Testing for significant differences between two spatial patterns using covariates

M.I. Borrajo[a,*], W. González-Manteiga[a], M.D. Martínez-Miranda[b]

[a] *Department os Statistics, Mathematical Analysis and Optimisation, University of Santiago de Compostela*
[b] *Department of Statistics and Operations Research, University of Granada*

Abstract

This paper addresses the problem of comparing the spatial distribution of two point patterns. A formal statistical test is proposed to decide whether two observed patterns share the same theoretical intensity model. This underlying model assumes that the first-order intensity function of the process generating the patterns may depend on covariate information. The test statistic consists of an L_2-distance between two kernel estimators for the corresponding relative density, which is shown to be asymptotically normal under the null hypothesis assuming that the underlying process is Poisson. In practice a suitable bootstrap method is proposed to calibrate the test. Simulations are used to explore the ability of the proposed test to identify different spatial patterns. An application to the analysis of wildfires in Canada shows the practicality of the proposal, with appealing conclusions regarding to the need of including covariate information.

Keywords: Spatial point processes, two-sample problem, covariates
2010 MSC: 62G10, 62G20, 60G55, 60-02

1. Introduction

In many real-world scenarios spatial point process data may be obtained from two populations. This is for example the case of plant locations, divided into two populations based upon the species [28]. The location of neurons within the brain depending on whether the individual suffers from

*mariaisabel.borrajo@usc.es

mental illness [14]. The morphologies of intracellular structures under different experimental conditions [16]. The spatial location of wildfires in a region classified according to their cause or their type ([18]). In these cases two point pattern distributions are observed within the same region, and one appealing question is how to quantitatively compare those two distributions. This translates the classical two-sample problem to the context of spatial point processes.

[18] discussed that testing whether two spatial point patterns share the same spatial structure can be done through comparison of first-order properties, since these properties describe the spatial distribution of events in the observation region. Assuming that the underlying process is Poisson, the authors suggest an extension of the two-sample multivariate density problem proposed in [16] where the comparison is based on the density of event locations. Recently different studies have introduced area-based tests to address this problem, see [2] and [1].

For a spatial point process X defined in a planar region $W \in \mathbb{R}^2$, the first-order intensity function is defined as

$$\lambda(x) = \lim_{|dx|\to 0} \frac{\mathbb{E}[N(dx)]}{|dx|},$$

where $|dx|$ denotes the area of an infinitesimal region containing the point x. It represents the expected number of events per unit area and it fully characterises the spatial distribution of a Poisson point process. Its estimation has been widely analysed assuming parametric models and using likelihood or pseudo-likelihood procedures, see for example [36], [29],[37], [21] and [22]. But also nonparametric methods have been proposed, see [11], [9], [17] and [4].

The natural spatial variation is sometimes not enough to describe some phenomena and considering additional covariate information is required. [13] described an example where the incidence of some type of cancer near nuclear installations is analysed. This paper formulates a model where the intensity function depends not only on the spatial location of the events but also on a covariate (the distance to a specified point source). It is a simple multiplicative model which belongs to the class of modulated Poisson processes introduced by [8].

The inclusion of covariates in the first-order intensity modelling was first formalised in [3] through the model:

$$\lambda(u) = \rho(Z(u)),\ u \in W \subset \mathbb{R}^2, \tag{1}$$

where $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ is a spatial continuous covariate that is exactly known in every point of the region of interest W. [3] proposed likelihood based and nonparametric kernel estimators for the ρ function. In [4] a consistent theoretical framework for this estimator was detailed for the first time, including a bootstrap procedure, some specifically designed bandwidth selection methods and an extension to deal with mutidimensional covariates. Under the Poisson assumption, the goodness-of-fit of model (1) has been addressed in [5], and the extension to case of multivariate covariates has also been described. It makes this model particularly appealing for many real data applications where the spatial distribution of the events can be better described using some important covariates. In some cases a covariate can completely describe the phenomenon (see [5]) and in general covariates add necessary information to the natural spatial variation.

The aim of this paper is to test whether two observed spatial point patterns share the same spatial structure. This was the same objective of [18] where the Poisson assumption was used to calibrate the test. We closely follow the approach of [18] and focus on comparing the first-order intensities of patterns assuming an inhomogeneous Poisson point process. However, different from this former paper, we assume that the underlying intensity may depend on covariates and it is estimated using the consistent kernel estimator of [4]. Our approach has some advantages with respect to [18]. There are situations where the spatial distribution can be described using a single one-dimensional covariate, in these cases assuming model (2) would lead to a dimension reduction that notably simplifies technical issues such as the bandwidth selection problem, and reduces the computational burden. Moreover the comparison problem may sometimes be better described using covariate information (possibly multidimensional), which might be crucial to distinguish between two observed spatial patterns.

The rest of the paper is organised as follows. In Section 2 we formulate the testing problem and describe our proposal. First we derive the test statistic and its asymptotic normality and then we suggest a convenient bootstrap method to accomplish its calibration in practice. For simplicity in the exposition we assume that the covariate is one dimensional but our proposal easily extends to more dimensions, as shown later in the data applications. Section 3 describes a simulation study to evaluate the finite-sample performance of the proposed test. In Section 4 we illustrate our proposal and some appealing extensions with a data application on wildfires in Canada. Final conclusions are drawn in Section 5. Proofs are deferred to the Appendix.

2. The proposed method

We propose a formal test to check whether two given patterns are originated by the same stochastic process, in terms of their spatial structure. To this goal we assume an intensity model such as the one formulated in (1), where the theoretical intensity depends on a known covariate, but within a multidimensional scenario for the covariates:

$$\lambda(u) = \rho(\mathbf{Z}(u)),\ u \in W \subset \mathbb{R}^2, \tag{2}$$

where $\mathbf{Z} = (Z^1, \ldots, Z^p)$ and every $Z^j : W \subset \mathbb{R}^2 \to \mathbb{R}$ is a one-dimensional covariate fulfilling the same conditions as the original Z in (1). We define an L_2-distance based test statistic to detect differences between two observed patterns under this more flexible model.

2.1. The test

Let X_i with $i = 1, 2$ be two point processes defined in a region $W \subset \mathbb{R}^2$, where W is assumed to have finite positive area. Let $X_{11}, \ldots, X_{1N_1}$ and $X_{21}, \ldots, X_{2N_2}$ be two realisations of the processes where N_i are the random variables counting the number of events. Let $\mathbf{Z} = (Z^1, \ldots, Z^p) : W \subset \mathbb{R}^2 \to \mathbb{R}^p$ be a multidimensional spatial continuous covariate which is exactly known in every point of W. In practice this covariate will commonly be known in an enough amount of points spread over the region, so the values for the rest of the points can be interpolated and that these values are indeed the real ones, see [3] for more details. We warn the reader at this point to be careful about the super and sub-index notation: while the sub-indexes are always referring to the two underlying processes, the super-indexes denote the components of a vector.

Under model (2), instead of looking at the natural spatial point processes X_i, we look at the transformed ones through the covariate, $\mathbf{Z}_i = \mathbf{Z}(X_i)$. Particularly, we observe two patterns $\mathbf{Z}_{1i} = \mathbf{Z}(X_{1i}) \in \mathbb{R}^p$ and $\mathbf{Z}_{2j} = \mathbf{Z}(X_{2j}) \in \mathbb{R}^p$ with $i = 1, \ldots, N_1$ and $j = 1, \ldots, N_2$, which consist of the values of the covariate measured at the events locations, leading to new processes. Let denote by $\lambda_i(x) = \rho_i(\mathbf{Z}(x))$ the intensity functions corresponding to the processes X_i, with $i = 1, 2$. [4] provides a useful result, in their Further Extensions section, stating that having a spatial point process X fulfilling condition (2) and a spatial covariate, $\mathbf{Z}$, then $\mathbf{Z}(X)$ is a point process with intensity of the form $\rho(\cdot)g^\star(\cdot)$, where $g^\star(\cdot) = |W|g(\cdot)$, with g being the derivative of the spatial cumulative distribution function of $\mathbf{Z}$, i.e.,

$G'(z) = g(z)$, with $G(z) = \frac{1}{|W|}\int_W 1_{\{\mathbf{Z}(u)\leq z\}}du$, where $\mathbf{Z}(u) \leq z$ refers to $(Z^1(u) \leq z^1) \cap \ldots \cap (Z^p(u) \leq z^p)$, with z^j the components of the vector $z \in \mathbb{R}^p$.

To compare the two spatial patterns we formulate the null hypothesis $H_0 : f_1(z) = f_2(z)$, $z \in \mathbb{R}^p$, versus a two-sided alternative, where $f_i(z) = \frac{g^\star(z)\rho_i(z)}{m_i}$ are the relative densities, and $m_i = \int_W \lambda_i(x)dx$ is the expected number of events in each of the processes. Remark that this does not really need to be in $\mathbb{R}^p$ but in a subset covering the range of values of the covariate $\mathbf{Z}$. Notice also that we are comparing the relative densities instead of the intensities because we are interested in the spatial structure and not the total number of events, i.e., two patterns one double size of the other might have the same spatial structure but it is impossible for them to have the same intensity because of the number of events. Moreover, the first-order intensity functions of spatial point patterns with the same structure are proportional and therefore they have the same relative density (see [9] and [18]).

