design defined for data collection is illustrated in Figure 1. It started with a coarse grid of 63 locations and ended up with 98 additional measurements within four fine grids. These locations are identified by time label t_0 and t_1, respectively. As this process involved two-stage of uniform and clustered samples, we wonder about the impact on conclusions from a standard analysis that does not account for either of these features. To proceed our analysis, the methods will be applied to transformed data with a constant variance as described next.

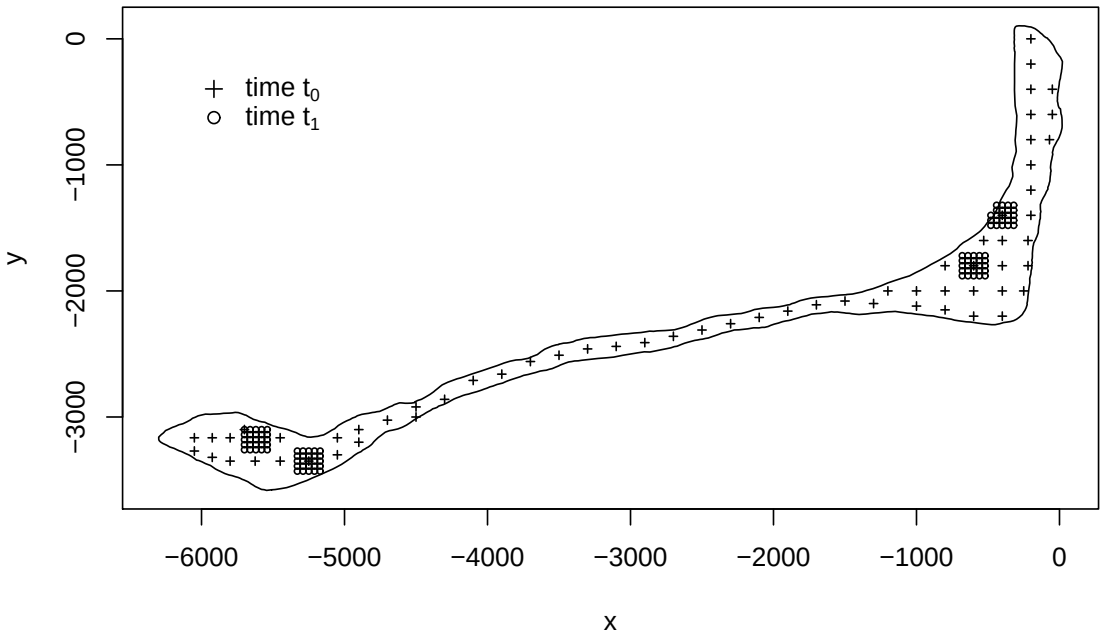


Figure 1: Rongelap's island: two-stage strategy of uniform and clustered samples.

Our variables of interest from Rongelap data set are the spatial coordinates $\mathbf{x_i}$, the counts Y_i of radioactive emissions at each location, the length l_i of time over which the counts are recorded, and the stage in sampling measurements were made. The total sample size is $n = 161$. Note that the Y_i are treated as realizations of mutually independent Poisson random variables with expectations $l_i\lambda(\mathbf{x_i})$, where $\lambda(\mathbf{x})$ measures the local radioactivity at location $\mathbf{x}$. We chose the data transformation $Z_i = \sqrt{Y_i/l_i}$ to make the variability more consistent and more Gaussian[1].

3 Assessing through simulation

After introducing our motivating data set, we move to simulated data to develop and study our diagnostic tools for data analysis. It is well known that simulation allows a level of knowledge and control that leads to more robust and defendable solutions. Using simulated data sets, where the characteristics of the data and the sampling designs are controlled and varied, will help the research of the technique's potential, and to assess its performance in specific situations. We can gain insight about what happens when assumptions are violated since the true model is known.

[1]According to Delta method used to estimate a variance of a transformed parameter, one has $\mathrm{Var}[G(T)] \simeq \mathrm{Var}[T] \times (G'(\mu))^2 = const$, where $T = Y/l$, $\mathrm{E}[T] = \mathrm{Var}[T] = \mu$ and $G(T) = \sqrt{T}$.

3.1 Sample generation algorithm

Typically, when one carries out some study of geostatistical data, the sample locations are uniformly spread over the observation region. Suppose now that one wishes to proceed with a multi-stage collection of data. If the goal is to better characterize the spatial variability for short distances, then one solution is to include some clusters of locations into later stages. Alternatively, suppose the goal is, as exemplified before, to pursue the maximum values of the spatial variable of interest, then the complete sample data set is expected to be mainly represented by large data values. These previous situations may condition the sampling design. In our simulation studies, we shall then consider four distinct sampling designs: complete spatial randomness (CSR); just clustered; biased but non-clustered; and, finally, biased and clustered.

The sample generation algorithm presented here considers the case of a two-stage approach for sampling collection, with the second stage potentially influenced by the first. It can be easily extended to more than two stages, even though our experience confirmed that similar results are obtained.

We consider spatial locations $\mathbf{x}$ within the unit square $[0,1] \times [0,1]$. A theoretical variogram is chosen to model the spatial dependency structure. Data sets are then generated with Gaussian data, $Z(\mathbf{x_1}), ..., Z(\mathbf{x_n})$, where $Z(\mathbf{x})$ denotes the spatial random process. The proposed algorithm allows the generation of K clusters, each one inside a sub-region R_k. For example, if one wishes to produce a biased sample with just one cluster, this can be done by restricting the sampling points from stage 2 to R_1 and around the maximum of measurements from stage 1. The total sample size will be n, with n_1 from stage 1 and n_2 from stage 2. The algorithm may be summarized as follows

