ESAIM: PS 17 (2013) 767–788
DOI: 10.1051/ps/2012027

ESAIM: Probability and Statistics
www.esaim-ps.org

TESTING RANDOMNESS OF SPATIAL POINT PATTERNS WITH THE RIPLEY STATISTIC

Gabriel Lang[1] and Eric Marcon[2]

Abstract. Aggregation patterns are often visually detected in sets of location data. These clusters may be the result of interesting dynamics or the effect of pure randomness. We build an asymptotically Gaussian test for the hypothesis of randomness corresponding to a homogeneous Poisson point process. We first compute the exact first and second moment of the Ripley K-statistic under the homogeneous Poisson point process model. Then we prove the asymptotic normality of a vector of such statistics for different scales and compute its covariance matrix. From these results, we derive a test statistic that is chi-square distributed. By a Monte-Carlo study, we check that the test is numerically tractable even for large data sets and also correct when only a hundred of points are observed.

Mathematics Subject Classification. 60G55, 60F05, 62F03.

Received June 22, 2011. Revised November 5, 2012.

Introduction

Analysis of point patterns is relevant in many sciences: cell biology, ecology or spatial economics. The observation of clusters in point locations is considered as a hint for non observable dynamics. For example the clustering of tree locations in a forest may come from better soil conditions or from spreading of seeds of a same mature individual; but clusters are also observed in random distribution as a Poisson point process sample. It is therefore essential to distinguish between clusters resulting from relevant interactions or from complete randomness. Ripley function [20,21] is a widely used tool to quantify the structure of point patterns, especially in ecology, and is well referenced in handbooks [7, 8, 15, 18, 23, 25]. Up to a renormalization by the intensity of the process, this statistic denoted here $\hat{K}(r)$ estimates the expectation $K(r)$ of the number of neighbors at distance less than r of a point in the sample. The observed $\hat{K}(r)$ is compared to the value of $K(r)$ for a homogeneous Poisson point process chosen as a null hypothesis: the Poisson point process is characterized by an independence of point locations, modelling an absence of interactions between individuals in ecosystems. In this case $K(r)$ is simply the mean number of points in a ball of radius r divided by the intensity, that is πr^2. If $\hat{K}(r)$ is significantly larger than πr^2 (respectively smaller), the process is considered as aggregated (respectively over-dispersed) at distance r.

Keywords and phrases. Central limit theorem, goodness-of-fit test, Höffding decomposition, K-function, point pattern, Poisson process, U-statistic.

[1] AgroParisTech, UMR 518 Mathématique et Informatique Appliquées, 19 avenue du Maine, 75732 Paris Cedex 15, France. gabriel.lang@agroparistech.fr

[2] AgroParisTech, UMR 745 Ecologie des Forêts de Guyane, Campus agronomique BP 316, 97379 Kourou Cedex, France. eric.marcon@agroparistech.fr

Article published by EDP Sciences

© EDP Sciences, SMAI 2013

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

In order to decide if the difference is statistically significant, we build a test of the Poisson process hypothesis; we need information on the distribution of $\hat{K}(r)$ for this process. But even the variance is not known and statistical methods generally rely on Monte-Carlo simulations. Ripley [22] used them to get confidence intervals. Starting from previous results [24], he also gave critical values for the L function, a normalized version of K introduced by [4]. These critical values are valid asymptotically, for a large number of points but low intensity, so that both edge effects and point-pair dependence can be neglected. Further computations of confidence interval bands based on simulation have been proposed in [16] and corrected in [5]. But the simulation is a practical issue for large point patterns, because computation time is roughly proportional to the square of the number of points (one has to calculate the distances between all pairs of points) multiplied by the number of simulations.

We propose here to compute the exact variance of the Ripley statistic. Ward and Ferrandino [30] studied this variance. But they ignored that point pairs are not independent even though points are (Eq. A8, p. 235), thus their derivation of the variance of $\hat{K}(r)$ was erroneous. A rigourous computation of the variance has been carried out in [27] for a independent sample of uniform variables on the unit square, that is for the Poisson process conditioned by a fixed number of points; for the Poisson process, we compute the exact covariance, considering the Ripley statistic as a U-statistic as remarked in [22] and using the Höffding decomposition. As the variance is not enough to build a test, we study the distribution of the statistic. We prove its asymptotic normality as the size of the observation window grows. It is then easy to build an asymptotically Gaussian test.

Another concern is to test simultaneously the aggregation/dispersion at different scales. This is rarely correctly achieved in practical computations with Monte-Carlo simulations. The confidence bands or test rejection zone are often determined without taking the dependence between the numbers of neighbors at different scales into account. Heinrich [14] proposed the first multiscale goodness-of-fit tests based on the Ripley function for Poisson processes. He considered a set of scales $(r_1, \dots, r_d)$, computed the covariance matrix of the estimates $\hat{K}(r_i)$ and proved the asymptotic normality for the vector $(\hat{K}(r_1), \dots, \hat{K}(r_d))$. He derived Kolmogorov–Smirnov, Cramer von Mises and chi-square goodness-of-fit tests from these results. Grabarnik and Chiu [10] proposed a similar test based on the k first neighbors of a point, that is more difficult to use in practice because the number of neighbours is an additional parameter to tune. The test that we propose is very similar to the chi-square test of Heinrich; the only difference lies in the correction of the bias due to edge effects. Our method of correction allows us to compute the exact value of the covariance matrix and not only its asymptotical value, as for the Heinrich test. This is a major improvement in practice because the level of the test is very sensitive to approximations in the computation of the covariance matrix. A similar exact computation of the variance matrix is untractable for the Heinrich test: only an estimation method based on subsampling of the data may be proposed as done in [12] for the inhomogeneous case.

The paper is built as follows: Section 1 introduces the precise definition of $K(r)$ and the current definition of $\hat{K}(r)$. In Section 2, after the definition of our statistics (no edge effects correction, known or unknown intensity), we list the main results of the paper: exact bias due to the edge effects and exact variance of $\hat{K}(r)$ for a homogeneous Poisson process with known or unknown intensity; covariance between $\hat{K}(r)$ and $\hat{K}(r')$ for two different distances r and r'. The main theorem contains the convergence of the vector $(\hat{K}(r_1), \dots, \hat{K}(r_d))$ to a Gaussian distribution with explicit covariance in the following asymptotic framework: data from the same process are collected on growing squares of observation. These results allow a simple, multiscale and efficient test procedure of the Poisson process assumption. Section 3 provides a Monte-Carlo comparison of the tests and Section 4 gives our conclusions. The last section contains the proofs.

1. Definition of the Ripley K-function

We recall the characterizations of the dependence of the locations for a general point process X over $\mathbb{R}^2$. We refer to the presentation of [18].

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

1.1. Definitions

For a point process X, define the point process $X^{(2)}$ on $\mathbb{R}^2 \times \mathbb{R}^2$ of all the couples of two different points of the original process. The intensity of this new process gives information on the simultaneous presence of points in the original process. Denote $\rho^{(2)}(x,y)$ its density (called the second-order product density). The Poisson process of density $\rho(x)$ is such that $\rho^{(2)}(x,y) = \rho(x)\rho(y)$.

The Ripley statistic is a way to estimate the density $\rho^{(2)}(x,y)$. Precisely it is an estimate of the integral on test sets of the ratio $\mathsf{g}(x,y) = \rho^{(2)}(x,y)/\rho(x)\rho(y)$. The function $\mathsf{g}(x,y)$ characterizes the fact that the points x and y appear simultaneously in the samples of X. If $\mathsf{g}(x,y) = 1$, the points appear independently. If $\mathsf{g}(x,y) < 1$, they tend to exclude each other; if $\mathsf{g}(x,y) > 1$, they appear more frequently together.

We assume the translation invariance of the point process: $\mathsf{g}(x,y) = \mathsf{g}(x-y)$. In order to estimate the function g, we define its integral as the set function $\mathcal{K}$. Let A be a Borel set:

$$\mathcal{K}(A) = \int_A \mathsf{g}(x)\mathrm{d}x.$$

If we also assume that the point process is isotropic, we define the Ripley K-function as

$$K(r) = \mathcal{K}(B(x,r)),$$

where $B(x,r)$ is the closed ball with center x and radius r. The translation invariance implies that $\mathcal{K}(B(x,r))$ does not depend on x. For example, if the process is a Poisson process then $\mathsf{g}(x) = 1$ and $K(r) = \pi r^2$. We define the Ripley statistic that estimates the K-function. Let A be a bounded Borel set of the plane $\mathbb{R}^2$, m the Lebesgue measure, $\widehat{\rho}(x)$ an estimator of the local intensity of the process and $\mathrm{I}\{\cdot\}$ denotes the indicator function of a set; for a realization S of the point process X, $S = \{X_1, \ldots, X_N\}$, a general form of the Ripley statistic is

$$\widehat{K}_A(r) = \frac{1}{m(A)} \sum_{X_i \neq X_j \in S} \frac{\mathrm{I}\{d(X_i, X_j) \leq r\}}{\widehat{\rho}\,(X_i)\ \widehat{\rho}\,(X_j)}.$$

Note that estimator $\widehat{K}_A(r)$ refers to a preexistent estimator of the local intensity $\widehat{\rho}(x)$, to make it unsensible to the inhomogeneity of the intensity. In practice, $\widehat{\rho}(x)$ is a local kernel estimator, that uses the only available local information contained in the locations of neighbors in a fixed ball around the considered point of the sample. This estimator is then very much dependent of the indicator function in the numerator, because they are based on the same information. It cannot be considered as a constant close to the true value of the local intensity. This is why we do not manage to compute the exact value of the two first moments of this statistic. We only address the problem of testing homogeneous Poisson processes and the Ripley statistic has simplified expressions given below.

2. Main results

This section presents the theoretical results on the Ripley statistic and the resulting test.

2.1. Definitions and assumptions

Throughout the paper, we refer to the expectation $e_{r,n}$, the centered indicator function h and its conditional expectation h_1. We gather here these definitions.

Let n be an integer; A_n denotes the square $[0,n]^2$; U is a random location in A_n with an uniform random distribution; its density is $1/n^2$ with respect to the Lebesgue measure $\mathrm{d}\xi_1\mathrm{d}\xi_2$ over A_n. V is an independent copy of U. We denote $d(x,y)$ the Euclidean distance between x and y in the plane. We define $e_{r,n} = \mathbb{E}(\mathrm{I}\{d(U,V) \leq r\})$, $h(x,y,r) = \mathrm{I}\{d(x,y) \leq r\} - e_{r,n}$ and $h_1(x,r) = \mathbb{E}(h(U,V,r)|\ V = x)$.

We assume that X is a homogeneous Poisson process on $\mathbb{R}^2$ with intensity ρ. We consider that the data are available on the square A_n. The setting of the asymptotics was suggested by practitioners in ecological modeling

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

and forestry: the accumulation of tree location data comes from measuring wider and wider sets of land and the inter-tree distances r do not vary with n.

