CSGB CENTRE FOR STOCHASTIC GEOMETRY AND ADVANCED BIOIMAGING

Ute Hahn

A Studentized Permutation Test for the Comparison of Spatial Point Patterns

No. 12, December 2010

A Studentized Permutation Test for the Comparison of Spatial Point Patterns

Ute Hahn

Centre for Stochastic Geometry and Advanced Bioimaging,
Department of Mathematical Sciences, University of Aarhus,
DK-8000 Århus C, Denmark

Abstract

A new test is proposed for the hypothesis that two (or more) observed point patterns are realizations of the same spatial point process model. To this end, the point patterns are divided into disjoint quadrats, on each of which an estimate of Ripley's K-function is calculated. The two groups of empirical K-functions are compared by a permutation test using a studentized test statistic. The proposed test performs convincingly in terms of empirical level and power in a simulation study, even for point patterns where the K-function estimates on neighboring subsamples are not strictly exchangeable. It also shows improved behavior compared to a test suggested by Diggle et al. (1991, 2000) for the comparison of groups of independently replicated point patterns. In an application to two point patterns from pathology that represent capillary positions in sections of healthy and tumorous tissue, our studentized permutation test indicates statistical significance, although the patterns cannot be clearly distinguished by eye.

Key words: Nonparametric test, K-function, quadrat, spatial point process, subsampling.

1 Introduction

Many fields of science deal with data that are point patterns, like maps of disease incidents, ore deposits, trees or galaxies, or locations of pores or cells in sections through material or tissue. Statistical analysis of these patterns aims at characterizing the spatial arrangement of the points. An example situation is given in Figure 1 which shows midpoints of capillary profiles on sections of healthy and cancerous prostate tissue. The two patterns have about the same number of points but they differ slightly in the mutual positions of the points.

Spatial arrangement can be captured and summarized using Ripley's (1976) K-function $K(r)$, which is proportional to the mean number of further points within distance r to a typical point of the process. The K-function is one of the main tools in the analysis of spatial point processes, in particular for hypothesis testing. The majority of tests proposed in the literature so far focus on the question whether or not an observed point pattern is a realization of a specified null model (e.g. Besag

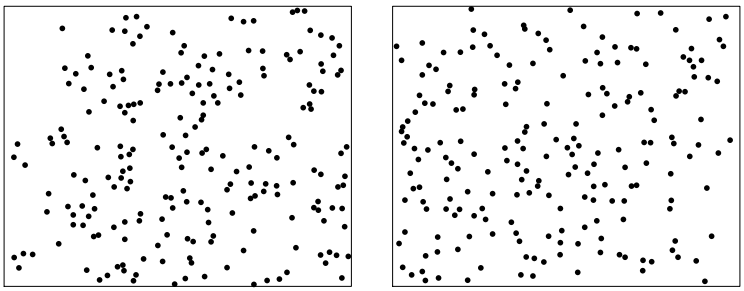


Figure 1: Capillary profiles marked as dots in sections of prostate tissue. The size of the fields of view is 1860 µm × 1500 µm. Left: healthy tissue, 212 points, right: cancerous tissue, 204 points. Data from Mattfeldt et al. (2006, 2007), by courtesy of T. Mattfeldt.

and Diggle (1977); Diggle (1979); Ripley (1979); Ho and Chiu (2006)), whereas tests for nonparametric hypotheses appear to be less common. This may be explained by the fact that observed data most often consist of a single point pattern, while groups of independent replicated samples form a rare exception. For the latter case, Diggle and coworkers (Diggle et al., 1991, 2000) have proposed bootstrap Monte Carlo tests for the comparison of empirical K-functions between groups.

To the best of our knowledge, a nonparametric test that permits a direct comparison of two single point patterns is not available in the literature. In the present paper, we develop a test of equality of K-functions estimated from two (or more) point patterns. Instead of independently collected samples, we use estimated K-functions based on subsamples of the point patterns on disjoint quadrats. Such subsampling has earlier been used by Loh and Stein (2004) to assess the variance of K-function estimates.

The bootstrap Monte Carlo tests of Diggle et al. seem to be good candidates for comparing subsamples of single patterns. However, a simulation study presented in this paper reveals that they may behave pronouncedly liberal (anti-conservative) if the samples are small in terms of the number of point patterns per group. In most cases, the number of disjoint quadrats obtained from a single point pattern will be quite limited, because they need to contain enough points to yield reasonably stable estimates of the K-function. We therefore propose a permutation test based on a studentized test statistic, since permutation tests by construction meet the nominal level of significance when comparing independent, identically distributed replicates, even when the group size is small.

The article is organized as follows. Section 2 introduces spatial point processes and the K-function. A brief overview of statistical tests for point processes is given in Section 3. The performance of Diggle et al.'s Monte Carlo test is examined in Section 4. In Section 5, the studentized permutation test is presented and its behavior in the case of independent replicated samples is studied by simulation. Section 6 explains the use of this test to compare subsamples of point patterns. We investigate the effect of size and number of quadrats on the empirical level and power. Practical recommendations are given in Section 7 alongside with an application to the data

shown in Figure 1. In Section 8, we review open questions, indicate directions for future research and suggest possible extensions. Some aspects of Diggle et al.'s test under heteroscedasticity are discussed in an Appendix.

2 Spatial Point Processes and the K-Function

A spatial point process $\boldsymbol{X}$ is a random variable taking values in the set of all locally finite point patterns, i.e. countable subsets $\boldsymbol{x} = \{x_1, x_2, \dots\} \subset \mathbb{R}^d$ such that every bounded set $B \subset \mathbb{R}^d$ contains only finitely many points. In the following, we use the notation $\boldsymbol{X}(B)$ for the number of points in a set $B \subset \mathbb{R}^d$, and we will concentrate on planar point processes.

The vast majority of theoretical results and statistical methods is confined to stationary point processes, i.e. point processes with translation invariant distribution. Stationary point processes have a constant intensity (mean number of points per unit area) which will be denoted by λ. The most basic model is the Poisson point process, also known as the complete spatial randomness (CSR) model. A stationary Poisson point process $\boldsymbol{X}$ with intensity λ is characterized by the property that $\boldsymbol{X}(B)$ is Poisson distributed with mean $\lambda A(B)$, where $A(B)$ denotes the area of B. The restrictions of a Poisson point process to pairwise disjoint sets $B_1, B_2, \dots$ are independent.

The K-function of a stationary point process is defined as the mean number of other points within distance r to a typical point of the process, divided by the intensity λ of the process (Ripley, 1976, 1977). Under CSR, $K(r)$ equals the area of a circle of radius r. Clustering, that is, attraction between points within distance r is reflected by a higher value of $K(r) > \pi r^2$, while lower values indicate dispersion or repulsion between points. One has to be aware that different point process models do not necessarily have different K-functions, in particular, the K-function is invariant under independent thinning of the point process, see e.g. Baddeley et al. (2000).

In order to estimate $K(r)$ from a given point pattern on a bounded region $B \subset \mathbb{R}^2$, one would average the number of other points v within a distance r around every point u in B and divide by the estimated intensity. In most cases, only points inside the window B are registered. This means that observation of other points inside the circular disk of radius r around u is censored whenever the point u lies within distance r to the boundary of B. To compensate for the unobservable part, usual estimators for $K(r)$ therefore include a Horvitz-Thompson correction term $w(u, v; B)$ and are of the form

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda}^2 A(B)} \sum_{u,v \in \boldsymbol{X} \cap B}^{\neq} \mathbf{1}(\|u - v\| \le r) w(u, v; B). \tag{1}$$

Note that $\hat{\lambda} A(B) = \boldsymbol{X}(B)$ is the number of points in B. Most common are

- the *translational edge correction* $w_t(u, v; B) = A(B)/A(B_u \cap B_v)$, where B_u stands for the set B shifted by the vector u, and
- Ripley's *isotropic edge correction* $w_i(u, v; B)$, which is the reciprocal of the fraction of the perimeter of the circle centred at u and passing through v which lies inside the sampling window (Ripley, 1976).