1. Sample n_1 points $\mathbf{x_i}$ at random on $[0,1] \times [0,1]$;
2. Generate $\mathbf{Z} = (Z(\mathbf{x_1}), ..., Z(\mathbf{x_{n_1}})) \sim MVN$;
3. For $k = 1, ..., K$
   (a) If biased=TRUE then select $Z(\mathbf{x_{m,k}}) = \max_i \{Z(\mathbf{x_i}) | \mathbf{x_i} \in R_k\}$
   else select $Z(\mathbf{x_{m,k}}) = \texttt{random}_{\,i} \{Z(\mathbf{x_i}) | \mathbf{x_i} \in R_k\}$;
   (b) Sample $n_{2,k}$ points at random on $[\mathbf{x_{m,k}} - \theta \ , \ \mathbf{x_{m,k}} + \theta]^2$;
4. Consider $n_2 = \sum_{k=1}^{K} n_{2,k}$;
5. Generate $\mathbf{Z}^* = (Z(\mathbf{x_{n_1+1}}), ..., Z(\mathbf{x_{n_1+n_2}}))$ where
$$\mathbf{Z}^* \mid \mathbf{Z} \sim MVN \left(\Sigma_{12}^T \Sigma_{11}^{-1} \mathbf{Z} \ , \ \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12} \right)$$
and $\Sigma_{22} = \text{var}\{\mathbf{Z}^*\}$, $\Sigma_{11} = \text{var}\{\mathbf{Z}\}$, $\Sigma_{12} = \text{cov}\{\mathbf{Z}, \mathbf{Z}^*\}$;

The θ parameter in step 3 defines the size of the cluster; moreover, points are rejected if not within the observation region. The conditional distribution from step 5 was derived from the joint distribution using properties of the multivariate Gaussian distribution. Additionally, bear in mind that a completely random sample can be obtained avoiding stage 2, i.e. $n_2 = 0$, or generating the n_2 points uniformly spread over all unit square. Moreover, the cluster effect tends to disappear for a large K.

3.2 Impact on variogram estimation

We now want to analyse the impact of clustered and biased multi-stage sampling on variogram estimation. We consider a stationary and isotropic spatial process $\mathrm{Z}(\mathbf{x})$, in which case the variogram reduces to $2\gamma(\|\mathbf{x_i}-\mathbf{x_j}\|) = \mathrm{E}[(Z(\mathbf{x_i})-Z(\mathbf{x_j}))^2]$. Consequently, we can estimate the variogram from sample data replacing the previous theoretical expectation by the corresponding sample average. The variogram estimator most commonly adopted was proposed by Matheron in [12] and it can be represented by the weighted average

$$2\widehat{\gamma}(u) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(u)[Z(\mathbf{x_i})-Z(\mathbf{x_j})]^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(u)}, \quad u \in I\!R$$

where $w_{ij}(u) = I_{\{\|\mathbf{x_i}-\mathbf{x_j}\|=u\}}$. In practice, this estimator is usually smoothed by taking a tolerance region T around u. Alternatively, one can take the weights as $w_{ij}(u) = K\left(\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right)$, where K is a symmetric, zero-mean and bounded density function, with compact support $[-C, C]$. The positive number h is usually called bandwidth. The resulting variogram estimator is commonly referred to as the kernel estimator. Bear in mind that the weights are at their maximum when the distance between two points is close to u, and zero values if $\left|\frac{u-\|\mathbf{x_i}-\mathbf{x_j}\|}{h}\right| > C \iff \|\mathbf{x_i}-\mathbf{x_j}\| \notin [u-hC, u+hC]$. Consequently, it offers a smoother estimation of the variogram.

Both Matheron and the kernel variogram estimators are included in most software available to practicing statisticians/geostatisticians. An example is the geoR library from R, described in [13], which offers a kernel estimator for exploratory purposes with the bandwidth being chosen by the user. In [14], it is suggested a transformed version of the kernel estimator, not restricted to exploratory aims but allowed to be used in kriging. They adapt the Nadaraya-Watson regression estimation to the context of spatial data and propose an asymptotically optimal bandwidth parameter. In [15], the performance of the NW kernel estimator and the Matheron one are compared, under different spatial correlation models; the results suggest the usual superiority of the former estimator.

In this work, we also take the Matheron and NW kernel estimators, aiming to analyse the effect of clustered and biased multi-stage sampling. The numerical study of these estimators' behaviour must be based on results from several independent cases. We then generate a total of 100 independent data sets and, for each one, derive the integrated square error (ISE) between the estimator and the theoretical variogram. The ISE, defined as $\int [\widehat{\gamma}(u)-\gamma(u)]^2\, du$, was approximated numerically through the trapezoid rule. We chose a Matérn model to model the spatial dependency, with a Bessel function of order $\kappa = 1$, a range equals to 0.2 (distance beyond which the correlation between variables is zero) and a partial sill equals to 2.25 (corresponds to $\mathrm{Var}[Z(\mathbf{x})]$).

In Table I, we summarize the mean value of the ISE, considering the 4 possible combinations of biased two-stage sampling and clustered sampling. To generate a biased but not strongly clustered sample, we split the sample grid into 25 sub-areas. We start to generate randomly 75 values and locations in the total area and then generate 5 more points clustered around the maximum of previous measurements in each sub-area. A final sample size of 200 is obtained. For standardization reasons, in the remaining cases, n_1 equals to 75 and n_2 equals to 125 is also chosen.

The results from Table I, for Matheron and NW kernel estimators, suggest a poor performance for both estimators under biased clustering. Note that Table I also includes the results of two other variogram estimators, *RobClust* and *Pooled*, that will be introduced in later Sections of this paper. We now want to restrict our attention to the Matheron and NW kernel estimators.

One may observe that there is a larger degradation for the Matheron's estimator. In fact, when all lags less than or equal to 0.6 are considered, this estimator and the kernel one grow worse 4.9 and 3.2 times, respectively. The worst results normally associated to Matheron's estimator tend to be not so obvious for smaller lags. This should be an indirect consequence of the typical less satisfactory behaviour of kernel estimators in boundaries. In any case, with respect to larger lags under *just biased* or *just clustered* sampling designs, note that the NW kernel estimator performs quite well. Finally, from Table I, the clustering issue seems to have a larger impact on variogram estimates than the sequential dependence issue.

