Construction of probability maps under local stationarity

P. García-Soidán[1,*] and R. Menezes[2]

[1] *Dept. Statistics and Operations Research, University of Vigo, Pontevedra (Spain); pgarcia@uvigo.es*
[2] *Dept. Mathematics and Applications, University of Minho, Guimarães (Portugal); rmenezes@math.uminho.pt*
[*] *Corresponding author*

Abstract. *In this paper, kernel-type estimators of the spatial distribution function are constructed, under non-constant trend, by first approximating the distribution at the sampled sites and then obtaining a weighted average of the resulting values. Unlike other alternatives, our proposals provide non-decreasing functions and do not require previous estimations of the indicator variogram or the trend function. However, appropriate bandwidths parameters are needed and selection of them in practice will be addressed.*

Keywords. *Distribution function; Kernel function; Stationarity.*

1 Introduction

For construction of probability maps, the distribution function of the underlying random process $\{Z(\mathrm{s}) : \mathrm{s} \in D \subset \mathbb{R}^d\}$ must be approximated, where $Z(\mathrm{s})$ represents the variable of interest and D is the observation region. We will write $F_\mathrm{s}(x)$ for the distribution function of $Z(\mathrm{s})$ at x, namely:

$$F_\mathrm{s}(x) = \mathcal{P}\left(Z(\mathrm{s}) \leq x\right) \tag{1}$$

Some approaches for approximation of the spatial distribution function are based on estimating the indicator variogram, through the sample one, and then deriving the required value of the distribution function either by computing the sill or by applying the indicator kriging techniques, as described in Journel [3] or Goovaerts [2], respectively. An alternative is introduced in García-Soidán [1], which suggests using a kernel-type estimator in the first step, as it provides a smoother approximation of the indicator variogram than the sample estimator. These methods allow us to obtain the probability required in an indirect way, which is strongly dependent on the appropriate characterization of the dependence structure of the underlying indicator process. In addition, the aforementioned techniques have been designed for stationary data, although application of them can be extended to a more general setting, where a deterministic trend $\mu(\mathrm{s}) = \mathrm{E}[Z(\mathrm{s})]$ is admitted, by previously detrending the available data. However, care must be taken when proceeding in this way, as these attempts can lead to biased results; see Papriz [4].

In this paper, we will focus our attention on the direct approximation of the distribution function, through kernel-type estimators. Among the main advantages of our proposals, we can highlight that neither knowledge of the trend is required for implementation of the proposed distribution estimator, nor a previous approximation of the dependence structure, through the indicator variogram, is necessary. Furthermore, our kernel approaches provide valid distribution functions, which are not affected by the order relation problem.

2 Main results

To develop our approaches, we will assume that the random process can be modeled as:

$$Z(\mathrm{s}) = \mu(\mathrm{s}) + Y(\mathrm{s}) \tag{2}$$

where $\{Y(\mathrm{s}) \in \mathbb{R} : \mathrm{s} \in D \subset \mathbb{R}^d\}$ is a zero-mean strictly stationary random process and $\mu(\cdot)$ represents the deterministic trend, namely, $\mathrm{E}[Z(\mathrm{s})] = \mu(\mathrm{s})$, for all $\mathrm{s} \in D$.

Suppose that n data, $Z(\mathrm{s}_1)$, $Z(\mathrm{s}_2)$, ..., $Z(\mathrm{s}_n)$, have been collected, at the respective spatial locations s_1, s_2, ..., s_n. A first attempt to derive a kernel-type estimator of $F_{\mathrm{s}}(x)$ can lead us to:

$$\hat{F}_{\mathrm{s},h}(x) = \frac{\sum_i K\left(\frac{\mathrm{s}-\mathrm{s}_i}{h}\right) I_{\{Z(\mathrm{s}_i)\leq x\}}}{\sum_i K\left(\frac{\mathrm{s}-\mathrm{s}_i}{h}\right)}$$

where K represents a d-variate kernel function, h is the bandwidth parameter and I_A denotes the indicator function of the set A.

Estimator $\hat{F}_{\mathrm{s},h}(x)$ converges in probability to the random variable $I_{\{Z(\mathrm{s})\leq x\}}$ rather than to the theoretical distribution $F_{\mathrm{s}}(x)$. Hence, we suggest making use of the kernel method in an alternative way, which departs from approximating the distribution at each sampled location s_i by:

$$\tilde{F}_{1,\mathrm{s}_i,h_1}(x) = \frac{\sum_j K_1\left(\frac{Z(\mathrm{s}_i)-Z(\mathrm{s}_j)}{h_1}\right) I_{\{Z(\mathrm{s}_j)\leq x\}}}{\sum_j K_1\left(\frac{Z(\mathrm{s}_i)-Z(\mathrm{s}_j)}{h_1}\right)} \tag{3}$$

and then combine the resulting terms, by incorporating weights proportional to the lags between locations, to obtain:

