Assessing spatial dependence for clustered data

Raquel Menezes[1], Pilar García-Soidán[2], and Manuel Febrero[3]

[1] Universidade do Minho, Dep. Matemática para a Ciência e Tecnologia, Campus de Azurém, 4800-058 Guimarães, Portugal, rmenezes@mct.uminho.pt
[2] Universidade de Vigo, Fac. Ciencias Sociais e da Comunicación, Campus A Xunqueira, C.P. 36005 Pontevedra, Spain, pgarcia@uvigo.es
[3] Universidade de Santiago, Fac. Matemáticas, Campus Sur, C.P. 15782 Santiago, Spain, mfebrero@usc.es

Summary. Variogram analysis provides a useful tool for measuring the dependence between spatial locations. Suppose that the nature of the sampling process leads to the presence of clustered data; the latter makes it advisable to use a variogram estimator that aims to adjust for clustering of samples. In this setting, the use of a nonparametric weighted estimator, obtained by considering an inverse weight to the neighborhood density combined with the kernel method, seems to have a satisfactory behavior in practice. Thus, we proceed in this work with the theoretical study of the latter estimator, by proving that it is asymptotically unbiased as well as consistent and by providing criteria for selection of the bandwidth parameter and the neighborhood radius.

Key words: Bandwidth, clustered data, consistency, neighborhood radius, variogram.

1 Introduction

Among existing geostatistical methods, variogram analysis may provide a useful tool for summarizing spatial data and a measure of spatial dependence between samples. Typically, one assumes that the sampling points are uniformly spread over the observation region. In this setting, we may use the empirical estimator, analyzed in [Mat63], or consider instead the Nadaraya-Watson semivariogram, detailed in [GFG04]; see also [MGF05] for a review of several approaches, which are put into comparison in a numerical study covering different spatial dependence situations.

However, the sampling strategy may originate unequal samples density, leading to the presence of clustered data. A possible reason might be related to the adoption of a denser sampling in areas that are deemed critical (e.g. maximum values search). Clustered locations may also be driven from external factors, like for example the existence of specific geographic or demographic

spots, or even they may be needed to better characterize short-range variability. In this context, the behavior of the traditional variogram estimators may significantly decay.

Declustering methods are quite intuitive, and their need is well recognized in the spatial statistics literature to estimate spatially representative mean trends for clustered data. On the contrary, the corresponding need for the reliable estimation of the second-order spatial structures is not normally considered. The presence of clustered sample data is, however, not negligible at all as analyzed in [KC04].

Consequently, a compensation for the unpopulated areas is proposed, by suggesting an inverse weight to a given neighborhood density and, simultaneously, joining the benefits outcome from a kernel estimator; see [Men05] for details. In this work, we shall prove that the variogram estimator proposed for clustered data, which will be designed as RobCluster estimator, enjoys good properties, such as asymptotic unbiasedness and consistency. In particular, the dominant terms of the bias and the variance will be established. A numerical study has also been included to give account of the better performance of the new kernel estimator when compared to the other estimators, in the presence of clustered data.

The RobCluster estimator requires the selection of two user-adjustable values: the kernel bandwidth and the neighborhood radius. The first will be treated via the MSE, i.e. the minimum square error. The latter will result from the analysis of the density estimation derived on the observation region.

2 Definitions and assumptions

A random process $\{Z(\mathrm{x}) : \mathrm{x} \in D \subset \boldsymbol{R}^d\}$ is defined as intrinsically stationary and isotropic, with semivariogram γ, if the following conditions are satisfied:

(i) $\mathrm{E}[Z(\mathrm{x_i}) - Z(\mathrm{x_j})] = 0$, for all $\mathrm{x_i}, \mathrm{x_j} \in D$.
(ii) $\mathrm{Var}[Z(\mathrm{x_i}) - Z(\mathrm{x_j})] = 2\gamma(\|\mathrm{x_i} - \mathrm{x_j}\|)$, for all $\mathrm{x_i}, \mathrm{x_j} \in D$, where $\|.\|$ denotes the euclidean norm.

To ensure consistency in the estimation of γ, we will follow the strategy proposed in [HFH94] so that the observation region D will be considered to be increasing and a random design will be assumed for the spatial locations.