4 Data exploratory methods

We have shown that the non-standard sampling designs described in Section 3.1 are responsible for a more difficult estimation of the spatial dependency structure. This suggests the need for detecting biased multi-stage sampling and, ideally, for correcting solutions.

Still using simulated data, we first investigate data exploratory tools to reveal hidden dependency patterns in a given sample data set and, then, we present an hypothesis test based on those tools. More precisely, the practical part of this research is concerned with the exploratory analysis of sampled data in order to understand if it is reasonable to assume dependency between data values and locations and if this dependency is indeed sequential.

4.1 Detection of sequential dependence

In a context of marked point processes, Schlather et al. in [2] investigate marks and locations interactions, by introducing functions of the inter-point distance u, under the assumption of stationary and isotropy. One of these functions denotes the conditional expectation of a mark, given that there is a further point of the process a distance u away. Writing $\Phi = \bigcup_i \mathbf{x_i}$ for the corresponding unmarked point process, it may be represented by:

$$E(u) = \mathrm{E}\left[Z(\mathbf{x}) \mid \mathbf{x}, \mathbf{x}' \in \Phi,\ ||\mathbf{x} - \mathbf{x}'|| = u\right].$$

These authors then present tests based on E for the hypothesis of dependency between the values of the marks and their locations.

In order to decide if existing dependency is sequential, we propose a new version for the conditional expectation function, denoted by $E_{seq}(u)$, restricted to the *latest* values of the spatial variable. This new function can be defined as:

$$E_{seq}(u) = \mathrm{E}\left[Z(\mathbf{x}) \mid \mathbf{x}, \mathbf{x}' \in \Phi,\ ||\mathbf{x} - \mathbf{x}'|| = u,\ t(\mathbf{x}) > t(\mathbf{x}')\right],$$

where $t(.)$ identifies the stage when data were collected, i.e. a time label. It is then assumed that the analyst is aware of how, or if one prefers when, the collection of the sample data occurred, making time labels available.

Both conditional expectation functions, $E(u)$ and $E_{seq}(u)$, can be approximated through a sample average, but the second considers a sub set of the total data values considered for the first. Allowing a tolerance region for lag u, our estimator can be defined as

$$\widehat{E}_{seq}(u) = N_u^{-1} \sum_{\substack{|\|\mathbf{x_i}-\mathbf{x_j}\|-u| \leq \frac{\varepsilon}{2} \\ t(\mathbf{x_i})>t(\mathbf{x_j})}} Z(\mathbf{x_i})$$

where $\varepsilon > 0$ is a fixed bin-width and N_u is the number of pairs $(\mathbf{x_i}, \mathbf{x_j})$ for which $|\|\mathbf{x_i}-\mathbf{x_j}\|-u| \leq \frac{\varepsilon}{2}$.

The behaviour of these functions was then investigated through a new simulation study aiming to analyse the influence of sequential biased and clustered sampling. We have considered 1000 independent data sets and the same features chosen for our previous study: same sampling designs and spatial dependency structure.

In Figures 2 and 3, we plot the mean of 1000 estimated conditional expectation functions, given by $\widehat{E}_{seq}(u) - \widehat{E}(u)$, $\widehat{E}_{seq}(u) - \widehat{E}(Z)$ and $\widehat{E}(u) - \widehat{E}(Z)$. The confidence intervals (CI) for the sampling distribution of differences constructed from 1000 samples were added to check the variability of these estimations.

We conclude that under the absence of a sequential biased sample, and just in this case, the difference functions $\widehat{E}_{seq}(u) - \widehat{E}(u)$ and $\widehat{E}_{seq}(u) - \widehat{E}(Z)$ are approximately zero. The corresponding CIs embrace the theoretical $E_{seq}(u) - E(u) = E_{seq}(u) - E(Z) = 0$. Otherwise, with or without the presence of strong clustering, our two difference functions are clearly non-zero, reflecting the existence of bias in *latest* data points (higher values in our simulation).

The plots illustrate the following dependency pattern of a biased multi-stage collection of sample data

$$E_{seq}(u) - E(u) \neq 0,$$

which we shall adopt in our proposal.

4.2 Monte Carlo tests

The widely used Monte Carlo (MC) significance testing was originally proposed in [16] and its basic idea is as follows. Suppose H_0 is the null hypothesis about the model which generates $Y = \{(\mathbf{x_i}, Z(\mathbf{x_i})) : i = 1, ..., n\}$, and r_1 is an observed value of a real valued statistic $R = h(Y)$, which has a distribution function F, possibly mathematically intractable. Moreover, suppose we agree to reject H_0 for a *large* value of r_1.

Hence, we can use pseudo-random numbers to simulate a random sample $r_2, ..., r_m$ of $m-1$ observations from distribution F and to construct a test by comparing these simulated values with r_1. If F is continuous and $k = 1 + \#\{j : j = 2, ..., m \; and \; r_1 > r_j\}$, then H_0 will be rejected at the k/m attained significance level, since the rank of r_1 is uniformly distributed on the integers $1, ..., m$ when H_0 is true. See [17] for a general discussion of Monte Carlo tests. Note that the parametric bootstrap techniques work in a similar way to those described in here (see e.g. [18]).

