AStA Advances in Statistical Analysis (2022) 106:499–524
https://doi.org/10.1007/s10182-021-00434-4

ORIGINAL PAPER

A spatial randomness test based on the box-counting dimension

Yolanda Caballero[1] · **Ramón Giraldo**[1] · **Jorge Mateu**[2]

Received: 27 February 2021 / Accepted: 8 December 2021 / Published online: 5 January 2022

Abstract
Statistical modelling of a spatial point pattern often begins by testing the hypothesis of spatial randomness. Classical tests are based on quadrat counts and distance-based methods. Alternatively, we propose a new statistical test of spatial randomness based on the fractal dimension, calculated through the box-counting method providing an inferential perspective contrary to the more often descriptive use of this method. We also develop a graphical test based on the log–log plot to calculate the box-counting dimension. We evaluate the performance of our methodology by conducting a simulation study and analysing a COVID-19 dataset. The results reinforce the good performance of the method that arises as an alternative to the more classical distances-based strategies.

Keywords Box-counting dimension · Complete spatial randomness Fractal dimension · Poisson distribution · Spatial point patterns

1 Introduction

Spatial statistics is the branch of statistics that deals with the modelling of realisations of spatially indexed stochastic processes (Schabenberger and Gotway, 2017). This field covers three acknowledged areas: geostatistics, areal data, and spatial point patterns (Cressie, 1991). The last one concerns the analysis of the spatial distribution of locations of events such as earthquakes, landslides or forest

✉ Ramón Giraldo
rgiraldoh@unal.edu.co

Yolanda Caballero
ybcaballero@unal.edu.co

Jorge Mateu
mateu@mat.uji.es

[1] Department of Statistics, Universidad Nacional de Colombia, Bogotá, Colombia

[2] Department of Mathematics, Universidad Jaume I, Castellón, Spain

fires (Baddeley et al., 2013). Other examples are patterns of towns in a region, trees in a forest or galaxies in space (Ripley, 1977). In all these cases, the relative position of points is compared with clustered, random, or regular generating processes (Bivand et al., 2013). For a theoretical review on spatial point patterns, the reader is referred to Daley and Vere-Jones (2008), Diggle (2013), Illian et al. (2008), and Møller and Waagepetersen (2004). A more practical overview can be found in Baddeley et al. (2015) and Gaetan and Guyon (2010). The implementation of methods and models to analyse point patterns with R software R Core Team (2020) is described in Baddeley et al. (2015), Bivand et al. (2013) and Plant (2012). The theory of spatial point patterns is an active research field with challenging theoretical problems and applications in a broad range of sciences such as agriculture, astronomy, biology, climatology, ecology, epidemiology, geology, among many others (Baddeley et al., 2006).

Complete spatial randomness (CSR) describes a point process whereby point events occur within a given study area in a completely random fashion. It is synonymous with a homogeneous spatial Poisson process. Usually, the first step in analysing a spatial point pattern is to test for CSR. If the hypothesis is not rejected, one can assume that the given point pattern is random, and we refer to it as a homogeneous Poisson point pattern (Illian et al., 2008). Generally, for this purpose, some tests based on quadrat counts and distances between locations of events are used (Banerjee et al., 2015).

There are indeed several distance-based functions that are often used in testing for CSR (Diggle, 2013). As widely used tools, we have the distribution of distances from an event to its nearest neighbour (function G), the distribution of distances from an arbitrary point of the plane to its nearest neighbour (F), the function J (calculated in terms of F and G), and the number of events encountered up to a given distance of any particular event (Ripley's K function) (Baddeley et al., 2013). However, these functions are only known under a few theoretical models and are mathematically unknown for many other types of spatial dependence. As functions that depend on distances, we have to choose a particular metric. The usual one is the Euclidean distance, but these functions have to be adapted when this distance is not realistic. These ones are highly time-consuming when the number of points increases. With the current technologies, we have point patterns with thousands of events, and it is often the case that we cannot calculate the K-function, say. All these sorts of drawbacks motivate our proposal for testing the hypothesis of CSR. Specifically, we propose a new alternative computationally efficient for testing this hypothesis of CSR, which is based on calculating the fractal dimension (Wiegand and Moloney, 2013) utilizing the box-counting method (Foroutan-pour et al., 1999). Several authors have used box-counting and the spectrum of generalised dimensions to analyse point patterns (see for example Salvadori et al. 1997; Tuia and Kanevski 2008, and Vega et al. 2015). However, all these contributions use this tool from a descriptive point of view. One intuitive advantage of the methodology considered here is that the statistic defined can summarise the information of the spatial point pattern in only one value. Also, since it does not depend on distances, it is not necessary to consider the edge effect. This makes computation more straightforward and much faster than the classical tests.

The notion of a fractal dimension was introduced by Mandelbrot (1967) who used it as an indicator of surface roughness. A shape with a higher fractal dimension is rougher than one with a lower dimension. Many methods exist for estimating the fractal dimension. Box-counting, R/S analysis and the variation method can be used for this purpose (Breslin and Belward, 1999). Fractal dimension and its estimation using the box-counting method have been used in different fields of statistics. We can find contributions, among other statistical contexts, in time series analysis (Kopytov et al., 2016), clustering analysis (Bones et al., 2016), principal components analysis (Mo and Huang, 2010) and geostatistics (Vidal et al., 2010). As mentioned before, we show how these concepts can be used in point pattern analysis from an inferential perspective. We then develop our test based on the box-counting dimension of a spatial point pattern. We also propose a graphical test in the line of the classical graphical tests based on G, F or K functions. We evaluate the performance of our methodology by conducting a simulation study through three known spatial structures that can be generated using the library spatstat (Baddeley et al., 2015) in `R` (R Core Team, 2020). In all cases, the results are consistent with those found by using the functions G, F and Ripley's K (Diggle, 2003).

The paper is organised as follows. Section 2 introduces the box-counting methodology. Section 3 presents the proposed test, an illustration through simulated and real data, and a power study. Section 4 describes a graphical approach to test for CSR (also based on the box-counting dimension). Section5 shows an application of the method to a real data set of COVID-19 cases in Cali, Colombia. The article ends with a brief discussion and suggestions for further research.

2 Box-counting estimation of the fractal dimension for spatial point patterns under CSR

The hypothesis of CSR for a spatial point pattern asserts that the number of events in any region follows a Poisson distribution with a given mean count per uniform subdivision. The events of a pattern are independently and uniformly distributed over space. In other words, the events are equally likely to occur anywhere and do not interact with each other. Here, we use uniform in the sense of following a uniform probability distribution across the study region, not in the sense of "evenly" dispersed across the study region. There are no interactions amongst the events, as the intensity of events does not vary over the plane. Thus, the independence assumption would be violated if the existence of one event either encouraged or inhibited the occurrence of other events in the neighbourhood. In this sense, CSR acts as a benchmark hypothesis to distinguish between randomness and clustering or regularity due to some form of interaction.

A fractal is a non-regular geometric shape with the same degree of non-regularity at all scales. It can be treated as a self-similar structure in the sense that even an indefinitely small part of a shape is geometrically similar to the whole (Debnath, 2006). The fractal dimension is a ratio providing a statistical index of complexity comparing how the details in a pattern change with the scale at which they are measured (Falconer, 2004). The dimension of self-similar fractals is given by

$$D_s = \frac{\log(M)}{\log\left(\frac{1}{\epsilon}\right)}, \tag{1}$$

where M is the number of self-similar pieces, and ϵ is a scale factor, such that $M\epsilon^{D_s} = 1$. In Eq. (1) log corresponds to the logarithm to the base 10. We use this same notation throughout the paper to be consistent with the related published literature. The use of D_s in Eq. (1) is quite limited in practice. An alternative is using the box-counting method (Liebovitch and Toth, 1989). Suppose the object of interest is covered with a number $\Gamma(\delta)$ of non-overlapping squares with sides of length δ. The box-counting estimation of the fractal dimension (hereinafter box-counting dimension) is given by (Addison, 1997)

$$D = \lim_{\delta \to 0} \frac{\log(\Gamma(\delta))}{\log\left(\frac{1}{\delta}\right)}. \tag{2}$$

In practice, D in (2) is calculated as the slope of a linear regression between $\log(\Gamma(\delta))$ and $\log\left(\frac{1}{\delta}\right)$. Given a number of δ_i values ($i = 1, 2, 3, \ldots$), D is defined by means of the linear model (see Addison (1997)).

$$\log(\Gamma(\delta_i)) = D\log\left(\frac{1}{\delta_i}\right). \tag{3}$$

2.1 Expected value of *D* under CSR

We now show how the box-counting dimension given in (3) can be adapted to the context of spatial point patterns and can be used to test the hypothesis of CSR. Under CSR, we have that the number of events in a square A, with area $|A|$ and sides of length k (without loss of generality, we can take $k = 1$), is Poisson distributed with mean $\lambda|A|$, where λ is the constant intensity of the point process; that is, the probability function of the number of events in A is

$$P(N(A) = x) = \frac{\exp^{-\lambda|A|}(\lambda|A|)^x}{x!}, x = 0, 1, 2, \ldots. \tag{4}$$

From (4), we have

$$P(N(A) > 0) = (1 - P(N(A) = 0)) = 1 - \frac{\exp^{-\lambda|A|}(\lambda|A|)^0}{0!} = 1 - \exp^{-\lambda|A|}.$$

Assume the original square A is divided into β_i non-overlapping squares $A_j, j = 1, \ldots, \beta_i$ with sides of length $\delta_i = \frac{1}{i}, i = 1, 2, 3, \ldots$ (see Fig. 1).