To define the test statistic we consider a discrepancy measure or distance between the two theoretical relative densities. Specifically we consider an L_2-distance defined as

$$\begin{aligned} D &= \int_{\mathbb{R}^p} \left(f_1(z) - f_2(z)\right)^2 dz \\ &= \int_{\mathbb{R}^p} f_1^2(z)dz + \int_{\mathbb{R}^p} f_2^2(z)dz - \int_{\mathbb{R}^p} f_1(z)f_2(z)dz - \int_{\mathbb{R}^p} f_2(z)f_1(z)dz \\ &= \mathbb{E}_{\mathbf{Z}_1}\left[f_1(\mathbf{Z}_1)\right] + \mathbb{E}_{\mathbf{Z}_2}\left[f_2(\mathbf{Z}_2)\right] - \mathbb{E}_{\mathbf{Z}_2}\left[f_1(\mathbf{Z}_2)\right] - \mathbb{E}_{\mathbf{Z}_1}\left[f_2(\mathbf{Z}_1)\right] \\ &= \psi_{11} + \psi_{22} - \psi_{12} - \psi_{21}. \end{aligned} \tag{3}$$

From the observed patterns the test statistic T is derived using the multivariate estimator defined in [4] which leads to the expressions:

$$T = \widehat{\psi}_{11} + \widehat{\psi}_{22} - \widehat{\psi}_{12} - \widehat{\psi}_{21}, \tag{4}$$

where

$$\widehat{\psi}_1 = \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\frac{g^\star(\mathbf{Z}_{1i})}{g^\star(\mathbf{Z}_{1j})}\mathbf{K}_{H_1}(\mathbf{Z}_{1i}-\mathbf{Z}_{1j})1_{\{N_1\neq 0\}},$$

$$\widehat{\psi}_2 = \frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2}\frac{g^\star(\mathbf{Z}_{2i})}{g^\star(\mathbf{Z}_{2j})}\mathbf{K}_{H_2}(\mathbf{Z}_{2i}-\mathbf{Z}_{2j})1_{\{N_1\neq 0\}},$$

$$\widehat{\psi}_{12} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{g^\star(\mathbf{Z}_{2j})}{g^\star(\mathbf{Z}_{1i})}\mathbf{K}_{H_1}(\mathbf{Z}_{2j}-\mathbf{Z}_{1i})1_{\{N_1\neq 0,N_2\neq 0\}},$$

$$\widehat{\psi}_{21} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{g^\star(\mathbf{Z}_{1i})}{g^\star(\mathbf{Z}_{2j})}\mathbf{K}_{H_2}(\mathbf{Z}_{1i}-\mathbf{Z}_{2j})1_{\{N_1\neq 0,N_2\neq 0\}},$$

with H_i being bandwidth matrices, $\mathbf{K}$ a p-dimensional kernel function, $\mathbf{K}(z) = |H|^{-1}\mathbf{K}(H^{-1/2}z)$ with $|H|$ denoting the determinant of the matrix H and $1_{\{\cdot\}}$ denoting the indicator function.

2.2. Asymptotic properties and calibration under the Poisson assumption

In this section we first derive the asymptotic distribution of the statistic (4) under a suitable framework. For simplicity we develop the asymptotic theory for $p = 1$. The result can be directly transferred to $p > 1$ due to the ideas and techniques applying to the multidimensional framework, taking into account that we would be working with bandwidth matrices and p-dimensional kernel functions instead of scalar ones. This will just increase the complexity of the notation and does not contribute to the understanding of the reader. Afterwards we propose a bootstrap method to calibrate the test in practice with multidimensional covariates.

In point processes we may find at least two different types of asymptotics: the increasing domain, see [20], where the expected number of events tends to infinity with the increasing size of the observation region (remark that with this idea, "new points", i.e., extra information, is only given in the boundary of the region and the estimated intensity at each point depends on an expected number of events tending to zero); and infill structure, initially proposed by [12], which overcomes the previously detailed problem stating that the expected number of events tends to infinity for a fixed bounded observation domain. In this work we consider the second option following [7], [9], [33], [17] and [4].

Assume now that $p = 1$, we need to introduce some regularity conditions common in the nonparametric field for the development of asymptotics in kernel techniques:

(A.1) $\int_{\mathbb{R}} K(z)dz = 1$; $\int_{\mathbb{R}} zK(z)dz = 0$ and $\mu_2(K) := \int_{\mathbb{R}} z^2K(z)dz < \infty$, with K a unidimensional kernel function.

(A.2) $\lim_{m\to\infty} h_i = 0$ and $\lim_{m\to\infty} \frac{A(m_i)}{h_i} = 0$, where h_i is a scalar bandwidth and $A(m_i) := \mathbb{E}\left[\frac{1}{N_i}1_{\{N_i\neq 0\}}\right]$.

(A.3) G and the densities of events location are three times differentiable.

(A.4) $Z(x)$ is a continuity point of ρ_i for all $x \in W$.

The following theorem establishes the asymptotic distribution of T under the null hypothesis.

Theorem 2.1. *Given inhomogeneous Poisson point patterns $X_i, i = 1, 2$ and denoting $f_i(z) = \frac{g^\star(z)\rho_i(z)}{m_i}$ the relative densities associated to each process. Under conditions (A.1) to (A.4) and assuming the null hypothesis $H_0 : f_1(z) = f_2(z)$ for all $z \in \mathbb{R}$, it holds*

$$\frac{T - \mu_T}{\sigma_T} \longrightarrow N(0, 1),$$

where

$$\mu_T = (A(m_1)h_1 + A(m_2)h_2)\, K(0) + o(A(m_1)) + o(A(m_2)) \text{ and}$$

$$\begin{aligned}\sigma_T = {} & 2B(m_1)\frac{1}{h_1}R(K)\psi + 2B(m_2)\frac{1}{h_2}R(K)\psi + A(m_1)A(m_2)\psi R(K)\left(\frac{1}{h_1} + \frac{1}{h_2}\right) \\ & + A(m_1)A(m_2)\psi\left(\frac{1}{h_1}\int K(u)K_{h_2/h_1}(u)du + \frac{1}{h_2}\int K(u)K_{h_1/h_2}(u)du\right) \\ & + O(B(m_1) + O(B(m_2))),\end{aligned}$$

with $\psi \equiv \psi_{11} \equiv \psi_{22} \equiv \psi_{12} \equiv \psi_{21}$ under the null, $K_l(\cdot) = \frac{1}{l}K\left(\frac{\cdot}{l}\right)$, $R(K) := \int K^2(u)du$ and $B(m_i) = \mathbb{E}\left[\frac{1}{N_i^2}1_{\{N_i\neq 0\}}\right]$.

The proof of this result is fully detailed in the Appendix.

In principle the asymptotic distribution provided in Theorem 2.1 can be used to calibrate the test in practice. To this aim we would need to estimate the unknown quantities involved in the above expression. To estimate μ_T and σ_T^2, m_i could be replaced by the sample size n_i, and $A(m_i)$ by $1/n_i$ (see [9]). However, the asymptotic distribution is not the best way to calibrate the test in practice due to the slow convergence rate, which might lead to wrong conclusions. There exists several examples in different context with a similar problem, see [35] for regression models in mean, [15] for the two-sample problem in density estimation, [19] for directional data, [18] for spatial point process without covariates, and [25] for quantile regression models. In this paper we suggest bootstrap to perform the calibration of the test, but other procedures have also been used for the same purpose, for example, the permutation test in [10].

Our bootstrap method is based on the one defined in [4], which was inspired in [6] and [7] for our multidimensional covariate scenario. The idea is to generate bootstrap patterns under the null hypothesis to derive the empirical distribution of the test statistic under the null. The resampling scheme assuming that the underlying process is Poisson consists of the following steps:

1. Consider $X_{11}, \ldots, X_{1N_1}$ and $X_{21}, \ldots, X_{2N_2}$ as a unique pattern coming from the same process.
2. Conditional on this joint pattern, $X_1, \ldots, X_{N_1+N_2}$, let $N^* \sim Pois\left(\int_W \hat{\rho}_I(\mathbf{Z}(x))dx\right)$, where $\hat{\rho}_I = \sum_{i=1}^{N_1+N_2} \frac{1}{g^\star(\mathbf{Z}(X_i))}\mathbf{K}_I\left(\mathbf{Z}(x) - \mathbf{Z}(X_i)\right)$ and I a bandwidth matrix.
3. Generate two realisations, n_1^* and n_2^*, from the random variable N^*.
4. Draw two independent patterns, $X_{11}^*, \ldots, X_{1n_1^*}^*$, and $X_{21}^*, \ldots, X_{2n_2^*}^*$, by sampling from the distribution with density $\frac{\hat{\rho}_I(\mathbf{Z}(x))}{\int_W \hat{\rho}_I(\mathbf{Z}(x))dx}$.

To perform the calibration of the test, we repeat this algorithm B times, with B large enough. For each pair of bootstrap samples we compute the test statistic (4), obtaining B values of it. Then we compute, for a level α of our test, the empirical $(1-\alpha)$-quantile of the B values of the statistic and this becomes our critical level. Hence, given the two original patterns, we compute the test statistic (4) and we reject the null hypothesis if this value is higher than the computed empirical quantile of the bootstrap replications.