(A1) $D = D_n = \lambda D_0$ where $\lambda = \lambda_n \overset{n\to\infty}{\longrightarrow} +\infty$ and $D_0 \subset \boldsymbol{R}^d$ is a bounded and fixed region.
(A2) $\mathrm{x_i} = \lambda \mathrm{v_i}$, $i = 1...n$, where $\mathrm{v_i}$ is a realization of a random sample $\mathrm{V_i}$ from f_0, the density function defined on D_0.
(A3) For all $\mathrm{v} \in D_0$ and for some positive constants d_1 and d_2, one has $d_1 < f_0(\mathrm{v}) < d_2$.

Additionally, some hypotheses will be imposed on the random process.

(A4) γ admits three continuous derivatives in a neighborhood of u, for all $u > 0$.

(A5) There is a bounded and continuously differentiable function $g : \boldsymbol{R}^{3d} \to \boldsymbol{R}$ satisfying that

$$\mathrm{Cov}\left[(Z(\mathrm{x_i}) - Z(\mathrm{x_j}))^2, (Z(\mathrm{x_k}) - Z(\mathrm{x_l}))^2\right] = g(\mathrm{x_i} - \mathrm{x_j}, \mathrm{x_i} - \mathrm{x_k}, \mathrm{x_i} - \mathrm{x_l})$$

together with

$$\lim_{\|\mathrm{x}_2\| \geq r \bigvee \|\mathrm{x}_3\| \geq r} |g(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3)| = 0$$

with $0 < r < +\infty$.

Bear in mind that, in the context of a Gaussian process, one has

$$\mathrm{Cov}\left[(Z(\mathrm{x_i}) - Z(\mathrm{x_j}))^2, (Z(\mathrm{x_k}) - Z(\mathrm{x_l}))^2\right] =$$

$$= 2\left[\gamma(\|\mathrm{x_i} - \mathrm{x_k}\|) + \gamma(\|\mathrm{x_j} - \mathrm{x_l}\|) - \gamma(\|\mathrm{x_i} - \mathrm{x_l}\|) - \gamma(\|\mathrm{x_j} - \mathrm{x_k}\|)\right]^2$$

and, afterwards, one may take

$$g(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3) = 2\left[\gamma(\|\mathrm{x}_2\|) + \gamma(\|\mathrm{x}_3 - \mathrm{x}_1\|) - \gamma(\|\mathrm{x}_3\|) - \gamma(\|\mathrm{x}_2 - \mathrm{x}_1\|)\right]^2$$

so that condition (A5) is satisfied provided that the semivariogram is bounded and has an asymptotic range.

In what respects to the convergence rates, the following conditions will be assumed.

(A6) $\{h + \lambda^{-1} + \lambda^d n^{-1} + (nh)^{-1}\} \overset{n\to\infty}{\longrightarrow} 0$.

(A7) Take $\delta = \lambda a$, for some positive and bounded a.

Here δ is the neighborhood radius in D and $a > 0$ is the equivalent in D_0.

3 Main results

Let $\{Z(\mathrm{x}) : \mathrm{x} \in D \subset \boldsymbol{R}^d\}$ be an intrinsic and isotropic random process. Denote by $Z(\mathrm{x}_1), ..., Z(\mathrm{x}_n)$ the values of the process observed at spatial locations $\mathrm{x}_1, ..., \mathrm{x}_n$, respectively.

The kernel semivariogram estimator proposed in the case of clustered data is defined as follows:

$$\widehat{\gamma}(u) = \frac{\sum_{i=1}^n \sum_{j=1}^n \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathrm{x_i} - \mathrm{x_j}\|}{h}\right) [Z(\mathrm{x_i}) - Z(\mathrm{x_j})]^2}{2\sum_{i=1}^n \sum_{j=1}^n \frac{1}{\sqrt{n_i \times n_j}} \times K\left(\frac{u - \|\mathrm{x_i} - \mathrm{x_j}\|}{h}\right)} \qquad (1)$$

for $u \geq 0$, where h and δ represent the bandwidth and neighborhood radius selectors, respectively, and $n_i = \sum_k I_{\{\|\mathrm{x_i} - \mathrm{x_k}\| \leq \delta\}}$.

The latter estimator will be referred to as RobCluster estimator. The asymptotic results for the expectation and variance of above estimator are presented in the following theorems. The derivation of the referred results requires the assumptions introduced in previous Section, leading us to a desirable consistent estimation. Additionally, one should note that:

- under isotropy, the variogram domain is restricted to non-negative values;
- the kernel function operates on the distances $\|\mathrm{x_i} - \mathrm{x_j}\| \in [u - Ch, u + Ch]$;
- it is assumed that interval $[u - Ch, u + Ch]$ is wholly contained within the domain of γ.