Then, denoting $\beta_i = i^2$, we have

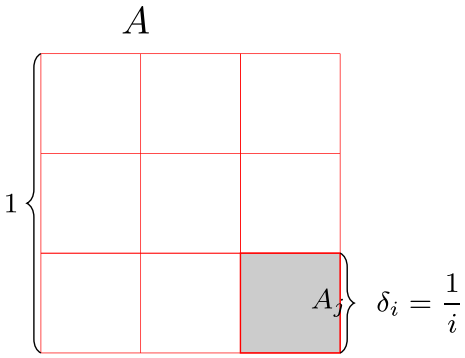

Fig. 1 The square A with sides of length 1 is divided into $\beta_3 = 9$ non-overlapping squares A_j with sides of length $\delta_3 = \frac{1}{3}$

$$|A| = \beta_i |A_j| = \beta_i \delta_i^2 = 1. \tag{5}$$

Under the CSR condition, $\mathbb{E}(N(A)) = \lambda|A| = \mu$, the mean of a homogeneous Poisson process. From Eq. (5)

$$\lambda = \frac{\mu}{|A|} = \frac{\mu}{\beta_i \delta_i^2} = \mu,$$

and consequently

$$P(N(A_j) > 0) = 1 - \exp^{-\lambda|A_j|} = 1 - \exp^{-\mu\delta_i^2}.$$

Define the random variable $\Gamma(\delta_i), i = 1, 2, \ldots$, as the number of squares of side δ_i containing at least one event, that is, $\Gamma(\delta_i)$ corresponds to the number of squares required to cover the point pattern (see Figs. 1 and 2). This variable can be defined as

$$\Gamma(\delta_i) = \sum_{j=1}^{\beta_i} Z_j, \text{ with } Z_j = \begin{cases} 1 & \text{if } N(A_j) > 0 \\ 0 & \text{other case} \end{cases}. \tag{6}$$

The expected value of $\Gamma(\delta_i)$ in (6) is

$$\mathbb{E}\big(\Gamma(\delta_i)\big) = \sum_{j=1}^{\beta_i} \mathbb{E}(Z_j) = \beta_i \mathbb{E}(Z_j) = \frac{1}{\delta_i^2} P(N(A_j) > 0) = \frac{1}{\delta_i^2}\left[1 - \exp^{-\mu\delta_i^2}\right].$$

Note that in order to define $\mathbb{E}(D)$ in (3), it is requiered to find $\mathbb{E}(\log(\Gamma(\delta_i)))$. Using the first-order Taylor expansion of $\log\big(\Gamma(\delta_i)\big)$ around $\mathbb{E}(\Gamma(\delta_i))$, we have

$$\mathbb{E}\Big(\log\big(\Gamma(\delta_i)\big)\Big) \approx \log(\mathbb{E}\big[\Gamma(\delta_i)\big]) = \log\left(\frac{1}{\delta_i^2}\left[1 - \exp^{-\mu\delta_i^2}\right]\right). \tag{7}$$

Taking expectation in (3) and using (7), we have under CSR

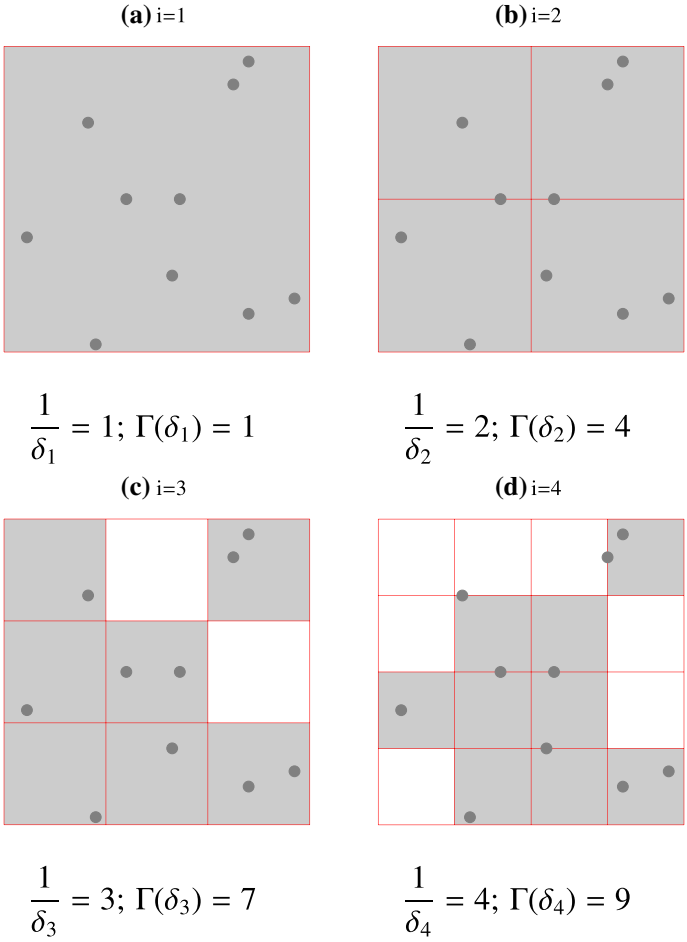

Fig. 2 Graphical representation of δ_i, $\Gamma(\delta_i)$, and $\beta_i = i^2$ (for $i = 1, \ldots, 4$)

$$\begin{aligned} \mathbb{E}\left(\log\left(\Gamma(\delta_i)\right)\right) &= \mathbb{E}(D)\log\left(\frac{1}{\delta_i}\right) \\ \log\left(\frac{1}{\delta_i^2}\left[1 - \exp^{-\mu\delta_i^2}\right]\right) &= \mathbb{E}(D)\log\left(\frac{1}{\delta_i}\right). \end{aligned} \tag{8}$$

When $\mu \to \infty$ in (8), we have

$$\begin{aligned} \log\left(\frac{1}{\delta_i^2}\right) &= \mathbb{E}(D)\log\left(\frac{1}{\delta_i}\right) \\ \mathbb{E}(D) = 2,\ \delta_i < 1. \end{aligned}$$

In general, if A (Fig. 1) is a square with side length $k \neq 1$, assuming again that $\lambda|A| = \mu$, we then have

$$\delta_i = \frac{k}{i},\ \beta_i = \left(\frac{k}{\delta_i}\right)^2 = i^2, \lambda = \frac{\mu}{k^2},$$
$$P(N(A_j) > 0) = 1 - \exp^{-\lambda|A_j|} = 1 - \exp^{\frac{-\mu\delta_i^2}{k^2}},$$
$$\mathbb{E}\big(\Gamma(\delta_i)\big) = \left(\frac{k}{\delta_i}\right)^2 P(N(A_i) > 0) = \left(\frac{k}{\delta_i}\right)^2 \left[1 - \exp^{\frac{-\mu\delta_i^2}{k^2}}\right].$$

Using again the first-order Taylor expansion of $\log\big(\Gamma(\delta_i)\big)$ around $\mathbb{E}(\Gamma(\delta_i))$, we have

$$\mathbb{E}\Big(\log\big(\Gamma(\delta_i)\big)\Big) \approx \log\big(\mathbb{E}(\Gamma(\delta_i)\big) = \log\left(\left(\frac{k}{\delta_i}\right)^2 \left[1 - \exp^{\frac{-\mu\delta_i^2}{k^2}}\right]\right). \tag{9}$$

Then, under CSR, the expected value of the fractal dimension for a square of side k calculated with the box-counting method is defined by the linear model

$$\begin{aligned} \mathbb{E}\Big(\log\big(\Gamma(\delta_i)\big)\Big) &= \mathbb{E}(D)\log\left(\frac{1}{\delta_i}\right) \\ \log\left(\left(\frac{k}{\delta_i}\right)^2 \left[1 - \exp^{\frac{-\mu\delta_i^2}{k^2}}\right]\right) &= \mathbb{E}(D)\log\left(\frac{1}{\delta_i}\right). \end{aligned} \tag{10}$$

Note that $\lambda|A| = \mu$ in (10) is usually unknown. Based on just one realisation of a homogeneous Poisson process, we can then estimate μ by n (the number of points of the observed point pattern) to estimate $\mathbb{E}(D)$. Taking $\lim_{\mu\to\infty}$ in (10) we obtain

$$\mathbb{E}(D) \approx \frac{\log\left(\frac{k}{\delta_i}\right)^2}{\log\left(\frac{1}{\delta_i}\right)}. \tag{11}$$

2.2 log–log relationship

The functional relationship between $\mathbb{E}(\log(\Gamma(\delta_i)))$ and $\log\left(\frac{1}{\delta_i}\right)$ defined in Eq. (10) allows to characterise the behaviour of $\mathbb{E}(D)$. $\mathbb{E}(\log(\Gamma(\delta_i))$ depends on k (the side length of the original square) and on $\mathbb{E}(N(A)) = \mu$ (the expected number of events of the spatial point pattern in A). Given μ, the shape of the curves does not change (Fig. 3). Note that the greater the value of k, the more the curve is shifted to the left. Likewise, given a fixed k, the effect of μ is reflected on the maximum of $\mathbb{E}(\log(\Gamma(\delta_i)))$. The greater μ, the greater the value at which $\mathbb{E}(\log(\Gamma(\delta_i)))$ becomes constant (Fig. 4).

The minimum number of boxes required to cover the point pattern is obtained when $i = 1$ (initial square). In this case $\log\left(\frac{1}{\delta_1}\right) = \log\left(\frac{1}{k}\right)$. The ordinate for this

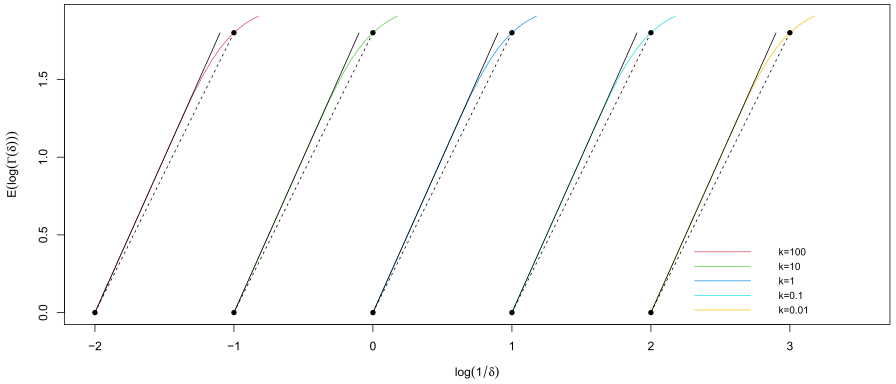

Fig. 3 Relation between $\mathbb{E}(\log(\Gamma(\delta_i)))$ and $\log\left(\frac{1}{\delta_i}\right)$, $i = 1, 2, 3, \ldots, 100$, when the initial square has sides of length k (0.01, 0.1, 1, 10, 100) and the expected number of events is $\lambda|A| = \mu = 100$. Black points at each curve correspond to coordinates $\left(\log\left(\frac{1}{k}\right), \log(1 - \exp^{-\mu})\right)$ and $\left(\frac{1}{2}\log(\mu) - \log(k), \log(\mu) - 0.199\right)$, respectively (see text for explanations on these values). The slopes of the dashed lines define $\mathbb{E}(D_1)$ (see Eq. 15). At each case $\mathbb{E}(D_1) = 1.80$. Black lines (slope 2) correspond to $\mathbb{E}(D_1) = 2$ (limit when $\mu \to \infty$)