3. Simulations

We analyse the performance of the test described in the previous section through Monte Carlo simulations. To reduce the computational cost of these simulations we are again showing them for $p = 1$. Moreover, to keep as close as possible to practice we have defined two simulation scenarios that fulfil model (1) and they are based on two real datasets.

The first dataset consists of 255 locations of gold deposits and the surrounding geological faults in a region of $330km \times 394km$ located in the Murchison area of Western Australia (see for example [3]). At this scale (1:500000) the gold deposits spatial extension is negligible and they can be considered as points without losing generality. Note that the real gold deposits and faults are three-dimensional while here we use a two-dimensional projection. Moreover, some geological faults may have been missed because they are not recorded by direct observation but in magnetic field surveys or geologically inferred from discontinuities in the rock sequences. A representation of the data can be seen in Figure 1.

Figure 1: Murchison geological survey data: gold deposits (points) and geological faults (lines).

The covariate is defined as the distance from every point in the observation region to the nearest fault (see Figure 2).

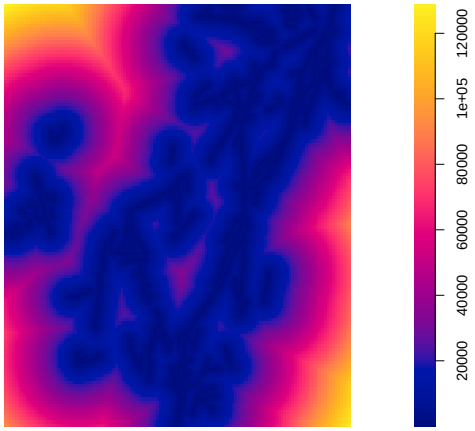

Figure 2: Covariate information for the Murchison dataset: distance to the nearest geological fault (in kilometres).

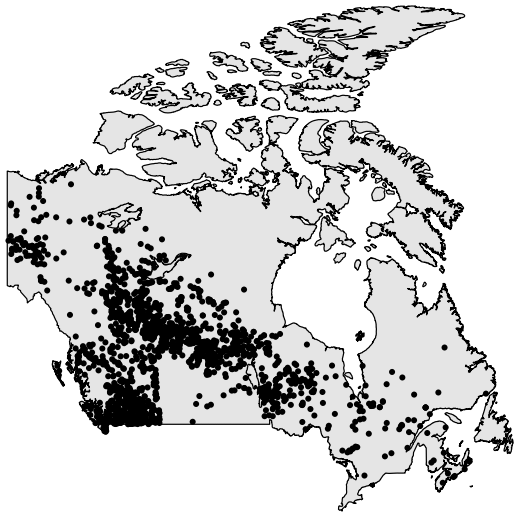

Figure 3: Wildfires in Canada during June 2015.

The second model is related to one of the most important natural disturbances, wildfires. We use the wildfire records in Canada during June 2015 that are available at the Canadian Wildland Fire Information System website (http://cwfis.cfs.nrcan.gc.ca/home), see Figure 3. The fire season in Canada lasts from late April until August, with a peak of activity in June and July; based on the existing literature the main cause of these fires relies

on meteorological conditions. Hence, we have obtained meteorological data of Canada, particularly daily temperature data for the whole month, and we have constructed a covariate for the model defined as the third quartile of the maximum daily temperature during the month of June (in order to avoid extreme values that may interfere in the analysis), see Figure 4.

Figure 4: Third quartile of the temperature registered in June 2015 in Canada, after a gaussian smoothing with $\sigma = 2$ (in Celsius degrees).

From the observed data and the defined covariates, the theoretical intensities for the simulations have been derived under (1) by the nonparametric kernel intensity estimator of [4], with the bootstrap bandwidth estimate defined in the same work. Other bandwidth estimates could be considered to this purpose such as cross-validation, or a simple rule-of-thumb. Here we have chosen the bootstrap bandwidth estimate because it is a consistent estimate that does not involve intense computations, and it has excellent finite-sample properties (see the simulation study in [4]). The resulting theoretical intensity functions for the two simulation models can be seen in Figure 5 and Figure 6, respectively. Remark that in both scenarios intensities are then functions of a single known covariate (in the first scenario the distance to the nearest fault and in the second the temperature).

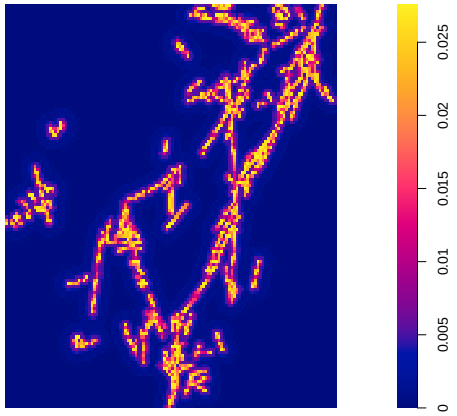

Figure 5: Theoretical intensity function for the first model analysed in the simulation study, that has been obtained applying a kernel intensity estimator to the Murchison dataset.

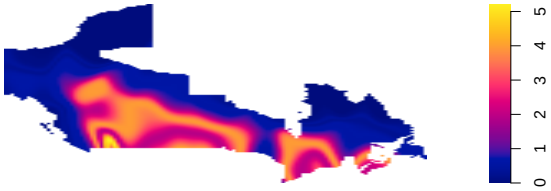

Figure 6: Theoretical intensity function for the second model analysed in the simulation study, that has been obtained applying a kernel intensity estimator to the Canada wildfire dataset.

The simulation study is divided in two parts, the first one devoted to analyse the level values of the test, i.e., the rejection proportions of the null hypothesis in a situation where the null hypothesis is true; and the second to evaluate the power, i.e., the rejection proportions of the null hypothesis in a situation where it is false. In the first part we have simulated $M = 10000$ samples to better estimate the low quantile $\alpha = 0.05$, while in the second we have found that $M = 5000$ is enough since the results are stable. We have considered, for each of the patterns in the comparison, six different pairs of expected sample sizes, $m_1 = m_2 = 50, 100, 200, 500$; $m_1 = 300$ and $m_2 = 700$; $m1 = 100$ and $m_2 = 900$. For the bootstrap calibration we have drawn $B = 500$ bootstrap samples. In all cases we simulate the samples under the Poisson assumption.

A technical problem that is common to all nonparametric tests based kernels is the choice of the bandwidth. Unfortunately this is still an open question and there is no bandwidth selection procedure specifically designed for testing. In the literature the problem has been usually treated in either of the two following ways: use an automatic data-driven bandwidth selector which is good for the involved kernel estimators (see [18] for example in spatial point processes); or compute the test using a suitable range of possible bandwidths (see [5]).

Our test involves the choice of three bandwidths: two of them to estimate the addends in (4), previously denoted in the paper as H_1 and H_2, and a third one to estimate the model under the null hypothesis and perform the bootstrap calibration detailed at the end of the previous section, denoted by I. In our simulations, as we are restricted to $p = 1$, H_1 and H_2 are scalar bandwidths, those are bandwidths used in the kernel estimation, so following our discussion above we have used the bootstrap bandwidth selector proposed in [4]. Remark that the use of a cross-validation in this stage becomes infeasible due to its computational burden. For the third one, I, which in our simulations is also a scalar value, we consider bandwidths which produce a test with the correct nominal level. For the Murchison dataset, again the bootstrap bandwidth selector of [4] seems to produce accurate values, while for the Canada wildfires model we have used a grid of bandwidths in the range $(0.005, 0.3)$, which contains the bootstrap bandwidth. The results obtained for the level of the test in each model can be seen respectively in Table 2 ($d_M = 0$) and Table 1. In the first case we show the results only for the bootstrap bandwidth selector and in the second, for different bandwidths in considered range. In both cases the test seems to be well calibrated in the sense that the observed proportion of rejections is close to the nominal level of 0.05. The only exception might be the scenario with a very unbalanced design in the sample sizes, where the proportions are a bit lower than desirable, not reaching the value 0.05. However this just happens when this design is very unbalanced, in the previous scenario, where one sample size doubles the other, the level values are still around 0.05 in both models.

	$I = 0.005$	$I = 0.05$	$I = 0.1$	$I = 0.3$
$m_1 = m_2 = 50$	0.0513	0.0522	0.0526	0.0552
$m_1 = m_2 = 100$	0.0497	0.0509	0.0510	0.0531
$m_1 = m_2 = 200$	0.0483	0.0487	0.0481	0.0534
$m_1 = m_2 = 500$	0.0473	0.0477	0.0486	0.0557
$m_1 = 300$, $m_2 = 700$	0.0415	0.0456	0.0420	0.0540
$m_1 = 100$, $m_2 = 900$	0.0318	0.0328	0.0331	0.0460

Table 1: Rejection proportions under the null hypothesis for the Canada model, with six different pairs of expected samples sizes, and bandwidth values I in a suitable range.

We now look at the power of the test. We have constructed alternatives that fulfil assumption (1) in the following way. First, for both Murchison and Canada models we have built a new covariate, function of the initial one and varying on a real parameter, $d_\bullet$. This parameter determines the distance between the null and the alternative hypotheses. As the parameter increases, the model is further away from the null, and as the parameter decreases to zero the model approaches the null. Second, we construct the theoretical intensity by computing the kernel estimator of [4] with the new covariate. In the Murchison example, the new covariate is defined as $d_M e^{-Z(x)} + Z(x)$, and in the Canada model as $d_C \frac{1}{Z(x)} + Z(x)$, where $Z(x)$ denote in each case the corresponding initial covariate, $d_M = 0.5, 2, 5, 10$ and $d_C = 50, 100, 200, 300$. Particularly, $d_M = 0$ and $d_C = 0$ lead to the null hypothesis.