Let N denote the random number of observed points and $S = \{X_1, \dots, X_N\}$ denote the sample of observed points. We consider two cases:

1. If the intensity ρ is known, the Ripley statistic is expressed as

$$\widehat{K}_{1,n}(r) = \frac{1}{n^2\rho^2} \sum_{X_i \neq X_j \in S} \mathrm{I}\{d(X_i, X_j) \leq r\}.$$

2. If the intensity ρ is unknown, we use the unbiased estimator $\widehat{\rho^2} = N(N-1)/n^4$ (see [26]) and define

$$\widehat{K}_{2,n}(r) = \frac{n^2}{N(N-1)} \sum_{X_i \neq X_j \in S} \mathrm{I}\{d(X_i, X_j) \leq r\}.$$

2.2. Bias

It is known that a large number of neighbors of the points located near the edges of A_n may lie outside A_n causing a bias in the estimation. We compute the bias due to this edge effect.

Proposition 2.1. *Assume that* $r/n < 1/2$.

$$\mathbb{E}\widehat{K}_{1,n}(r) - K(r) = r^2 \left(-\frac{8r}{3n} + \frac{r^2}{2n^2} \right).$$
$$\mathbb{E}\widehat{K}_{2,n}(r) - K(r) = r^2 \left(-\frac{8r}{3n} + \frac{r^2}{2n^2} \right) - r^2 \left(1 + \rho n^2\right) \mathrm{e}^{-\rho n^2} \left(\pi - \frac{8r}{3n} + \frac{r^2}{2n^2} \right).$$

Notes.

- The assumption $r/n < 1/2$ means that at least some balls of radius r are included in the square A_n.
- The additional term for $K_{2,n}$ corresponds to the probability to draw a sample with zero or one point in the square. This term gives a zero contribution as soon as the mean number of points ρn^2 is larger than 20.
- The proof may be adapted for a convex polygon of perimeter Ln to compute the first order term of the bias; for $u = 1$ or 2:

$$\mathbb{E}\widehat{K}_{u,n}(r) - K(r) = -\frac{2Lr^2}{3}\frac{r}{n} + O\left(\frac{r^2}{n^2}\right).$$

2.3. Variance

We compute the covariance matrix of $\widehat{K}_{u,n}(r)$ for $u = 1$ or 2. We get an exact computation for the variance, that can be used for any value of n.

Proposition 2.2. *For* $0 < r < r'$,

$$\operatorname{var}\left(\widehat{K}_{1,n}(r)\right) = \frac{2e_{r,n}}{\rho^2} + \frac{4n^2 e_{r,n}^2}{\rho} + \frac{4n^2}{\rho}\mathbb{E}h_1^2(U, r),$$
$$\operatorname{cov}\left(\widehat{K}_{1,n}(r), \widehat{K}_{1,n}(r')\right) = \frac{2e_{r,n}}{\rho^2} + \frac{4n^2 e_{r',n} e_{r,n}}{\rho} + \frac{4n^2}{\rho}\operatorname{cov}\left(h_1(U, r'), h_1(U, r)\right),$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

$$\begin{aligned}\operatorname{var}(\widehat{K}_{2,n}(r)) &= 2n^4\mathbb{E}\left(\frac{I\{N>1\}}{N(N-1)}\right)(e_{r,n}-e_{r,n}^2) = 4n^4\mathbb{E}\left(\frac{I\{N>1\}(N-2)}{N(N-1)}\right)\mathbb{E}h_1^2(U,r)\\ &\quad + n^4\mathrm{e}^{-\rho n^2}\left(1+\rho n^2\right)\left(1-\mathrm{e}^{-\rho n^2}-\rho n^2\mathrm{e}^{-\rho n^2}\right)\mathrm{e}_{r,n}^2,\\ \operatorname{cov}(\widehat{K}_{2,n}(r),\widehat{K}_{2,n}(r')) &= 2n^4\mathbb{E}\left(\frac{I\{N>1\}}{N(N-1)}\right)(e_{r,n}-e_{r',n}e_{r,n})\\ &\quad + 4n^4\mathbb{E}\left(\frac{I\{N>1\}(N-2)}{N(N-1)}\right)\operatorname{cov}(h_1(U,r'),h_1(U,r))\\ &\quad + n^4\mathrm{e}^{-\rho n^2}\left(1+\rho n^2\right)\left(1-\mathrm{e}^{-\rho n^2}-\rho n^2\mathrm{e}^{-\rho n^2}\right)e_{r',n}e_{r,n},\end{aligned}$$

where $e_{r,n}=\dfrac{\pi r^2}{n^2}-\dfrac{8r^3}{3n^3}+\dfrac{r^4}{2n^4}$ *and* $\mathbb{E}h_1^2(U,r)=\dfrac{r^5}{n^5}\left(\dfrac{8}{3}\pi-\dfrac{256}{45}\right)+\dfrac{r^6}{n^6}\left(\dfrac{11}{48}\pi-\dfrac{56}{9}\right)+\dfrac{8}{3}\dfrac{r^7}{n^7}-\dfrac{1}{4}\dfrac{r^8}{n^8}$

Notes.

- The variances of both estimators are exact and can be computed with any precision, as inverse moments of the Poisson variable correspond to fast converging series.
- The covariances are not explicit because the terms $\operatorname{cov}(h_1^2(U,r'),h_1^2(U,r))$ involve parts that have to be numerically integrated.
- The leading terms of the variances of $K_{1,n}(r)$ and $K_{2,n}(r)$ as n tends to infinity are $2\pi r^2/n^2\rho^2+4\pi r^4/n^2\rho$ and $2\pi r^2/n^2\rho^2$.

2.4. Central Limit Theorem

We show that a normalized vector of Ripley statistics for different r converges in distribution to a normal vector. Let $\mathcal{N}(0,\Sigma)$ denote the Gaussian multivariate centred distribution with covariance matrix Σ.

Theorem 2.3. *Let d be an integer, $0<r_1<\ldots<r_d$ a set of reals and for $u=1$ or 2, define $\mathcal{K}_{u,n}=(\widehat{K}_{u,n}(r_1),\ldots,\widehat{K}_{u,n}(r_d))$. Then $n\sqrt{\rho}(\mathcal{K}_{u,n}-\pi(r_1^2,\ldots r_d^2))$ converges in distribution to $\mathcal{N}(0,\Sigma)$ as n tends to infinity, where for s and t in $\{1,\ldots,d\}$*

- *if* $u=1$, $\Sigma_{s,t}=\dfrac{2\pi(r_s^2\wedge r_t^2)}{\rho}+4\pi^2r_s^2r_t^2$.
- *if* $u=2$, $\Sigma_{s,t}=\dfrac{2\pi(r_s^2\wedge r_t^2)}{\rho}$.

Note. The first term of the variance corresponds to a case where the couples of points are independent from each others; this was used as an approximation without proof in [30]; our work proves that the actual variance and limit process are different in the first case and that the approximation holds only in the second case.

2.5. Applications to test statistics

From Theorem 2.3, we deduce that $T_u=\Sigma^{-1/2}\mathcal{K}_{u,n}$ is asymptotically $\mathcal{N}(0,I_d)$ distributed. For the hypothesis

$$H_0\text{: } X \text{ is a homogeneous Poisson process of intensity } \rho$$

we use $T^2=\|T_u\|_2^2$ as a test statistic with rejection zone for the level α:

$$T^2>\chi_\alpha^2(d).$$

where $\chi_\alpha^2(d)$ is the $(1-\alpha)$-quantile of the $\chi^2(d)$ distribution.

Note. the covariance matrix Σ depends on the intensity parameter ρ, so that in the case of the unknown parameter we have to use an estimate of ρ in the formula defining Σ.

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

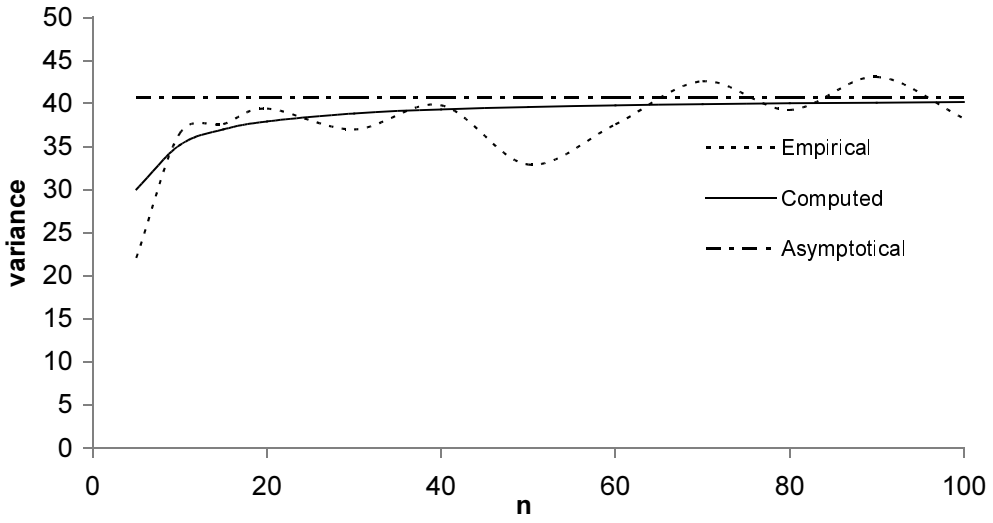


FIGURE 1. Comparison of normalized variances for $K_1(1)$, $\rho = 5$.

3. SIMULATIONS

We study the empirical variance of the proposed statistics by a Monte-Carlo simulation. Then we apply the test procedure to simulated data sets, observe the number of rejections and compare it to the level of the test.

3.1. Variance

We simulate a sample of 1000 repetitions with $\rho = 5$ and compare (after renormalization by $n\sqrt{\rho}$) the empirical variance and the exact computed variance with the limit variance for different value of n (Fig. 1). With 1000 repetitions, the oscillations of the empirical variance are still large; we will use a larger number of repetitions in the following study of the test.

The convergence of the computed variance to the limit value is not so fast and for applications with hundreds of points (corresponding in Fig. 1 to $n < 15$) the distance between the variances is still large. A preliminary study, not presented here, showed that the test procedure is perturbed by any small error in the covariance matrix, as we tried simplified versions of the covariance by ignoring the contribution of points in the corner of the observation window. It is crucial to use an accurate computation of the covariance matrix to have a correct approximation of the square root inverse matrix $\Sigma^{-1/2}$. Therefore we will use the exact variance formula instead of the asymptotic formula in the test procedure.

3.2. Test level

In the known parameter case, the computation of the test statistic T_1 is straightforward. In the unknown parameter case, the computation of the test statistic T_2 is done by replacing the unknown parameter ρ by the estimator N/n^2. We also choose to replace the expectation $\mathbb{E}\left(\mathrm{I}\{N > 1\}/(N(N-1))\right)$ by the observed value $1/(N(N-1))$ and $\mathbb{E}\left(\mathrm{I}\{N > 1\}(N-2)/(N(N-1))\right)$ by $(N-2)/(N(N-1))$, because the dispersion of a Poisson variable is low with respect to the expectation when its parameter is large. For comparison, a chi-square test T_3 based on the unbiased Ripley estimator $\widehat{K}_{3,n}$ is given using the asymptotic variance as proposed in [14]. The correction of the bias consists in dividing the indicator function not by the constant area of the

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

TABLE 1. Percentile of rejection over 10 000 repetitions of the test with level $\alpha = 0.05$.