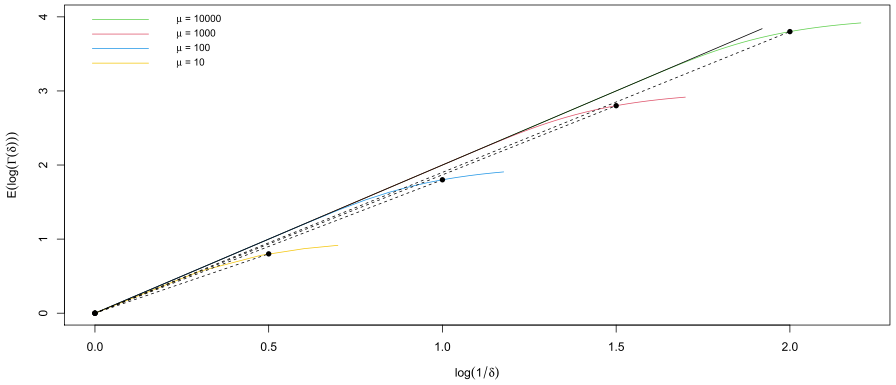

Fig. 4 Relation between $\mathbb{E}(\log(\Gamma(\delta_i)))$ and $\left(\frac{1}{\delta_i}\right)$, $i = 1, 2, 3, \ldots, 100$, according to the expected number of points of the pattern (μ), when initial square has sides of length $k = 1$. Black points at each curve corresponds to coordinates $\left(\log\left(\frac{1}{k}\right), \log(1 - \exp^{-\mu})\right)$ and $\left(\frac{1}{2}\log(\mu) - \log(k), \log(\mu) - 0.199\right)$, respectively (see text for explanations on these values). The slopes of the dashed lines define $\mathbb{E}(D_1)$ (see Eq. 15). These are, respectively, 1.60 ($\mu = 10$), 1.80 ($\mu = 100$), 1.87 ($\mu = 1000$) and 1.90 ($\mu = 10000$). In general when $\mu \to \infty, \mathbb{E}(D_1) \to 2$ (black line)

value is $\mathbb{E}(\log(\Gamma(\delta_i))) = \log(1 - \exp^{-\mu})$. On the other hand, the maximum number of partitions (corresponding to the minimum size of δ) is found when the expected number of events in A_j is $\lambda|A_j| = 1$. Under this scenario, we have

$$\begin{aligned}\lambda|A_j| = \frac{\mu\delta_i^2}{k^2} &= 1\\ \delta_i &= \frac{k}{\sqrt{\mu}}\\ \log\left(\frac{1}{\delta_i}\right) &= \log\left(\frac{\sqrt{\mu}}{k}\right)\\ &= \frac{1}{2}\log(\mu) - \log(k).\end{aligned} \tag{12}$$

Replacing (12) into (10) with $\lambda|A_j| = \frac{\mu\delta_i^2}{k^2} = 1$, we obtain

$$\begin{aligned}\mathbb{E}\left(\log\left(\Gamma(\delta_i)\right)\right) &= 2\log\left(\frac{k}{\delta_i}\right) + \log\left(1-\exp^{-1}\right)\\ &= 2\log(k) + 2\log\left(\frac{1}{\delta_i}\right) - 0.199\\ &= 2\log(k) + 2\left(\frac{1}{2}\log(\mu) - \log(k)\right) - 0.199\\ &= \log(\mu) - 0.199.\end{aligned} \tag{13}$$

The log–log plots (Figs. 3 and 4) show a multifractal behaviour, i.e. the dependence between $\mathbb{E}\left(\log\left(\Gamma(\delta_i)\right)\right)$ and $\log\left(\frac{1}{\delta_i}\right)$ is non-linear. The box-counting dimension D in (3) is usually calculated with the portion of the data that allows to fit a linear model (see, for example, Kenkel, 2013; Mou and Wang, 2014; Vega et al., 2015, and Jaquette and Schweinhart, 2013). This option might not be appropriate to discriminate between the different types of spatial point patterns. In this context, it is important to take into account the minimum and maximum values of the log–log curves. Thus, here, we propose to characterise the relationship between $\mathbb{E}\left(\log\left(\Gamma(\delta_i)\right)\right)$ and $\log\left(\frac{1}{\delta_i}\right)$ in Eq. (10) with the slope of the straight line defined by the points $\left[\min\left(\log\left(\frac{1}{\delta_i}\right)\right), \min\left(\mathbb{E}\left(\log\left(\Gamma(\delta_i)\right)\right)\right)\right]$ and $\left[\max\left(\log\left(\frac{1}{\delta_i}\right)\right), \max\left(\mathbb{E}\left(\log\left(\Gamma(\delta_i)\right)\right)\right)\right]$ (see black points in Figs. 3 and 4), that is, the slope calculated with the coordinates

$$\begin{aligned}(x_1, y_1) &= \left(\log\left(\frac{1}{\delta_1}\right), \log\left(\left(\frac{k}{\delta_1}\right)^2\left[1-\exp^{\frac{-\mu\delta_i^2}{k^2}}\right]\right)\right) = \left(\log\left(\frac{1}{k}\right), \log(1-\exp^{-\mu})\right), \text{ and}\\ (x_2, y_2) &= \left(\frac{1}{2}\log(\mu) - \log(k), \log(\mu) - 0.199\right).\end{aligned} \tag{14}$$

We denote this slope as $\mathbb{E}(D_1)$ instead of $\mathbb{E}(D)$ to emphasise that we do not employ the traditional linear fitting used in box-counting estimation. Under CSR, we have

$$\mathbb{E}(D_1) = \frac{y_2 - y_1}{x_2 - x_1} = \frac{(\log(\mu) - 0.199) - (\log(1 - \exp^{-\mu}))}{\left(\frac{1}{2}\log(\mu) - \log(k)\right) - \left(\log\left(\frac{1}{k}\right)\right)} = 2\frac{(\log(\mu) - 0.199) - (\log(1 - \exp^{-\mu}))}{\log(\mu)}. \tag{15}$$

From Eq. (15), $\lim_{\mu\to\infty} \mathbb{E}(D_1) = 2$ (see Fig. 4).

3 CSR testing using the statistic $\hat{\mathbb{E}}(D_1)$

In practice with real data, μ in Eq. (15) is unknown. In this case in order to test for CSR, we can take $\hat{\mu} = n$ with n the number of points of the observed pattern, namely we assume that $N(A) \sim \text{Poisson}(\lambda|A| = n)$. In this scenario, the expected value of D_1 under CSR is defined as

$$\hat{\mathbb{E}}(D_1) = \frac{\hat{y}_2 - \hat{y}_1}{\hat{x}_2 - x_1} = \frac{(\log(n) - 0.199) - (\log(1 - \exp^{-n}))}{\left(\frac{1}{2}\log(n) - \log(k)\right) - \left(\log\left(\frac{1}{k}\right)\right)} = 2\frac{(\log(n) - 0.199) - (\log(1 - \exp^{-n}))}{\log(n)}. \tag{16}$$

and its estimation is given by

$$\hat{\hat{\mathbb{E}}}(D_1) = \frac{\hat{\hat{y}}_2 - \hat{y}_1}{\hat{x}_2 - x_1}, \tag{17}$$

where x_1, $\hat{x}_2$, and $\hat{y}_1$ are defined similarly as in (16), and $\hat{\hat{y}}_2$ is calculated from the scatter plot between $\log\left(\Gamma(\delta_i)\right)$ and $\log\left(\frac{1}{\delta_i}\right)$. Specifically, $\hat{\hat{y}}_2$ is the ordinate corresponding to the abscissa $\left(\frac{1}{2}\log(n) - \log(k)\right)$, with k the side length of the square. In practice, some mathematical interpolation procedure (linear, polynomial, etc) can be required to calculate $\hat{\hat{y}}_2$. By way of illustration, we show the results found with a simulation from $N(A) \sim \text{Poisson}(\mu = \lambda|A| = 100)$, $|A| = 1$. Figure 5 shows the spatial distribution of n=114 simulated events in the unit square, the number of events per cell for each one of the three partitions $\left(\left(\frac{1}{\delta_i}\right), i = 5, 10, \text{and } 15\right)$, and the value of $\Gamma(\delta_i)$ at each case.

We observe that the smaller the size of the partition, the greater the number of boxes without events (the number of boxes with zeros). Calculating $\log\left(\frac{1}{\delta_i}\right)$ and $\log\left(\Gamma(\delta_i)\right)$ for $i = 1, \ldots, 20$, we obtain the log–log scatter plot (white circles) shown in Fig. 6. Its behaviour, as expected, is similar to the theoretical log–log curve $\left(\log\left(\frac{1}{\delta_i}\right) \text{versus } \mathbb{E}(\log(\Gamma(\delta_i)))\right)$ under CSR (red line). The black points in this plot are the coordinates used to calculate the expected box-counting dimension $\hat{\mathbb{E}}(D_1)$ under the null hypothesis (Eq. 16). In this case $\hat{\mathbb{E}}(D_1) = 1.806$. The intersection of the blue lines corresponds to the coordinate $(\hat{x}_2, \hat{\hat{y}}_2)$ ($\hat{\hat{y}}_2$ is found by linear

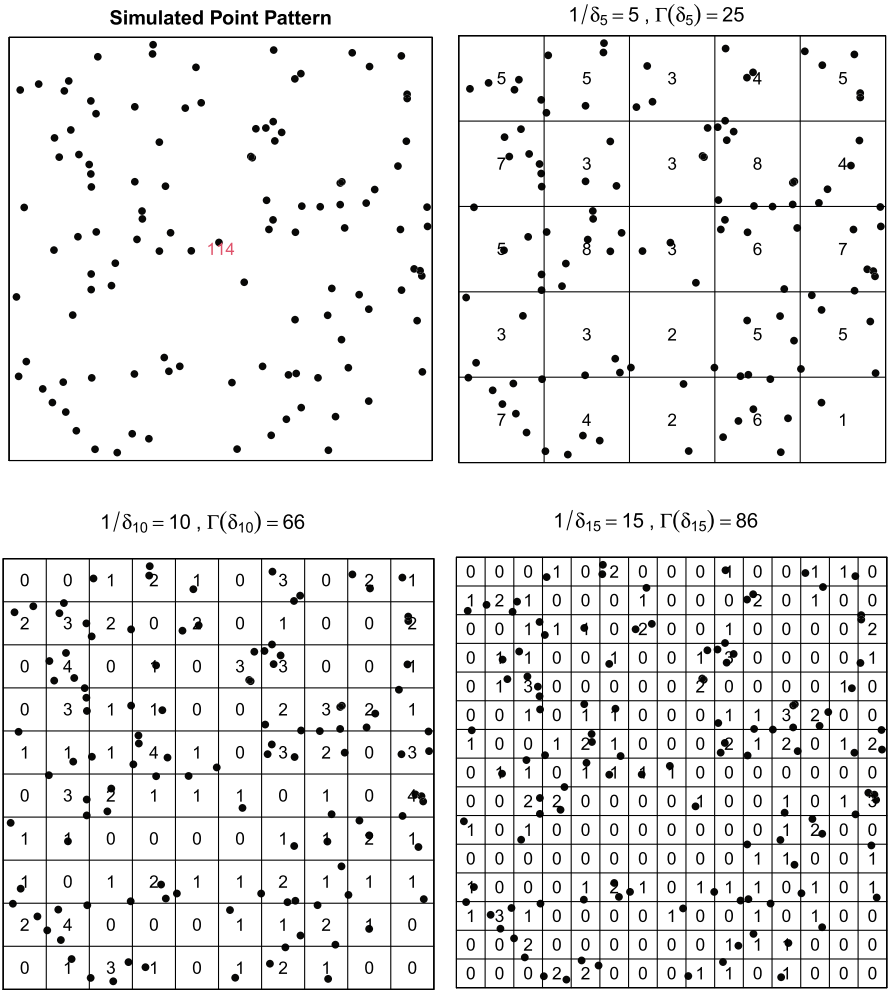

Fig. 5 Simulation size $n = 114$ of a spatial point pattern with $N(A) \sim \text{Poisson}(\mu = \lambda|A| = 100)$ in a square of side $k = 1$ (top left). The numbers at each panel indicate how many events are falling into each box. $\Gamma(\delta_i)$ corresponds to the number of boxes with one or more events for $\left(1/\delta_i = 5\right)$ (top right), $\left(1/\delta_i = 10\right)$ (bottom left), and $\left(1/\delta_i = 15\right)$ (bottom right), respectively