To derive the power of the test we generate at each iteration two-samples, one from the model with parameter $d = 0$ and another from the model with $d > 0$. These two samples satisfy the alternative hypothesis so we expect that the test would reject the null hypothesis. We repeat this $M = 5000$ times, computing from each generated samples the proposed test and computing the number of rejections which gives the empirical power of the test. The results are shown in Table 2 ($d_M > 0$) and Table 3 ($d_C > 0$). Notice that in the first model we keep using the bootstrap bandwidth for the test, and in the second one we have used a range of possible values which provided a well calibrated test (see the discussion of the level above). However for simplicity we report only one of the considered bandwidth values for the second model since the rest provide similar results. From the results in these tables we can observe better power values for the first model, where even for $d_M = 2$ (a

situation near to the null) the proportion of rejections is high for medium and large sample sizes. In the second model, the values do not reach those levels. However, when we are further enough from the null hypothesis the power increases notably, reaching the 100% for large sample sizes. Remark that again, in the very unbalanced sample size scenario, the power values are around a half, or slightly more than a half, than the ones that we obtained with the same total sample size in the less unbalanced design or in the case where both samples have the same size.

	$d_M = 0$	$d_M = 0.5$	$d_M = 1$	$d_M = 1.5$	$d_M = 2$
$m_1 = m_2 = 50$	0.0518	0.0944	0.1990	0.3406	0.4823
$m_1 = m_2 = 100$	0.0521	0.1267	0.3750	0.7083	0.9234
$m_1 = m_2 = 200$	0.0528	0.2001	0.7296	0.9946	1
$m_1 = m_2 = 500$	0.0512	0.4651	0.9993	1	1
$m_1 = 300$, $m_2 = 700$	0.0435	0.4482	0.9909	0.9910	1
$m_1 = 100$, $m_2 = 900$	0.0384	0.2889	0.8516	0.9996	1

Table 2: Rejection proportions for the Murchison model under the null hypothesis ($d_M = 0$) and the alternative hypothesis ($d_M > 0$), were the different values of the parameter d_M control the discrepancy from the null, and six pairs of expected sample sizes.

	$d_C = 0$	$d_C = 50$	$d_C = 100$	$d_C = 200$	$d_C = 300$
$m_1 = m_2 = 50$	0.0526	0.0604	0.0661	0.0918	0.3695
$m_1 = m_2 = 100$	0.0510	0.0576	0.0714	0.2462	0.8848
$m_1 = m_2 = 200$	0.0481	0.0612	0.0794	0.8940	1
$m_1 = m_2 = 500$	0.0486	0.0691	0.1128	1	1
$m_1 = 300$, $m_2 = 700$	0.0420	0.0632	0.0836	1	1
$m_1 = 100$, $m_2 = 900$	0.0331	0.0618	0.0725	0.4234	0.8726

Table 3: Rejection proportions for the Canada model under the null hypothesis ($d_C = 0$) and alternative hypothesis ($d_C > 0$), were the different values of the parameter d_C control the discrepancy from the null, and six pairs of expected sample sizes.

To provide the reader with a quick visualisation of the simulation results, we have included two graphics showing the rejection proportions in the considered scenarios with balanced designs. These go from the case where the null hypothesis is true to the scenario that is the furthest away from it, see Figure 7.

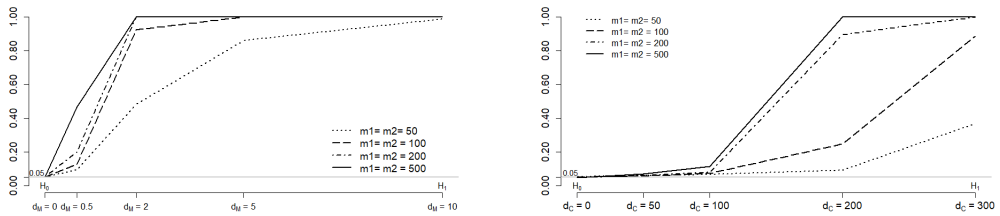

Figure 7: Representation of the rejection proportions for the different simulated scenarios of the Murchison model (left) and the Canada model (right).

4. Application to data on wildfires

In the simulation study detailed in the previous section we have defined two models based on real data: gold deposits and wildfires. The latest represents one of the most important natural disturbances in the world, which affects and damages several regions around the globe. Here we describe a data application of our proposal using the Canada wildfire dataset.

We are interested in determining whether wildfires in Canada have suffered a change in their spatial distribution in the last decades. Hence, we have selected the wildfire occurrences in June 1980 (a total number of 1207) and June 2015 (a total number of 1841); we have particularly focused on the ones taking place during the month of June because this is a peak of activity in the fire season in Canada. Moreover, we have also included in this analysis wildfire occurrences during June 2014 to compare two consecutive recent years and determine whether the spatial distribution is now changing. We suspect that this could be the case since in June 2014 there were "only" 950 wildfires and in the same month one year later this quantity has almost doubled up to 1841, but recall that the difference in the sample size does not necessary imply a change in the spatial distribution. The three resulting datasets are represented in the simplified Canada map in Figure 8. And our aim is to perform two comparisons: first to compare the spatial structure of

wildfires in June 1980 and June 2015, and second the two consecutive years, June 2014 and June 2015.

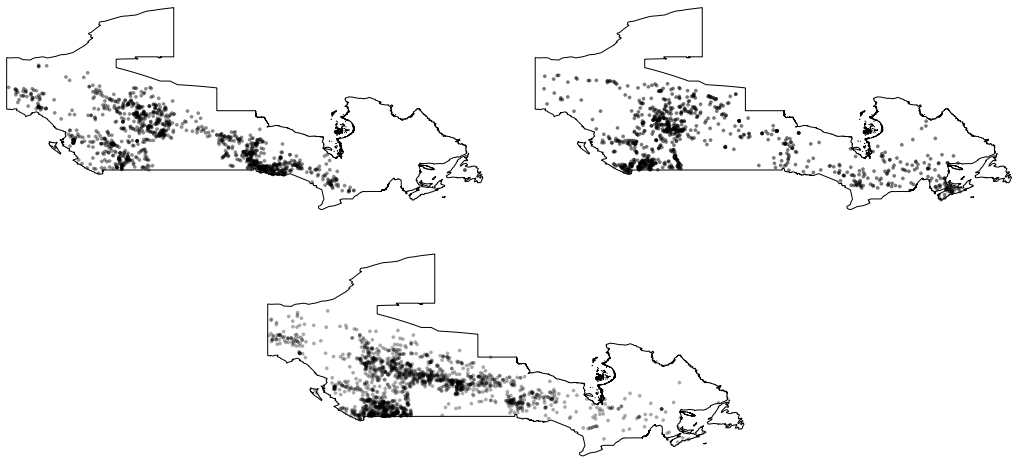

Figure 8: Wildfire locations in Canada during June 1980 (top left), June 2014 (top right) and June 2015 (bottom).

.

To perform the comparisons we first define the underlying intensity model of the type specified in (1), under which we have formulated the test. To start with, we look for a model with one single covariate. As Canadian forest fires are mainly caused by meteorological reasons, the temperature seems to be relevant to describe the wildfires occurrences. So we start defining a model where the intensity is explained just with this covariate, specifically the third quartile of the temperature for the observed patterns in chronological order, see Figure 9.

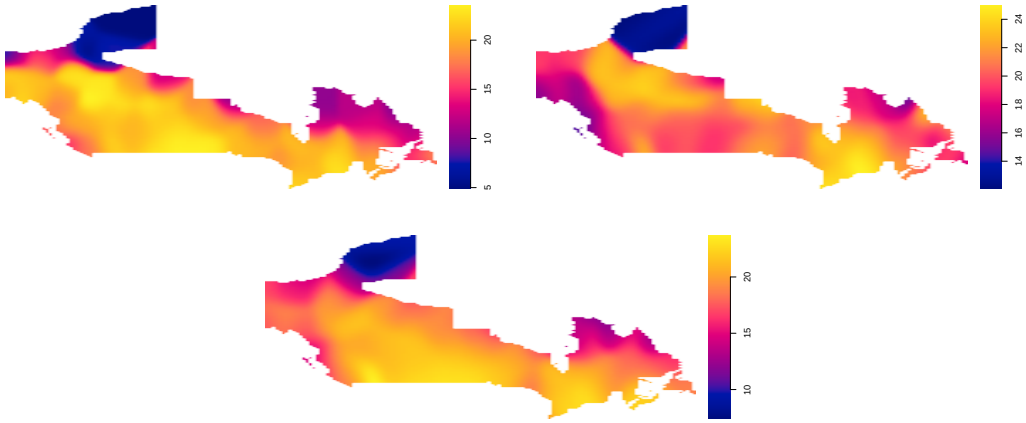

Figure 9: Third quartile of the temperature measured during June 1980 (top left), June 2014 (top right) and June 2015 (bottom).