Since $\widehat{\lambda^2}\widehat{K}(r)$ is an unbiased estimator for $\lambda^2 K(r)$, see Ohser and Stoyan (1981), division by the estimated squared intensity $\widehat{\lambda^2}$ introduces a bias, but it also reduces the estimation variance, as discussed in detail by Stoyan and Stoyan (2000).

The variance of the estimator $\widehat{K}(r)$ depends on the point process model itself as well as on the argument r and size and shape of the sampling window B. Typically, the variance is higher for cluster point processes than for regular models with the same intensity. According to asymptotic results of Ripley (1984, 1988) for a Poisson point process of intensity λ, it can be approximated by

$$\operatorname{var}\widehat{K}(r) \approx \frac{2\pi r^2}{\lambda^2 A(B)}\left(1 + c_1\frac{U(B)}{A(B)}r + c_2\frac{U(B)}{A(B)}\lambda r^3\right) \tag{2}$$

for small r, where $U(B)$ stands for the perimeter of B. The constants c_1 and c_2 depend on the type of edge correction used, Ripley (1988) gives $c_1 = 0.305$ and $c_2 = 0.0415$ in the case of isotropic edge correction, and similar values for the other correction methods. Conditional on the number $\boldsymbol{X}(B) = n$ of points in the observation window, the estimation variance is approximately

$$\operatorname{var}\left(\widehat{K}(r)\,\middle|\,\boldsymbol{X}(B) = n\right) \approx \frac{2\pi r^2 A(B)}{n^2}\left(1 + c_1\frac{U(B)}{A(B)}r + c_2 n\frac{U(B)}{A(B)^2}r^3\right) \tag{3}$$

In all cases, $\operatorname{var}\widehat{K}(r)$ is roughly proportional to r^2 if r is small. Thus, the variance of $\widehat{K}(r)/r$ is approximately independent of r.

3 State of the Art

The majority of the statistical tests for spatial point processes that have been proposed so far are based on Ripley's K-function (Ripley, 1976) or the closely related L-function (Besag, 1977), see e.g. Ho and Chiu (2009) and Yamada and Rogerson (2003). Various tests are compared in Ripley (1979), Diggle (1979), Gignoux et al. (1999).

The statistical analysis of summary functions such as the K-function is hampered by the fact that their estimates are functional data. It would of course be possible to inspect $K(r)$ for some fixed argument r, but this rises the problem of choosing an appropriate r. The best choice of r in terms of power depends heavily on the alternative to the null model. As estimation variance increases with r, the signal to noise ratio concomitantly worsens. Virtually all authors therefore recommend to compare the estimated second order functions as a whole, that is by establishing some distance between the functional data, mostly the supremum distance

$$d_\infty(f_1, f_2; r_0) = \sup_{r \le r_0} |f_1(r) - f_2(r)| \tag{4}$$

or the L_2-distance

$$d_2(f_1, f_2; r_0) = \int_{r \le r_0} \big(f_1(r) - f_2(r)\big)^2 \mathrm{d}r, \tag{5}$$

where f_1 and f_2 typically are the empirical and theoretical (or simulated) K- or L-function, respectively.

Only little is known about the distribution of these distance statistics. Some asymptotic formulae have been derived e.g. by Heinrich (1988), but these are of limited use in the case of relatively small point patterns. The majority of other results are based on simulation studies.

The power of goodness of fit tests based on the distances d_∞ or d_2 of K- or L-functions depends on the upper limit r_0. Ripley (1979) recommends to set $r_0 = 1.25/\sqrt{\lambda}$, a choice that has proven to yield powerful tests, see e.g. Diggle (1979). Ho and Chiu (2006) show that the power can be further improved by choosing adapted estimators for the intensity as suggested by Stoyan and Stoyan (2000). Another possibility to increase the power against specific alternatives is to use weighted distances instead of the plain integrals in (4) or (5), see Ho and Chiu (2009).

All the above mentioned studies concentrate on model tests against a null model that can be simulated, typically the Poisson model. To our knowledge, the first model free test was proposed by Diggle et al. (1991) who suggest to use bootstrap methods to determine the distribution of the test statistic in the situation of data sets consisting of several independently sampled point patterns. The same principle was then applied by Baddeley et al. (1993) in the analysis of independent replicates of three dimensional point patterns. Diggle et al. (2000) present a newer version of these bootstrap or Monte Carlo tests, which has influenced other authors later, e.g. Schladitz et al. (2003). These tests are studied in detail in the following section.

4 Diggle et al.'s Monte Carlo Test for Comparing Independent Samples of K-Functions

4.1 Diggle et al.'s Test Procedure

Diggle et al. (1991, 2000) suggest to test the difference between group means of independent replicates of empirical K-functions by a bootstrap test. They generate bootstrap samples $(\hat{K}^*_{ij})$ from the original sample $(\widehat{K}_{ij})_{i=1,\dots,g,\ j=1,\dots,m_i}$ as follows: First, residual functions $\hat{R}_{ij}(r)$ are calculated from the empirical K-functions $\widehat{K}_{ij}(r)$ that were estimated from the jth point pattern in the ith group:

$$\hat{R}_{ij}(r) := n_{ij}^{1/2}(\hat{K}_{ij}(r) - \bar{K}_i(r)), \tag{6}$$

where n_{ij} stands for the number of points in the pattern and $\bar{K}_i(r)$ for the group mean. Subsequently, a random random sample $(\hat{R}^*_{ij})$ from the set of residual functions is rescaled and combined with the overall mean $\bar{K}$:

$$\hat{K}^*_{ij}(r) = \bar{K}(r) + n_{ij}^{-1/2}\hat{R}^*_{ij}(r). \tag{7}$$

The choice of the weight $n_{ij}^{1/2}$ is motivated by the assumption that the variance of $\widehat{K}_{ij}(r)$ is inversely proportional to n_{ij}, for which authors refer to Cressie (1993, p. 642). Under this assumption, the residuals (6) are approximately exchangeable, and the distribution of $\hat{K}^*_{ij}$ approximates the distribution of $\hat{K}_{ij}$ under the null hypothesis.

In order to determine a bootstrap p-value, the observed value of a test statistic is ranked among the corresponding bootstrap values of the test statistic. Diggle et al.

(1991) use the statistic

$$D = \sum_{i=1}^{g} \int_0^{r_0} \left(\sqrt{\bar{K}_i(r)} - \sqrt{\bar{K}(r)}\right)^2 \mathrm{d}r, \tag{8}$$

and advise to draw the bootstrap residuals R_{ij}^* with replacement from all $\hat{R}_{ij}$. This design is changed later (Diggle et al., 2000) in favor of sampling without replacement, i.e. performing a (random) bootstrap permutation test, with the test statistic

$$D = \sum_{i=1}^{g} n_i \int_0^{r_0} \frac{1}{r^2} \left(\bar{K}_i(r) - \bar{K}(r)\right)^2 \mathrm{d}r, \tag{9}$$

that was chosen in analogy to classical ANOVA. The square root transformation in the first case resp. dividing by r^2 in the second case serves to achieve roughly constant variance over the integration interval. The group and overall means are obtained by inverse weighting with the variance, as usual in heteroscedastic ANOVA. With the above assumption on the variance,

$$\bar{K}_i(r) = \frac{1}{n_i} \sum_{j=1}^{m_i} n_{ij} \hat{K}_{ij}(r) \text{ and } \bar{K}(r) = \frac{1}{n} \sum_{i=1}^{g} n_i \bar{K}_i(r), \text{where } n_i = \sum_{j=1}^{m_i} n_{ij} \text{ and } n = \sum_{i=1}^{g} n_i.$$

In the following subsection, a simulation study of the statistical properties of the proposed bootstrap test is presented.