In our work, we are interest in a test for the hypothesis that a given data set does not incorporate sequential biasing, so that we shall define

$$H_0 : \; E_{seq}(u) - E(u) = 0.$$

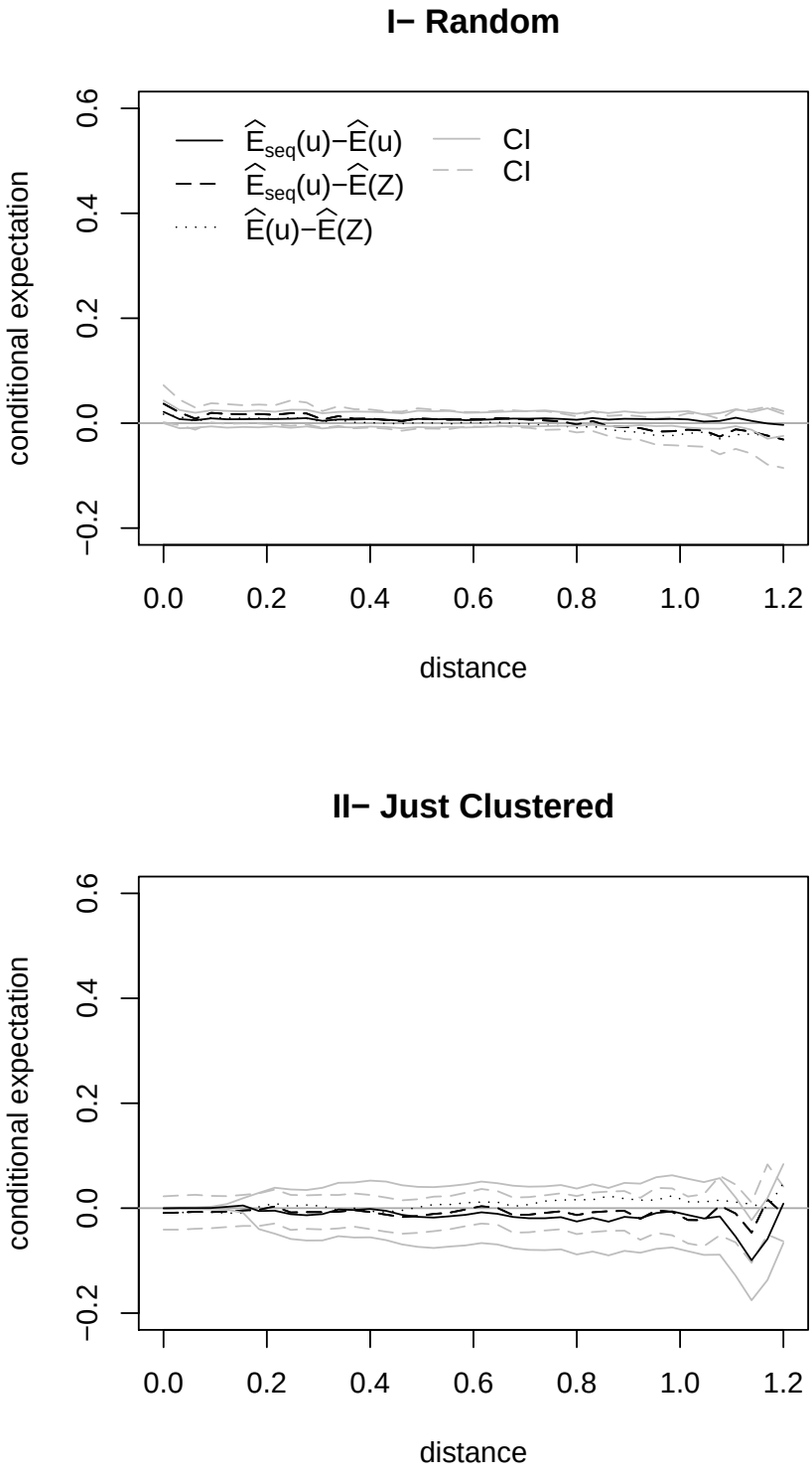
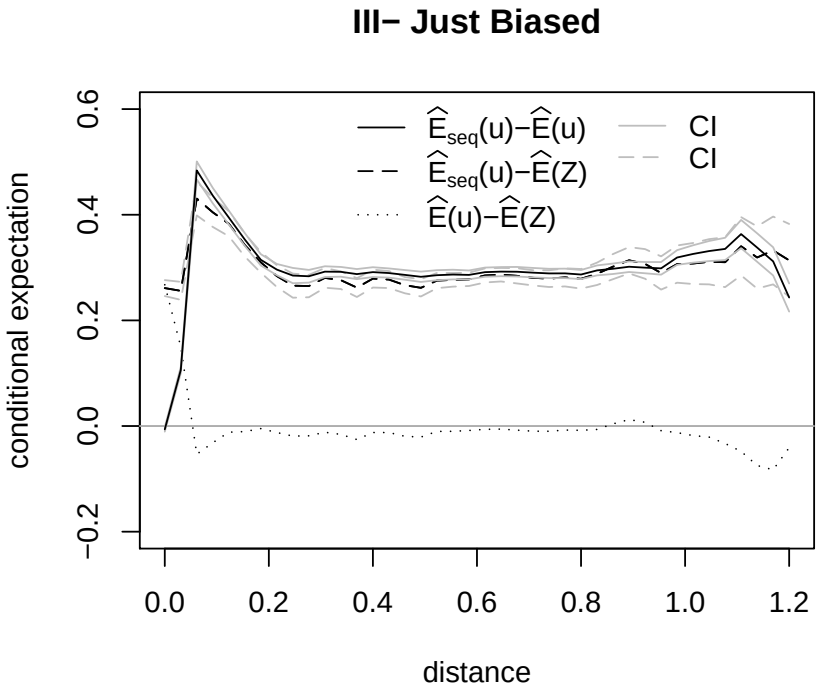


Figure 2: Part I - mean values of estimated conditional expectation functions. Total of replicas equals to 1000.