Under the Poisson assumption we can check the goodness-of-fit of the formulated model (1) using the test in [5]. In this paper an L_2-distance based test statistic is proposed to check the $H_0 : \lambda(u) = \rho(Z(u)), u \in W \subset \mathbb{R}^2$ versus a general alternative where the intensity is not defined through the covariate Z. It turns out that this model is not appropriate with a p-value about zero for the three datasets. Our conclusion is that the temperature is not enough to explain the wildfires occurrence. So we add more information and consider the spatial coordinates as additional covariates. Model (2) is now needed for the case of a three-dimensional covariate $\mathbf{Z} : W \subset \mathbb{R}^2 \to \mathbb{R}^3$, with particularly $\mathbf{Z}(x, y) = (x, y, Z(x, y))$, where x is the latitude, y is the longitude and Z is the third quartile of the temperature. As before, in [5], they propose also a multidimensional version on the previous procedure, so we use this goodness-of-fit test to check the new intensity model. The null hypothesis is now $H_0 : \lambda(x, y) = \rho(\mathbf{Z}(x, y)) = \rho(x, y, Z(x, y))$, versus a general alternative where the intensity depends only on the locations. In this case the model is fulfilled for the three point patterns with p-values of 0.946, 0.984 and 0.854, respectively in chronological order. Our conclusion is that the temperature contributes with new useful information in the modelling of the wildfires distribution in Canada for the three considered months, moreover the temperature is necessary to describe the wildfire occurrences.

We can now apply the two-sample test defined in Section 2 under the just defined three-dimensional intensity model. The test is generalized to this case considering the multivariate version of the kernel intensity estimator

proposed in [4], and a suitable multivariate version of the test statistic given in (4). Structurally the statistic remains the same, the only difference relies on using multivariate kernel estimators and bandwidth matrices for each of the addends instead of the one-dimensional version. Hence, we are dealing now with multivariate kernel functions and bandwidth matrices, where the theoretical developments can be replicated under the appropriate smoothing assumptions. In practice, we are using the Gaussian kernel and the plug-in bandwidth selector by [38] for dimension 3, available in the R-package `ks`.

With this multivariate extension of the test we compare the wildfire spatial structure in 1980 and 2015 and we obtain a p-value of zero. We reject the null hypothesis and conclude that the spatial structure is different in both years. When comparing 2014 and 2015 we obtain a p-value of 0.716, so we cannot reject the null hypothesis. The spatial distribution of wildfires has not changed significantly in these two consecutive years, in spite of the difference in the observed number of wildfires. Recall that the test statistic (4) measures differences between spatial distributions in terms of relative densities, without considering the number of occurrences. The conclusion of the test can be intuitively visualized comparing the corresponding relative densities represented in Figure 10 (left column). These are the estimated relative densities of the wildfire patterns using longitude, latitude and temperature as covariates. Notice that the relative density of the pattern from 1980 (top) is different from the pattern from 2014 (middle) and 2015 (bottom), while 2014 and 2015 seems to be more similar.

To complete this analysis we compare our results with the test proposed in [18], which does not consider covariate information. The test statistic is based on [16] and uses an L_2-distance to compare the relative densities obtained with a kernel estimator proposed in [17]. The estimated relative densities for the three datasets are represented in Figure 10 (right column) next to our estimates considering the covariate. Notice that relative densities with and without the covariate look different, in particular for the pattern from 2014. The test in this case gives p-values of zero for the two comparisons (1980 versus 2015 and 2014 versus 2015), confirming that the differences shown in the plots in Figure 10 are indeed significant.

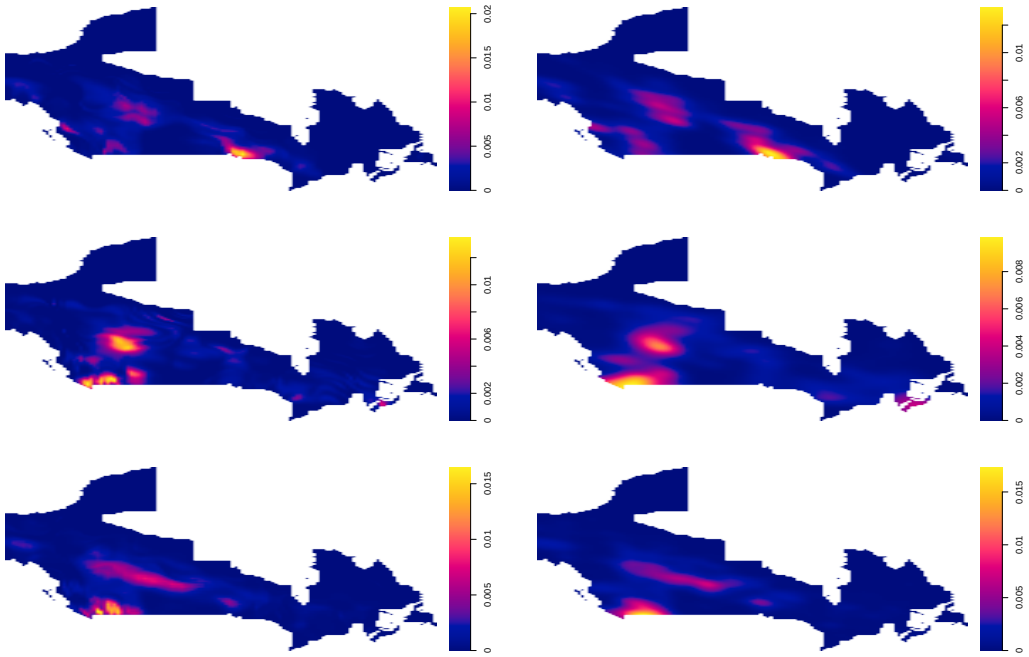

Figure 10: *Left column:* relative density estimations of the wildfire pattern in June 1980 (top), June 2014 (centre) and June 2015 (bottom), using longitude, latitude and temperature as covariates with the multivariate extension of the estimator presented in [4]. *Right column:* relative density estimations of the wildfire pattern in June 1980 (top), June 2014 (centre) and June 2015 (bottom), using the kernel intensity estimator in [17].

Hence, the two tests yield to the same result for the longer term comparison, while we see a different output comparing 2014 and 2015. We analyse in depth the origin of this difference. First recall that our test uses the spatial location plus the covariate information (in this case the temperature). This seems to be important since we have concluded that the temperature is relevant to better describe the intensity of the wildfires (p-value is around 0.9 for the three patterns). This immediately implies that the estimations of the intensity (or more precisely the relative density) derived for our test are more accurate than those of [17], without the covariate. And it can be visualized looking again at the graphs of the relative densities in Figure 10. The estimates on the right (without the temperature) find it harder to represent all the areas covered by the events (see the representation of the wildfire ignition points in Figure 8).

Moreover, going deeper into the fact that the conclusion of both tests provide with different results for a similar problem, we analyse what is hap-

pening with the estimates under the null hypothesis. We have represented in Figure 11 the estimation of the relative density under the null hypothesis (i.e. we gather both patterns, the one from 2014 and the one from 2015, into a unique sample) using our estimator (including temperature information) and the one in [17] (just with the locations), left and right respectively. Comparing the graphs in Figure 11 with the two lower rows in Figure 10, we conclude that the estimates without the temperature information tends to represent more the relative density corresponding to 2015, while the one using the temperature lies in between the estimation of 2014 and 2015. This is probably due to the difference in the sample sizes, recall that in June 2014 we have 950 events while in 2015 we almost double the quantity with 1841. Hence, the estimator proposed in [17] tends to represent the bigger sample not gathering the variability provided by the smaller pattern. This means that the statistic under the null hypothesis is less variable than desirable, leading to the rejection conclusion in the test. Meanwhile, our proposal reduces this effect of the difference in the sample sizes by using common information in both patterns through the covariate information. Comparing Figure 11 (left) with Figure 10 (centre and bottom left) we can see that our estimate represents the joint pattern which is between the two original ones.

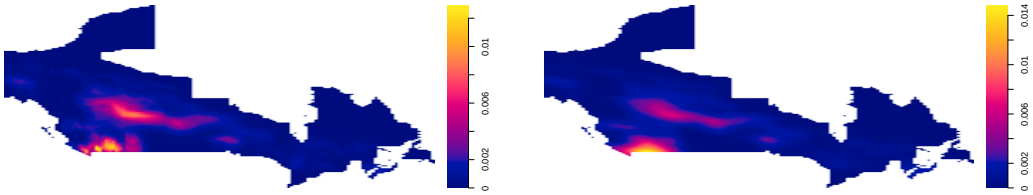

Figure 11: Relative densities under the null hypothesis for the datasets in 2014 and 2015 including covariate information in the estimation (left) and using only location information (right).

An important fact along this application section is the inhomogeneous Poisson assumption. We have used the tools available in the R-package *spatstat* to perform the test described in [24] (`dclf.test`) and [31, 32] (`mad.test`), where the inhomogeneous intensities are the ones in Figure 10 (left column). For our three patterns, i.e., wildfires in June 1980, June 2014 and June 2015, the conclusion is that the null hypothesis of inhomogeneous Poisson point process can not be rejected at 5% of significance, see the p-values for the different tests and processes in Table 4.

	June 1980	June 2014	June 2015
dclf.test	0.25	0.23	0.07
mad.test	0.09	0.12	0.07

Table 4: Resulting p-values on testing inhomogeneous Poisson assumption for the three processes: wildfires in June 1980, wildfires in June 2014 and wildfires in June 2015.