4.2 A simulation study of Diggle et al.'s test

We determined the empirical level of the Monte Carlo test in the version of Diggle et al. (2000) by simulation for the case of Poisson point processes. The parameters were chosen to mimic the subsampling situation we are finally interested in, namely relatively small observation windows containing about 20 to 30 points and small sample sizes m_i around ten. 10 000 replications of an experiment with two independent samples of $m_1 = m_2 = 9$ point patterns each were generated. We considered three versions of the null hypothesis "the two underlying point processes have the same K-function", namely

a) the homoscedastic case: both samples were generated from a Poisson point process with intensity $\lambda = 100$ on a 0.5×0.5 square window,

b) same window, different intensities: the first sample was from a Poisson point process with intensity $\lambda = 100$, while the intensity was $\lambda = 200$ in the second sample, both on the same 0.5×0.5 square window,

c) same intensity, different windows: both samples were from the same Poisson point process with $\lambda = 100$, but on a unit square window in the second sample.

With an expected number of 25 points per window in the homoscedastic setting, the study is comparable to the simulations by Diggle et al. (2000) who considered tests of two samples of $m_1 = m_2 = 10$ patterns from Markov point processes with intensity 30 on the unit square. Due to limited computing time their study comprised only

20 to 50 replicates of the test. They report empirical Kolmogorov-Smirnov distances to uniformity of the observed p-value distribution with values between 0.09 and 0.48 (i.e. the maximal difference between observed and theoretical distribution function). However many of the observed distances are not significant due to the small study size, and it is not known if the discrepancies to uniformity occur at the important small p-values.

In our simulation study, the K-function was estimated using Ripley's isotropic edge correction, see Section 2. The integral in (9) was approximated by a sum on discrete values of r, with an upper limit $r_0 = 0.2$. Unfortunately, the test appears to be very liberal (anti-conservative) when group sizes are as small as in the present simulation study, in particular in the heteroscedastic situations b) and c), see Table 1.

Table 1: Observed rejection rates in Diggle et al.'s test, applied to two groups of 9 realizations each of a Poisson point process, for the homoscedastic case a) and the heteroscedastic cases b) and c), see text. Observed rejection rates are based on 10 000 replications.

Nominal significance level α	0.01	0.05	0.10
a) same intensity and window	0.025	0.072	0.121
b) different intensities	0.038	0.100	0.156
c) different windows	0.036	0.096	0.159

The empirical level of the test approaches the nominal level if larger sample sizes are used, but not if the point patterns contain more points. Table 2 lists the empirical level for tests with small group size $m_1 = m_2 = 9$ and large windows (100 points on average) as well as for test with group size $m_1 = m_2 = 18$ and small windows (25 points on average).

Table 2: Observed rejection rates in Diggle et al.'s test, in the homoscedastic case of Poisson point processes with intensity $\lambda = 100$. Large windows: two groups of 9 replicates each on the unit square, large samples: two groups of 18 replicates each on square windows with side length 0.5.

Nominal significance level α	0.01	0.05	0.10
large window	0.029	0.079	0.130
large samples	0.014	0.060	0.110

The simulation study indicates that this way of bootstrapping by permutation of empirical residuals may fail to reproduce the distribution of the test statistic closely enough when sample sizes are small. This seems to be a general problem with bootstrap based on small samples, in particular when non pivotal statistics are used. Already Schenker (1985) reports that confidence intervals for the variance of the normal distribution tend to be too narrow when they are based on samples of size 50 or less. In the case of Diggle et al.'s test, the distribution of the statistic D obtained from permutations of a given small sample of residuals tends to be less variable than the distribution of the same statistic based on independent small samples. Estimated residuals are negatively correlated and have a smaller variance than true residuals.

As a result, the variance between groups, and thus the distance statistic D, is likely to be smaller in the bootstrapped sample than in the original sample.

The even more pronounced nonuniformity of p-values in the heteroscedastic cases b) and c) is presumably due to an unlucky choice of the weights for the residuals, as detailed in the Appendix. The choice was motivated by the assumption that the variance of $\widehat{K}$ is inversely proportional to the number of points in the pattern. However, for a binomial point process on a fixed window with n points one can show that the variance of $\widehat{K}(r)$ is rather roughly proportional to n^{-2} (or $n(n-1)$), see the end of Section 2.

It should be noted that the bootstrapped K-functions can become negative, and that the mean of the bootstrapped K-functions differs from $\bar{K}$,

$$\begin{aligned}\bar{K}^*(r) &= \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{m_i} n_{ij}\big(\bar{K}(r) + n_{ij}^{-1/2}\hat{R}^*_{ij}(r)\big) \\ &= \bar{K}(r) + \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{m_i}\sqrt{n_{ij}n^*_{ij}}(\hat{K}^*_{ij}(r) - \bar{K}^*_i(r)),\end{aligned}$$

where $\bar{K}^*_i$ is the mean of the group K^*_{ij} originally belonged to. The difference between $\bar{K}$ and $\bar{K}^*$ vanishes if the residuals are weighted with the number n_{ij} of points instead of the square root $n_{ij}^{1/2}$.

5 A Studentized Permutation Test

5.1 Test Procedure

Nonuniformity of p-values under the null hypothesis as observed with Diggle et al.'s Monte Carlo test is apparently a general problem of bootstrap tests based on small samples. As an alternative to bootstrapping, we construct a "pure" permutation test in the sense of Fisher (1935, 1966, section 21) and Pitman (1937). Such tests have uniformly distributed rejection rates by construction even when sample sizes are small, as long as the samples are exchangeable.

In order to achieve robustness of the test towards heteroscedasticity, we suggest to use a statistic related to the Behrens-Fisher-Welch t-statistic, or alternatively the corresponding F-statistic, which have proven very robust in permutation tests for a large range of distributions (Janssen and Pauls, 2005). This is generalized to the functional data case by considering the L_2-norm of the t-statistic, i.e. the integral over the squared studentized differences between the group means:

$$T = \sum_{1\le i<j\le g}\int_0^{r_0}\frac{\big(\bar{K}_i(r) - \bar{K}_j(r)\big)^2}{\frac{1}{m_i}s_i^2(r) + \frac{1}{m_j}s_j^2(r)}\,\mathrm{d}r, \tag{10}$$

with

$$\bar{K}_i(r) = \frac{1}{m_i}\sum_{j=1}^{m_i}\widehat{K}_{ij}(r) \quad \text{and} \quad s_i^2(r) = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(\widehat{K}_{ij}(r) - \bar{K}_i(r))^2.$$

Note that this statistic is in fact very similar to the statistic used in Diggle et al.'s Monte Carlo test. For $g = 2$ groups, (9) reduces to

$$D = \frac{1}{2} \int_0^{r_0} \frac{\left(\bar{K}_1(r) - \bar{K}_2(r)\right)^2}{r^2/n_1 + r^2/n_2} \mathrm{d}r.$$

Under the assumption made by Diggle et al. (1991, 2000) that the variance of $\hat{K}_{ij}(r)$ is asymptotically proportional to r^2/n_{ij}, the denominator in T could be considered as an estimator for the denominator in D, up to a scale factor.