According to the above, note that next results are attained on $u \geq Ch$.

Theorem 1. *Let $\{Z(\mathrm{x}) : \mathrm{x} \in D \subset \boldsymbol{R}^d\}$ be an intrinsic and isotropic random process with semivariogram γ. Assume that conditions (A1)-(A4) are satisfied. Additionally, suppose the convergence rates stated in (A6). Then, for $u \geq Ch$, one has*

$$\mathrm{E}\left[\hat{\gamma}(u)\right] = \gamma(u) + \frac{1}{2} c_K \gamma''(u) h^2 + o\left(h^2\right),$$

where $c_K = \int z^2 K(z) dz$.

This theorem also shows that the proposed estimator is asymptotically unbiasedness.

Remark 1. According to Theorem 1, the bias of $\hat{\gamma}(u)$ is of the exact order h^2, for $u \geq Ch$; however, near the endpoint 0, $u < Ch$, an order h rather than h^2 is expected, due to the boundary effect. As suggested in [GFG04], the adoption of a specific combination of boundary kernels is a possible solution to keep the same rate of convergence. Although, Theorem 1 would remain valid in practice for any $u > 0$ and large n, since the bandwidth parameter h tends to 0 as n increases.

For the analysis of the asymptotic efficiency, we proceed with the variance result of the proposed variogram estimator. A decreasing variance estimate means a growing efficiency of the estimator, as it will tend to be more accurate.

Theorem 2. *Assume the hypotheses required in Theorem 1. Additionally, suppose that assumptions (A5) and (A7) are satisfied. Then, for $u \geq Ch$, one has*

$$\begin{aligned}
\mathrm{Var}\left[\hat{\gamma}(u)\right] = & \frac{\int \frac{f_0(w_1)^2}{H(a,w_1)^2} dw_1}{\left(\int \ldots \int J_d\left(\theta_1, \ldots, \theta_{d-1}\right) d\theta_1 \ldots d\theta_{d-1} \int \frac{f_0(w_1)^2}{H(a,w_1)} dw_1\right)^2} \cdot \\
& \cdot \left(\frac{A_d(u)\ d_K}{2u^{d-1}}\ n^{-2}\lambda^d h^{-1} \ + \ B_d(u)\ n^{-1} \ + \ \frac{C_d(u)}{4}\ \lambda^{-d}\right) + \\
& + \ o(n^{-2}\lambda^d h^{-1} \ + \ n^{-1} \ + \ \lambda^{-d} + h^4) \ = \\
= \ & D(a,d,u)\ n^{-2}\lambda^d h^{-1} \ + \ E(a,d,u)\ n^{-1} \ + \ \ F(a,d,u)\lambda^{-d} \ + \\
& + \ o(n^{-2}\lambda^d h^{-1} \ + \ n^{-1} \ + \ \lambda^{-d} + h^4)
\end{aligned}$$

where $d_K = \int (K(z))^2 dz$, $J_d(\theta_1, ..., \theta_{d-1}) = (\sin\theta_1)^{d-2}(\sin\theta_2)^{d-3}... \sin\theta_{d-2}$ *and* $H(a, w_1) = \int_{\|w\| \le a} f_0(w_1 - w)dw$ *, together with:*

$$
\begin{aligned}
A_d(u) = & \int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} J_d(\theta_1, ..., \theta_{d-1}) \cdot \\
& \cdot g \left(u(\cos\theta_1, ..., \prod_{j=0}^{d-1} \sin\theta_j),\ 0,\ u(\cos\theta_1, ..., \prod_{j=0}^{d-1} \sin\theta_j) \right) d\theta_1 ... d\theta_{d-1} \\
B_d(u) = & \int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} \int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} J_d\left(\theta_{1,1}, ..., \theta_{d-1,1}\right) J_d\left(\theta_{1,2} ... \theta_{d-1,2}\right) \cdot \\
& \cdot g \left(u(\cos\theta_{1,1}, ..., \prod_{j=0}^{d-1} \sin\theta_{j,1}),\ 0,\ u(\cos\theta_{1,2}, ..., \prod_{j=0}^{d-1} \sin\theta_{j,2}) \right) \cdot \\
& \cdot d\theta_{1,1} ... d\theta_{d-1,1} d\theta_{1,2} ... d\theta_{d-1,2} \\
C_d(u) = & \int_0^{\delta} \int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} \int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} \int_0^{\pi} ... \int_0^{\pi} \int_0^{2\pi} J_d\left(\theta_{1,1}, ..., \theta_{d-1,1}\right) \cdot \\
& \cdot J_d\left(\theta_{1,2} ... \theta_{d-1,2}\right) J_d\left(\theta_{1,3} ... \theta_{d-1,3}\right)\ t^{d-1} \cdot \\
& \cdot g \left(u(\cos\theta_{1,1}, ..., \prod_{j=0}^{d-1} \sin\theta_{j,1}), \right. \\
& \quad (t\cos\theta_{1,3} - u\cos\theta_{1,2}, ..., t\prod_{j=0}^{d-1} \sin\theta_{j,3} - u\prod_{j=0}^{d-1} \sin\theta_{j,2}), \\
& \quad \left. t(\cos\theta_{1,3}, ..., \prod_{j=0}^{d-1} \sin\theta_{j,3}) \right) \cdot \\
& \cdot dt d\theta_{1,1} ... d\theta_{d-1,1} d\theta_{1,2} ... d\theta_{d-1,2} d\theta_{1,3} ... d\theta_{d-1,3}
\end{aligned}
$$