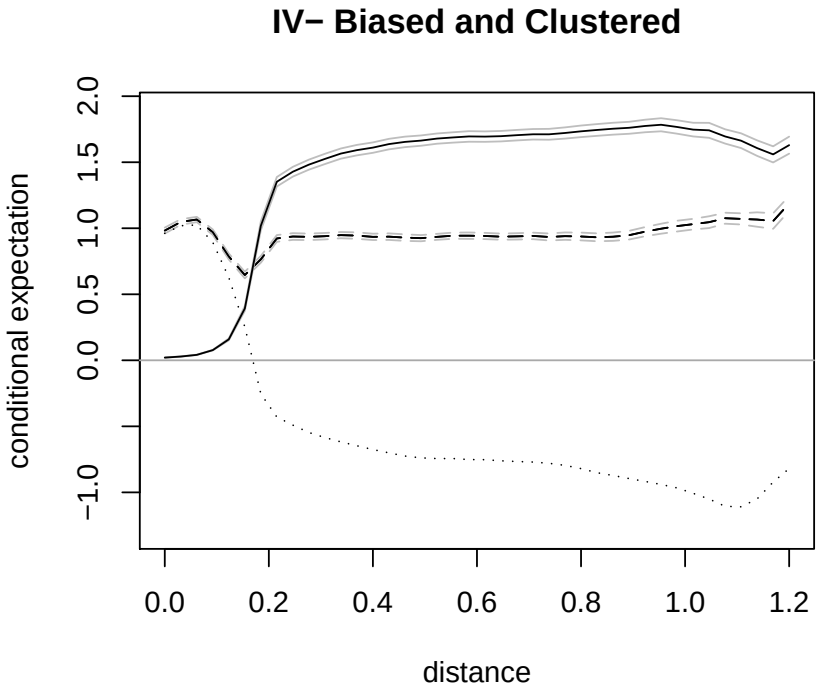


Figure 3: Part II - mean values of estimated conditional expectation functions. Total of replicas equals to 1000.

Under this hypothesis, the spatial process can be generated by sampling a random field Z at the given locations $\mathbf{x_i}$, $i = 1, ..., n$, with no sequential dependence. In this way, we can simulate $m - 1$ further data sets under H_0, and define r_j to be a measure of discrepancy between $\widehat{E}^j_{seq}(u)$ and $\widehat{E}^j(u)$ over the whole range of u. For example, our test statistic can be given by the integrated squared difference

$$r_j = \int \{\widehat{E}^j_{seq}(u) - \widehat{E}^j(u)\}^2 du.$$

We can then proceed to a formal test based on the rank of r_1 amongst r_j, because under H_0 all ranking of r_1 are equiprobable. Bear in mind that m is rather smaller than might perhaps be expected, in contrast with the much larger sample which would be needed for accurate estimation of F, the distribution function of R. According to [19], for a one-sided test at the conventional 5% level of significance, $m = 100$ is suitable.

Additionally, a preliminary rough visual guide to address the problem being investigated can be provided by means of the well-known "simulation envelopes" plot. Testing involves comparing an observed test statistic with samples from the model under consideration. Consequently, this visual approach is based directly on the variation in estimates obtained from data generated from the model. The maximum and minimum of the total $m-1$ independent simulations allow the definition of *upper* and *lower envelopes*. See Figure 4, for an example.

Diggle in [5] emphasises the use of such a plot as a visual aid to interpretation. Comparison of the observed curve $\widehat{E}^1_{seq}(u) - \widehat{E}^1(u)$ with that expected from a random arrangement of $\widehat{E}^j_{seq}(u) - \widehat{E}^j(u)$, $j = 2, ..., m$ allows an assessment of the overall degree of coverage. If the observed curve lies between the two envelopes, this suggests the acceptance of hypothesis H_0. If the observed curve exceeds the envelopes for some distances u, this is an initial and informal indication of the possibility of H_0 rejection. Anyway, in our case, we prefer to deepen analysis and to proceed with a formal Monte Carlo test.

4.3 Rongelap island's data

First, we need to simulate $j = 2, ..., m$ datasets under H_0, i.e. "Rongelap island's data does not incorporate sequential biasing". It was found convenient to start with the maximum likelihood estimation of the spatial dependency structure. So, we consider a variogram estimator, obtained by using the coarse data and, derived through restricted maximum likelihood (REML).

In Figure 4, we present the results of our Monte Carlo test. We generate 99 simulations of a random field over the total 157 distinct[2] locations of Rongelap's island. From this plot, we would not reject the null hypothesis, as confirmed through a formal test. So, we would tend to refuse the existence of sequential bias. However, when one replaces this variogram estimator for one of those proposed in Section 5, this tendency is not that clear.

The previous approach requires model assumptions, like Gaussianity of data. If one wishes to avoid it, an alternative Monte Carlo method can be supported by the theory of randomization tests. The basic idea is to calculate a test statistic from the observed data, and then reshuffle the data a large number of times, recalculating the test statistic for each

[2]Four locations were overlapped in fine and coarse grids.

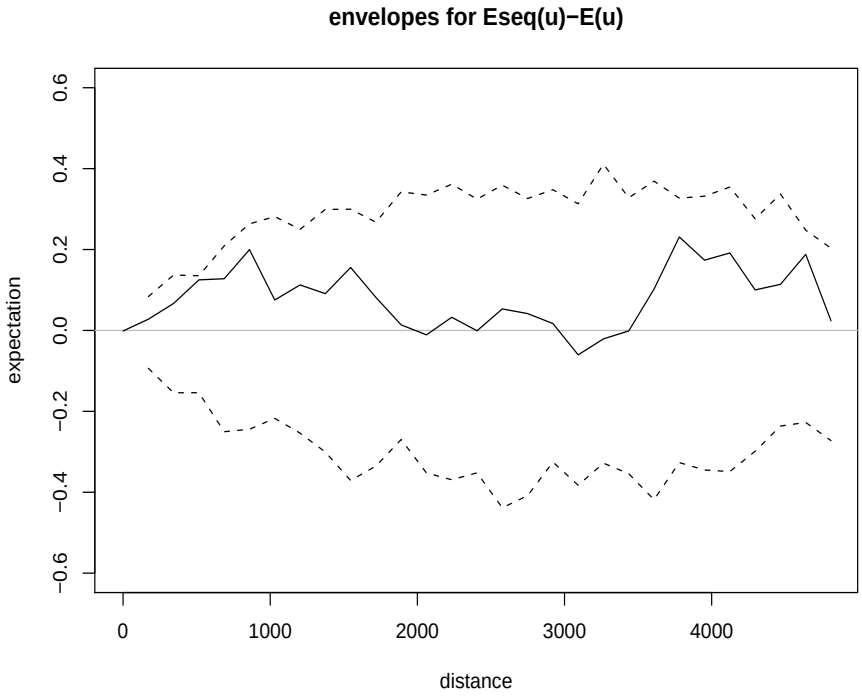


Figure 4: Simulation envelopes of $E_{seq}(u) - E(u)$ for Rongelap island's data, with $\widehat{\gamma}$ obtained through REML: data (solid curve); upper and lower envelopes from 99 simulations of a random field (dashed curves).