As we will see later, tests using the statistic T are still sensitive to pronounced heteroscedasticity. In these cases, where quadrats are known to be of markedly different size, or point patterns have very different intensity, we propose to use the following statistic instead:

$$U = \sum_{1 \le i < j \le g} \int_0^{r_0} \frac{\left(\bar{K}_i(r) - \bar{K}_j(r)\right)^2}{r^2 a_{ij}} \mathrm{d}r, \qquad a_{ij} = \frac{1}{r_0} \int_0^{r_0} \left(\tfrac{1}{m_i} s_i^2(r) + \tfrac{1}{m_j} s_j^2(r)\right) / r^2 \mathrm{d}r \tag{11}$$

This statistic is motivated by the fact that the variance of $\hat{K}(r)$ is roughly proportional to r^2, see Equation (2). Being based on the estimates $\hat{K}(r)$ for all r in the interval, the variance estimator used in U is more stable than the individual denominators in the statistic T. According to the recommendations that Hall and Wilson (1991) give for bootstrap tests, U should therefore be preferred to T.

5.2 Empirical Level of the Studentized Permutation Test

We investigated the empirical level of the studentized permutation test by simulation of Poisson point processes. Table 3 summarizes the results for tests comparing two groups of 9 patterns each, based on the statistic T given by (10) and U, given by (11), respectively. The same scenarios for the homoscedastic and the heteroscedastic case were used as in Section 4.2.

Table 3: Observed rejection rates in 10 000 replications of the studentized permutation test using test statistic T or U, applied to two groups of 9 realizations each of a Poisson point process. a) Both point processes with intensity $\lambda = 100$ on a 0.5×0.5 square, b) same window, different intensity ($\lambda_1 = 100, \lambda_2 = 200$), and c) same intensity ($\lambda = 100$), different windows (squares with side lengths 0.5 and 1.0, resp.).

Nominal significance level α	0.01		0.05		0.10	
Test statistic	T	U	T	U	T	U
a) same intensity and window	0.011	0.010	0.050	0.048	0.098	0.099
b) different intensity	0.018	0.013	0.070	0.058	0.131	0.110
c) different windows	0.021	0.015	0.074	0.062	0.133	0.113

As expected, both studentized permutation tests adhere to the nominal rejection rate if the extended null hypothesis, i.e. exchangeability of the estimated K-functions, is met, that is in the homoscedastic situation a). The rejection rates in

the heteroscedastic cases b) and c) are somewhat larger than the nominal rejection rate. However compared to Diggle et al.'s test, both tests show improved conformity with the nominal level of significance, with U being clearly more robust towards heteroscedasticity than T.

5.3 Power of the Test

The power of the test has been exemplarily studied by testing regular and clustered point patterns against realizations of a Poisson point process. The point patterns were obtained by simulating from Matérn's (1960) hard core and cluster point processes with parameters that lead to different degrees of regularity or clustering. Both types of point processes are derived from a parent Poisson point process. The Matérn hard core model used here is obtained by dependent thinning. To this end, the points are first marked with independent, identically distributed "arrival times", and all points that have a higher mark than any neighbors within the hardcore distance h are removed. The Matérn cluster point process consists of independent clusters of daughter points around each parent. The numbers of daughter points per cluster are Poisson distributed with mean μ, and the points are positioned independently uniformly random in a circular disk of radius r around the corresponding parent point. The parent points do not belong to the resulting point process. Example realizations of both models on a unit square are shown in Figure 2.

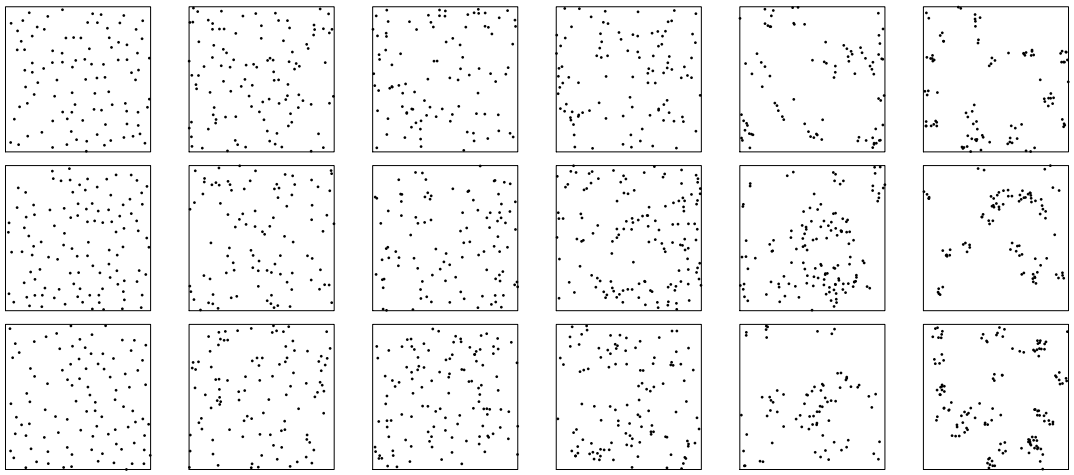


Figure 2: Sets of three independent realizations of Matérn hard core (a and b), Poisson (c) and Matérn cluster point processes (d, e, f), with intensity $\lambda = 100$ on a unit square. Model parameters: hard core radius a) $h = 0.05$, b) $h = 0.02$; mean number μ of points per cluster and cluster radius r d) $\mu = 1, r = 0.1$, e) $\mu = 4, r = 0.1$, f) $\mu = 4, r = 0.05$. The short names for the models given below the patterns are used later in the text.

For the cluster models used here, the K-functions lie entirely above the K-function of the Poisson point process, while the K functions of the hard core models lie entirely below, see Figure 3.

Figure 4 depicts the results for tests with significance level $\alpha = 0.05$, applied on groups of $m_1 = m_2 = 9$ patterns on square windows of edge length 0.5, that is with 25

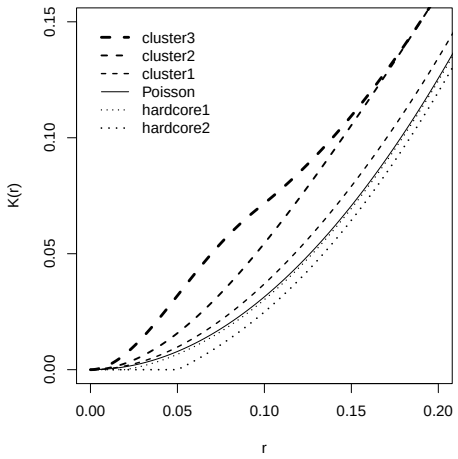


Figure 3: K-functions of Matérn cluster, Poisson and Matérn hard core point processes as shown in Figure 2.

points per pattern on average. Already with these small data sets and point pattern windows, the strongly clustered and strongly regular models are discriminated from CSR with probability 1 for virtually any choice of the upper limit r_0. However, the power of testing the weakly clustered or regular models (cluster-1 and hardcore-1) against CSR depends strongly on the parameter r_0.

Interestingly, tests based on the statistic T and tests based on U are virtually equally powerful except in the comparison of the hardcore-1 model against CSR, where T outperforms U. Apparently, increased robustness of U is paid for with a slight loss in power when testing regular point patterns. This may be explained with the fact that, for hardcore models, low values of $\mathrm{var}(\hat{K}(r)/r)$ coincide with large differences $K(r)-\pi r^2$ to the K-function under CSR, namely for small r. Therefore the integrand contributing to the statistic T, Equation (10), becomes disproportionately large for small r as compared to the corresponding integrand in U, Equation (11).