3.1 Bandwidth parameter

We now intend to use the information available in the sampled data to make guesses about the optimal kernel bandwidth h. One of the most common criteria is that of minimizing the Mean Square Error, or MSE, which is defined as

$$\mathrm{MSE}[\hat{\gamma}(u)] = \mathrm{E}\left[(\widehat{\gamma}(u) - \gamma(u))^2\right]$$

Then, according to previous results, one has

$$\mathrm{MSE}\left[\hat{\gamma}(u)\right] = \left(\mathrm{Bias}\left[\hat{\gamma}(u)\right]\right)^2 \ + \ \mathrm{Var}\left[\hat{\gamma}(u)\right] \simeq \frac{c_K^2 \gamma''(u)^2}{4} h^4 \ +$$

$$D(a,d,u)\ n^{-2}\lambda^d h^{-1} \ + \ E(a,d,u)\ n^{-1} \ + \ \ F(a,d,u)\lambda^{-d}$$

From here, the bandwidth parameter that asymptotically minimizes the MSE $[\hat{\gamma}(u)]$ becomes

$$h_{opt}(u) = \left[\frac{D(a,d,u)}{2\ c_K^2 \gamma''(u)^2}\right]^{1/5} n^{-2/5}\lambda^{d/5}$$

Remark 2. Alternatively, one might deal with a global bandwidth parameter, by minimizing the Mean Integrated Square Error, or MISE, defined as

$$\text{MISE}[\hat{\gamma}(u)] = \int_R (\text{MSE}[\hat{\gamma}(u)])\ du =$$

$$= \int_R (\text{Bias}\,[\hat{\gamma}(u)])^2\ du\ +\ \int_R \text{Var}\,[\hat{\gamma}(u)]\ du$$

for some $R \subset [0, +\infty)$. For instance, we may take $R = [m_0, m]$, where $m = \sup\{\|\mathrm{x_i} - \mathrm{x_j}\| : \mathrm{x_i}, \mathrm{x_j} \in D\}$ and some constant m_0, $0 < m_0 < m$. The resulting optimal bandwidth would be

$$h_{opt} = \left[\frac{\int_R D(a,d,u)du}{2\ c_K^2 \int_R \gamma''(u)^2 du}\right]^{1/5} n^{-2/5}\lambda^{d/5}$$

Both found local and global bandwidth expressions involve the unknown function $\gamma(u)$. For this purpose, a simple parametric approach may be used to estimate $\gamma(u)$ or, even, the latter procedure can be improved by being incorporated into an iterated non-parametric procedure.

3.2 Neighborhood radius selector

Most methodologies for cluster analysis are directly motivated from those derived for density estimation, supporting the natural idea that clusters correspond to modes or peaks in the underlying density function f on $\boldsymbol{R}^d$. In here, we want to suggest a value for the neighborhood radius δ in (1); thus, we are more interested on the density estimation of the distances between locations than on the density estimation of the locations themselves.

We will start proposing two different approaches on the density estimation of the distances. Suppose $\{\mathrm{x_i}\}_{i=1}^n$ locations in $\boldsymbol{R}^d$, then define

$$d_j = \|\mathrm{x_i} - \mathrm{x_k}\|,\ j = 1, ..., \frac{n(n-1)}{2}$$

Firstly, a kernel estimation may be applied on equispaced distances, ranging from the lowest to the largest sampled distance d_j. An alternative may be that of applying the estimation on the sampled distances themselves. The δ quantity may then be derived from the maximum of these functions or even from, for instance, the 10% highest values. The final results, from these two approaches, tend to be very similar.