We have also double checked these results by representing the MonteCarlo envelopes for the inhomogeneous K and L functions, showing that in the three scenarios, the empirical functions lie within them, see Figure 12.

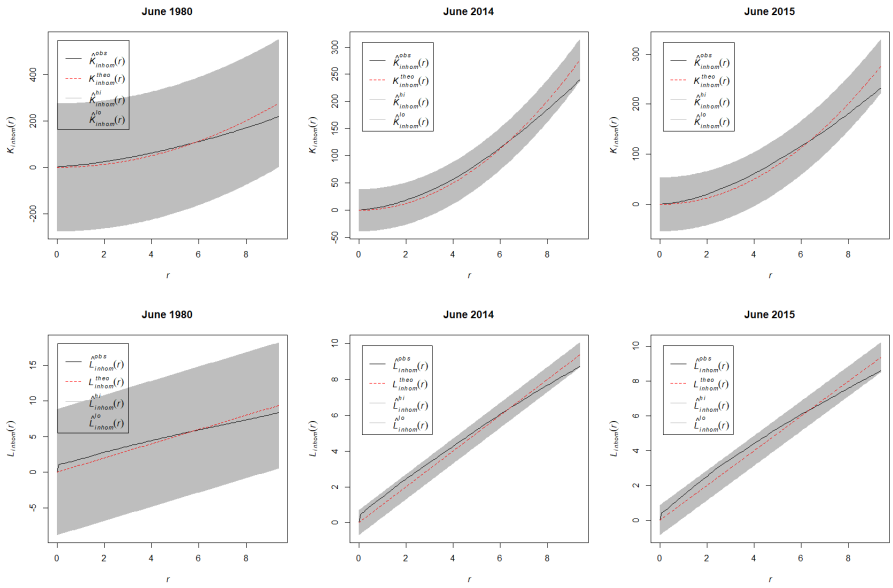

Figure 12: Envelopes and empirical estimates of the inhomogeneous K-function (upper row) and L-function (lower row) computed using MonteCarlo simulations.

5. Conclusions

In this work we have addressed the classical two-sample problem in the context of point processes with covariates. We have assumed an appealing model for the underlying process where the first order intensity depends on known spatial covariates, and we propose an L_2-distance based test statistic to measure discrepancies between two given spatial patterns. Under the

theoretical framework detailed in [4], which includes the Poisson assumption for the underlying process, we have proved the asymptotic normality of the test statistic, and we have proposed a bootstrap procedure to accomplish in practice its calibration. We have carried out a simulation study based on real-data mo¡dels confirming the good performance of our test, which reaches competitive values in terms of level and power. Finally we have applied our proposal to a real life problem with data on wildfires, concluding that the inclusion a specific covariate seems to be crucial to distinguish the spatial distribution of two patterns of wildfires.

In this work the Poisson assumption has been used to prove the asymptotic null distribution of the test statistic as well as the bootstrap resampling scheme. However the proposed bootstrap method can be extended to other types of point processes assuming some additional knowledge. It is not straightforward to define it globally for a general situation covering all the possible types of processes because of the additional information needed for every specific scenario. Further research is still needed in that direction.

A possible next step following this paper is addressing a k-sample problem in this context, where k might be bigger than two. Testing the equality of k distributions from independent random samples is a classical problem where most commonly used test are based on the empirical distribution function: [23] proposed an extension of the Kolmogorov-Smirnov and Cramer-von-Mises tests, while [34] detailed a generalisation of the Anderson-Darling test. To the extent of our knowledge, the first paper in the literature comparing kernel density estimates was [26], where the authors proposed a new measure to determine the distance between the k kernel density estimates. Even though it is out of the scope of our paper, this later idea could be introduced in our context, however the theoretical developments supporting that approach are not straightforward generalisable into our framework. Several considerations need to be taken into account, such as the randomness of the sample size that would increase notably the complexity of the theoretical results. Hence, an extra effort seems to be needed to adapt the k-sample test to the context of point processes.

6. Acknowledgements

The authors are very grateful for constructive comments from the associate editor and two reviewers which helped to improve the paper. They also acknowledge the support from the Spanish Ministry of Economy and

Competitiveness, through grant number MTM2016-76969P, which includes support from the European Regional Development Fund (ERDF). Support from the IAP network StUDyS from Belgian Science Policy, is also acknowledged. M.I. Borrajo has been supported by FPU grant (FPU2013/00473) from the Spanish Ministry of Education. The authors acknowledge as well the Canadian Wildland Fire Information System for their activity in recording and freely providing part of the real data used in this paper.

7. References

[1] Alba-Fernández, M., Ariza-López, F., Jiménez-Gamero, M. D., and Rodríguez-Avi, J. (2016). On the similarity analysis of spatial patterns. *Spatial Statistics*, 18:352–362.

[2] Andersen, M. A. (2009). Testing for similarity in area-based spatial patterns: A nonparametric monte carlo approach. *Applied Geography*, 29:333–345.

[3] Baddeley, A., Chang, Y. M., Song, Y., and Turner, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and Its Interface*, 5:221–236.

[4] Borrajo, M., González-Manteiga, W., and Martínez-Miranda, M. (2019a). Bootstrapping kernel intensity estimation for nonhomogeneous point processes depending on spatial covariates. *(Submitted).*

[5] Borrajo, M. I., González-Manteiga, W., and Martínez-Miranda, M. D. (2019b). Testing first-order intensity model in non-homogeneous poisson point processes with covariates. *(Submitted).*

[6] Cao, R. (1993). Bootstrapping the mean integrated squared error. *Journal of Multivariate Analysis*, 45(1):137–160.

[7] Cowling, A., Hall, P., and Phillips, M. J. (1996). Bootstrap confidence regions for the intensity of a poisson point process. *Journal of the American Statistical Association*, 91(436):1516–1524.

[8] Cox, D. R. (1972). Regression models and life tables. In Samuel, K. and Johnson, N. L., editors, *Breakthroughs in Statistics Volume II*, 527–543. Springer.

[9] Cucala, L. (2006). *Espacements bidimensionnels et données entachés d'erreurs dans l'analyse des procesus ponctuels spatiaux*. PhD thesis, Université des Sciences de Toulouse I.

[10] Cucala, L., Genin, M., Occelli, F. and Soula, J. (2019). A multivariate nonparametric scan statistic for spatial data. *Spatial Statistics*, 29:1–14.

[11] Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(2):138–147.

[12] Diggle, P. and Marron, J. S. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. *Journal of the American Statistical Association*, 83:793–800.

[13] Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):349–362.

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

[15] Duong, T. (2013). Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics*, 25(3):635–645.

[16] Duong, T., Goud, B., and Schauer, K. (2012). Closed-form density-based framework for automatic detection of cellular morphology changes. *Proceedings of the National Academy of Sciences*, 109(22):8382–8387.

[17] Fuentes-Santos, I., González-Manteiga, W., and Mateu, J. (2015). Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial poisson point processes. *Scandinavian Journal of Statistics*, 43(2):416–435.

[18] Fuentes-Santos, I., González-Manteiga, W., and Mateu, J. (2017). A nonparametric test for the comparison of first-order structures of spatial point processes. *Spatial Statistics*, 22:240–260.

[19] García-Portugués, E., Van Keilegom, I., Crujeiras, R., and González-Manteiga, W. (2016). Testing parametric models in linear-directional regression. *Scand. J. Statist.*, 43(4):1178–1191.

[20] Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, 103(483):1238–1247.

[21] Guan, Y. and Loh, J. M. (2007). A thinned block bootstrap variance estimation procedure for inhomogeneous spatial point patterns. *Journal of the American Statistical Association*, 102(480):1377–1386.

[22] Guan, Y. and Shen, Y. (2010). A weighted estimating equation approach for inhomogeneous spatial point processes. *Biometrika*, 97(4):867–880.

[23] Kiefer, J. (1959). k-Sample analogues of the Kolmogorov-Smirnov, Cramr von Mises test. *Annals of Mathematical Statistics*, 30:420–447.

[24] Loosmore, N.B. and Ford, E.D. (2006). Statistical inference using the G or K point pattern spatial statistics *Ecology*, 87:1925–1931.

[25] Maistre, S., Lavergne, P., and Patilea, V. (2017). Powerful nonparametric checks for quantile regression. *Journal of Statistical Planning and Inference*, 180:13–29.

[26] Martínez-Camblor, P., de Uña-Álvarez, J. and Corral, N. (2008). k-Sample test based on the common area of kernel density estimator. *Journal of Statistical Planning and Inference*, 138(12):4006–4020.

[27] Martínez-Camblor, P. and de Uña-Álvarez, J. (2009). Non-parametric k-sample tests: Density functions vs distribution functions. *Computational Statistics and Data Analysis*, 53:3344–3357.

[28] Mateu, J., Schoenberg, F. P., Diez, D. M., González, J. A., and Lu, W. (2015). On measures of dissimilarity between point patterns: Classification based on prototypes and multidimensional scaling. *Biometrical Journal*, 57(2):340–358.

[29] Moller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.

[30] Reitzner, M. and Schulte, M. (2013). Central limit theorems for u-statistics of poisson point processes. *The Annals of Probability*, 41(6):3879–3909.

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

[32] Ripley, B.D. (1981) *Spatial statistics*. John Wiley and Sons.