iteration. These statistics are used as before to generate a distribution of values. The observed value can be compared to the distribution to see whether the observed case is a tail value, i.e. an event that is unlikely to occur through chance. The latter tests are sometimes referred to as permutation ones, because the randomization can be done by reordering the positions of elements in an array.

In geostatistics, a natural permutation test can be derived when the actual data values are maintained, but they are randomly permuted in order to obtain the distribution of the test statistic. Exactly how they are permuted depends on the null hypothesis to be tested.

In our case, for testing $H_0 : E_{seq}(u) - E(u) = 0$ on Rongelap data, we suggest the following non-parametric approach. Suppose the locations and values for the first stage of the sampling were fixed a priori, then we can assess the variation in the test statistic over randomisation of the second stage sampling. In here, to keep avoiding the assumption of a model for the spatial process, we can select at random over all the locations from the two stages and using the observed values at these selected sites, this would avoid the need for a model.

The results plotted in Figure 5 were derived following this type of approach. We fixed as true the 63 sampled locations and values from the first stage

$$\{(\mathbf{x_i}, Z(\mathbf{x_i})) : \ i = 1, ..., 161, \quad t(\mathbf{x_i}) = t_0\}.$$

For each simulation, we chose randomly 98 extra data points $\mathbf{x_k}$, among the total 161 avail-

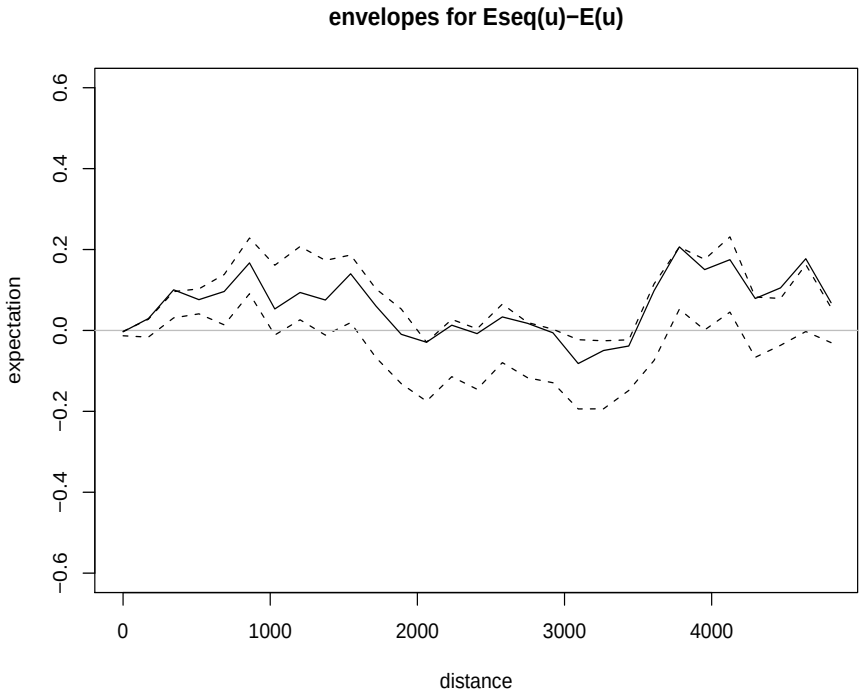


Figure 5: Simulation envelopes of $E_{seq}(u) - E(u)$ for Rongelap island's data, using a non-parametric approach: data (solid curve); upper and lower envelopes from 99 simulations of a random field (dashed curves).

able and we got a new data set representative of the second stage

$$\{(\mathbf{x_k}, Z(\mathbf{x_k})) : t(\mathbf{x_k}) = t_0 \text{ or } t(\mathbf{x_k}) = t_1\}.$$

We could then derive the conditional expectation functions $\widehat{E}^j_{seq}(u)$ and $\widehat{E}^j(u)$ for $j = 1, ..., 99$. According to Figure 5, we realize that this non-parametric approach gives a narrow envelope interval when compared to the one obtained through REML in Figure 4, probably because of the smaller variability associated to a permutation test. These simulation envelopes actually suggest a possible rejection of H_0. However, this rejection was not confirmed with the formal test. The observed test statistic r_1 was the 92^{th} largest of all values r_j, so H_0 should be accepted with an attained significance level of 0.92.

5 Non-standard sampling correctors

In Section 3.2, when analysing the Matheron's and Nadaraya-Watson kernel variogram estimators under different sampling designs, we have concluded that they may produce poor estimates of the spatial variability. This can happen when the sampling strategy causes later samples to be located in areas with atypical, usually high or low, data values. While, it is likely that these samples give good information on the spatial variance within the clusters, they are not representative of the remaining area. The naive approach of discarding clustered biased data would force us to lose useful information, as well as, it may not always be

possible to identify those that should be kept and those that should be discarded. To obtain a good estimate of the global spatial variance, one may claim a method of weighting individual samples and clustered ones, in such a way the latter do not have an undue influence on the estimate.

5.1 Method to adjust for clustering

Our first concern is then the clustering issue. We propose to modify the NW kernel estimator $2\widehat{\gamma}(u)$ in Section 3.2 trying to adjust for clustering of samples and minimize the negative impact on variogram estimation.