Poisson			T_1	T_2	T_3
$n = 30$	$\rho = 1$	$r = (0.2, 0.5, 1)$	5.14	5.17	5.78*
$n = 10$	$\rho = 5$	$r = (0.2, 0.5, 1)$	4.66	4.74	12.31*
$n = 10$	$\rho = 5$	$r = (0.1, 0.2, \ldots, 1)$	5.37	5.10	10.78*
$n = 10$	$\rho = 1$	$r = (1, 2, 5)$	5.62*	5.09	56.30*
$n = 10$	$\rho = .2$	$r = (1, 1.5, 2)$	6.74*	5.27	9.22*
$n = 10$	$\rho = .2$	$r = (0.2, 0.5, 1)$	6.47*	6.59*	7.73*

square $m(A_n) = n^2$, but by the area of the intersection of the translated squares $A_n + X_i$ and $A_n + X_j$. The corresponding unbiased estimator (see [14]) is:

$$\widehat{K}_{3,n}(r) = \frac{n^4}{N(N-1)} \sum_{X_i \neq X_j \in S} \frac{\mathrm{I}\{d(X_i, X_j) \leq r\}}{m((A_n + X_i) \cap (A_n + X_j))}.$$

Concerning the choice for the range for distances r, there are two situations. From the theoretical point of view, all the scales are of the same interest. One may plot the statistic K and choose the range where the empirical values depart from expected and investigate if the difference is significative. From the practical point of view, when observing real data, practitioners often know in advance the scale they are interested in: range from 2 to 50 meters for tree locations for example, or ten meters to one kilometer for locations of shops in a city; ... Concerning the number d of different distances in a fixed range, it is theoretically not very useful to compute K for a lot of them, because K is a step function so that there is a limit to the information one gathers by refining the distances.

The test output is a Bernoulli random variable with parameter α. With a sufficient index of repetition m, the mean number of rejection is close to a normal variable with expectation α and variance $\alpha(1-\alpha)/m$. We consider that the test works correctly when the observed frequency of rejection is in the 95% Gaussian confidence interval $[\alpha - 1.96\sqrt{\alpha(1-\alpha)/m}, \alpha + 1.96\sqrt{\alpha(1-\alpha)/m}]$. With $m = 10\,000$ and $\alpha = 0.05$, the interval is $[0.0457; 0.0543]$. Percentiles of rejection in Table 1 should lie in $[4.57; 5.43]$. Stars indicate values outside this confidence interval. The performances of T_1 (known parameter ρ) are good except when the number of points is less than 100. The test T_2 (unknown parameter ρ) performs better than T_1 for small data sets. The comparison of line 5 and 6 in Table 1 shows that T_2 has a bad level if the distances are so small that the corresponding balls have a large probability to be void. The test T_3 is systematically affected by edge effects. This is due to the use of an asymptotic formula for the variance that is not sufficiently accurate even for samples with 500 points.

3.3. Test power against dependence

We investigate the power of the test T_2 against the alternative of dependent point processes. In Table 2, we simulate six Thomas cluster processes [28] and two Hardcore Strauss processes. A Thomas process is a clustered Neyman-Scott process; the germs of the clusters are drawn as a sample of a homogeneous Poisson process of intensity κ. For each germ, an inhomogeneous Poisson process is drawn with intensity measure μf, where f is the density of the Gaussian two-dimensional vector centered on the germ and with independent coordinates of standard error σ. The Thomas process results from the superimposition of these inhomogeneous Poisson processes. The germs are not conserved. Note that this process is homogeneous, with resulting intensity $\rho = \kappa\mu$. Figure 2 presents a sample of these Thomas processes compared to a sample of a Poisson process of the same intensity. It shows that the visual inspection is not sufficient to distinguish between the processes, especially when the number of points is more than 250. The Hardcore Strauss process is a over-dispersed Markov process defined by a density with respect to a homogeneous Poisson process. The density of a point set is equal to zero when two points are at distance less than a constant radius R and constant for other point sets.

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

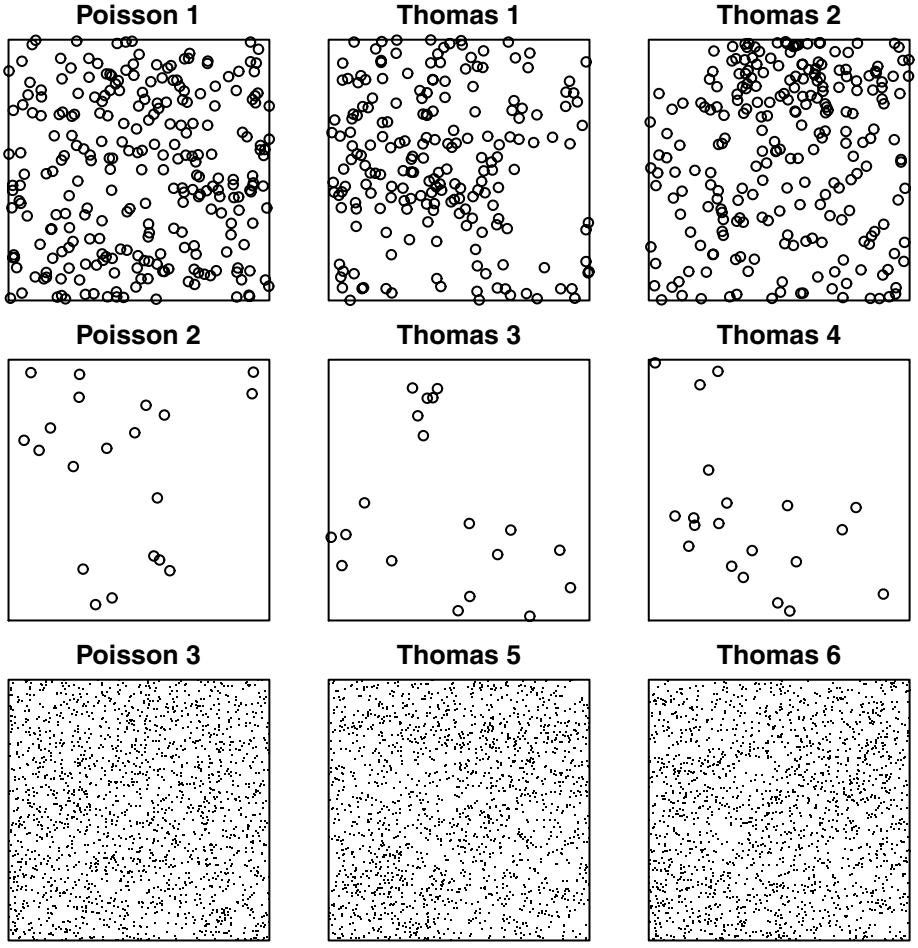


FIGURE 2. Samples of processes of table 2. Expected number of points: first row 250 , second row 20, third row 2000. For the processes Thomas 1 to 6 , the expected size of the clusters are respectively 10, 5, 4, 2, 10 and 4.

We compare test T_2 with Heinrich test T_3 and with the Monte Carlo test Lm based on the uncorrected Besag function $L(r)$ estimated in the band $r \in [0 \quad 2.5]$ (see [11]); the test statistic is $Lm = \max|\hat{L}(r) - \mathbb{E}(\hat{L}(r))|$ computed on this band. The value of $\mathbb{E}(L(r))$ is estimated by a first Monte-Carlo sampling of size 10 000 and the distribution of Lm is estimated by a second Monte-Carlo sampling of the same size. The rejection zone corresponds to the largest values corresponding to 5% of the Monte-Carlo sample.

Edge effects. The first row of Table 2 shows that test T_2 rejects a bit less than Heinrich test T_3; this is mainly due to the incorrect level of the test T_3 as shown in the third line (28.9% rejection instead of 5% expected for the reference Poisson process). This means that edge effects are too strong for T_3 when the ratio $\max(r/n)$ is equal to 0.2. The test T_2 has a correct level, detects perfectly the large clusters of model Thomas 1 and quite well (67%) the clusters in model Thomas 2. It performs better than the Lm test. The second row displays tests with lesser edge effects ($\max(r/n)$ is equal to 0.1), for a sample with very few points (20 points expected). The third line shows that the level of T_2 and T_3 are acceptable even for small samples. The power of the two tests

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

TABLE 2. Percentile of rejection over 10 000 repetitions of the test with level $\alpha = 0.05$ for dependent processes compared to Poisson processes.

$n = 10$				T_2	T_3	Lm
250 points	$r = (0.5, 1, 2)$	Thomas 1	$(\kappa, \mu, \sigma) = (0.25, 10, 1)$	94.8	97.6	89.7
		Thomas 2	$(\kappa, \mu, \sigma) = (0.5, 5, 1)$	67.2	83.3	56.8
		Poisson 1	$\rho = 2.5$	4.9	28.9	5.1
20 points	$r = (0.2, 0.5, 1)$	Thomas 3	$(\kappa, \mu, \sigma) = (0.05, 4, 1)$	60.8	62.5	52.0
		Thomas 4	$(\kappa, \mu, \sigma) = (0.1, 2, 1)$	32.0	33.8	23.0
		Poisson 2	$\rho = 0.2$	6.7	7.5	5.4
2000 points	$r = (0.2, 0.5, 1)$	Thomas 5	$(\kappa, \mu, \sigma) = (2, 10, 1)$	72.8	87.5	24
		Thomas 6	$(\kappa, \mu, \sigma) = (5, 4, 1)$	32.0	63.7	16.9
		Poisson 3	$\rho = 20$	5.0	30.3	5.4
2000 points	$r = (0.5, 1, 3)$	Thomas 6	$(\kappa, \mu, \sigma) = (5, 4, 1)$	42.6		16.9
		Poisson 3	$\rho = 20$	5.3		5.4
2000 points	$r = (1, 2, 5)$	Thomas 6	$(\kappa, \mu, \sigma) = (5, 4, 1)$	46.2		16.9
		Poisson 3	$\rho = 20$	5.1		5.4
2000 points	$r = (1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5)$	Thomas 6	$(\kappa, \mu, \sigma) = (5, 4, 1)$	25.6		16.9
		Poisson 3	$\rho = 20$	5.5		5.4
10 000 points	$r = (1, 2, 5)$	Thomas 7	$(\kappa, \mu, \sigma) = (10, 10, 1)$	69.6		47.5
		Thomas 8	$(\kappa, \mu, \sigma) = (25, 4, 1)$	30.9		24.5
		Poisson 4	$\rho = 100$	5.3		5.1
69 points	$r = (0.1, 0.5, 2)$	Strauss 1	$R = 0.4$	44.2	53.2	100
		Poisson 5	$\rho = 0.69$	6.2	12.5	4.9
74 points	$r = (0.1, 0.5, 2)$	Strauss 2	$R = 0.35$	21.2	29.9	100
		Poisson 6	$\rho = 0.74$	5.4	11.9	4.8

are similar and quite good for a sample of Thomas 3 (clusters of expected size 4); they still detect around 30% of the samples of Thomas 4, that is much closer to a Poisson process (clusters of expected size 2).