interpolation between the two nearest values), which is replaced in Eq. (17) to find the estimated box-counting dimension ($\hat{\mathbb{E}}(D_1) = 1.784$). Generating m simulations from $N(A) \sim \text{Poisson}(n = \lambda|A| = 114)$ and repeating the procedure above described, we can find m estimations under the null hypothesis of CSR. A value of $\hat{\mathbb{E}}(D_1)$ at the extreme of the tail of the null distribution would indicate that the spatial randomness hypothesis should be rejected. Analogously, defining $B = (\hat{\mathbb{E}}(D_1) - \hat{\hat{\mathbb{E}}}(D_1))$ we reject the randomness hypothesis if this value is at the extreme of the tail of the corresponding null distribution. Using B may be preferable because in all cases (regardless of the type of pattern considered), the zero will be the reference value of the

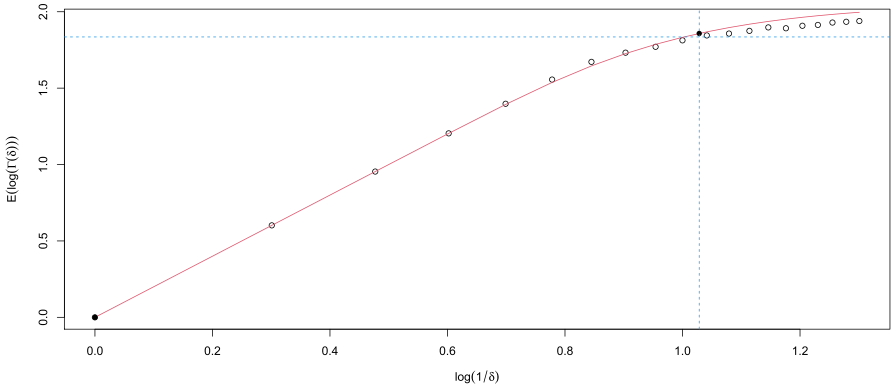

Fig. 6 Scatter plot (white circles) obtained from the pairs $\left(\log\left(\frac{1}{\delta_i}\right), \log(\Gamma(\delta_i))\right)$, $i = 1, \ldots, 20$, calculated with a simulation size $n = 114$ from $N(A) \sim \text{Poisson}(\mu = \lambda|A| = 100)$. The red line is the theoretical log–log curve $\left(\log\left(\frac{1}{\delta_i}\right)\right.$versus $\left.\mathbb{E}(\log(\Gamma(\delta_i)))\right)$ under CSR (assuming $N(A) \sim \text{Poisson}(n = \lambda|A| = 114)$). Black circles are the coordinates used to calculate the box-counting dimension ($\hat{\mathbb{E}}(D_1)$) under the null hypothesis (Eq. (16)). The intersection of the blue lines corresponds to the coordinate $(\hat{x}_2, \hat{\hat{y}}_2)$ used to obtain the estimated box-counting dimension ($\hat{\hat{\mathbb{E}}}(D_1)$)(Eq. (17)). $\hat{\hat{y}}_2$ is calculated by linear interpolation of the two nearest points $\left(\log\left(\frac{1}{\delta_{10}}\right), \log\left(\Gamma(\delta_{10})\right)\right)$ and $\left(\log\left(\frac{1}{\delta_{11}}\right), \log\left(\Gamma(\delta_{11})\right)\right)$

centre of the distribution (see Fig. 7). This is illustrated in Sect. 3.1 with a simulation study. The procedure to test for CSR based on $\hat{\mathbb{E}}(D_1)$, $\hat{\hat{\mathbb{E}}}(D_1)$, and B above described is summarised in Algorithm 1, where in addition to presenting in a schematic way, the steps required to perform the spatial randomness test using the box-counting method, it is shown how to estimate the corresponding p value.

Algorithm 1 Complete spatial randomness test based on $\hat{\mathbb{E}}(D_1)$, $\hat{\hat{\mathbb{E}}}(D_1)$, and B.

1. Given the observed point pattern (size n) find, under the assumption of CSR, the expected box-counting dimension $\hat{\mathbb{E}}(D_1)$ (Equation (16)), its estimation $\hat{\hat{\mathbb{E}}}(D_1)$ (Equation (17)), and $B = (\hat{\mathbb{E}}(D_1) - \hat{\hat{\mathbb{E}}}(D_1))$.

for j = 1 to m **do**

(a) Simulate a sample from $N(A) \sim \text{Poisson}(\lambda|A| = n)$ and obtain n_j, the number of events generated. Note that a conditional simulation (fixing n) can be also carried out.

(b) Calculate $\hat{\mathbb{E}}(D_1)$, $\hat{\hat{\mathbb{E}}}(D_1)$ and $B = (\hat{\mathbb{E}}(D_1) - \hat{\hat{\mathbb{E}}}(D_1))$ based on the point pattern generated in (a), i.e., replace n by n_j in Equations (16) and (17). Call these values $\hat{\mathbb{E}}(D_1)_j$, $\hat{\hat{\mathbb{E}}}(D_1)_j$ and $B_j = (\hat{\mathbb{E}}(D_1)_j - \hat{\hat{\mathbb{E}}}(D_1)_j)$.

end for

2. Decision rule

- Obtain from $B_1, \ldots, B_m$ the quantiles B_α, $B_{1-\alpha}$, $B_{\frac{\alpha}{2}}$, and $B_{1-\frac{\alpha}{2}}$ (α the level of significance)
- Using the statistic B (calculated in the step 1) reject the null hypothesis of CSR against the alternative of regularity if $B < B_\alpha$. Likewise reject the null hypothesis of CSR against the alternative of clustering if $B > B_{1-\alpha}$. When the direction of the alternative is not specified, reject the hypothesis of CSR if $B < B_{\frac{\alpha}{2}}$ or $B > B_{1-\frac{\alpha}{2}}$.
- Alternatively, defining the dichotomous variables (according to the alternative) as

$$\psi_j = \begin{cases} 1 \text{ if } B_j < B \\ 0 \text{ other case} \end{cases}, \quad \psi_j = \begin{cases} 1 \text{ if } B_j > B \\ 0 \text{ other case} \end{cases}, \text{ or } \psi_j = \begin{cases} 1 \text{ if } B_j < B \text{ or } B_j > B \\ 0 \text{ other case} \end{cases},$$

calculate the corresponding empirical p-values

$$as\text{p-value} = \frac{\sum_{j=1}^{m} \psi_j}{m},$$

and reject the hypothesis of CSR if p-value $< \alpha$.

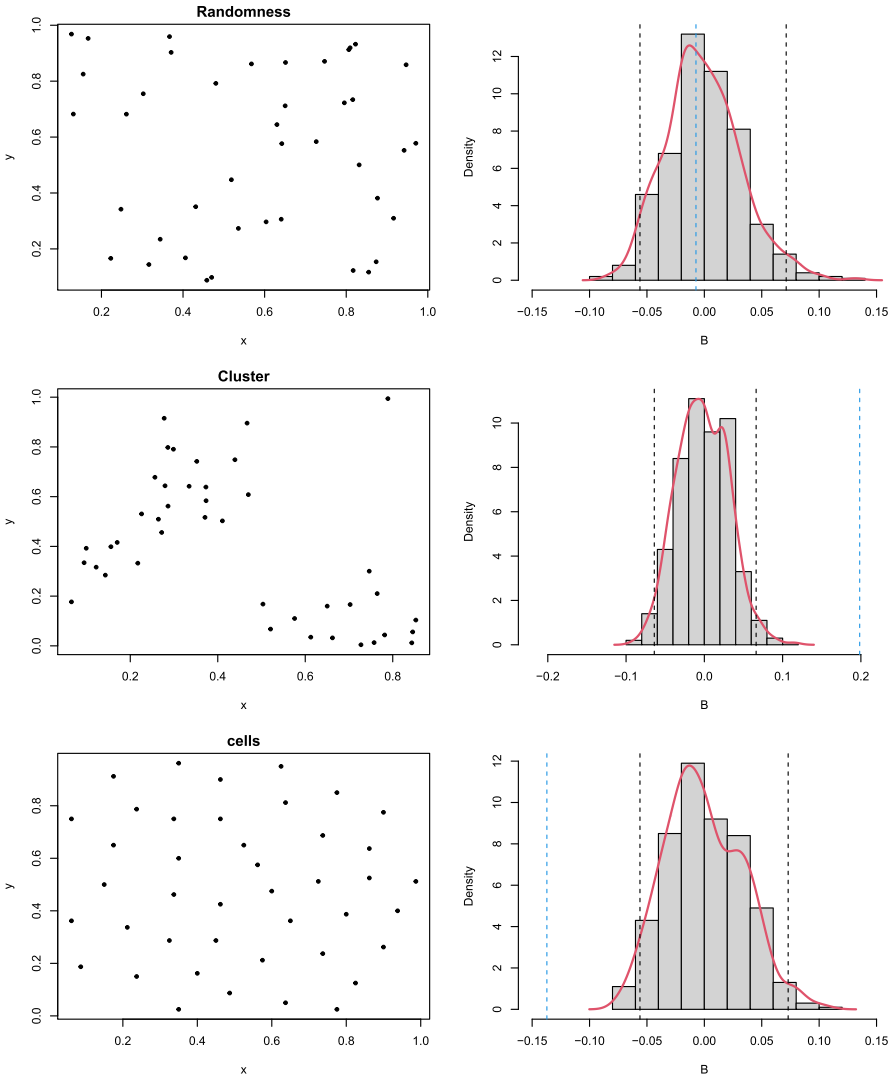

Fig. 7 Simulated point patterns under spatial randomness (top left), clustering (centre left) and point pattern cells (bottom left). The number of events is $n = 42$ in all cases. On the right, we show the null distributions of the statistic B generated by Monte Carlo simulation (histogram density estimation (grey) and kernel density estimation (red)). Dashed black lines correspond to the quantiles $B_{\frac{\alpha}{2}}$ and $B_{1-\frac{\alpha}{2}}$ ($\alpha = 5\%$) of the B values simulated. The dashed blue line corresponds to the B calculated with the point pattern given on the left panel

3.1 An illustration of the test

As an initial review of the goodness of fit of the test proposed in Sect. 3, we present the results of a Monte Carlo simulation study to describe the test behaviour under the three types of points structures generally considered in point pattern analysis.