Consequently, a compensation for the unpopulated areas is proposed, by suggesting an inverse weight to a given neighbourhood density and, simultaneously, joining the benefits outcome from a kernel estimator. A possible way to extend weights $w_{ij}(u) = K\left((u - \|\mathbf{x_i} - \mathbf{x_j}\|)/h\right)$ to adjust for clustering is to use

$$w_{ij}(u) = \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right), \qquad n_i = \sum_k I_{\{\|\mathbf{x_i} - \mathbf{x_k}\| \leq \delta\}}$$

where n_i represents the number of points that fall within the circle of radius δ and center $\mathbf{x_i}$.

In [20], this new variogram estimator is proved to enjoy good properties, such as asymptotically unbiasedness and consistency. Additionally, it is also proposed optimal values the neighbourhood radius δ and the kernel bandwidth h. The first results from the analysis of the density estimation derived on the observation region. The latter is treated via the MSE, i.e. the minimum square error.

5.2 Sequential biased corrector

The second concern is about the bias possibly present in the final sample data set, when some type of multi-stage sampling design is adopted. Here, we propose a very naive approach. If the exploratory analysis from previous Sections suggests the presence of sequential bias, we propose to slightly modify the adjust for clustering's method in such a way that those data values included and those not included into a region regarded as sequential biased would not be mixed. Once more, the implementation of this proposal assumes certainly that one keeps track of time labels associated to each data value, i.e., that knowledge about how the multi-stage collection of data occurred is made available. Then, the previous weight expression changes to

$$w_{ij}(u) = \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathbf{x_i} - \mathbf{x_j}\|}{h}\right) \times I_{\{t(\mathbf{x_i}) = t(\mathbf{x_j})\}}, \quad n_i = \sum_k I_{\{\|\mathbf{x_i} - \mathbf{x_k}\| \leq \delta\}}$$

Under biased clustered samples, this resulting *pooled* variogram estimator can be roughly described as using latest data values for small lags' estimations and the remaining data values otherwise. Under the absence of sequential bias, this estimator should produce very similar results to estimator from Section 5.1.

5.3 Simulation study re-visited

The comparison study of variogram estimators described in Section 3.2 can now be concluded. In Table I, we include the results achieved by the variogram estimators proposed in the two previous Sections, tagged *RobClust* and *Pooled*, respectively. Please remember that, for this simulation study, we have considered a two-stage approach for sampling collection, with the second stage possibly influenced by the first, suggesting four possible combinations of biased sampling and clustered sampling.

Under random sampling, the four estimators present similar results, with just a slightly better performance for the three kernel estimators. The best improvement accomplished by the *Pooled* estimator occurs under simultaneously biased and clustered sampling, when the errors decrease 4.5 times and 7.2 times, when compared to NW kernel and Matheron estimators, respectively. Under just clustered sampling, these same values decrease 1.2 and 2.0 times, respectively. For the last combination, bias but no strong clustering, the *Pooled* estimator origins values 1.9 and 1.4 smaller than NW kernel and Matheron estimators.

As expected, the new estimators *RobClust* and *Pooled* present very similar ISE values under the absence of sequential sampling. However, under simultaneously biased and clustered sampling, there is some practically relevant improvement of our naive *Pooled* estimator over existing methods. This suggests to consider time labels in the estimates whenever possible.

6 Discussion on Rongelap island's data

The direct observation of the Rongelap island's data, in Figure 1, allow us to conclude that some sub-areas were more intensively sampled than others. This spatial sampling design was adopted to better characterize short-range variability, which requires a denser sampling, but sometimes too costly to cover the whole study region. However, if later sample locations concentrated in sub-regions corresponding to atypical high (or low) values of the measurements made at earlier sample locations, then the total data points are not equally representative of the overall data. Thus, the traditional methods of geostatistics must be carefully adopted or even avoided.

In this closing Section, we do the assessment of the proposed variogram estimators on Rongelap data, when testing for the presence of sequential dependency. In Section 4.3, the Monte Carlo test suggested for the Rongelap data has employed a variogram estimator derived through maximum likelihood. This estimation was required to model the spatial dependency and to generate 99 simulations of a random field according to the null hypothesis (i.e. absence of sequential dependency). We now investigate the influence of adopting the new variogram estimators instead. Bear in mind that conditional negative-definite versions of these estimators must be used, which are obtained by fitting the empirical estimators introduced in Sections 5.1 and 5.2 to a permissible variogram given by Bochner's theorem as described in [21]. Both estimators, with similar outcomes, were applied to all data from coarse and fine grids.

In Figure 6, we choose to illustrate the Monte Carlo test related to the *Pooled* estimator. The observed curve $\widehat{E}_{seq}(u) - \widehat{E}(u)$ is outside the simulation envelopes for small and large lags. According to these results, the presence of sequential dependency in the samples should not be totally excluded. Actually, the rejection of the null hypothesis was confirmed with the proposed formal test at the 5% level of significance.