Comparison with the Lm estimator. In [11], the authors claim that estimators based on maximum absolute deviation as Lm work better for small samples when edge effects are not corrected by the band correction, that discards the data in the band of width r from the edges of the square of observation. We see in row 1 and 2 that the test T_2 performs better than the Monte-Carlo test Lm, even for small sample. This is not contradictory with the conclusions of [11], as our edge effects correction does not discard data. The simulation is also different because we do not fix the number of points as it was done in [11]. For small samples the variation of the number of points is significative and may explain the different conclusions. We think that for very small sample, test Lm is advantageous because it is very easy to compute with a perfect level by construction. But for medium size sample of 250 points, T_2 is easier to compute, with a perfect level and better power.

Effect of the number of points. In row 3, we study samples with a larger number of points (2000 points expected) in the same space and with the same range of distance. First notice that is not easier to distinguish between Poisson and cluster models when the number of points increases because the clusters are forced to overlap. The performance of T_2 is a bit lower than for 250 points but still acceptable. T_3 has an incorrect level and Lm has a weak power.

Effect of the range of distance. In row 4 and 5, we change the range of distance in T_2 for the same process Thomas 6. We do not compute T_3 because its level is worse than in row 1. We see that the performance of T_2 increases when the range of distance is larger with the best result for the value $r = 5$ corresponding to the maximal ratio $r/n = 1/2$. In row 3, we see that T_2 outperforms the Lm estimator even if they are computed on a similar range of distance.

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

Effect of the number of distances. In row 6, we keep the same process Thomas 6 and the same range of distance but we increase the number of distances. The level of the test is correct, but the power is lower. Notice that increasing the number of distances does not necessarily increase the information (only the jumps in the function $\hat{K}$ are informative), but it is still surprising that the performances decrease so fast. This could be a consequence of instabilities in the computation of the inverse square root of the covariance matrix as its dimension increases.

Larger set of points. In row 7, we study a small sample (100 repetitions) of processes with a larger number of points (10 000 points expected) with the largest range of distance. Comparing the three processes with cluster size equal to 4, (Thomas 3, 6 and 8), we confirm that the power decreases when the number of clusters increases. Clusters of size 10 of Thomas 7 are still well detected. Lm has a low power.

Test of over-dispersion. In row 8 and 9, we study a sample (10 000 repetitions) of hardcore Strauss processes with intensity $\beta = 1$ and hardcore radius equal to 0.4 and 0.35. The Kolmogorov Smirnov test Lm is considered as very powerful in this context [15]. We observe that Lm rejects perfectly the sample where T_2 and T_3 have poor performances. Here again, T_3 seems to have a better power, but its level is not correct as can be seen with Poisson simulation with the corresponding mean number of points.

3.4. Test power against inhomogeneity

In [13], Ho and Chiu propose to test inhomogeneity versus homogeneity with goodness-of-fit tests for the uniform distribution. A general study of those tests is [6]. The advantage is that the distribution is free of edge effects. As our test is also free of edge effects, we investigate here its power against inhomogeneous Poisson processes. Function K being the same for homogeneous and heterogeneous Poisson process, how could the method work for testing homogeneous Poisson versus heterogeneous? Simply because our definition of $\hat{K}$ was adapted to the homogeneous case and has a different distribution under inhomogeneous Poisson assumptions.

We derive our models from those of [9]. In this paper, the authors consider five types of intensity functions $s_i(x)$ for inhomogeneous Poisson processes on the segment $[0, 1]$. We simulate inhomogeneous Poisson processes on the square $[0, 1] \times [0, 1]$ with intensity $100 s_i(x) s_i(y)$. The functions $s_i(x)$ are:

$$
\begin{aligned}
s_1(x) &= (1+\varepsilon)\ \mathrm{I}\{0 \le x < 0.125\} + (1-\varepsilon)\ \mathrm{I}\{0.125 \le x < 0.25\} + \ \mathrm{I}\{0.25 \le x \le 1\} \\
s_2(x) &= \frac{1}{1+1.27\varepsilon}\left(1 + \varepsilon \sum_{j=1}^{11} h_j\ \mathrm{I}\{x < p_j\}\right) \\
s_3(x) &= (1-\varepsilon)\ \mathrm{I}\{0 \le x \le 1\} + \frac{\varepsilon}{0.284}\left(\sum_{j=1}^{11} g_j \left(1 + \frac{|x-p_j|}{w_j}\right)^{-4}\right) \\
s_4(x) &= (1-\varepsilon)\ \mathrm{I}\{0 \le x < 0.75\} + (1+3\varepsilon)\ \mathrm{I}\{0.75 \le x \le 1\} \\
s_5(x) &= (1-\varepsilon) + \varepsilon\beta x^{\beta-1}
\end{aligned}
$$

Parameters p_j, h_j, g_j and w_j are constant parameters defining the different functions (their values are the same than in [9]). Parameter ε corresponds to the strength of heterogeneity within a model. Parameter β modifies the shape of function s_5. Functions s_1, s_2 and s_4 are step functions, function s_3 shows steep pikes and function s_5 is smooth. All functions have integral equal to 1, so that the expected number of points in the samples is 100. The last column corresponds to the level, that is the homogeneous Poisson process with intensity 100.

The powers of the test T_2 are comparable to those of the tests proposed in [9], for the same expected number of points. They are better for models s_2 and s_5, the same for model s_4 and worse for models s_1 and s_3. Notice that the comparison can not be made rigorous, as the Poisson processes in [9] were defined on the line. The range of distance r has been lowered for Heinrich test T_3 to keep the level acceptable (for $r = (0.1, 0.2, 0.5)$ the observed level is 55%). Even then T_3 performs worse than T_2. The Lm test performs better than T_2 for step functions. This may come from the fact that it uses a finer grid of distances r.

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

TABLE 3. Percentile of rejection over 10 000 repetitions of the test with level $\alpha = 0.05$ for 12 inhomogeneous Poisson processes and the reference homogeneous Poisson process.

model	s_1	s_1	s_1	s_2	s_2	s_3	s_3	s_3	s_4	s_4	s_5	s_5	ρ
ε	0.5	0.8	1	0.5	2	0.2	0.4	0.6	0.2	0.4	1	0.6	
β											1.5	2	
T_2, $r = (0.1, 0.2, 0.5)$	18.3	55.8	84.3	86.5	100	20.9	61.7	90.1	67.6	100	85.9	73.0	5.5
T_3, $r = (0.03, 0.05, 0.1)$	13.7	39.3	71.1	76.4	99.5	20.3	67.4	94.1	45.6	99.8	70.7	61	6.8
Lm	28.9	72.6	95.8	88.7	100	18.3	67.1	97.2	29.7	99.7	88.0	73.6	4.6

4. Conclusion

We provide an efficient test of the null hypothesis of a homogeneous Poisson process for point patterns in a square domain, by proposing a new correction of edge effects. Sample correction (for each point of the data) has rarely been questioned since Ripley's original paper, except by authors claiming test statistics with no correction as more powerful (see [1,11]). Instead of correcting on each sample to reduce or cancel the bias, we compute the exact bias, so that we avoid to increase of the variance by discarding some of the observed points. The resulting test is efficient on samples with a few dozens of points as encountered in actual data sets.

This is a theoretical and practical improvement on Monte-Carlo methods as it is quicker and often more powerful. Monte-Carlo simulation of the distribution is a good method for small samples but becomes tedious when the number of points increases. Marcon and Puech [17] computed K for a 36,000-point data set (the largest ever published as far as we know), but had to limit the number of simulations to 20. With a personal computer, calculating the distribution of Lm with 10 000 simulations of a 10 000-points set is 2 days long. But it takes approximatively 3 minutes to compute T_2 for three distances with optimized C++ code and 5 minutes with a R routine [19].

Our work should be extended in two directions: to other domain shapes that are of interest for the practitioners and to 3-dimensional data for high resolution medical imaging. A further study of the asymptotics of the distribution of $\hat{K}(r)$ for dependent point process models such as Markov or Cox processes should also be achieved to inform on the power of the test.

5. Proofs

5.1. Proof of Proposition 2.1

Let U and V be two independent uniform variables on A_n. The expectations of the Ripley statistics are

$$\mathbb{E}\widehat{K}_{1,n}(r) = \frac{1}{n^2\rho^2}\mathbb{E}\left(\sum_{X_i\neq X_j\in S} \mathrm{I}\{d(X_i,X_j)\leq r\}\right) = \frac{\mathbb{E}\left(N(N-1)\right)}{n^2\rho^2}\mathbb{E}(\,\mathrm{I}\{d(U,V)\leq r\}) = n^2 e_{r,n}.$$

$$\begin{aligned}\mathbb{E}\widehat{K}_{2,n}(r) &= n^2\mathbb{E}\left(\frac{1}{N(N-1)}\sum_{X_i\neq X_j\in S} \mathrm{I}\{d(X_i,X_j)\leq r\}\right) = n^2\mathbb{P}\left(N>1\right)\mathbb{E}(\,\mathrm{I}\{d(U,V)\leq r\})\\ &= n^2\left(1-(1+\rho n^2)\mathrm{e}^{-\rho n^2}\right)e_{r,n}.\end{aligned}$$

The following lemma allows to conclude:

Lemma 5.1.

$$e_{r,n} = \frac{\pi r^2}{n^2} - \frac{8r^3}{3n^3} + \frac{r^4}{2n^4}.$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

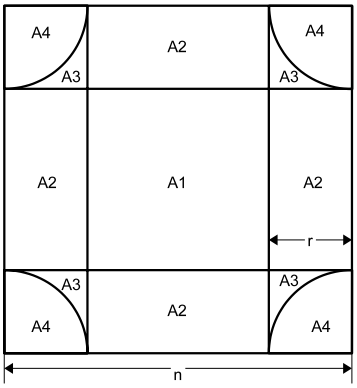


FIGURE 3. Zones in the square.

Proof. We split A_n into four parts to compute $e_{r,n}$:

$$e_{r,n} = \int_{\xi\in A_n^1}\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\le r\}\frac{1}{n^4}\mathrm{d}\xi\mathrm{d}\eta \tag{5.1}$$

$$+ \int_{\xi\in A_n^2}\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\le r\}\frac{1}{n^4}\mathrm{d}\xi\mathrm{d}\eta \tag{5.2}$$

$$+ \int_{\xi\in A_n^3}\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\le r\}\frac{1}{n^4}\mathrm{d}\xi\mathrm{d}\eta \tag{5.3}$$

$$+ \int_{\xi\in A_n^4}\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\le r\}\frac{1}{n^4}\mathrm{d}\xi\mathrm{d}\eta \tag{5.4}$$

where (see Fig. 2)

- (interior) $A_n^1 = \{\xi,\ \xi$ is at distance larger than r from the boundary$\}$
- (one edge) $A_n^2 = \{\xi,\ \xi$ is at distance less than r from an edge, larger than r from the others$\}$
- (two edges) $A_n^3 = \{\xi,\ \xi$ is at distance less than r from two edges and larger than r from the corner$\}$
- (corner) $A_n^4 = \{\xi,\ \xi$ is at distance less than r from the corner$\}$

Note that A_n^2, A_n^3 and A_n^4 are composed of four parts that contribute identically. We establish formulas only for one of these parts.

Lemma 5.2. *Define function* $g(x) = \arccos(x) - x\sqrt{1-x^2}$.