6 Comparing Two Point Patterns by Subsampling

6.1 Test Procedure

In the previous section, we introduced tests that allows to compare groups of empirical K-functions obtained on independently sampled replicates of two (or more) stationary point processes. These tests can also be used for the null hypothesis that two single observed patterns are realizations from stationary point processes with the same K-function. This is accomplished by subdividing the two observation windows into disjoint quadrats (which need not be quadratic) and applying the permutation test to the artificial sample of K-functions estimated from the subpatterns. Under the null hypothesis, the empirical K-functions are identically distributed if the quadrats are congruent, and they still have the same mean, if the quadrats differ in size or shape, or if the patterns are independently thinned.

The test procedure requires large enough point patterns in order to be able to estimate K on quadrats obtained by subdividing the original observation windows.

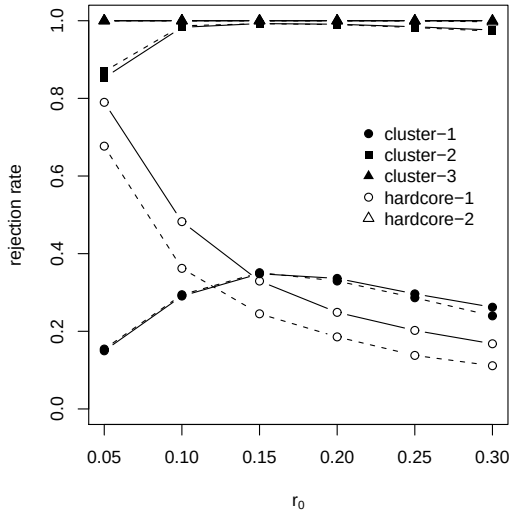


Figure 4: Power when testing the five non Poisson models against CSR, as a function of the upper limit r_0. Test at significance level $\alpha = 5\%$ based on $m_1 = m_2 = 9$ patterns on a square of side length 0.5. Solid lines: tests based on the statistic T, dashed lines: tests based on U.

On the basis of the simulation experiments described below, we recommend to only include estimates that were obtained on quadrats with at least $n_{\min} = 10$ points. The whole test procedure to compare two patterns observed on windows B_1 and B_2 can be summarized as follows:

1. For each window B_i, $i = 1, 2$, generate a set $\{B_{ij}\}, j = 1, \ldots, m_i$, of pairwise disjoint quadrats, of at least roughly the same size and shape.

2. Calculate the empirical K-functions $\hat{K}_{ij}$ on the B_{ij}. Only include quadrats B_{ij} with at least $n_{\min}$ points, thus possibly decreasing the initial sample sizes m_i.

3. Apply the studentized permutation test, using one of the statistics T or U given by (10) and (11), respectively.

Subdivision of the initial observation windows, the first step of the test, bears some freedom of choice — should the number m_i of quadrats be small or rather the size of the quadrats? We explore the effect of choosing different scenarios, including the upper integration bound r_0 in the test statistics, by a simulation study in the next subsection.

It has to be noted that subpatterns observed on disjoint windows are not independent in general, with the important exception of the Poisson point process. This could be problematic, since dependence between the $\hat{K}_{ij}$, more precisely violation of the exchangeability assumption, may distort the empirical level of any permutation test. However, most of the popular point process models have mixing properties, meaning that the dependence between subpatterns decreases as the distance between the quadrats increases. In fact, asymptotic tests for point processes, such as Heinrich (1991) and Guan (2008), rely on this property. The asymptotic independence of subpatterns on distant windows naturally carries over to asymptotic independence of any derived random variable, such as the empirical K-function. In the simulation

study described below, the observed point patterns were exhaustively partitioned into quadrats, that is, some of the quadrats were adjacent. Although this constitutes a worst case situation, the deviation from the nominal significance level was small even for the strongly clustered or strongly regular point processes where patterns on adjacent windows cannot be considered independent.

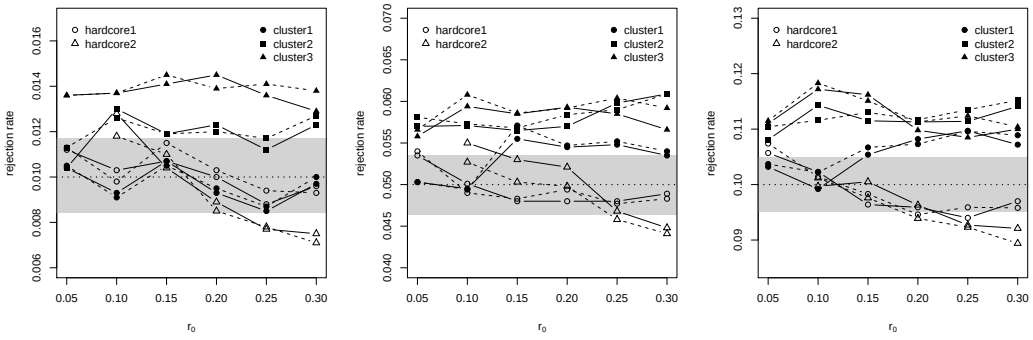
6.2 Empirical Level and Power of the Test

In this section, we present the results of a simulation study of the empirical rejection rate of the above described tests, both under the null hypothesis and under various alternatives. For this purpose, independent realizations of the models depicted in Figure 2 were generated on square observation windows of various sizes with edge lengths ranging from 1.5 to 3.0, corresponding to mean number of points between 225 and 900. The windows were partitioned into $3 \times 3, 4 \times 4, 5 \times 5$ and 6×6 square quadrats, though with minimal edge length 0.5 corresponding to 25 points on average. Thus it is possible to study both the effect of window size and sample size. For the upper limit r_0 in the test statistics T and U, we chose the values $r_0 = 0.05, 0.10, \ldots, 0.30$. The integrals were approximated by sums on discrete values of r with $r = 0, 0.001, 0.002, \ldots$. Each simulation experiment was repeated 10 000 times. In the settings with 3×3 quadrats, that is with initial sample size $m_1 = m_2 = 9$, all possible $\binom{18}{9}/2 = 24310$ levels of the test statistic under permutation were generated. For larger sample size, we calculated p-values using 4000 random permutations.

Empirical level under the null hypothesis Figure 5 shows the observed rejection rates under the non Poisson null hypotheses, that is for the Matérn cluster and Matérn hard core models, for the significance levels $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$, and two sampling windows of side length 1.5 divided into 3×3 quadrats. For the Poisson model, the empirical level of the test corresponds to the nominal level, since the estimated K-functions on different quadrats are exchangeable, and deviations would be purely random. To give an impression of the magnitude of such random deviations, a central 90% interval is also indicated in the graphs.

While the test behaves slightly liberal for the strongly clustered models (cluster-2 and cluster-3), it behaves conservative with the regular models, at least for larger values of r_0. These findings may be ascribed to the correlation structure of the models. For the cluster models, the empirical K-functions on neighboring quadrats are positively correlated, which entails that the within-subsampling-group variance is smaller than under independent sampling or under resampling by permutation. Since this variance appears in the denominators of the test statistics T and U, the result for the original samples tends to be larger than the results for the permutation samples, hence small p-values are observed more frequently than under full exchangeability. Conversely, negative correlation of empirical K-functions on neighboring quadrats in hard core models leads to smaller values of the test statistic in the original sample as compared to the permutated samples, and thus to conservative behavior of the test.

The moderate deviation from the nominal level of significance becomes smaller as the window size increases, see Figure 6, which is due to a better mixing, i.e., decreasing dependence of the subpatterns. By contrast, increasing the number of



$\alpha = 0.01$ $\alpha = 0.05$ $\alpha = 0.10$

Figure 5: Empirical rejection rate of the test under the null hypothesis as a function of the upper limit r_0, for nominal significance levels $\alpha = 0.01, \alpha = 0.05$ and $\alpha = 0.10$. Results for the five non Poisson models, using 3×3 quadrats of side length 0.5 with an average of 25 points per quadrat. Dashed lines: tests using U, solid lines: tests using T, grey area: central 90% interval of observed frequencies around the true significance level under uniformity of p-values.

quadrats does not improve the empirical level of the test, as evident from the second row in Figure 6.