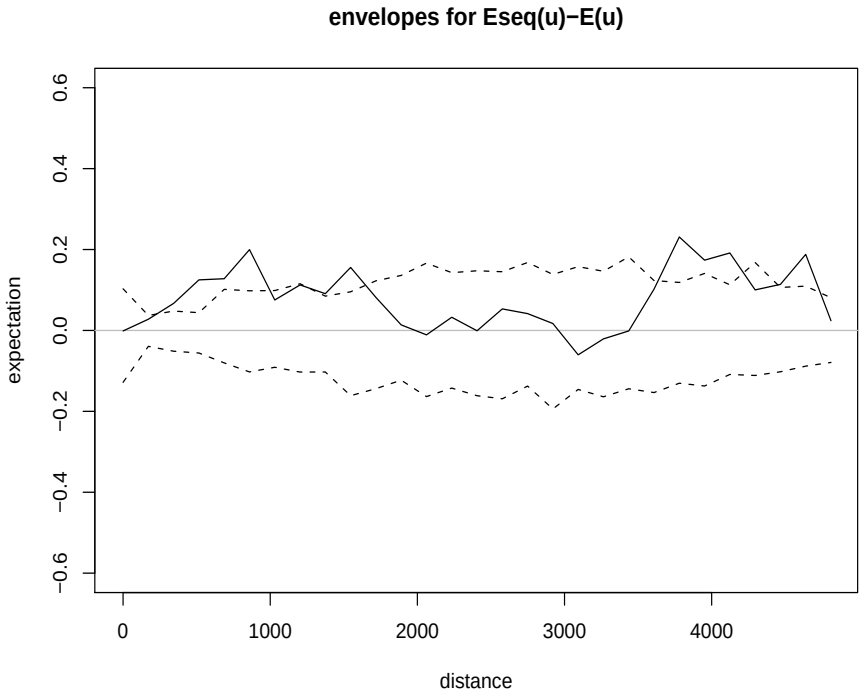


Figure 6: Simulation envelopes of $E_{seq}(u) - E(u)$ for Rongelap island's data, using the *Pooled* variogram estimator: data (solid curve); upper and lower envelopes from 99 simulations of a random field (dashed curves).

It is also noteworthy that this data set underlies some characteristics, like a low spatial variance and locations almost forming a straight line due to the island's layout, requiring a careful estimation. Consequently, corrector methods like the ones proposed are advised.

Bibliography

[1] Diggle PJ and Ribeiro Jr PJ and Christensen O (2003). An introduction to model-based geostatistics, *in* J.Moller, ed. 'Spatial Statistics and Computational Methods', **173**, Lecture Notes in Statistics, Springer, 43–86.

[2] Schlather M, Diggle PJ and Ribeiro Jr PJ (2004). Detecting dependence between marks and locations of marked point processes. *JRSS, Series B*, **66(1)**, 79–93.

[3] Mateu J and Ribeiro Jr PJ (1999). Geostatistical data versus point process data: analysis of second-order characteristics. *in* J. Goméz-Hernandéz and R. Froidevaux. eds 'GeoENV-II - Geostatistics for environmental applications', Vol.10 of *Quantitative Geology and Geostatistics*, Kluwer Academic, pp.213–224.

[4] Ripley BD (1981). *Spatial statistics.* Wiley, New York.

[5] Diggle PJ (2003). *Statistical analysis of spatial point patterns.* Arnold, London.

[6] Diggle PJ, Tawn JA and Moyeed RA (1998). Model-based geostatistics, *JRSS, Series C*, **47(3)**, 299–350.

[7] Walder O and Stoyan D (1996). On variograms in point processes statistics, *Biometrical Journal*, **38(8)**, 895–905.

[8] Schlather M (2002). Characterization of point processes with Gaussian marks independent of locations. *Mathematische Nachrichten*, **240**, 204–214.

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

[10] Isaaks EH and Srivastava RM (1989). *An introduction to applied geostatistics*. Oxford University Press.

[11] Diggle PJ, Harper L and Simon SL (1997). Geostatistical analysis of residual contamination from nuclear weapons testing, *in* V.Barnett and K.Feridun-Turkman eds, *Statistics for the Environment 3: Pollution Assessment and Control*, John Wiley & Sons Ltd, 89–107.

[12] Matheron G (1963). Principles of geostatistic. *Economic Geology*, **58**, 1246–1266.

[13] Ribeiro Jr P and Diggle P (2001). geoR: A package for geostatistical analysis. *R News*, bf 1, 2, ISSN 1609–3631.

[14] Garcia-Soidán P Febrero-Bande M and Gonzalez-Manteiga W (2004). Nonparametric kernel estimation of an isotropic variogram. *J. Statist. Plann. Inference*, **121**, 65–92.

[15] Menezes R, Garcia-Soidán P and Febrero-Bande M (2005). A comparison of approaches for valid variogram achievement. *J.Comput.Stat.*, **20**, 4, 623–642.

[16] Barnard GA (1963). Discussion of Professor Bartlett's paper. *JRSS, Series B*, **25**, 294.

[17] Besag J and Diggle PJ (1977). Simple Monte Carlo tests for spatial pattern, *JRSS, Series C*, **26**, 327–333.

[18] Gentle JE (2002). *Elements of computational statistics*, Springer Verlag.

[19] Hope ACA (1968). *A simplified Monte Carlo significance test procedure*. JRSS, Series B, **30**, 582–598.

[20] Menezes R, Garcia-Soidán P and Febrero-Bande M (2006). A kernel variogram estimator for clustered data. *Scand.J.Stat.*, **35**, 1, 18–37.

[21] Shapiro A and Botha JD (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Comput. Statist. Data Anal.*, **11**, 87-96.

Table 1: Comparison of four distinct variogram estimators, through the mean values of the evaluated ISE. Four distinct sampling designs were considered, from simultaneously biased and clustered sample to a completely random sample. Total of replicas equals to 100 and each replica total sample size equals to 200.

Sampling Design	$\widehat{\gamma}(u)$	$u \leq 0.6$	$u \leq 0.3$	$u \leq 0.2$	$u \leq 0.1$
Random	Matheron	0.608	0.239	0.123	0.032
	NW kernel	0.580	0.275	0.161	0.043
	RobClust	0.571	0.265	0.154	0.042
	Pooled	0.574	0.264	0.153	0.041
Just Clustered	Matheron	0.957	0.453	0.315	0.070
	NW kernel	0.567	0.248	0.159	0.063
	RobClust	0.512	0.260	0.164	0.059
	Pooled	0.472	0.238	0.152	0.059
Just Biased	Matheron	0.685	0.350	0.246	0.111
	NW kernel	0.507	0.268	0.164	0.048
	RobClust	0.496	0.255	0.154	0.044
	Pooled	0.352	0.177	0.105	0.027
Biased and Clustered	Matheron	2.989	0.766	0.336	0.090
	NW kernel	1.882	0.338	0.176	0.071
	RobClust	1.102	0.402	0.218	0.068
	Pooled	0.415	0.212	0.142	0.061