If $\xi \in A_n^1$,

$$\int_{\eta\in A_n} I\{d(\xi,\eta)\le r\}\mathrm{d}\eta = \pi r^2.$$

If $\xi \in A_n^2$, *with* $n - r < \xi_1 < n$, $x_1 = \frac{1}{r}(n-\xi_1)$,

$$\int_{\eta\in A_n} I\{d(\xi,\eta)\le r\}\mathrm{d}\eta = r^2(\pi - g(x_1))$$

If $\xi \in A_n^3$, *with* $n - r < \xi_1 < n$, $n - r < \xi_2 < n$ *and* $(x_1, x_2) = \frac{1}{r}(n-\xi_1, n-\xi_2)$,

$$\int_{\eta\in A_n} I\{d(\xi,\eta)\le r\}\mathrm{d}\eta = r^2(\pi - g(x_1) - g(x_2)).$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

If $\xi \in A_n^4$, *with* $n-r<\xi_1<n$, $n-r<\xi_2<n$ *and* $(x_1,x_2)=\frac{1}{r}(n-\xi_1,n-\xi_2)$,

$$\int_{\eta\in A_n} I\{d(\xi,\eta)\le r\}\mathrm{d}\eta = r^2\left(\frac{3\pi}{4}+x_1x_2-\frac{g(x_1)+g(x_2)}{2}\right).$$

Note. For $0\le x\le 1$, function $g(x)$ is the area of the intersection of a ball of radius 1 with a half plane, when the center of the ball lies outside the half plane at a distance x from its boundary.

Proof. Let $B(\xi,r)$ denote the ball of center ξ and radius r. For the interior points $\xi\in A_n^1$, $B(\xi,r)\subset A_n$.

Let $\xi\in A_n^2$. We compute the area of $B(\xi,r)\cap A_n$.

$$\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\le r\}\mathrm{d}\eta = \frac{\pi r^2}{2}+2r^2\int_0^{x_1}\sqrt{1-t^2}dt = r^2\left(\pi-\arccos(x_1)+x_1\sqrt{1-x_1^2}\right)=r^2\left(\pi-g(x_1)\right).$$

Note that $r^2g(x)$ is the part of the ball that lies out of the square A_n if the center is at distance xr from the edge of the square.

Let $\xi\in A_n^3$. Here the ball intersects two edges of the square and the area of $B(\xi,r)\cap A_n$ is

$$\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\le r\}\mathrm{d}\eta = r^2\left(\pi-g(x_1)-g(x_2)\right).$$

Let $\xi\in A_n^4$. Divide the ball into four quarters along axes parallel to the coordinate axes. One of the quarter is inside the square, two intersect the edges, leaving outside an area equal to $(g(x_1)+g(x_2))/2$. The area of the intersection of the last quarter with the square is x_1x_2 so that the area of $B(\xi,r)\cap A_n$ is

$$\int_{\eta\in A_n} \mathrm{I}\{d(\xi,\eta)\le r\}\mathrm{d}\eta = r^2\left(\frac{3\pi}{4}+x_1x_2-\frac{g(x_1)+g(x_2)}{2}\right). \qquad \square$$

Proof of Lemma 5.1(continued). The left-hand side of (5.1) is $m(A_n^1)\pi r^2=\pi(n-2r)^2r^2$. Recall that A_n^2 is composed of four parts that contribute identically. Using the integration formula of the arccos function, we get the integral of g:

$$G(x)=\int_0^x g(u)du = x\arccos(x)-\sqrt{1-x^2}+\frac{1}{3}(1-x^2)^{3/2}+\frac{2}{3}.$$

Then the contribution (5.2) is equal to

$$4r\int_r^{n-r}\mathrm{d}\xi_2\int_0^1 r^2(\pi-g(x))\mathrm{d}x = 4r^3(n-2r)(\pi-G(1)) = \left(4\pi-\frac{8}{3}\right)r^3(n-2r).$$

We consider A_n^3; the domain of integration is symmetric in (x_1,x_2) so that the contribution (5.3) is equal to

$$4r^4\int_0^1\mathrm{d}x_1\int_{\sqrt{1-x_1^2}}^1(\pi-2g(x_1))\mathrm{d}x_2 = r^4\left(4\pi\left(1-\frac{\pi}{4}\right)-8G(1)+8\int_0^1 g(x_1)\sqrt{1-x_1^2}\mathrm{d}x_1\right).$$

But $\displaystyle\int_0^1 g(x_1)\sqrt{1-x_1^2}\mathrm{d}x_1=\frac{\pi^2}{16}$, so that contribution (5.3) is equal to $\displaystyle r^4\left(4\pi-\frac{\pi^2}{2}-\frac{16}{3}\right)$.

We consider A_n^4; the contribution (5.4) is equal to

$$4r^4\int_0^1\mathrm{d}x_1\int_0^{\sqrt{1-x_1^2}}\left(\frac{3\pi}{4}+x_1x_2-g(x_1)\right)\mathrm{d}x_2 = r^4\left(\frac{3\pi^2}{4}+\frac{1}{2}-4\int_0^1 g(x_1)\sqrt{1-x_1^2}\mathrm{d}x_1\right)=r^4\left(\frac{\pi^2}{2}+\frac{1}{2}\right).$$

Gathering the four contributions, we get

$$e_{r,n}=\frac{r^2}{n^2}\left(\pi\left(1-\frac{2r}{n}\right)^2+\left(4\pi-\frac{8}{3}\right)\frac{r}{n}\left(1-\frac{2r}{n}\right)+\left(4\pi-\frac{29}{6}\right)\frac{r^2}{n^2}\right)=\frac{r^2}{n^2}\left(\pi-\frac{8}{3}\frac{r}{n}+\frac{1}{2}\frac{r^2}{n^2}\right). \qquad \square$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

5.2. Proof of Proposition 2.2

For $u = 1$ or 2, we decompose the variance of $K_{u,A_n}(r)$ by conditioning the variable with respect to the number N of points in the sample. Conditionally to N, $K_{u,A_n}(r)$ has the form of a U-statistic. Then we apply the Höffding decomposition to this U-statistic. We use the relation

$$\operatorname{var}(\widehat{K}_{u,A_n}(r)) = \operatorname{var}\mathbb{E}(\widehat{K}_{u,A_n}(r)|N) + \mathbb{E}\operatorname{var}(\widehat{K}_{u,A_n}(r)|N).$$

We first consider the conditional expectation of $\widehat{K}_{u,A_n}(r)$.

$$\mathbb{E}(\widehat{K}_{1,n}(r)|N) = \frac{1}{n^2\rho^2}\left(\sum_{i\neq j=1}^{N} \mathbb{E}\ \mathrm{I}\{d(X_i,X_j)\le r\}\right) = \frac{N(N-1)e_{r,n}}{n^2\rho^2},$$

$$\mathbb{E}(\widehat{K}_{2,n}(r)|N) = \frac{n^2}{N(N-1)}\sum_{i\neq j=1}^{N} \mathbb{E}\ \mathrm{I}\{d(U_i,U_j)\le r\} = n^2 e_{r,n}\ \mathrm{I}\{N>1\}.$$

Because N is a Poisson variable with intensity ρn^2

$$\begin{aligned}\mathbb{E}N^2(N-1)^2 &\quad \rho^4n^8 + 4\rho^3n^6 + 2\rho^2n^4.\\ \operatorname{var}N(N-1) &= 4\rho^3n^6 + 2\rho^2n^4.\end{aligned} \tag{5.5}$$

Then

$$\operatorname{var}\mathbb{E}(\widehat{K}_{1,n}(r)|N) = \frac{(4\rho n^2+2)\mathrm{e}_{r,n}^2}{\rho^2}\,. \tag{5.6}$$

$$\operatorname{var}\mathbb{E}(\widehat{K}_{2,n}(r)|N) = n^4\mathbb{P}\{N>1\}(1-\mathbb{P}\{N>1\})\mathrm{e}_{r,n}^2 = n^4\mathrm{e}^{-\rho n^2}(1+\rho n^2)\left(1-\mathrm{e}^{-\rho n^2}\left(1+\rho n^2\right)\right)\mathrm{e}_{r,n}^2. \tag{5.7}$$

We compute the conditional variances.

$$\operatorname{var}(\widehat{K}_{1,n}(r)|N) = \frac{1}{n^4\rho^4}\operatorname{var}\left(\sum_{i\neq j=1}^{N} h(X_i,X_j,r)\right),$$

$$\operatorname{var}(\widehat{K}_{2,n}(r)|N) = \frac{n^4}{N^2(N-1)^2}\operatorname{var}\left(\sum_{i\neq j=1}^{N} h(X_i,X_j,r)\right).$$

Conditionally to N, the locations of the points are independent and uniformly distributed variables U_i over A_n. We introduce the Höffding decomposition of the U-statistic kernel h:

$$h(x,y,r) = h_1(x,r) + h_1(y,r) + h_2(x,y,r),$$

where $h_1(x) = \mathbb{E}(h(U,V,r)|V=x)$, (U,V) being two independent uniform random variables on A_n.

Then $\mathbb{E}h_1(U,r) = 0$ and $\mathbb{E}(h_2(U,V,r)|U) = \mathbb{E}(h_2(U,V,r)|V) = 0$, so that

$$\operatorname{var}h(U,V,r) = \operatorname{var}h_1(U,r) + \operatorname{var}h_1(V,r) + \operatorname{var}h_2(U,V,r) = 2\mathbb{E}h_1^2(U,r) + \operatorname{var}h_2(U,V,r).$$

From

$$\sum_{i\neq j=1}^{N} h(U_i,U_j,r) = 2(N-1)\sum_{i=1}^{N} h_1(U_i,r) + \sum_{i\neq j=1}^{N} h_2(U_i,U_j,r).$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

we get

$$\begin{aligned}\operatorname{var}(\widehat{K}_{1,n}(r)|N) &= \frac{4(N-1)^2}{n^4\rho^4}\operatorname{var}\left(\sum_{i=1}^{N} h_1(U_i,r)\right) + \frac{1}{n^4\rho^4}\operatorname{var}\left(\sum_{i\neq j=1}^{N} h_2(U_i,U_j,r)\right)\\ &= \frac{4N(N-1)^2}{n^4\rho^4}\mathbb{E}h_1^2(U,r) + \frac{2}{n^4\rho^4}\sum_{i\neq j=1}^{N}\operatorname{var} h_2(U_i,U_j,r)\\ &= \frac{4N(N-1)^2}{n^4\rho^4}\mathbb{E}h_1^2(U,r) + \frac{2N(N-1)}{n^4\rho^4}(\operatorname{var} h(U,V,r) - 2\mathbb{E}h_1^2(U,r))\\ &= \frac{4N(N-1)(N-2)}{n^4\rho^4}\mathbb{E}h_1^2(U,r) + \frac{2N(N-1)}{n^4\rho^4}\operatorname{var} h(U,V,r),\end{aligned}$$

Note that the factor 2 in the second line may be surprising in the variance of a sum of independent variables, but each variance term appears four times in the expansion of the variance of the sum over $i \neq j$. Now $\operatorname{var} h(U,V,r) = e_{r,n} - e_{r,n}^2$ and using factorial moments of the Poisson distribution

$$\mathbb{E}\operatorname{var}(\widehat{K}_{1,n}(r)|N) = \frac{4n^2}{\rho}\mathbb{E}h_1^2(U,r) + \frac{2}{\rho^2}\left(e_{r,n} - e_{r,n}^2\right). \tag{5.8}$$

Lemma 5.3 gives the exact value of $\mathbb{E}h_1^2(U,r)$. Its proof is postponed at the end of the paper.