Based on simulations from Poisson homogeneous and Matérn cluster point patterns and a real dataset (point pattern cells), we display the performance of the statistic B, and implicitly of the statistic $\hat{\mathbb{E}}(D_1)$, under randomness, clustering, and inhibition.

We show in Fig. 7 realisations of three spatial point patterns with different underlying structures: a spatial point pattern size n=42 simulated from a homogeneous Poisson model $N(A) \sim \text{Poisson}(\mu = \lambda|A| = 42)$(top left), a cluster point pattern of $n = 42$ events (centre left) generated from a Matérn cluster process with parameters $\kappa = 5, r = 0.2, \nu = 8$, and a regular point pattern (bottom left), which corresponds to the database `cells` widely known and used as example in many works on point patterns (Ripley, 1977, 1981; Diggle, 1983). Note that in the case of the Matérn process, we use ν instead of μ to avoid confusion with the notation in Eq. (16). The intensity of the Matérn cluster process is $\kappa\nu$ (Waagepetersen, 2007), and the level of aggregation is determined by parameter r. Fixed κ and ν, the aggregation level increases when r decreases (Fig. 8).

We particularly looked for simulations of size 42 to generate the point patterns under randomness and clustering (top left and centre left of Fig. 7) so that the results were more easily comparable with those of the cells pattern (which has 42 events). The distributions of the statistic B on the right of Fig. 7 were generated assuming a fixed n, although the results do not change significantly if an unconditional simulation is considered. The functions `rpoispp` and `rMatClust` of the `spatstat` library (Baddeley et al., 2015) of `R` (R Core Team, 2020) were used to simulate the random and clustered patterns. The point pattern cells are also available in

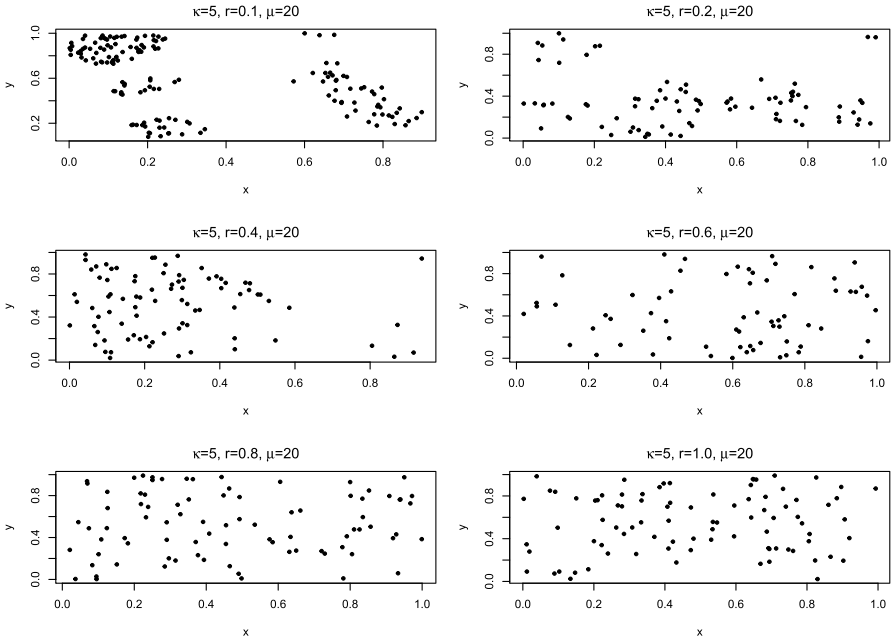

Fig. 8 Simulations of Matérn cluster point patterns with parameters (κ, r, ν). The mean at each case is $\kappa\nu$. The number of events simulated according to the r value are: 182 (r=0.1), 111 (r=0.2), 65 (r=0.4), 77 (r=0.6), 73 (r=0.8), and 79 (r=1.0)

Table 1 Expected number of events (μ), number of events recorded (n), expected box-counting dimension conditional to n ($\hat{\mathbb{E}}(D_1)$), and estimates ($\hat{\hat{\mathbb{E}}}(D_1)$ and B) for each one of the three types of point patterns considered

Point Pattern	μ	n	$\hat{\mathbb{E}}(D_1)$	$\hat{\hat{\mathbb{E}}}(D_1)$	B	$B_{0.025}$	$B_{0.0975}$
Poisson homogeneous	42	42	1.754	1.762	−0.007	−0.056	0.070
Matérn Cluster	40	42	1.754	1.556	0.198	−0.074	0.093
Cells		42	1.754	1.897	−0.132	−0.065	0.067

$B_{0.025}$ and $B_{0.975}$ are the quantiles $B_{\frac{\alpha}{2}}$ and $B_{1-\frac{\alpha}{2}}$ of B distribution ($\alpha = 5\%$)

`spatstat`. We apply the methodology presented in Sect. 3 to test the hypothesis of CSR with each one of these datasets. Employing the n values in Table 1 and Eqs. (16) and (17), we calculate for each one of the patterns in Fig. 7, $\hat{\mathbb{E}}(D_1)$, $\hat{\hat{\mathbb{E}}}(D_1)$, and B (Table 1).

A quick inspection of the results in Table 1 reveals that the value of $\hat{\hat{\mathbb{E}}}(D_1)$ found with the point pattern simulated under CSR (top left of Fig. 7) is very close to the expected value of $\hat{\mathbb{E}}(D_1)$ under complete spatial randomness, while in the other two cases, $\hat{\hat{\mathbb{E}}}(D_1)$ is relatively far from this value of reference (below when the pattern is cluster and above if it is inhibitory). The same information is taken considering the B statistic. (In this case, the reference is zero.) The value of B under the Poisson process is close to zero, while the B values of the Matérn cluster and cells patterns are far from zero (above when the pattern is cluster and below if it is inhibitory). The distribution of the statistic B under the null hypothesis was estimated generating 500 simulations from $N(A) \sim$Poisson$(\lambda|A| = 42)$ (see the histograms in right panel of Fig. 7), that is, for $j = 1 \ldots, 500$, we obtained $\hat{\hat{\mathbb{E}}}(D_1)_j$ and $B_j = (\hat{\mathbb{E}}(D_1)_j - \hat{\hat{\mathbb{E}}}(D_1)_j)$. A kernel density estimation (Sheater, 2004) of the B distribution is also obtained at each case (red curves in right panel of Fig. 7). We use a Gaussian kernel, and the bandwidth is defined using the Silverman's rule (Sheater, 2004). Note in Fig. 7 that we obtain three different distributions of B under randomness. Only one of these distributions could have been used. However, to present the results in more detail, we include three sets of independent simulations. Using the $B_j, j = 1 \ldots, 100$, and the function `quantile` of the library `stats` of `R` (R Core Team, 2020), the percentiles $B_{0.025}$ and $B_{0.975}$ of the B distribution (black dashed lines in Fig. 7) were calculated. The null hypothesis of CSR is rejected at each case if the B values are lower or greater than the estimated percentiles $B_{0.025}$ and $B_{0.975}$, respectively. The kernel density estimates (histograms and red curves) in Fig. 7 suggest that the distributions of B under CSR are symmetric around zero. A large value of B (in the upper tail of the distribution of B) will indicate that the pattern under study is clustered. On the contrary, a very low value of B (lower tail of the distribution of B) will give evidence that the process of interest follows an inhibition model.

Two aspects are noted from Table 1 and Fig. 7. On the one hand, the B value calculated with the point pattern simulated under randomness (−0.007) (dashed blue line in the top right panel of Fig. 7) is around the centre of the null distribution, i.e. as expected, the test indicates that there is not evidence to reject the null hypothesis of CSR. On the other hand, for the Matérn cluster point pattern (centre

left) and cells (bottom left), the values of the statistic B (Table (1)) are on the tails of the corresponding distributions under randomness (on the right in the case of the Matérn cluster process and the opposite for the inhibition pattern (Fig. 7)), that is, these indicate that the hypothesis of randomness should be rejected. In summary, the plots on the right panel of Fig. 7 show that the test proposed (based on B or $\hat{\hat{\mathbb{E}}}(D_1)$) in all the three cases came to the correct decision. If the hypothesis of spatial randomness is rejected, it indicates whether the pattern is cluster or inhibitory. From Table 1, it is important to note that conditional on n there is a value of reference ($\hat{\hat{\mathbb{E}}}(D_1)$) for the randomness hypothesis. The simulation-based distributions allow to establish whether the estimate $\hat{\hat{\mathbb{E}}}(D_1)$ is significantly different from this value. The value of B allows measuring the strength of inhibition or clustering. The smaller or larger (further from zero) B is, the greater the degree of inhibition or clustering, respectively, of the point pattern under consideration.

3.2 Power of the test

Algorithm 2 Power of the test based on $\hat{\mathbb{E}}(D_1)$, $\hat{\hat{\mathbb{E}}}(D_1)$, and B.

Repeat the following steps for each value of r with $r = 0.10, 0.15, 0.20, \ldots, 0.90, 0.95, 1.00$. Specify a simulation size s
for i = 1 to s **do**

1. Simulate a Matérn cluster process $(5, r, 20)$
2. Given the simulated spatial point pattern (size n) find, under the assumption of CSR, the expected box-counting dimension $\hat{\mathbb{E}}(D_1)$ (Equation (16)), its estimation $\hat{\hat{\mathbb{E}}}(D_1)$ (Equation (17)), and $B = (\hat{\mathbb{E}}(D_1) - \hat{\hat{\mathbb{E}}}(D_1))$.

for j = 1 to m **do**

(a) Simulate a sample from $N(A) \sim \text{Poisson}(\lambda|A| = n)$ and obtain n_j, the number of events generated. Note that a conditional simulation (fixing n) can also be carried out.