[33] Rodríguez-Cortés, F. J. and Mateu, J. (2015). Second-order smoothing of spatial point patterns with small sample sizes: a family of kernels. *Stochastic environmental research and risk assessment*, 29(1):295–308.

[34] Scholz, F.W. and Stephens, M.A. (1987). k-Samples Anderson-Darling test. *Journal of American Statistics Association*, 82:918–924.

[35] Stute, W., González-Manteiga, W., and Presedo-Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, 93(441):141–149.

[36] Van Lieshout, M. N. M. (2000). *Markov point processes and their applications*. World Scientific.

[37] Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous neyman–scott processes. *Biometrics*, 63(1):252–258.

[38] Wand, M.P. and Jones, M.C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, 9:97–116.

Appendix A. Proof of Theorem 2.1

Here we obtain the mean and variance of the statistic T and show that it is asymptotic normal under the null hypothesis., and remind the reader this calculations has been done for $p = 1$, i.e., one-dimensional covariate.

Using common properties of the mean and variance operators we can see that

$$\mathbb{E}[T] = \mathbb{E}\left[\widehat{\psi}_{11}\right] + \mathbb{E}\left[\widehat{\psi}_{22}\right] - \mathbb{E}\left[\widehat{\psi}_{12}\right] - \mathbb{E}\left[\widehat{\psi}_{21}\right]$$

and

$$\begin{aligned} Var[T] = {} & Var\left[\widehat{\psi}_{11}\right] + Var\left[\widehat{\psi}_{22}\right] + Var\left[\widehat{\psi}_{12}\right] + Var\left[\widehat{\psi}_{21}\right] + 2Cov\left[\widehat{\psi}_{11}, \widehat{\psi}_{12}\right] \\ & + 2Cov\left[\widehat{\psi}_{11}, \widehat{\psi}_{21}\right] + 2Cov\left[\widehat{\psi}_{22}, \widehat{\psi}_{12}\right] + 2Cov\left[\widehat{\psi}_{22}, \widehat{\psi}_{21}\right] - 2Cov\left[\widehat{\psi}_{12}, \widehat{\psi}_{21}\right]. \end{aligned}$$

Remark that we haven't considered the covariance between $\widehat{\psi}_{11}$ and $\widehat{\psi}_{22}$ since it is zero due to the independence between the processes X_1 and X_2.

Mean of T

$$\begin{aligned} \mathbb{E}\left[\widehat{\psi}_{11}\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\psi}_{11}|N_1\right]\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}K_{h_1}(Z_{1i}-Z_{1j})1_{\{N_1\neq 0\}}|N_1\right]\right] \\ &= \sum_{l=1}^{\infty}\frac{\mathbb{P}(N_1=l)}{l^2}\mathbb{E}\left[\sum_{i=1}^{l}K_{h_1}(0) + \sum_{i=1}^{l}\sum_{j\neq i}\frac{g^\star(Z_{1i})}{g^\star(Z_1 j)}K_{h_1}(Z_{1i}-Z_{1j})\right] = A(m_1)K_{h_1}(0) \\ &+ \left(1-e^{-m_1}-A(m_1)\right)\left(\psi_{11} + \frac{1}{2}h_1^2\mu_2(K)\int\frac{f_1}{g^\star}(y)(g^\star f_1)''dy + \int\frac{f_1}{g^\star}(y)dy\; o(h_1^2)\right), \end{aligned} \tag{A.1}$$

where we have used a second order Taylor expansion to obtain

$$\begin{aligned} &\mathbb{E}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12})\right] = \int\int\frac{g^\star(x)}{g^\star(y)}K_{h_1}(x-y)f_1(x)f_1(y)dxdy \\ &= \int\frac{f_1}{g^\star}(y)\int K(u)(g^\star f_1)(y-h_1u)dudy = \psi_{11} + \frac{1}{2}h_1^2\mu_2(K)\int\frac{f_1}{g^\star}(y)(g^\star f_1)''dy \\ &+ \int\frac{f_1}{g^\star}(y)dy\; o(h_1^2). \end{aligned}$$

Similarly, we obtain

$$\mathbb{E}\left[\widehat{\psi}_{22}\right]A(m_2)K_{h_2}(0) + \left(1 - e^{-m_2} - A(m_2)\right)\left(\psi_{22} + \frac{1}{2}h_2^2\mu_2(K)\int \frac{f_2}{g^\star}(y)(g^\star f_2)^{''}dy + \int \frac{f_2}{g^\star}(y)dy\; o(h_2^2)\right). \tag{A.2}$$

Now, we compute the expectations of the other two terms:

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}_{12}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\widehat{\psi}_{12}|N_1,N_2\right]\right]\right] = \sum_{l=1}^{\infty}\sum_{k=1}^{\infty}\mathbb{E}\left[\widehat{\psi}_{12}|N_1,N_2\right]\mathbb{P}(N_1=l)\mathbb{P}(N_2=k)\\ &= (1-e^{-m_1})(1-e^{-m_2})\\ &\left(\psi_{12} + \frac{1}{2}\frac{h_1^2}{m_1}\mu_2(K)\int (g^\star f_2)(y)\rho_1^{''}(y)dy + \int (g^\star f_2)(y)dy\; o\left(\frac{h_1^2}{m_1}\right)\right),\end{aligned} \tag{A.3}$$

where we have taken into account that

$$\begin{aligned}&\mathbb{E}\left[\frac{g^\star(Z_{21})}{g^\star(Z_{12})}K_{h_1}(Z_{21}-Z_{12})\right] = \int\int \frac{g^\star(y)}{g^\star(x)}K_{h_1}(y-x)f_1(x)f_2(y)dxdy\\ &= \int \frac{f_2}{g^\star}(y)\int K(u)\left(\frac{f_1}{g^\star}\right)(y-h_1u)dudy = \psi_{12} + \frac{1}{2}\frac{h_1^2}{m_1}\mu_2(K)\int (g^\star f_2)(y)\rho_1^{''}(y)dy\\ &+ \int (g^\star f_2)(y)dy\; o\left(\frac{h_1^2}{m_1}\right).\end{aligned}$$

And again, similarly, we get

$$\begin{aligned}\mathbb{E}\left[\widehat{\psi}_{21}\right] &= (1-e^{-m_1})(1-e^{-m_2})\\ &\left(\psi_{21} + \frac{1}{2}\frac{h_2^2}{m_2}\mu_2(K)\int (g^\star f_1)(y)\rho_2^{''}(y)dy + \int (g^\star f_1)(y)dy\; o\left(\frac{h_2^2}{m_2}\right)\right).\end{aligned} \tag{A.4}$$

Hence, gathering equations (A.1) – (A.4), and taking into account that we are under the null hypothesis, i.e., $f_1 = f_2 := f$ and $\psi_{11} = \psi_{22} = \psi_{12} = \psi_{21} := \psi$, we have that

$$\mathbb{E}[T] = \left(A(m_1)h_1 + A(m_2)h_2\right)K(0) + o(A(m_1)) + o(A(m_2)).$$

Variance of T

We need to compute the variances of every addend in the statistic as well as the

covariances detailed at the beginning of this proof. First we work on the variances:

$$Var\left[\widehat{\psi}_{11}\right] = \mathbb{E}\left[Var\left[\widehat{\psi}_{11}|N_1\right]\right] + Var\left[\mathbb{E}\left[\widehat{\psi}_{11}|N_1\right]\right] = 4A(m_1)\left(\int f_1^3(x)dx - \psi_{11}^2\right)$$
$$+ 2B(m_1)\frac{1}{h_1}R(K)\psi_{11} + O(A(m_1)h_1^2) + O(B(m_1)). \tag{A.5}$$

To obtain this result we use the following calculation in the first addend of the variance decomposition

$$\begin{aligned}
&\mathbb{E}\left[Var\left[\widehat{\psi}_{11}|N_1\right]\right] \\
&= \sum_{l=1}^{\infty} Cov\left[\frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}K_{h_1}(Z_{1i}-Z_{1j}), \frac{1}{N_1^2}\sum_{s=1}^{N_1}\sum_{t=1}^{N_1}\frac{g^\star(Z_{1s})}{g^\star(Z_{1t})}K_{h_1}(Z_{1s}-Z_{1t})\right] \\
&= B(m_1)Var\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12})\right] \\
&+ B(m_1)Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{12})}{g^\star(Z_{11})}K_{h_1}(Z_{12}-Z_{11})\right] \\
&+ A(m_1)Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{11})}{g^\star(Z_{13})}K_{h_1}(Z_{11}-Z_{13})\right] \\
&+ A(m_1)Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{13})}{g^\star(Z_{11})}K_{h_1}(Z_{13}-Z_{11})\right] \\
&+ A(m_1)Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{12})}{g^\star(Z_{13})}K_{h_1}(Z_{12}-Z_{13})\right] \\
&+ A(m_1)Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{13})}{g^\star(Z_{12})}K_{h_1}(Z_{13}-Z_{12})\right] \\
&= 4A(m_1)\left(\int f_1^3(x)dx - \psi_{11}^2\right) + O(A(m_1)h_1^2) + O(B(m_1)h_1)
\end{aligned}$$

where after some second-order Taylor expansions and reducing the negligible terms,

we have

$$Var\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12})\right]=\frac{1}{h_1}R(K)\psi_{11}-\psi_{11}^2-\mu_1(K^2)\int\frac{f_1}{g^{\star^2}}(g^{\star^2}f_1)'(x)dx+O(h_1^2),$$

$$Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{12})}{g^\star(Z_{11})}K_{h_1}(Z_{12}-Z_{11})\right]=\frac{1}{h_1}R(K)\psi_{11}-\mu_1(K^2)\int\frac{f_1}{g^{\star^2}}(g^{\star^2}f_1)'$$
$$-\psi_{11}^2+O(h_1),$$

$$Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{11})}{g^\star(Z_{13})}K_{h_1}(Z_{11}-Z_{13})\right]=\int f_1^3(x)dx-\psi_{11}^2+O(h_1),$$

$$Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{13})}{g^\star(Z_{11})}K_{h_1}(Z_{13}-Z_{11})\right]=\int f_1^3(x)dx-\psi_{11}^2+O(h_1),$$

$$Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{12})}{g^\star(Z_{13})}K_{h_1}(Z_{12}-Z_{13})\right]=\int f_1^3(x)dx-\psi_{11}^2+O(h_1)\text{ and}$$

$$Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{13})}{g^\star(Z_{12})}K_{h_1}(Z_{13}-Z_{12})\right]=\int f_1^3(x)dx-\psi_{11}^2+O(h_1).$$