Lemma 5.3.

$$\mathbb{E}h_1^2(U,r) = \frac{r^5}{n^5}\left(\frac{8}{3}\pi - \frac{256}{45}\right) + \frac{r^6}{n^6}\left(\frac{11}{48}\pi - \frac{56}{9}\right) + \frac{8}{3}\frac{r^7}{n^7} - \frac{1}{4}\frac{r^8}{n^8}.$$

With relations (5.6) and (5.8), we get

$$\begin{aligned}\operatorname{var}(\widehat{K}_{1,n}(r)) &= \frac{2e_{r,n}}{\rho^2} + \frac{4n^2\mathrm{e}_{r,n}^2}{\rho} + \frac{4n^2}{\rho}\mathbb{E}h_1^2(U_j,r)\\ &= \frac{1}{n^2}\left(\frac{2\pi r^2}{\rho^2} + \frac{4\pi^2 r^4}{\rho}\right) - \frac{1}{n^3}\left(\frac{16}{3}\frac{r^3}{\rho^2} + \left(\frac{32\pi}{3} + \frac{1024}{45}\right)\frac{r^5}{\rho}\right) + \frac{1}{n^4}\left(\frac{r^4}{\rho^2} + \left(\frac{59\pi}{12} + \frac{32}{9}\right)\frac{r^6}{\rho}\right).\end{aligned}$$

Similarly

$$\begin{aligned}\operatorname{var}(\widehat{K}_{2,n}(r)|N) &= \frac{4n^4\ \mathrm{I}\{N>1\}(N-2)}{N(N-1)}\mathbb{E}h_1^2(U,r) + \frac{2n^4\ \ \mathrm{I}\{N>1\}}{N(N-1)}\operatorname{var} h(U,V,r),\\ \mathbb{E}\operatorname{var}(\widehat{K}_{2,n}(r)|N) &= 4n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}(N-2)}{N(N-1)}\right)\mathbb{E}h_1^2(U,r) + 2n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}}{N(N-1)}\right)\left(e_{r,n} - e_{r,n}^2\right).\end{aligned}$$

From this and relation (5.7), we get

$$\begin{aligned}\operatorname{var}(\widehat{K}_{2,n}(r)) = {}& 2n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}}{N(N-1)}\right)\left(e_{r,n} - e_{r,n}^2\right) + 4n^4\mathbb{E}\left(\frac{\mathrm{I}\{N>1\}(N-2)}{N(N-1)}\right)\mathbb{E}h_1^2(U_j,r)\\ &+ n^4\mathrm{e}^{-\rho n^2}\left(1+\rho n^2\right)\left(1 - \mathrm{e}^{-\rho n^2} - \rho n^2\mathrm{e}^{-\rho n^2}\right)\mathrm{e}_{r,n}^2.\end{aligned}$$

We now apply the same decomposition to $\operatorname{cov}(\widehat{K}_{1,n}(r), \widehat{K}_{1,n}(r'))$,

$$\operatorname{cov}(\mathbb{E}(\widehat{K}_{1,n}(r')|N), \mathbb{E}(\widehat{K}_{1,n}(r)|N)) = \frac{(4\rho n^2+2)e_{r',n}e_{r,n}}{\rho^2}. \tag{5.9}$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

$$\begin{aligned}\operatorname{cov}(\widehat{K}_{1,n}(r'),\widehat{K}_{1,n}(r)|N) &= \frac{4(N-1)^2}{n^4\rho^4}\operatorname{cov}\left(\sum_{i=1}^{N} h_1(U_i,r'),\sum_{i=1}^{N} h_1(U_i,r)\right)\\ &\quad+\frac{1}{n^4\rho^4}\operatorname{cov}\left(\sum_{i\neq j=1}^{N} h_2(U_i,U_j,r'),\sum_{i\neq j=1}^{N} h_2(U_i,U_j,r)\right)\\ &= \frac{4N(N-1)(N-2)}{n^4\rho^4}\operatorname{cov}(h_1(U,r'),h_1(U,r))\\ &\quad+\frac{2N(N-1)}{n^4\rho^4}\operatorname{cov}(h(U,V,r'),h(U,V,r)).\end{aligned}$$

$$\mathbb{E}\operatorname{cov}(\widehat{K}_{1,n}(r'),\widehat{K}_{1,n}(r)|N) = \frac{4n^2}{\rho}\operatorname{cov}(h_1(U,r'),h_1(U,r)) + \frac{2}{\rho^2}(e_{r,n}-e_{r',n}e_{r,n})$$

To compute $\operatorname{cov}(h_1(U,r'),h_1(U,r))$, the square A_n should now be split into 16 different zones according to the 4 zones of the preceding section with respect to r and the 4 zones with respect to r'. Because of inclusions, the actual number of zones to consider reduces to 9. The corresponding computation is easy in the center zone, but can not be achieved in a close form in the edge bands and in the corner. We consider the following zones:

- (interior) $A_n^{1,1}=\{\xi,\ \xi$ is at distance larger than r' from the boundary$\}$,
- (interior-edge) $A_n^{1,2}=\{\xi,\ \xi$ is at distance between r and r' from an edge, larger than r' from the others$\}$,
- (edge) $A_n^{2,2}=\{\xi,\ \xi$ is at distance less than r from an edge, larger than r' from the others$\}$,
- (corner) $A_n^{3,3}=\{\xi,\ \xi$ is at distance less than r' from two edges$\}$.

Denoting $x_1=\frac{1}{r}(n-\xi_1)$ and $x_1'=\frac{1}{r'}(n-\xi_1)$ we get

$$\begin{aligned}h_1(X_j,r')h_1(X_j,r) &= \left(\frac{\pi r'^2}{n^2}-e_{r',n}\right)\left(\frac{\pi r^2}{n^2}-e_{r,n}\right)\ \text{on } A_n^{1,1},\\ &= \left(\frac{\pi r'^2}{n^2}-e_{r',n}-\frac{r'^2}{n^2}g(x_1')\right)\left(\frac{\pi r^2}{n^2}-e_{r,n}\right)\ \text{on } A_n^{1,2},\\ &= \left(\frac{\pi r'^2}{n^2}-e_{r',n}-\frac{r'^2}{n^2}g(x_1')\right)\left(\frac{\pi r^2}{n^2}-e_{r,n}-\frac{r^2}{n^2}g(x_1)\right)\ \text{on } A_n^{2,2}.\end{aligned}$$

Denote $b_{r,n}=\left(\pi-\frac{n^2}{r^2}e_{r,n}\right)=\frac{8r}{3n}-\frac{r^2}{2n^2}$.

$$\operatorname{cov}(h_1(X_j,r'),h_1(X_j,r)) = C(A_n^{1,1})+C(A_n^{1,2})+C(A_n^{2,2})+C(A_n^{3,3})$$

$$\begin{aligned}C(A_n^{1,1}) &= \frac{r'^2r^2}{n^4}\left(1-\frac{2r'}{n}\right)^2 b_{r',n}b_{r,n}\\ C(A_n^{1,2}) &= 4\left(1-\frac{2r'}{n}\right)\frac{r'^3r^2}{n^5}b_{r,n}\int_{r/r'}^{1}(b_{r',n}-g(x_1'))\mathrm{d}x_1'\\ C(A_n^{2,2}) &= 4\left(1-\frac{2r'}{n}\right)\frac{r^3r'^2}{n^5}\int_0^1(b_{r',n}-g(rx_1/r'))(b_{r,n}-g(x_1))\mathrm{d}x_1.\end{aligned}$$

The first integral may be expressed in terms of function G, the second integral is elliptic and has to be numerically evaluated; as the integrand is bounded and very smooth this can be achieved without difficulties. To compute

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

the term $C(A_n^{3,3})$, we rewrite the different values of function h_1 with the help of indicator functions:

$$h_{A1}(x,r) = b_{r,n}\ \mathrm{I}\{x_1 \geq 1; x_2 \geq 1\}$$
$$h_{A2}(x,r) = (b_{r,n} - g(x_2))\ \mathrm{I}\{x_1 \geq 1;\ x_2 < 1\} + (b_{r,n} - g(x_1))\ \mathrm{I}\{x_2 \geq 1;\ x_1 < 1\}$$
$$h_{A3}(x,r) = (b_{r,n} - g(x_1) - g(x_2))\ \mathrm{I}\{x_1 < 1;\ x_2 < 1;\ x_1^2 + x_2^2 \geq 1\}$$
$$h_{A4}(x,r) = (b_{r,n} - \pi/4 + x_1 x_2 - (g(x_1) + g(x_2))/2)\ \mathrm{I}\{x_1^2 + x_2^2 < 1\}$$

For $x' = \frac{1}{r'}(n - \xi_1, n - \xi_2)$, $C(A_n^{3,3}) = 4\dfrac{r^2 r'^4}{n^6} \displaystyle\int_0^1 \int_0^1 \sum_{i=1}^{4} h_{Ai}(r'x'/r, r) \times \sum_{i=3}^{4} h_{Ai}(x', r')\mathrm{d}x_1'\mathrm{d}x_2'$

and this integral also can be numerically evaluated.

Note. The whole computation of this term of the covariance could be numerically achieved, but we gain some useful precision with an exact computation whenever it is possible.

The case of the covariance of $K_{2,n}(r)$ is analogous:

$$\mathrm{cov}\,(\mathbb{E}(\widehat{K}_{2,n}(r')|N), \mathbb{E}(\widehat{K}_{2,n}(r)|N)) = n^4 \mathrm{e}^{-\rho n^2}\left(1 + \rho n^2\right)(1 - \mathrm{e}^{-\rho n^2}(1 + \rho n^2)) e_{r',n} e_{r,n}.$$

$$\begin{aligned}\mathbb{E}\ \mathrm{cov}\,(\widehat{K}_{2,n}(r'), \widehat{K}_{2,n}(r)|N) = {} & 4n^4 \mathbb{E}\left(\frac{\mathrm{I}\{N > 1\}(N-2)}{N(N-1)}\right) \mathrm{cov}\,(h_1(U, r'), h_1(U, r)) \\ & + 2n^4 \mathbb{E}\left(\frac{\mathrm{I}\{N > 1\}}{N(N-1)}\right)(e_{r,n} - e_{r',n} e_{r,n})\,.\end{aligned}$$

5.3. Proof of Theorem 2.3

We show that any linear combination of the $K_{1,n}(r_t)$ is asymptotically normal. Let $\Lambda = (\lambda_1, \ldots \lambda_d)$ be a vector of real coefficients. Define $Z_1 = \sum_{t=1}^{d} \lambda_t K_{1,n}(r_t)$. We use the Bernstein blocks technique [3]: we divide the square A_n into squares of side p with $p = o(n)$. These squares are separated by gaps of width $2r_d$ so that the sums over couples of points in each square are independent. The couples of points with at least one point in the gaps give a negligible contribution, so that the statistic Z_1 is equivalent to a sum of independent variables and asymptotically normal.