$$\hat{F}_{1,\mathrm{s},h,h_1}(x) = \frac{\sum_i K\left(\frac{\mathrm{s}-\mathrm{s}_i}{h}\right) \tilde{F}_{1,\mathrm{s}_i,h_1}(x)}{\sum_i K\left(\frac{\mathrm{s}-\mathrm{s}_i}{h}\right)} = \sum_i \sum_j \frac{K\left(\frac{\mathrm{s}-\mathrm{s}_i}{h}\right) K_1\left(\frac{Z(\mathrm{s}_i)-Z(\mathrm{s}_j)}{h_1}\right) I_{\{Z(\mathrm{s}_j)\leq x\}}}{\sum_{i'} K\left(\frac{\mathrm{s}-\mathrm{s}_{i'}}{h}\right) \sum_{j'} K_1\left(\frac{Z(\mathrm{s}_i)-Z(\mathrm{s}_{j'})}{h_1}\right)} \tag{4}$$

where K and K_1 denote a d-variate kernel and a univariate kernel functions, respectively, and the bandwidth parameters are represented by h and h_1. We should remark that $\hat{F}_{1,\mathrm{s},h,h_1}$ is a non-decreasing function, as it involves indicator functions satisfying this property, and consequently it avoids the order relation problem.

Selection of the bandwidth parameters h and h_1 could be addressed by asymptotically minimizing the corresponding mean squared error (MSE) or mean integrated squared error (MISE), although the resulting optimal bandwidths would be dependent on unknown terms. A different alternative can be derived from the proposal developed in Terrell and Scott [5] for density estimation, based on selecting a balloon estimator. Proceeding in this way, we can take h as the m-th percentile of the distances $\|\mathrm{s}-\mathrm{s}_i\|$, for all i and some $m \in (0,1)$. With regard to h_1, a local bandwidth $h_1(\mathrm{s}_i)$ could be constructed by considering the percentile of the order $m_1(\mathrm{s}_i)$ of the positive values $|Z(\mathrm{s}_i)-Z(\mathrm{s}_j)|$, for each s_i and some $m_1(\mathrm{s}_i) \in (0,1)$. On the other hand, the maximum of the selectors $h_1(\mathrm{s}_i)$, derived in the manner previously described, would provide us with a global bandwidth h_1.

Since $\hat{F}_{1,\mathrm{s},h,h_1}$ is a discrete distribution function, a continuous approach can be derived by applying the integral of a density in (3), rather than an indicator function, and by replacing $\tilde{F}_{1,\mathrm{s}_i,h_1}$ in (4) by the resulting estimator. In other words, we could construct a distribution estimator at each sampled site:

$$\tilde{F}_{2,\mathrm{s}_i,h_1,h_2}(x) = \frac{\sum_j K_1\left(\frac{Z(\mathrm{s}_i)-Z(\mathrm{s}_j)}{h_1}\right) \mathcal{K}_2\left(\frac{x-Z(\mathrm{s}_j)}{h_2}\right)}{\sum_j K_1\left(\frac{Z(\mathrm{s}_i)-Z(\mathrm{s}_j)}{h_1}\right)} \tag{5}$$

and then obtain a weighted average of the values achieved as follows:

$$\begin{aligned} \hat{F}_{2,\mathrm{s},h,h_1,h_2}(x) &= \frac{\sum_i K\left(\frac{\mathrm{s}-\mathrm{s}_i}{h}\right) \tilde{F}_{2,\mathrm{s}_i,h_1,h_2}(x)}{\sum_i K\left(\frac{\mathrm{s}-\mathrm{s}_i}{h}\right)} = \\ &= \sum_i \sum_j \frac{K\left(\frac{\mathrm{s}-\mathrm{s}_i}{h}\right) K_1\left(\frac{Z(\mathrm{s}_i)-Z(\mathrm{s}_j)}{h_1}\right) \mathcal{K}_2\left(\frac{x-Z(\mathrm{s}_j)}{h_2}\right)}{\sum_{i'} K\left(\frac{\mathrm{s}-\mathrm{s}_{i'}}{h}\right) \sum_{j'} K_1\left(\frac{Z(\mathrm{s}_i)-Z(\mathrm{s}_{j'})}{h_1}\right)} \end{aligned} \tag{6}$$

where $\mathcal{K}_2(x) = \int_{-\infty}^{x} K_2(y)dy$, K_2 is a univariate kernel function and h_2 is a new bandwidth parameter, which must be estimated through the local or global balloon approaches.

Acknowledgments. The first author's work has been supported by the Spanish Government Grants CONSOLIDER-INGENIO 2010 CSD2008-00068 and TEC2011-28683-C02-02, as well as by Xunta de Galicia Grant CN2012/279. The second author acknowledges financial support by FEDER Funds through "Programa Operacional Factores de Competitividade COMPETE" and by Portuguese Funds through FCT-"Fundação para a Ciência e a Tecnologia", within the Project PEst-OE/MAT/UI0013/2014.

References

[1] García-Soidán, P. and Menezes, R. (2012). Estimation of the spatial distribution through the kernel indicator variogram. *Environmetrics* **23**, 535-548.

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

[3] Journel, A. G. (1983). Nonparametric estimation of spatial distribution. *Mathematical Geology* **15**, 445–468.

[4] Papritz, A. (2009). Why indicator kriging should be abandoned, *Pedometron* **26**, 4–7.

[5] Terrell, G. and Scott, D. W. (1992). Variable kernel density estimation. *Annals of Statistics* **20**, 1236–1265.