Power of the test We investigated the power of the test against CSR for the clustered and hardcore models as before by simulation. The results for the test based on subsampled quadrats were virtually indistinguishable from the results for independent realizations on windows of the same size as the quadrats. For windows or quadrats of edge length 0.5 and sample size 9, they can be found in Section 5.2 and are therefore not listed here.

When the observed point patterns are larger than just about 200 points, one might ask the question whether it is better to increase the sample size in terms of the number of quadrats, or rather to increase the size of the quadrats. In order to find an answer, the simulations were extended to different quadrat size and number. The results shown in Figure 7 indicate that tests based on few large quadrats reach a (slightly) higher power than tests based on many small quadrats, at least for clustered point processes and in the investigated range of 3×3 up to 6×6 quadrats.

At the first glance, this finding may seem counterintuitive. In traditional statistics, one would not pool or average data before carrying out a t-test, because the decrease in degrees of freedom would implicate a decrease in power. The seemingly paradoxical behavior of the studentized permutation tests may be explained by the fact that the estimation variance of the within group means $\bar{K}_i(r), i = 1, 2$, becomes smaller if few large quadrats are used rather than many small ones - as one can see from the approximations (2) or (3), the variance of $\widehat{K}(r)$ becomes larger if the ratio window perimeter to area is large.

7 Practical Application

The two point patterns shown in Figure 1 represent the positions of capillary profiles on sections of healthy and tumorous prostate tissue, respectively. They were

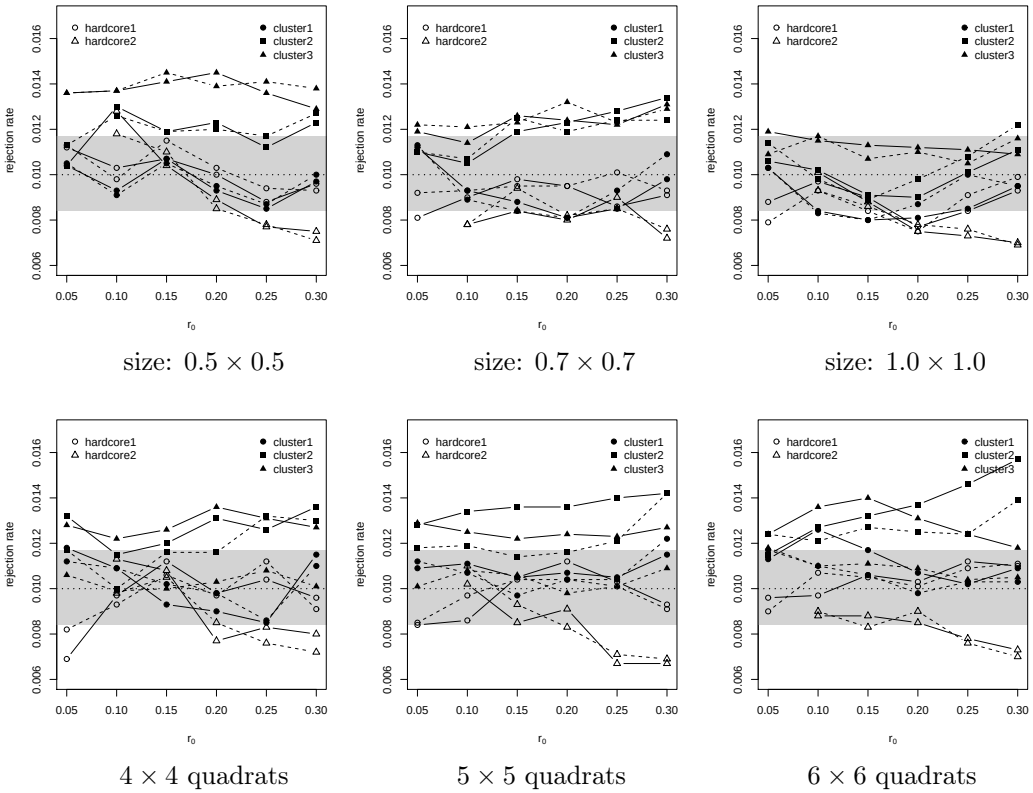
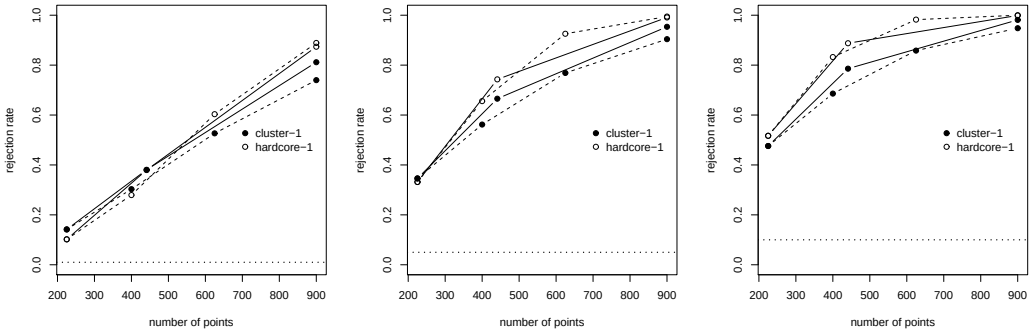


Figure 6: Effect of quadrat size (upper row) and number (lower row) on the empirical level of the test. Results based on 10 000 replications, in the upper row for 3×3 quadrats, in the lower row for quadrat size 0.5×0.5. In all cases, the intensity of the point processes studied is $\lambda = 100$. Remaining legend as in Figure 5.

taken from larger studies by Mattfeldt et al. (2006, 2007), who concluded that capillary profile patterns are more clustered in healthy tissue than in tumorous tissue by comparing two groups of 12 independently sampled patterns. The first study used the empirical pair correlation functions $g(r)$, which is related to the K-function by $g(r) = \mathrm{d}K(r)/\mathrm{d}r/(2\pi r)$. In the second study, a point process model (the Strauss hard core model) was fit to the data that allows for various degrees of interactions, and fitted parameters were compared.

We could confirm the findings of Mattfeldt et al. (2006, 2007) by testing the difference between the empirical K-functions of the two point patterns. According to the simulation study in Section 6, the studentized permutation test is slightly more powerful and also closer to the nominal level of significance if only few quadrates are used. We therefore divided the two patterns into 3×3 quadrates. In both cases, one of the quadrates contained less than 15 points, which was considered not sufficient to estimate the K-function. These two subpatterns were therefore discarded, and two groups of eight point patterns remained. The estimated K-function for the two patterns, as shown in the left part of Figure 8 below, indicates clustering in the pattern from healthy tissue for values of r up between 70 µm and 140 µm, where $\widehat{K}(r)$ exceeds the CSR-K-function $K(r) = \pi r^2$, while for values of r exceeding $r = 230$ µm, $\widehat{K}(r)$ lies below the πr^2. The point pattern from tumorous tissue shows the opposite behavior.



$\alpha = 0.01$ $\alpha = 0.05$ $\alpha = 0.10$

Figure 7: Power of the test using statistic T and fixed $r_0 = 0.15$, of the models cluster-1 and hardcore-1 against CSR, as a function of the total number of points in the sample. Different ways to partition the observation window were used. Solid lines: 3 quadrats of increasing size (side length 0.5,0.7 and 1.0). Dashed lines: increasing number of quadrats (3×3, 4×4, 5×5, 6×6).