Set $p = n^{1/4}$. Assume that the Euclidean division of n by $(p + 2r_d)$ gives a quotient a and a remainder q. For $l = 0, \ldots, a$, we define the segment $I_l = [(p + 2r_d)l, (p + 2r_d)l + p - 1]$. We order the set $\{0, \ldots, a\}^2$ by the lexicographic order. To any integer i such that $1 \leq i \leq k = (a+1)^2$, corresponds an element (j_1, j_2) of this set; we define the block $P_{i,n} = I_{j_l} \times I_{j_2}$ and $Q = A_n \backslash \cup_i P_{i,n}$ the set of points that are in none of the $P_{i,n}$'s. For each block $P_{i,n}$ and Q, we define the partial sums:

$$u_{i,n} = \frac{1}{n\rho^{3/2}} \sum_{X_l \neq X_m \in P_{i,n}} \sum_{t=1}^{d} \lambda_t\ \ \mathrm{I}\{d(X_l, X_m) \leq r_t\},$$

$$v_{i,n} = \frac{1}{n\rho^{3/2}} \sum_{X_l \in P_{i,n}, X_m \in Q} \sum_{t=1}^{d} \lambda_t\ \ \mathrm{I}\{d(X_l, X_m) \leq r_t\}$$

$$w_n = \frac{1}{n\rho^{3/2}} \sum_{X_l \neq X_m \in Q} \sum_{t=1}^{d} \lambda_t\ \ \mathrm{I}\{d(X_l, X_m) \leq r_t\}.$$

then

$$n\sqrt{\rho}(Z_1 - \mathbb{E}Z_1) = \sum_{i=1}^{k}(u_{i,n} - \mathbb{E}u_{i,n}) + \sum_{i=1}^{k}(v_{i,n} - \mathbb{E}v_{i,n}) + w_n - \mathbb{E}w_n,$$

We show that the sum of the $u_{i,n}$ converges in distribution to a Gaussian variable and that the other term are negligible in L^2. We check the conditions of the following CLT adapted from [2].

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

Theorem 5.4. *Let $(z_{i,n})_{0\leq i\leq k(n)}$ be an array of random variables satisfying*

1. *There exists $\delta > 0$ such that $\sum_{i=0}^{k(n)} \mathbb{E}|z_{i,n}|^{2+\delta}$ tends to 0 as n tends to infinity,*
2. *$\sum_{i=0}^{k(n)} \operatorname{var} z_{i,n}$ tends to σ^2 as n tends to infinity,*

then $\sum_{i=0}^{k(n)} z_{i,n}$ tends in distribution to $\mathcal{N}(0,\sigma^2)$ as n tends to infinity.

To check Condition 1, we compute the fourth order moment of $u_{i,n} - \mathbb{E}u_{i,n}$. Let N_i be the number of points of S that fall in $P_{i,n}$. Denote $f(x,y) = \sum_{t=1}^{d} \lambda_t \left(\mathrm{I}\{d(x,y) \leq r_t\} - e_{r,p}\right) = \sum_{t=1}^{d} \lambda_t \; h(x,y,r_t)$, then

$$\mathbb{E}((u_{i,n} - \mathbb{E}u_{i,n})^4 | N_i) = \frac{1}{n^4\rho^6}\mathbb{E}\left(\sum_{l\neq m=1}^{N_i} f(U_l, U_m)\right)^4$$

Denote f_1 and f_2 the decomposing functions of f: $\mathbb{E}(f_1(U_l)) = 0$, $\mathbb{E}(f_1(U_l)f_2(U_l,U_m)) = \mathbb{E}(f_1(U_m)f_2(U_l,U_m)) = 0$, for U_l and U_m two independent uniform variables on $P_{i,n}$.

$$\sum_{l\neq m=1}^{N_i} f(U_l,U_m) = 2(N_i-1)\sum_{l=1}^{N_i} f_1(U_l) + \sum_{l\neq m=1}^{N_i} f_2(U_l,U_m).$$

Lemma 5.5. *f_1 is bounded by Cp^{-2}. f_2 is bounded by a constant and $f_2(x,y) \leq Cp^{-2}$ as soon as $\|x-y\| > r$.*

Proof. All the quantities computed in Lemma 5.2 for the four different cases are bounded by a constant so that $\mathbb{E}(\{\mathrm{I}\{d(U_1,U_2)\leq r\}|U_1) = O(p^{-2})$. As $e_{r,p} = O(p^{-2})$, this is also true for $h_1(x,r)$ for any r and then for f_1. Because $f_2(x,y) = f(x,y) - f_1(x) - f_1(y)$, it is bounded by a constant and $f_2(x,y) = O(p^{-2})$ as soon as the indicator function vanishes.

Define $M_1 = \mathbb{E}\left(\sum_{l=1}^{N_i} f_1(U_l)\right)^4$. Then $M_1 = N_i E(f_1^4(U)) + 6N_i(N_i-1)E(f_1^2(U))^2$ and $\mathbb{E}(N_i-1)^4 M_1 = O(p^2)$. Define $M_2 = \mathbb{E}\left(\sum_{l\neq m=1}^{N_i} f_2(U_l,U_m)\right)^4$. Because f_2 is zero mean with respect to one coordinate, only the products where variables appear at least two times contribute.

$$\begin{aligned}
M_2 = {} & 8\sum_{l\neq m=1}^{N_i} \mathbb{E}f_2^4(U_l,U_m) + 48\sum_{l\neq m\neq u=1}^{N_i} \mathbb{E}f_2^2(U_l,U_u)f_2^2(U_u,U_m) \\
& + 96\sum_{l\neq m\neq u=1}^{N_i} \mathbb{E}f_2^2(U_l,U_m)f_2(U_m,U_u)f_2(U_u,U_l) \\
& + 12\sum_{l\neq m\neq u\neq v=1}^{N_i} \mathbb{E}f_2^2(U_l,U_m)f_2^2(U_u,U_v) \\
& + 48\sum_{l\neq m\neq u\neq v=1}^{N_i} \mathbb{E}f_2(U_l,U_m)f_2(U_m,U_u)f_2(U_u,U_v)f_2(U_v,U_l).
\end{aligned}$$

Consider the first sum

$$\sum_{l\neq m=1}^{N_i} \mathbb{E}f_2^4(U_l,U_m) \leq \sum_{l\neq m=1}^{N_i} \mathbb{P}\{d(U_l,U_m)\leq r\} + C\mathbb{P}\{d(U_l,U_m) > r\}p^{-8} \leq CN_i(N_i-1)p^{-2}.$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

In all the sums, the main term comes from sets of points with all interdistance less than r and the resulting magnitude of the expectation is $O(p^2)$, so that

$$\sum_{i=0}^{k} \mathbb{E}(u_{i,n} - \mathbb{E}u_{i,n})^4 = O(n^{-2}).$$

Thus condition 1 is realised. To check condition 2, note that the vector $(K_{1,P_i}(r_1), \ldots, K_{1,P_i}(r_d))$ has a covariance matrix Σ_p defined by Proposition 2.2 by substituting p to n in the expressions. The $u_{i,n} = \frac{p^2\sqrt{\rho}}{n} \sum_{t=1}^{d} \lambda_t (K_{1,P_i}(r_t) - \mathbb{E}K_{1,P_i}(r_t))$ are i.i.d variables with variance equal to $\frac{p^4\rho}{n^2} \Lambda^t \Sigma_p \Lambda$. But $p^2\rho\Sigma_p$ tends to Σ as p tends to infinity and

$$\sum_{i=0}^{k} \operatorname{var} u_{i,n} = \frac{kp^4\rho}{n^2} \Lambda^t \Sigma_p \Lambda \longrightarrow \Lambda^t \Sigma \Lambda$$

so that $\sum_{i=1}^{k} u_{i,n}$ tends in distribution to $\mathcal{N}(0, \Lambda^t \Sigma \Lambda)$.

Note that the $v_{i,n}$ are k independent variables. Denote N_{i,r_d} the number of points X_l in the boundary region P_{i,r_d} of $P_{i,n}$ such that the ball $B(X_l, r_d)$ intersects Q and let $D(X_l)$ denote this intersection. Note that

$$\mathbb{E}N_{i,r_d} = \rho m(P_{i,r_d}) \leq Cpr_d.$$

$$\operatorname{var} v_{i,n} \leq \frac{C}{n^2} \mathbb{E} \left(\sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_Q} \mathrm{I}\{X_m \in D(X_l)\} \right)^2 \leq \frac{C}{n^2}(T_1 + T_2),$$

where

$$T_1 = \mathbb{E} \sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_Q} \sum_{u=1}^{N_Q} \mathrm{I}\{X_m \in D(X_l)\}\, \mathrm{I}\{X_u \in D(X_l)\}$$

$$T_2 = \mathbb{E} \sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_{i,r_d}} \sum_{u=1}^{N_Q} \mathrm{I}\{X_u \in D(X_l) \cap D(X_m)\}.$$

$$T_1 \leq \mathbb{E}N_{i,r_d} \mathbb{E}N_Q^2 \mathbb{P}^2\{X_m \in D(X_l) | X_m \in Q\} \leq \rho^3 m(P_{i,r_d})(m^2(Q) + m(Q)) \left(\frac{\pi r_d^2}{2m(Q)} \right)^2 = O(p).$$

$$\begin{aligned} T_2 &= \mathbb{E} \sum_{l=1}^{N_{i,r_d}} \sum_{m=1}^{N_{i,r_d}} \sum_{u=1}^{N_Q} \mathrm{I}\{X_m \in B(X_l, 2r_d)\}\, \mathrm{I}\{X_u \in D(X_l) \cap D(X_m)\} \\ &\leq \mathbb{E}N_{i,r_d}^2 \mathbb{P}\{X_m \in B(X_l, r_d) | X_m \in P_{i,r_d}\} \mathbb{E}N_Q \mathbb{P}\{X_u \in D(X_l) | X_u \in Q\} \\ &\leq \rho^3 (m^2(P_{i,r_d}) + m(P_{i,r_d})) \left(\frac{\pi r_d^2}{m(P_{i,r_d})} \right) m(Q) \left(\frac{\pi r_d^2}{2m(Q)} \right) = O(p) \end{aligned}$$

and $\operatorname{var}\left(\sum_{i=1}^{k} v_{i,n}\right) = O\left(kp/n^2\right) = O\left(p^{-1}\right)$, so that this sum is negligible in L^2. Similarly

$$\operatorname{var}(w_n) \leq \frac{C}{n^2} \mathbb{E} \left(\sum_{l \neq m=1}^{N_Q} \mathrm{I}\{X_m \in B(X_l, r_d)\} \right)^2 \leq \frac{C}{n^2}(T_1 + T_2),$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

where

$$T_1 = \mathbb{E}\sum_{l=1}^{N_Q}\sum_{m=1}^{N_Q} \mathrm{I}\{X_m \in B(X_l, r_d)\}$$
$$\leq \mathbb{E}N_Q(N_Q-1)\mathbb{P}\{X_m \in B(X_l, r_d)|X_m \in Q\} \leq m^2(Q)\frac{\pi r_d^2}{m(Q)}.$$
$$T_2 = \mathbb{E}\sum_{l=1}^{N_Q}\sum_{m=1}^{N_Q}\sum_{u=1}^{N_Q} \mathrm{I}\{X_m \in B(X_l, r_d)\}\,\mathrm{I}\{X_u \in B(X_l, r_d)\}$$
$$\leq \mathbb{E}N_Q^2(N_Q-1)\mathbb{P}^2\{X_m \in B(X_l, r_d)|X_m \in Q\} \leq (m^3(Q)+2m^2(Q))\left(\frac{\pi r_d^2}{m(Q)}\right)^2.$$

Note that $m(Q) = O(\sqrt{k}n)$. Then $\mathrm{var}\,(w_n) = O\left(m(Q)/n^2\right) = O\left(p^{-1}\right)$ and w_n is negligible in L^2.