(b) Calculate $\hat{\mathbb{E}}(D_1)$, $\hat{\hat{\mathbb{E}}}(D_1)$ and $B = (\hat{\mathbb{E}}(D_1) - \hat{\hat{\mathbb{E}}}(D_1))$ based on the point pattern generated in (a), i.e., replace n by n_j in Equations (16) and (17). Call these values $\hat{\mathbb{E}}(D_1)_j$, $\hat{\hat{\mathbb{E}}}(D_1)_j$ and $B_j = (\hat{\mathbb{E}}(D_1)_j - \hat{\hat{\mathbb{E}}}(D_1)_j)$.

end for

3. Decision rule

- Obtain from $B_1, \ldots, B_m$ the quantile $B_{1-\alpha}$.
- Using the statistic B (calculated in step 2), reject the null hypothesis against the alternative of clustering if $B > B_{1-\alpha}$.
- Alternatively, defining the dichotomous variable

$$\psi_j = \begin{cases} 1 \text{ if } B_j > B \\ 0 \text{ other case} \end{cases},$$

Calculate the corresponding p-value as

$$\text{p-value}[i] = \frac{\sum_{j=1}^{m} \psi_j}{m}.$$

- Reject the hypothesis of CSR if p-value$[i] < \alpha$.

end for
Define

$$\pi_i = \begin{cases} 1 \text{ if p-value}[i] < \alpha \\ 0 \text{ other case} \end{cases},$$

Calculate the power of the test for each value of r as

$$\pi[r] = \frac{\sum_{i=1}^{s} \pi_i}{s}.$$

We generate realisations of a Matérn cluster process with parameters (κ, r, ν) (Waagepetersen, 2007). The method used generates a uniform Poisson point process of "parent" points with intensity κ. Then, each parent point is replaced by a random cluster of "offspring" points, the number of points per cluster being Poisson distributed, and their positions being placed and uniformly inside a disc of radius scale (r) centred on the parent point (Waagepetersen, 2007). We use the function `rMatClust` of the library `spatstat` (Baddeley et al., 2015) to generate the simulations. For six selected values of r (0.1, 0.2, 0.4, 0.6, 0.8, and 1.0), one resulting simulated process is shown in Fig. 8. From these plots, it gets clear that the smaller r, the greater the aggregation, and therefore more evidence to reject the hypothesis of spatial randomness. On the contrary, if r increases the configuration of points look more similar to a realisation of a random process under CSR. With this result in mind, in order to estimate the rejection probability of the test under different levels of spatial aggregation, we decided to propose a simulation study considering a more extensive set of values of r between 0.1 and 1 (0.1, 0.15, 0.20, ..., 0.90, 0.95, 1). Point patterns with a high level of aggregation are initially generated (using small r values), and then, (increasing r), we simulate others with point configurations similar to those obtained under spatial randomness. The procedure used is analogous to that described in Algorithm 1. Specifically, for each r value, the rejection probability of the CSR hypothesis is estimated using the iterative procedure given in Algorithm 2. The rejection probabilities for each r are shown in Table 2. According to the values from this table, it is clear that there is an inverse relationship between r (column 1) and the probability of rejecting the null hypothesis (column 13). The lower the r value, the greater $P(\text{Reject } H_0)$, i.e. the more evident the spatial aggregation, the greater the rejection probability of the complete spatial randomness hypothesis. On the contrary, when the value of r tends to one, the corresponding rejection probabilities of the randomness hypothesis tend to zero. We include in Table 2 the first 10 values of B (of the total of 500) with the corresponding associated empirical p values. It is clear from these values that there is (in general) a transition in the B values. When r is small ($r = 0.1, 0.15, 0.20$), the values of B tend to be relatively large, and therefore, the simulation-based p values are close to zero, while when r is large ($r = 0.9, 0.95, 1.00$) the opposite occurs, the values of B tend to be relatively small (close to zero or even negative), and consequently, the corresponding empirical p values are greater than α. The table results suggest that the proposed test is unbiased, i.e. the power of the test increases when the level of spatial aggregation increases.

4 CSR testing using the log–log plot

In the analysis of spatial point patterns, the test for CSR is often based on graphical methods. Generally, the distribution functions of the event–event distance (function G (Clark and Evans, 1954)), point–event distance (function F (Bartlett, 1964)), and the number of events encountered up to a given distance of any particular event (function Ripley's K Ripley (1977)) are employed for this purpose. These

functions are typically inspected by plotting the empirical function calculated from the data, together with the theoretical function of the homogeneous Poisson process with the same average intensity Baddeley et al. (2015). To assess the statistical significance of deviations between the observed and theoretical functions, it is required to know the expected variability when the pattern is completely random. To this purpose, simulated realisations under CSR are generated, and pointwise envelopes based on the minimum and maximum are calculated. In this Section, we show how the log–log plot defined in Sect. 2.2 can be used as an alternative to the functions G, F, and K to test graphically for CSR. The steps to define the graphical test based on the log–log plot are the following. Initially, calculate the log–log plot defined in Sect. 2.2 with the observed dataset. Then simulate m realisations from $N(A) \sim \text{Poisson}(\lambda|A| = n)$, and for each simulated point pattern obtain the log–log plot. From the generated m curves, define pointwise envelopes as in the case of the G, F, and K functions mentioned above. We illustrate the use of the log–log function based on the same point patterns considered in Sect. 3.1 (Fig. 7). The results obtained are compared with those found with the G, F, and K functions. We use the library `spatstat` (Baddeley et al., 2015) to generate the envelopes. Specifically, the functions `Gest`, `Fest`, `Kest` and `envelope` of the same library were used for carrying out the graphical tests. In all cases, the simulations to obtain the envelopes were conditioned to have the same number of events as the original point pattern ($n = 121$ (random), $n = 42$ (regular), and $n = 156$ (clustered)). Figures 9, 10 and 11 show the corresponding envelopes (grey shading) for the G (top left), F (top right), K (bottom left) and log–log functions (bottom right) generated from the point patterns in Fig. 7. The obtained results with the log–log function in all cases are in accordance with those given by the functions G, F, and K (Figs. 9, 10 and 11), that is, the estimated log–log function is inside the envelopes in the case of the Poisson pattern (Fig. 9) and outside of envelopes in the case clustering and inhibition (cells) (Figs. 10 and 11, respectively). From an empirical point of view, we can note that the log–log plot has the same performance as the traditional G, F, and K functions. The log–log function has an analogous interpretation to the F function. There is clustering when the estimated function is below the envelopes and inhibition when it is above (Figs. 10 and 11). The results based on the log–log plot are also consistent with those described in Sect. 3.1. Recall that under inhibition, the estimated box-counting dimension ($\hat{\mathbb{E}}(D_1)$) is greater than expected under randomness ($\hat{\mathbb{E}}(D_1)$), or the opposite if the process is clustered ($\hat{\mathbb{E}}(D_1) < \hat{\mathbb{E}}(D_1)$). A similar result can be identified from Figs. 10 and 11. The log–log plot for the pattern cells (Fig. 11) is above the envelopes, that is, it is greater than the expected log–log curve under CSR. Likewise, we can see in Fig. 10 (Matérn cluster process) that the estimated log–log function (black line) is below the envelopes, that is, the log–log plot for a clustered point pattern is lower than the expected under CSR. These results suggest a direct relationship between these two approaches.

Table 2 Statistic B and empirical p values obtained with the first 10 simulations (out of 500) of Matérn cluster point patterns ($\kappa = 5, r, \nu = 20$). The rejection probability of the null hypothesis ($P(\text{Reject } H_0)$) is calculated based on the 500 simulations

r		Results from 10 simulations										$P(\text{Reject } H_0)$
0.10	B calculated	0.37	0.50	0.39	0.43	0.24	0.46	0.39	0.28	0.28	0.45	1.00
	p−value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.15	B calculated	0.18	0.29	0.25	0.30	0.22	0.44	0.47	0.24	0.26	0.13	1.00
	p−value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.20	B calculated	0.18	0.16	0.10	0.19	0.09	0.16	0.08	0.13	0.19	0.20	1.00
	p−value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
0.25	B calculated	0.14	0.23	0.06	0.25	0.10	0.13	0.13	0.13	0.01	0.06	0.96
	p−value	0.06	0.04	0.08	0.09	0.09	0.11	0.20	0.11	0.08	0.13	
0.30	B calculated	0.12	0.00	0.03	0.06	0.01	0.04	0.09	0.19	0.11	0.21	0.78
	p−value	0.00	0.52	0.02	0.00	0.41	0.00	0.01	0.00	0.00	0.00	
0.35	B calculated	0.00	0.01	0.07	0.03	0.13	0.23	0.00	0.17	0.05	0.02	0.64
	p−value	0.00	0.00	0.01	0.00	0.12	0.24	0.00	0.00	0.01	0.20	
0.40	B calculated	0.02	0.06	0.05	0.15	0.02	0.02	0.00	0.03	0.03	0.01	0.46
	p−value	0.09	0.00	0.08	0.00	0.06	0.28	0.44	0.02	0.03	0.37	
0.45	B calculated	0.06	0.00	0.09	0.05	−0.01	0.02	0.02	−0.02	0.00	0.02	0.43
	p−value	0.01	0.40	0.03	0.01	0.75	0.11	0.22	0.82	0.52	0.14	
0.50	B calculated	0.00	0.04	0.03	0.05	0.04	0.06	0.03	0.05	0.00	0.00	0.23
	p−value	0.41	0.01	0.08	0.01	0.07	0.00	0.22	0.01	0.37	0.47	
0.55	B calculated	0.00	0.04	0.00	0.00	0.10	−0.02	0.01	0.01	0.00	0.04	0.20
	p−value	0.44	0.01	0.46	0.48	0.00	0.81	0.43	0.19	0.46	0.6	
0.60	B calculated	0.01	0.00	0.02	−0.01	−0.01	−0.01	−0.02	0.02	0.01	−0.01	0.17
	p−value	0.29	0.56	0.18	0.62	0.77	0.70	0.84	0.09	0.30	0.64	
0.65	B calculated	0.00	0.03	−0.02	−0.01	0.00	−0.05	0.02	−0.02	−0.01	−0.1	0.10
	p−value	0.49	0.10	0.92	0.72	0.30	0.98	0.14	0.97	0.59	0.74	

Table 2 (continued)

r		Results from 10 simulations										P(Reject H_0)
0.70	B calculated	0.04	0.04	0.05	−0.01	0.00	0.03	0.01	0.01	0.03	0.00	0.08
	p−value	0.06	0.06	0.05	0.64	0.64	0.58	0.06	0.36	0.21	0.58	
0.75	B calculated	0.04	0.01	−0.03	0.00	−0.01	0.01	0.03	−0.02	0.00	−0.05	0.08
	p−value	0.04	0.16	0.92	0.50	0.66	0.26	0.10	0.86	0.43	0.98	
0.80	B calculated	−0.04	0.03	0.01	−0.01	0.00	0.00	0.02	0.01	−0.01	0.00	0.06
	p−value	0.98	0.06	0.34	0.62	0.48	0.50	0.10	0.13	0.74	0.44	
0.85	B calculated	0.00	0.01	−0.02	−0.02	0.01	0.00	0.01	0.01	0.01	0.02	0.04
	p−value	0.56	0.30	0.88	0.84	0.35	0.52	0.24	0.28	0.28	0.18	
0.90	B calculated	−0.01	−0.01	−0.01	0.00	−0.02	−0.01	0.03	0.03	−0.01	0.01	0.04
	p−value	0.74	0.70	0.80	0.50	0.88	0.60	0.08	0.02	0.71	0.32	
1.00	B calculated	0.01	0.05	0.01	−0.01	0.00	−0.02	0.02	0.02	−0.02	0.03	0.02
	p−value	0.35	0.01	0.22	0.76	0.56	0.83	0.12	0.12	0.83	0.04	