The right part of Figure 8 illustrates the variability of the individual estimates $\widehat{K}(r)$ by means of the pointwise minimum and maximum of $\widehat{K}(r) - \pi r^2$ in the groups.

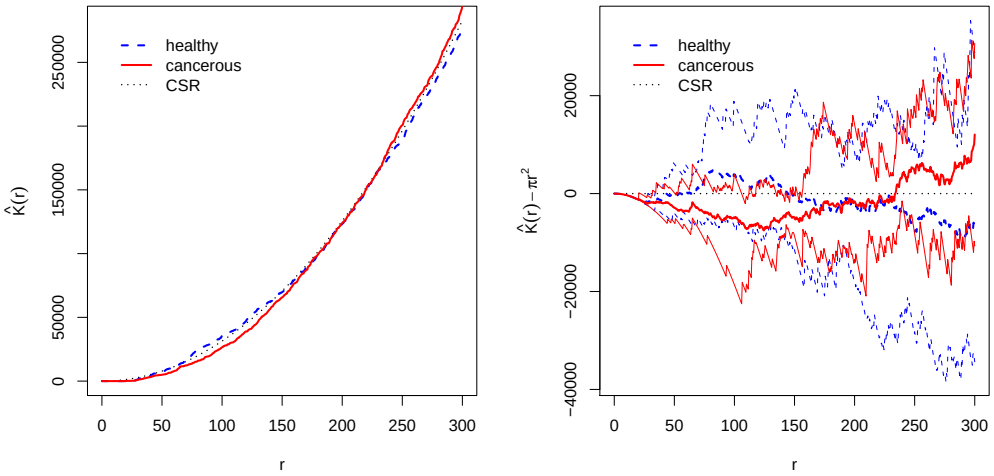


Figure 8: Estimated K-functions for the two point patterns. Left panel: mean values of $\widehat{K}(r)$ estimated on quadrates with at least 15 points. Right panel: mean values of $\widehat{K}(r) - \pi r^2$, strong lines: mean values, thin lines: range (pointwise minimum and maximum) of the individual estimates on the quadrats.

The tests based on the statistics T and U were carried out using integration bounds r_0 between $r_0 = 50\,\mu\text{m}$ and $r_0 = 300\,\mu\text{m}$. As it can be seen from Figure 9, the test based on U always yielded smaller p-values, i.e. higher significance, than the test based on T. In both cases, the first minimum, namely $p = 0.013$ for the test using T and $p = 0.009$ for the test using U, was attained for $140 \leq r_0 \leq 144\,\mu\text{m}$. Note that Ripley (1979) and Diggle (1979) recommend to use $r_0 = 1.25/\sqrt{\lambda}$ with similar statistics used for testing the CSR assumption, which amounts to $r_0 \approx 145\,\mu\text{m}$ in the present case. While the test based on T gave less significant results for higher values of r_0, the test based on U reached $p = 0.007$ for $r_0 \geq 299\,\mu\text{m}$.

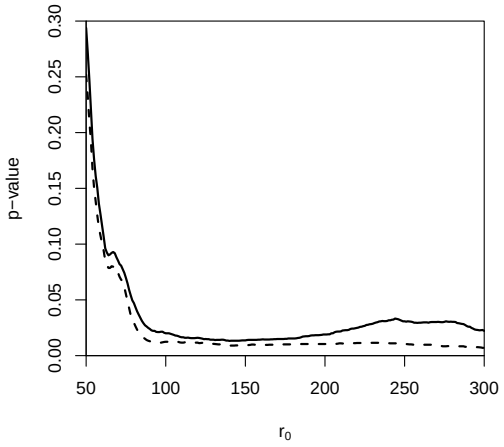


Figure 9: Resulting p-values of the tests using statistic T (solid line) and U (dashed line), as a function of the upper integration bound r_0.

For very small values of r_0, the test result is not statistically significant. Apparently, the two point patterns behave similar at short distances seen from a typical, i.e. randomly chosen, point. This corresponds to an observation reported by Mattfeldt et al. (2007), who found almost the same hard-core distance of about 25 µm in healthy and tumorous tissue. They explain their finding by the fact that the profile midpoints cannot come closer than the capillary diameter which is the same in both cases.

8 Conclusions

We have suggested a permutation test of equality of the K-functions in two groups of independently sampled point patterns, and used this test to distinguish between two single point patterns by subsampling quadrats. The permutation test requires that estimated K-functions on disjoint quadrats which belong to the same pattern are approximately exchangeable. A similar assumption is underlying the bootstrap method proposed by Loh and Stein (2004) to assess the variance of K-function estimates. The authors state that the estimates obtained on disjoint quadrats are approximately independent. The degree to which this is true depends on mixing properties of the point process — under the Complete Spatial Randomness (CSR) model, restrictions to disjoint quadrats are always independent. But even in cases where the null model is far from CSR, our simulation results showed that the empirical level of the tests was in good accordance with the theoretical significance level. One could criticize the lack of a formal method to check underlying mixing assumptions, however this also affects asymptotic tests for point processes such as Heinrich (1991) or Guan (2008) that rely heavily on mixing. Our test can always be made more robust towards long range dependence by leaving "guard zones" between the quadrats, which would however require correspondingly larger point patterns.

Our proposed test statistics use integral distances between the estimates of $K(r)$

over a certain interval $(0, r_0]$. The power of the test depends on the upper integration bound r_0. In the simulation study, this became particularly apparent for the test of slightly regular or slightly clustered point patterns against CSR, where the maximum power was reached for values r_0 less than the recommended $r_0 = 1.25/\sqrt{\lambda}$ for regular point patterns, and for values larger than $r_0 = 1.25/\sqrt{\lambda}$ for clustered point patterns. One possibility to improve the power of the tests consists in using a hierarchical procedure with a decision rule for r_0 based on a different diagnostic than the K-function, for example the Clark-Evans index (Clark and Evans, 1954) that allows to detect clustering or regularity. An alternative is to try to improve the test statistic by weighting the integrands depending on r. This was studied in detail by Ho and Chiu (2009) for simulation tests against the CSR hypothesis based on estimated L-functions, $L(r) = \sqrt{K(r)/\pi}$. They found that the optimal choice of weights depends on the range of interaction in the alternative model, that is, no general recommendation can be given.

The test principle can be generalized in many ways. Analogous permutation tests can be based on other point process characteristics such as the nearest-neighbor distance distribution $F(r)$, the empty space function $G(r)$ or the J-function (van Lieshout and Baddeley, 1996). By using tailor-made K-functions for inhomogeneous point processes, a similar test can be set up to distinguish between different types of inhomogeneity, as will be presented in a forthcoming paper by Baddeley, Hahn and Jensen (2011).

Acknowledgements

This research has been financially supported by the Danish Council for Strategic Research and the Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. The author cordially thanks Professor Eva B. Vedel Jensen for her encouragement and support, as well as for helpful discussions and valuable comments on the manuscript.

References

Baddeley, A., Hahn, U., and Jensen, E. B. V. (2011). Adaptive second order stationarity of spatial point patterns. In preparation.

Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3):329–350.

Baddeley, A. J., Moyeed, R. A., Howard, C. V., and Boyde, A. (1993). Analysis of a three-dimensional point pattern with replication. *Applied Statistics*, 42(4):641–668.

Besag, J. (1977). Discussion on Dr Ripley's paper. *Journal of the Royal Statistical Society, Series B*, 35(2):193–195.

Besag, J. and Diggle, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Applied Statistics*, 26:327–333.

Clark, P. J. and Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, revised edition.

Diggle, P. J. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics*, 35(1):87–101.