Consider now $K_{2,n}(r)$. Define $Z_2 = \sum_{t=1}^d \lambda_t K_{2,n}(r_t) = A_{N,n}Z_1$ where $A_{N,n} = \frac{n^4\rho^2}{N(N-1)}\cdot$ We have $\mathbb{E}(A_{N,n}^{-1}) = 1$ and from (5.5), $\mathrm{var}\,(A_{N,n}^{-1}) = \dfrac{4}{n^2\rho} + \dfrac{2}{n^4\rho^2}\cdot$ For $\delta > 0$, the Markov inequality gives

$$\mathbb{P}(|A_{N,n}^{-1}-1| > \delta) \leq \frac{\mathrm{var}\,(A_{N,n}^{-1})}{\delta^2}\cdot$$

Then, with $\delta = n^{-1/4}$, $\displaystyle\sum_{n=1}^{\infty}\mathbb{P}(|A_{N,n}^{-1}-1| > n^{-1/4}) < \sum_{n=1}^{\infty}\frac{4}{n^{3/2}\rho}+\frac{2}{n^{7/2}\rho^2} < \infty$. From the Borel–Cantelli lemma, we get that $A_{N,n}^{-1}$ converges a.s. to 1. By the Slutsky lemma, $A_{N,n}Z_1$ converges in distribution to $\mathcal{N}(0, \Lambda^t\Sigma\Lambda)$. □

5.4. Proof of Lemma 5.3

This lemma is equivalent to Result 1 of [27], substituting r/n to the parameter h and substracting $e_{r,n}^2$. From the computation of the bias, denoting $x_i = \frac{1}{r}(n-\xi_i)$, we get

$$\begin{aligned} h_1(\xi, r) &= \frac{\pi r^2}{n^2} - e_{r,n} \text{ on } A_n^1 \\ &= \frac{r^2}{n^2}(\pi - g(x_1)) - e_{r,n} \text{ on } A_n^2 \\ &= \frac{r^2}{n^2}(\pi - g(x_1) - g(x_2)) - e_{r,n} \text{ on } A_n^3 \\ &= \frac{r^2}{n^2}\left(\frac{3\pi}{4} + x_1x_2 - \frac{g(x_1)+g(x_2)}{2}\right) - e_{r,n} \text{ on } A_n^4 \end{aligned}$$

Integrating on the four zones, we get

$$\mathbb{E}(h_1(X_j, r))^2 = \pi^2\left(1-\frac{2r}{n}\right)^2\frac{r^4}{n^4} - \mathrm{e}_{r,n}^2 + T_1 + T_2 + T_3$$
$$T_1 = 4\left(1-\frac{2r}{n}\right)\frac{r^5}{n^5}\int_0^1(\pi - g(x_1))^2\mathrm{d}x_1$$
$$T_2 = 4\frac{r^6}{n^6}\int_0^1 \mathrm{d}x_1\int_{\sqrt{1-x_1^2}}^1(\pi - g(x_1) - g(x_2))^2\mathrm{d}x_2$$
$$T_3 = 4\frac{r^6}{n^6}\int_0^1 \mathrm{d}x_1\int_0^{\sqrt{1-x_1^2}}\left(\frac{3\pi}{4} + x_1x_2 - \frac{g(x_1)+g(x_2)}{2}\right)^2\mathrm{d}x_2.$$

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

To rewrite these three terms with the notations of [27], we denote $\theta = \arccos(x_1)$ and $\phi = \arccos(x_2)$.

$$T_1 = 4\left(1 - \frac{2r}{n}\right)\frac{r^5}{n^5}\int_0^{\pi/2}(\pi - \theta + \cos(\theta)\sin(\theta))^2\sin(\theta)\mathrm{d}\theta$$

$$T_2 = 4\frac{r^6}{n^6}\int_0^{\pi/2}\sin(\theta)\mathrm{d}\theta\int_0^{\pi/2-\theta}(\pi - \theta + \cos(\theta)\sin(\theta) - \phi + \cos(\phi)\sin(\phi))^2\sin(\phi)\mathrm{d}\phi$$

$$T_3 = \frac{r^6}{n^6}\int_0^{\pi/2}\sin(\theta)\mathrm{d}\theta\int_{\pi/2-\theta}^{\pi/2}(3\pi/2 + 2\cos(\theta)\cos(\phi) - \theta - \phi + \cos(\theta)\sin(\theta) + \cos(\phi)\sin(\phi))^2\sin(\phi)\mathrm{d}\phi$$

$$= \frac{r^6}{n^6}\int_0^{\pi/2}\cos(\theta')\mathrm{d}\theta'\int_0^{\theta'}(\pi/2 + 2\sin(\theta')\sin(\phi') + \theta' + \phi' + \cos(\theta')\sin(\theta') + \cos(\phi')\sin(\phi'))^2\cos(\phi')\mathrm{d}\phi'.$$

changing variables by $\theta' = \pi/2 - \theta$ and $\phi' = \pi/2 - \phi$. Then formulas (5), (6) and (7) in [27] give respectively

$$T_1 = \left(1 - \frac{2r}{n}\right)\frac{r^5}{n^5}\left(4\pi^2 - \frac{256}{45} - \frac{8\pi}{3}\right) \tag{5.10}$$

$$T_2 = \frac{r^6}{n^6}\left(-\frac{\pi^3}{4} + 4\pi^2 - \frac{9\pi}{2} - \frac{512}{45}\right) \tag{5.11}$$

$$T_3 = \frac{r^6}{n^6}\left(\frac{\pi^3}{4} + \frac{19\pi}{48} + \frac{8}{9}\right). \tag{5.12}$$

Note that the upper bound of the second integral in formula (7) of [27] is a mistyping. Gathering the expression of $e_{r,n}$, (5.10)–(5.12) gives the result.

Acknowledgements. We are thankful to Michel Koskas for his help in accelerating the computation of K with C^{++}. We also wish to thank the anonymous referees for their suggestions to improve the section concerning the power of the test.

References

[1] A.J. Baddeley, M. Kerscher, K. Schladitz and B.T. Scott, Estimating the J function without edge correction. *Research report of the department of mathematics*, University of Western Australia (1997).
[2] J-M. Bardet, P. Doukhan, G. Lang and N. Ragache, Dependent Lindeberg central limit theorem and some applications. *ESAIM: PS* **12** (2008) 154–172.
[3] S. Bernstein, Quelques remarques sur le théorème limite Liapounoff. *C.R. (Dokl.) Acad. Sci. URSS* **24** (1939) 3–8.
[4] J.E. Besag, Comments on Ripley's paper. *J. Roy. Statist. Soc. Ser. B* **39** (1977) 193–195.
[5] S.N. Chiu, Correction to Koen's critical values in testing spatial randomness. *J. Stat. Comput. Simul.* **77** (2007) 1001–1004.
[6] S.N. Chiu and K.I. Liu, Generalized Cramér-von Mises goodness-of-fit tests for multivariate distributions. *Comput. Stat. Data Anal.* **53** (2009) 3817–3834.
[7] N.A. Cressie, *Statistics for spatial data.* John Wiley and Sons, New York (1993).
[8] P.J. Diggle, *Statistical analysis of spatial point patterns.* Academic Press, London (1983).
[9] M. Fromont, B. Laurent and P. Reynaud-Bouret, Adaptive tests of homogeneity for a Poisson process. *Ann. I.H.P. (B)* **47** (2011) 176–213.
[10] P. Grabarnik and S.N. Chiu, Goodness-of-fit test for complete spatial randomness against mixtures of regular and clustured spatial point processes. *Biometrika* **89** (2002) 411–421.
[11] J. Gignoux, C. Duby and S. Barot, Comparing the performances of Diggle's tests of spatial randomness for small samples with and without edge effect correction: application to ecological data. *Biometrics* **55** (1999) 156–164.
[12] Y. Guan, On nonparametric variance estimation for second-order statistics of inhomogeneous spatial point Processes with a known parametric intensity form. *J. Am. Stat. Ass.* **104** (2009) 1482–1491.
[13] L.P. Ho and S.N. Chiu, Testing Uniformity of a Spatial Point Pattern. *J. Comput. Graph. Stat.* **16** 2 (2007) 378–398.
[14] L. Heinrich, Goodness-of-fit tests for the second moment function of a stationary multidimensional Poisson process. *Statistics* **22** (1991) 245–268.
[15] J. Illian, A. Penttinen, H. Stoyan and D. Stoyan, *Statistical analysis and modelling of spatial point patterns.* Wiley-Interscience, Chichester (2008).
[16] C. Koen, Approximate confidence bounds for Ripley's statistic for random points in a square. *Biom. J.* **33** (1991) 173–177.

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press

[17] E. Marcon and F. Puech, Evaluating the geographic concentration of industries using distance-based methods. *J. Econom. Geogr.* **3** (2003) 409–428.
[18] J. Møller and R.P. Waagepetersen, Statistical inference and simulation for spatial point processes, vol. 100 of *Monographs on statistics and applied probability*. Chapman and Hall/CRC, Boca Raton (2004).
[19] R Development Core Team (2012). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. http://www.R-project.org.
[20] B.D. Ripley, The second-order analysis of stationary point processes. *J. Appl. Probab.* **13** (1976) 255–266.
[21] B.D. Ripley, Modelling spatial patterns. *J. Roy. Statist. Soc. Ser. B* **39** 2 (1977) 172–212.
[22] B.D. Ripley, Tests of randomness for spatial point patterns. *J. Roy. Statist. Soc. Ser. B* **41** 3 (1979) 368–374.
[23] B.D. Ripley, *Spatial statistics*. John Wiley and Sons, New York (1981).
[24] R. Saunders and G.M. Funk, Poisson limits for a clustering model of Strauss. *J. Appl. Probab.* **14** (1977) 776–784.
[25] D. Stoyan, W.S. Kendall and J. Mecke, *Stochastic geometry and its applications*. Akademie-Verlag, Berlin (1987).
[26] D. Stoyan and H. Stoyan, *Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics*. John Wiley and Sons, New York (1994).
[27] C.C. Taylor, I.L. Dryden and R. Farnoosh, The K function for nearly regular point processes. *Biometrics* **57** (2000) 224–231.
[28] M. Thomas, A generalization of Poisson's binomial limit for use in ecology. *Biometrika* **36** (1949) 18–25.
[29] E. Thönnes and M.-C. van Lieshout, A comparative study on the power of van Lieshout and Baddeley's J function. *Biom. J.* **41** (1999) 721–734.
[30] J.S. Ward and F.J. Ferrandino, New derivation reduces bias and increases power of Ripley's L index. *Ecological Modelling* **116** (1999) 225–236.

https://doi.org/10.1051/ps/2012027 Published online by Cambridge University Press