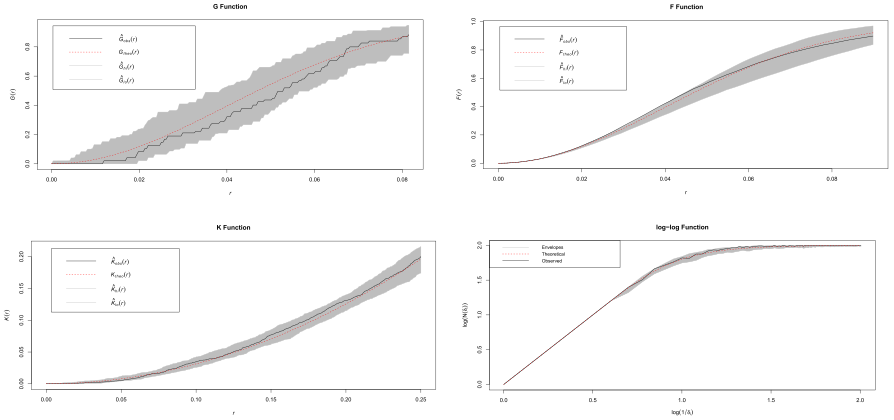

Fig. 9 Envelopes of G (top left), F (top right), K (bottom left), and log–log (bottom right) functions calculated with a Poisson process

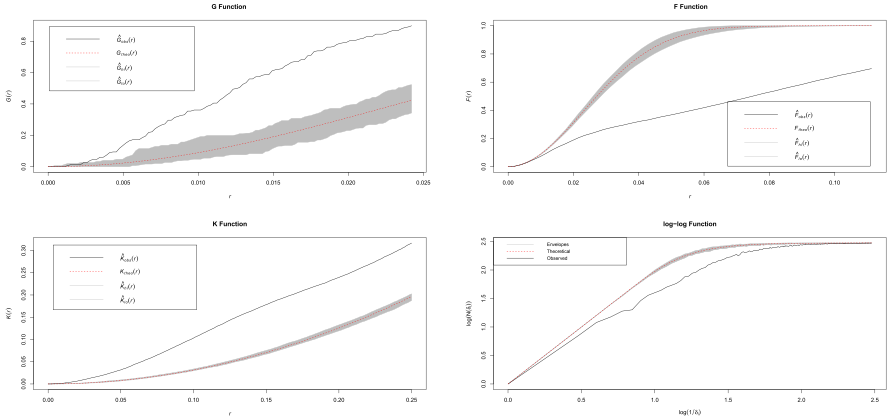

Fig. 10 Envelopes of G (top left), F (top right), K (bottom left), and log–log (bottom right) functions calculated with a Matérn cluster process

5 Spatial randomness test of COVID-19 cases in Cali, Colombia

Spatial statistics has emerged as a helpful tool in epidemiology to describe the spatial and spatio-temporal spread and incidence of different pathogens. This area of statistics is commonly used today in the study of the COVID-19 spread (a disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Chhikara et al. (2020)). Spatial statistics allows an understanding of how the COVID-19 outbreak is spatially distributed (Ramírez-Aldana et al., 2020). Studying the spatial behaviour (at the local and regional level) of the spread by COVID-19 is essential for the formulation of control and mitigation measures by government and health authorities. For this reason, there has been a growing number of academic and scientific works related to the spatial modelling of its spread patterns (Kang et al. 2020;

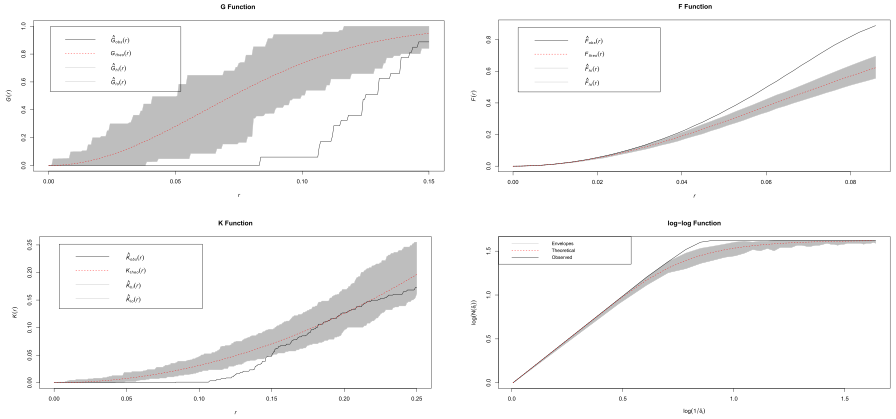

Fig. 11 Envelopes of G (top left), F (top right), K (bottom left), and log–log (bottom right) functions calculated with the point pattern `cells`

Miller et al. 2020). In this section, we show how the methodology given in Sect. 3 can be used for this purpose.

In particular, we apply the test proposed to a dataset corresponding to COVID-19 cases recorded in March 2020 in the metropolitan area of Cali city, located in the southwest region of Colombia (Fig. 12). The virus was confirmed to have reached Colombia in March 2020. Between March 2 and 31, 2020, there were 443 reports of COVID-19 infections in Cali. As input for our analysis, we take 405 spatial coordinates corresponding to the spatial residence locations of the infected people in this municipality. We exclude the duplicate coordinates. (The infections of several people in the same place are considered a single event.) In Fig. 13, it is shown the spatial distribution of the events in this month. The southernmost part of the city is rural and unpopulated, so we carry out the analysis by delimiting the perimeter to the inhabited area.

Observing both the right panel in Fig. 12 and the point pattern in Fig. 13, we identify zones with high cases burden. The most significant aggregation of cases is given in the city's south. However, other minor hot spots are placed to the west, the east, and the north. A detailed description of this respect is given in Cuartas (2020). Based on the coordinates of the spatial point pattern in Fig. 13, we estimated the functions G , F, and K (Fig. 14). The three plots are concordant and confirm the above; they allow us to conclude that the specific pattern of COVID-19 cases in Cali city during the first month of the pandemic was clustered. We also found the distribution (under CSR) of the statistics B defined in Sect. 3 (the top left panel of Fig. 14) and its calculated value $B = (\hat{\mathbb{E}}(D_1) - \hat{\mathbb{E}}(D_1)) = 0.1549$ (dashed blue line in Fig. 14) with the point pattern in Fig. 13. The B value is on the right tail of the distribution. Consequently, it indicates that the null hypothesis of CSR must be rejected, (The same conclusion given by the classic graphical tests.)

We have analysed just one dataset of COVID-19 cases. The four strategies allow us to reach the same conclusion. However, there are implicitly advantages in using the method based on the box-counting estimation. On the one hand, we have a p

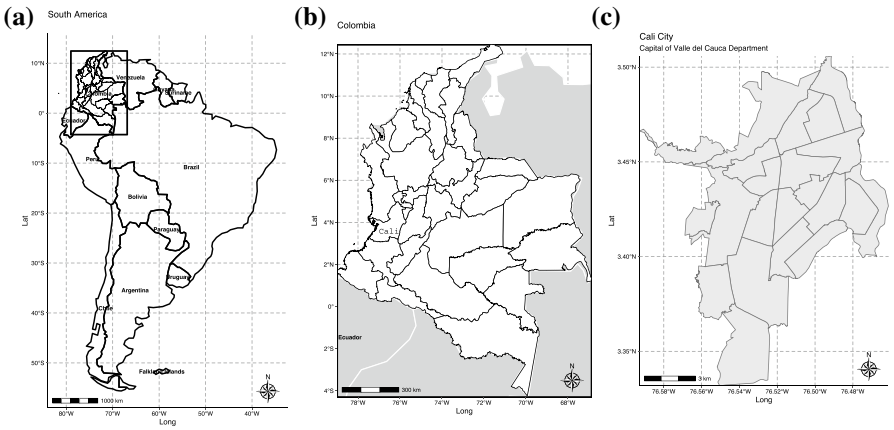

Fig. 12 Geographical location of the study area. Cali is the capital of the Department of Valle del Cauca

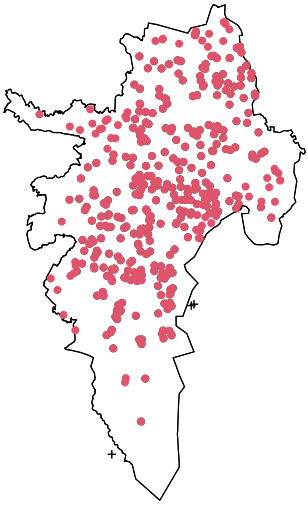

Fig. 13 COVID 19 infection sites (March 2020) in the urban area of Cali, Colombia. The sites that are outside the perimeter of the urban area (symbol +) are not considered in the analysis

value (see Algorithm 2 for its estimation), which allows being conclusive. (Sometimes the graphical tests are not.) On the other hand, using B (equivalently in $\hat{\mathbb{E}}(D_1)$) the point pattern under study is characterised with just one value. This opens the doors to the application of many traditional techniques (regression, ANOVA, longitudinal data analysis, time series, etc.) in those situations in which there is a collection of point patterns to be analysed simultaneously (obtained, for example, in different periods or under various experimental conditions).