Diggle, P. J., Lange, N., and Benes, F. M. (1991). Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association*, 86(415):618–625.

Diggle, P. J., Mateu, J., and Clough, H. E. (2000). A comparison between parametric and non-parametric approaches to the analysis of replicated spatial point patterns. *Advances in Applied Probability*, 32(2):331–343.

Fisher, R. A. (1966). *Design of Experiments*. Oliver and Boyd, Edinburgh, 8th edition.

Gignoux, J., Duby, C., and Barot, S. (1999). 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(1):156–164.

Guan, Y. (2008). A KPSS test for stationarity for spatial point processes. *Biometrics*, 64(3):800–806.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47(2):757–762.

Heinrich, L. (1988). Asymptotic Gaussianity of some estimators for reduced factorial moment measures and product densities of stationary Poisson cluster processes. *Statistics*, 19(1):87–106.

Heinrich, L. (1991). Goodness-of-fit tests for the second moment function of a stationary multidimensional poisson process. *Statistics*, 22(2):245–268.

Ho, L. P. and Chiu, S. N. (2006). Testing the complete spatial randomness by Diggle's test without an arbitrary upper limit. *Journal of Statistical Computation and Simulation*, 76(7):585–591.

Ho, L. P. and Chiu, S. N. (2009). Using weight functions in spatial point pattern analysis with application to plant ecology data. *Communications in Statistics. Simulation and Computation*, 38(2):269–287.

Janssen, A. and Pauls, T. (2005). A Monte Carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems. *Computational Statistics*, 20(3):369–383.

Loh, J. M. and Stein, M. L. (2004). Bootstrapping a spatial point process. *Statistica Sinica*, 14:69–101.

Matérn, B. (1960). Spatial variation. *Meddelanden från Statens skogsforskningsinstitut*, 49(5):1–144. Second edition: Matérn (1986).

Mattfeldt, T., Eckel, S., Fleischer, F., and Schmidt, V. (2006). Statistical analysis of reduced pair correlation functions of capillaries in the prostate gland. *Journal of Microscopy*, 223(2):107–119.

Mattfeldt, T., Eckel, S., Fleischer, F., and Schmidt, V. (2007). Statistical modelling of the geometry of planar sections of prostatic capillaries on the basis of stationary Strauss hard-core processes. *Journal of Microscopy*, 228(3):272–281.

Ohser, J. and Stoyan, D. (1981). On the second-order and orientation analysis of planar stationary point processes. *Biometrical Journal*, 23(6):523–533.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:172–212.

Ripley, B. D. (1979). Tests of 'randomness' for spatial point patterns. *Journal of the Royal Statistical Society, Series B (Methodological)*, 41(3):368–374.

Ripley, B. D. (1984). Spatial statistics: Developments 1980-3, correspondent paper. *International Statistical Review /Revue Internationale de Statistique*, 52(2):141–150.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association*, 390:360–361.

Schladitz, K., Särkkä, A., Pavenstädt, I., Haferkamp, O., and Mattfeldt, T. (2003). Statistical analysis of intramembranous particles using freeze fracture specimens. *Journal of Microscopy*, 211(2):137–153.

Stoyan, D. and Stoyan, H. (2000). Improving ratio estimators of second order point process characteristics. *Scandinavian Journal of Statistics*, 27:641–656.

van Lieshout, M. N. M. and Baddeley, A. J. (1996). A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50(3):344–361.

Yamada, I. and Rogerson, P. A. (2003). An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical Analysis*, 35(2):97–109.

Appendix: Diggle et al.'s Test Under Heteroscedasticity

In a simulation study not shown here, we modified the test by Diggle et al. to be based on plain, i.e. non reweighted, residuals. That simplification resulted in improved behavior under heteroscedasticity. This finding suggests that reweighting of the residuals is one reason for the markedly increased deviation from uniformity of the p-values — recall that the heteroscedastic situations (different window size or different intensity) always implied big differences in the number of points, thus also very different weights. Apparently the distribution of the reweighted residuals does not reflect the true distribution of residuals, thus contributing to the poor performance of the test under heteroscedasticity.

As Equation (3) shows, the variance of the K-function estimator and thus also the variance of the residuals is rather inversely proportional to the squared number n of points than to n. Under this assumption, the bootstrapped test statistic based on residuals that are reweighted with $n^{-1/2}$ has a smaller mean than the corresponding statistic using the "proper weights" n^{-1}, unless if n is constant, that is if all point patterns contain the same number of points. It makes therefore no wonder that the rejection rate under the null hypothesis is higher when Diggle's weights are used.

For a proof of this assertion we would need to know the covariance structure of the empirical K-function, since the test statistic is built on integral distances. The idea can however be conveyed in a simpler situation, corresponding to evaluating $K(r)$ at only one r and using the theoretical mean for calculating the residuals. These "true" residuals are represented by independent random variables $X_1, \ldots, X_{m_1}, X_{m_1+1}, \ldots, X_{m_1+m_2}$ with zero mean. For sake of simplicity, we set $m_1 = m_2 = m$ in the following. Consider the test statistic

$$D = m \sum_{i=1}^{2} (M_i - \bar{M})^2 = \frac{m}{2}(M_1 - M_2)^2, \quad M_1 = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad M_2 = \frac{1}{m}\sum_{i=m+1}^{2m} X_i.$$

Its mean is

$$\mathbf{E}\, D = \frac{1}{2m}\sum_{i=1}^{2m} \operatorname{Var} X_i.$$

The bootstrap method proposed by Diggle et al. (1991) consists in random permutation of weighted residuals $w_i X_i, i = 1, \ldots, 2m$ that are reweighted after having been permuted. From the permutation $\pi : \{1, \ldots, 2m\} \to \{1, \ldots, 2m\}$ of the index set we obtain a bootstrapped statistic

$$D^*(\pi) = \frac{m}{2}(M_1^* - M_2^*)^2, \quad M_1^* = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{w_i} X_{\pi(i)} w_{\pi(i)}, \quad M_2^* = \frac{1}{m}\sum_{i=m+1}^{2m} \frac{1}{w_i} X_{\pi(i)} w_{\pi(i)}$$

with mean

$$\mathbf{E}\, D^*(\pi) = \frac{1}{2m}\sum_{i=1}^{2m} \frac{w_{\pi(i)}^2}{w_i^2} \operatorname{Var} X_{\pi(i)}.$$

The weights chosen by Diggle et al. were meant to be proportional to $\operatorname{Var} X_i^{-1/2}$, but since $\operatorname{Var} \hat{K}(r)$ is rather inversely proportional to the squared number of points than to the number of points, as supposed by the authors, the weights are in fact

approximately proportional to $\operatorname{Var} X_i^{-1/4}$. Assuming that $w_i = c \operatorname{Var} X_i^{-1/4}$ for some c, we get

$$\begin{aligned}
\mathbf{E}\, D^*(\pi) &= \frac{1}{2m} \sum_{i=1}^{2m} \frac{\left(\operatorname{Var} X_{\pi(i)}\right)^{-1/2}}{\left(\operatorname{Var} X_i\right)^{-1/2}} \operatorname{Var} X_{\pi(i)} \\
&= \frac{1}{2m} \sum_{i=1}^{2m} \sqrt{\operatorname{Var} X_{\pi(i)} \operatorname{Var} X_i} \\
&\leq \frac{1}{2m} \sum_{i=1}^{2m} \frac{\operatorname{Var} X_{\pi(i)} + \operatorname{Var} X_i}{2} = \mathbf{E}\, D,
\end{aligned}$$

the inequality being strict iff $\operatorname{Var} X_{\pi(i)} \neq \operatorname{Var} X_i$ for some $i \in \{1 \dots 2m\}$.