Other possible approach for δ derivation is based on counts of distances. For a given point and for a list of equispaced distances, one must count how the remaining $n-1$ points are spread within that list of distances. After repeating this for all n points, the partial sum organized by the distances, give us the distance for the maximum count, i.e the proposal for δ value.

According to our experience, the estimates of the semivariogram function given in (1) seem not to be significantly affected by the selection of any of the previous approaches for δ derivation. So, we argue that they all are good candidates to be used as a neighborhood radius selector.

4 Numerical study

In order to analyze the performance of the proposed semivariogram estimator for clustered data, simulations of spatial data in $\boldsymbol{R}^2$ were carried out. Gaussian data were generated on the observation region $D \subset \boldsymbol{R}^2$, with D_0 bounded and fixed square unit and $\lambda = n^{4/9}$. We considered samples of size $n = 100$ and a theoretical exponential variogram with a nugget effect of 0.6, a sill of 1.336 and the corresponding range equal to 5.0.

The RobCluster estimator is compared against the estimator of Matheron and the one using the Nadaraya-Watson kernel. The symmetric Epanechnikov kernel was employed in the two previous kernel-type estimators. A conclusive analysis of the behavior of these semivariogram estimators 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 semivariogram. The ISE, defined as $\int_{\alpha}^{\beta} \left[\widehat{\gamma}(u) - \gamma(u)\right]^2 du$, was approximated numerically through the trapezoid rule. In Table 1, the mean values of the resulting ISEs are compared for two distinct sampling designs:

- A CSR design, where spatial locations are uniformly distributed on D;
- A clustered design, where 40% of the total spatial locations are gathered together into one sub-region of D.

As the observation region D depends on λ, we decided to group the mean values of the ISEs into four classes of lags: $(0, 0.6\lambda)$, $(0, 0.3\lambda)$, $(0, 0.2\lambda)$ and $(0, 0.1\lambda)$. To easily compare columns, all ISE values were standardized by dividing them by the corresponding integral interval, $\beta - \alpha$.

According to Table 1, the RobCluster estimator, defined in (1), offers a better performance than the others in the presence of clustered data. Under a CSR model, the Nadaraya-Watson kernel estimator and the new estimator present similar results, and better than those from Matheron's proposal.

Table 1. Mean values of the standardized ISEs, from the empirical estimators. Total of replicas equals to 100 and each replica total sample size equals to 100.

Design	Estimator	$u \leq 0.6\lambda$	$u \leq 0.3\lambda$	$u \leq 0.2\lambda$	$u \leq 0.1\lambda$
CSR	Matheron	1.270	0.943	0.819	0.763
	NW kernel	0.527	0.314	0.276	0.291
	RobCluster	0.500	0.307	0.276	0.298
Cluster	Matheron	1.519	1.141	0.889	0.568
	NW kernel	0.582	0.525	0.488	0.400
	RobCluster	0.392	0.294	0.243	0.245

References

[Dig03] Diggle, P.J.: Statistical analysis of spatial point patterns. Arnold, London (2003)

[GFG04] García-Soidán, P., Febrero-Bande, M., González-Manteiga, W.: Nonparametric kernel estimation of an isotropic semivariogram. J. Statist. Plann. Inference, **121**, 65–92 (2004)

[KC04] Kovitz J.L., Christakos G.: Spatial statistics of clustered data. Stochastic Environmental Research, **18**, 147–166 (2004)

[HFH94] Hall, P., Fisher, N.I., Hoffmann, B.: On the nonparametric estimation of covariance functions. Ann. Statist., **22**, 399–424 (2115–2134)

[Mat63] Matheron, G.: Principles of Geostatistics. Economic Geology., **58**, 1246–1266 (1963)

[Men05] Menezes, R.: Assessing spatial dependence under non-standard sampling. PhD Thesis, Universidade de Santiago de Compostela, Santiago de Compostela (Spain) (2005)

[MGF05] Menezes, R., García-Soidán, P., Febrero-Bande, M.: A comparison of approaches for valid variogram achievement. J. Comput. Stat., **20**, 4, 623–642 (2005)

[Ste99] Stein, M.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York (1999)

[WL83] Wong M.A., Lane T.: A kth nearest neighbour clustering procedure. Royal Statistical Society Ser. B, **45**, 362–368 (1983)