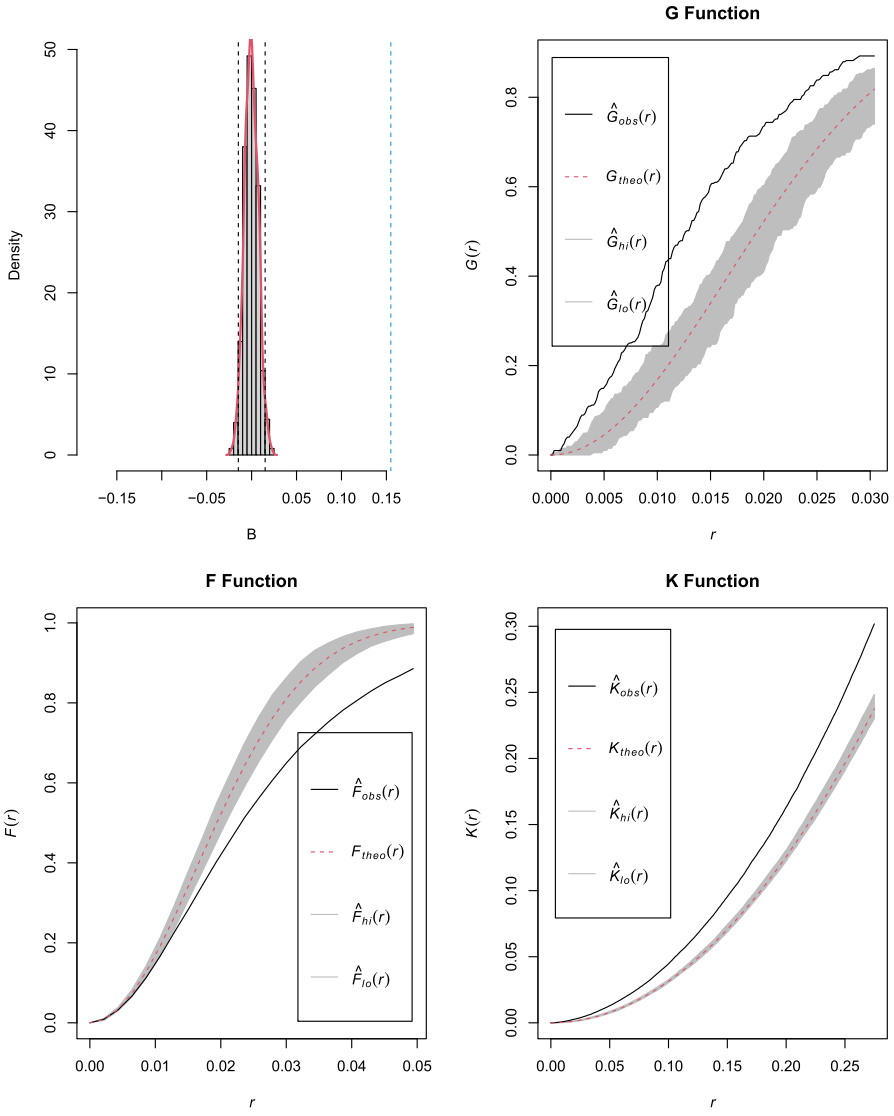

Fig. 14 Tests based on the expected value of the box-counting dimension (top left) and the functions G (top right), F (bottom left), and K (bottom right)

6 Conclusions and further research

We have proposed a test to evaluate the hypothesis of complete spatial randomness based on the fractal dimension and its estimation by the box-counting methodology. Also, a graphical test is derived. Using simulated point patterns under randomness, inhibition and clustering, we found that the two approaches have a good performance. The results are concordant and coherent with those obtained employing

classical graphical tests (G, F, and K functions). The graphical interpretation of the proposed test is similar to that obtained with the F function. The tests are not based on distances, and therefore, it is not necessary to consider the edge effect. A simulation study was carried out to show the behaviour of the test proposed under the null hypothesis (randomness) and the classical alternatives (inhibition and clustering). The simulation results were satisfactory. A detailed study about the power of the test was also conducted. This allows us to conclude that the test has a good performance under different levels of clustering. An advantage of the methodology considered is that a statistic is calculated ($\hat{\mathbb{E}}(D_1)$ or equivalently B), which allows summarising the information of the point pattern in just one value. This can be useful from many inferential perspectives. For example, for modelling spatio-temporal point patterns or comparing groups of point patterns through ANOVA.

Acknowledgements This work is part of the research project "Modelación Espacio-Temporal del Covid-19 en Colombia" financed by Dirección de Investigación of Universidad Nacional de Colombia. We thank the epidemiological surveillance group of the Secretary of Health of Cali for providing us with the analysed information.

References

Addison, P.: Fractals and Chaos: an illustrated course. CRC Press, London (1997)
Baddeley, A., Gregori, P., Mateu, J., Stoica, R., Stoyan, D.: Case studies in spatial point process modeling. Springer, Berlin (2006)
Baddeley, A., Turner, R., Mateu, J., Bevan, A.: Hybrids of Gibbs point process models and their implementation. J. Stat. Softw. **55**(11), 1–43 (2013)
Baddeley, A., Rubak, E., Turner, R.: Spatial point patterns: methodology and applications with R. Chapman and Hall/CRC, Boca Raton (2015)
Banerjee, S., Carlin, B., Gelfand, A.: Hierarchical modeling and analysis for spatial data. CRC Press, Boca Raton (2015)
Bartlett, M.: The spectral analysis of two-dimensional point processes. Biometrika **51**(3/4), 299–311 (1964)
Bivand, R., Pebesma, E., Gomez-Rubio, V.: Applied spatial data analysis with R. Springer, Berlin (2013)
Bones, C., Romani, L., de Sousa, E.: Clustering multivariate data streams by correlating attributes using fractal dimension. J. Inf. Data Manag. **7**(3), 249–249 (2016)
Breslin, M., Belward, J.: Fractal dimensions for rainfall time series. Math. Comput. Simul. **48**(4–6), 437–446 (1999)
Clark, P., Evans, F.: Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology **35**(4), 445–453 (1954)
Chhikara, B., Rathi, B., Singh, J., Poonam, F.: Corona virus SARS-CoV-2 disease COVID-19: infection, prevention and clinical advances of the prospective chemical drug therapeutics. Chem. Biol. Lett. **7**(1), 63–72 (2020)
Cressie, N.: Statistics for spatial data. Wiley, Hoboken (1991)
Cuartas, et al.: SARS-coV-2 spatio-temporal analysis in Cali. Colombia. Revista de Salud Pública **22**(2), 1–6 (2020)
Daley, D., Vere-Jones, D.: An introduction to the theory of point processes. Springer, Berlin (2008)
Debnath, L.: A brief historical introduction to fractals and fractal geometry. Int. J. Math. Educat. Sci. Technol. **37**(1), 29–50 (2006)
Diggle, P.: Statistical analysis of spatial point patterns. Academic Press, Cambridge (1983)
Diggle, P.: Statistical analysis of spatial point patterns. Edward Arnold (2003)
Diggle, P.: Statistical analysis of spatial and spatio-temporal point patterns. CRC Press, Boca Raton (2013)
Falconer, K.: Fractal geometry: mathematical foundations and applications. Wiley, Hoboken (2004)

Foroutan-pour, K., Dutilleul, P., Smith, D.: Advances in the implementation of the box-counting method of fractal dimension estimation. Appl. Math. Comput. **105**(2–3), 195–210 (1999)
Gaetan, C., Guyon, X.: Spatial statistics and modeling. Springer, Berlin (2010)
García, L., Bravo, L., Collazos, P., Ramírez, O., Carrascal, E., Nuñez, M., Portilla, Millan, E.: Métodos del Registro de Cáncer en Cali. Colombia. Revista Colombia Médica **49**(1), 109–120 (2018)
Illian, J., Penttinen, A., Stoyan, H., Stoyan, D.: Statistical analysis and modelling of spatial point patterns. Wiley, Hoboken (2008)
Jaquette, J., Schweinhart, B.: Fractal dimension estimation with persistent homology: a comparative study. Commun. Ecol. **84**, 105163 (2013)
Kang, D., Choi, H., Kim, J., Choi, J.: Spatial epidemic dynamics of the COVID-19 outbreak in China. Int. J. Infect. Dis. **94**, 96–102 (2020)
Kenkel, N.: Sample size requirements for fractal dimension estimation. Commun. Ecol. **14**(2), 144–152 (2013)
Kopytov, V., Petrenko, V., Tebueva, F., Streblianskaia, N.: An improved brown's method applying fractal dimension to forecast the load in a computing cluster for short time series. Indian J. Sci. Technol. **9**(19), 93909 (2016)
Liebovitch, L., Toth, T.: A fast algorithm to determine fractal dimensions by box counting. Phys. Lett. A **141**(8–9), 386–390 (1989)
Mou, D., Wang, Z.: Fractal dimension of well logging curves associated with the texture of volcanic rocks. In: 2014 international conference on mechatronics, electronic, industrial and control engineering (MEIC-14), (2014)
Mandelbrot, B.: How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science **156**, 636–644 (1967)
Mandelbrot, B.: The fractal geometry of nature. Freeman, New York (1982)
Miller, L., Bhattacharyya, R., Miller, A.: Spatial analysis of global variability in Covid-19 burden. Risk Manag. Healthc. Policy **13**, 519–522 (2020)
Mo, D., Huang, S.: Fractal-based intrinsic dimension estimation and its application in dimensionality reduction. IEEE Trans. Knowl. Data Eng. **24**(1), 59–71 (2010)
Møller, J., Waagepetersen, R.: Statistical inference and simulation for spatial point processes. Chapman and Hall/CRC, London (2004)
Plant, R.: Spatial data analysis in ecology and agriculture using R. CRC Press, London (2012)
R Core Team. (2020): R: A Language and Environment for Statistical Computing. R foundation for statistical computing, Vienna, Austria, https://www.R-project.org/
Ramírez-Aldana, R., Gomez-Verjan, J., Bello-Chavolla, O.: Spatial analysis of COVID-19 spread in Iran: insights into geographical and structural transmission determinants at a province level. PLoS Neglect. Trop. Dis. **14**(1), e0008875 (2020)
Ripley, B.: Modelling spatial patterns. J. R. Stat. Soc. Ser. B **39**(2), 172–192 (1977)
Ripley, B.: Spatial statistics. Wiley, Hoboken (1981)
Salvadori, G., Ratti, S., Belli, G.: Modelling spatial patterns. Environ. Sci. Pollut. Res. **4**(2), 91–98 (1997)
Schabenberger, O., Gotway, C.: Statistical methods for spatial data analysis. Chapman and Hall/CRC, London (2017)
Sheater, S.: Density estimation. Stat. Sci. **19**(4), 588–597 (2004)
Tuia, D., Kanevski, M.: Environmental monitoring network characterization and clustering. Geostatistics, machine learning and Bayesian maximum entropy, advanced mapping of environmental data (2008) pp. 19–46
Vega, C., Golay, J., Kanevski, M.: Multifractal portrayal of the Swiss population. Cybergeo: Eur. J. Geogr., (2015) http://journal.openedition.org/cybergeo/26829
Vidal, E., Vieira, S., Clerici, I., Paz, A.: Fractal dimension and geostatistical parameters for soil microrelief as a function of cumulative precipitation. Scientia Agricola **67**(1), 78–83 (2010)
Wiegand, T., Moloney, K.: Handbook of spatial point-pattern analysis in ecology. CRC Press, London (2013)
Waagepetersen, R.P.: An estimating function approach to inference for inhomogeneous Neyman-Scott processes. Biometrics **63**, 252–258 (2007)