The second addend of the variance decomposition is computed as follows:

$$Var\left[\mathbb{E}\left[\widehat{\psi}_{11}|N_1\right]\right]=Var\left[\frac{1}{N_1}K_{h_1}(0)1_{\{N_1\neq0\}}+\frac{N_1-1}{N_1}1_{\{N_1\neq0\}}\mathbb{E}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12})\right]\right]=$$
$$=K_{h_1}^2(0)Var\left[\frac{1}{N_1}1_{\{N_1\neq0\}}\right]+\mathbb{E}^2\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})K_{h_1}(Z_{11}-Z_{12})}\right]Var\left[\frac{N_1-1}{N_1}1_{\{N_1\neq0\}}\right]+$$
$$+2K_{h_1}(0)\mathbb{E}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})K_{h_1}(Z_{11}-Z_{12})}\right]Cov\left[\frac{1}{N_1}1_{\{N_1\neq0\}},\frac{N_1-1}{N_1}1_{\{N_1\neq0\}}\right]=$$
$$=K_{h_1}^2(0)\left(B(m_1)-A^2(m_1)\right)+\left(\psi_{11}^2+O(h_1)\right)\left(e^{-m_1}+B(m_1)+A^2(m_1)-2e^{-m_1}A(m_1)\right).$$

And gathering both of them we obtain the expression in (A.5). Using the same developments we get

$$Var\left[\widehat{\psi}_{22}\right]=4A(m_2)\left(\int f_2^3(x)dx-\psi_{22}^2\right)+2B(m_2)\frac{1}{h_2}R(K)\psi_{22}+O(A(m_2)h_2^2)+O(B(m_2)).\tag{A.6}$$

Now, we compute the variance of $\widehat{\psi}_{12}$,

$$\begin{aligned} Var\left[\widehat{\psi}_{12}\right] &= \mathbb{E}_{N_2}\left[\mathbb{E}_{N_1}\left[Var\left[\widehat{\psi}_{12}|N_1,N_2\right]\right]\right] + \mathbb{E}_{N_2}\left[Var_{N_1}\left[\mathbb{E}\left[\widehat{\psi}_{12}|N_1,N_2\right]\right]\right] \\ &+ Var_{N_2}\left[\mathbb{E}_{N_1}\left[\mathbb{E}\left[\widehat{\psi}_{12}|N_1,N_2\right]\right]\right] \\ &= A(m_1)A(m_2)\left(\frac{1}{h_1}\psi_{12}R(K) - \mu_2(K^2)\int\frac{(g^{\star^2}f_2)'(x)f_1(x)}{g^{\star^2}}dx\right. \\ &\left. - \psi_{12}^2 + O(h_1)\right) + A(m_1)(1-e^{-m_2})\left(\int f_1(x)f_2^2(x)dx - \psi_{12}^2 + O(h_1^2)\right) \\ &+ A(m_2)(1-e^{-m_1})\left(\int f_1^2(x)f_2(x)dx - \psi_{12}^2 + O(h_1^2)\right). \end{aligned} \tag{A.7}$$

To reach this final expression we have developed separately each of the three addends, but only the first one is not null. We have used the development of the conditional variance in terms of covariances, as well as the independence between N_1 and N_2.

$$\begin{aligned} Var\left[\widehat{\psi}_{21}\right] &= A(m_1)A(m_2)\left(\frac{1}{h_2}\psi_{21}R(K) - \mu_2(K^2)\int\frac{(g^{\star^2}f_1)'(x)f_2(x)}{g^{\star^2}}dx\right. \\ &\left. - \psi_{21}^2 + O(h_2)\right) + A(m_2)(1-e^{-m_1})\left(\int f_1^2(x)f_2(x)dx - \psi_{21}^2 + O(h_2^2)\right) \\ &+ A(m_1)(1-e^{-m_2})\left(\int f_1(x)f_2^2(x)dx - \psi_{21}^2 + O(h_2^2)\right). \end{aligned} \tag{A.8}$$

Finally we obtain the expressions of the covariances:

$$\begin{aligned} Cov\left[\widehat{\psi}_{11},\widehat{\psi}_{12}\right] &= \mathbb{E}_{N_1}\left[\mathbb{E}_{N_2}\left[Cov\left[\widehat{\psi}_{11},\widehat{\psi}_{12}|N_1,\ N_2\right]\right]\right] \\ &+ \mathbb{E}_{N_1}\left[Cov_{N_2}\left[\mathbb{E}\left[\widehat{\psi}_{11}|N_1,\ N_2\right],\mathbb{E}\left[\widehat{\psi}_{12}|N_1,\ N_2\right]\right]\right] + \\ &+ Cov_{N_1}\left[\mathbb{E}_{N_2}\left[\mathbb{E}\left[\widehat{\psi}_{11}|N_1,\ N_2\right]\right],\mathbb{E}\left[\mathbb{E}\left[\widehat{\psi}_{12}|N_1,\ N_2\right]\right]\right] = \\ &= 2A(m_1)\left(\int f_1^2(x)f_2(x)dx - \psi_{11}\psi_{12} + O(h_1^2)\right), \end{aligned} \tag{A.9}$$

where we have used that the two last addends are negligible and the first one is obtained as:

$$
\begin{aligned}
&Cov\left[\widehat{\psi}_{11},\widehat{\psi}_{12}|N_1,\ N_2\right] \\
&= \frac{1}{N_1^3 N_2}\sum_{i,j,k=1}^{N_1}\sum_{l=1}^{N_2} Cov\left[\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}K_{h_1}(Z_{1i}-Z_{1j}),\frac{g^\star(Z_{1k})}{g^\star(Z_{2l})}K_{h_1}(Z_{1k}-Z_{2l})\right] = \\
&= \frac{N_1-1}{N_1^2}Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{11})}{g^\star(Z_{21})}K_{h_1}(Z_{11}-Z_{21})\right] + \\
&+ \frac{N_1-1}{N_1^2}Cov\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{12})}{g^\star(Z_{21})}K_{h_1}(Z_{12}-Z_{21})\right].
\end{aligned}
$$

After applying expectations we easily get (A.9).

Following the same steps we obtain:

$$
Cov\left[\widehat{\psi}_{11},\widehat{\psi}_{21}\right] = 2A(m_1)\left(\int f_1^2(x)f_2(x)dx - \psi_{11}\psi_{21} + O(h_1^2) + O(h_2^2)\right), \tag{A.10}
$$

$$
Cov\left[\widehat{\psi}_{22},\widehat{\psi}_{12}\right] = 2A(m_2)\left(\int f_1(x)f_2^2(x)dx - \psi_{22}\psi_{12} + O(h_1^2) + O(h_2^2)\right) \text{ and} \tag{A.11}
$$

$$
Cov\left[\widehat{\psi}_{22},\widehat{\psi}_{21}\right] = 2A(m_2)\left(\int f_1(x)f_2^2(x)dx - \psi_{22}\psi_{21} + O(h_2^2)\right) \tag{A.12}
$$

The only term left is the covariance involving $\widehat{\psi}_{12}$ and $\widehat{\psi}_{21}$. The computation of this one is closely to the ones above but with more non-zero terms in the expression of the conditional covariance. Applying the same mathematical tools we get:

$$
\begin{aligned}
Cov\left[\widehat{\psi}_{12},\widehat{\psi}_{21}\right] &= A(m_1)A(m_2)\left(\frac{\psi_{12}}{2h_1}\int K(u)K_{h_2/h_1}(u)du\right) \\
&+ A(m_1)\left(\int f_1(x)f_2^2(x)dx - \psi_{12}\psi_{21}\right) \qquad (A.13) \\
&+ A(m_2)\left(\int f_1^2(x)f_2(x)dx - \psi_{12}\psi_{21}\right) \\
&+ O(A(m_1)h_1^2) + O(A(m_1)h_2^2) + O(A(m_2)h_1^2) + O(A(m_2)h_2^2). \qquad (A.14)
\end{aligned}
$$

Finally, gathering (A.5) – (A.13), taking into account that following [5] the order of the optimal bandwidths h_j is $A(m_j)^{1/5}$, that we are under the null hypothesis, i.e., $f_1 = f_2 := f$, and $\psi_{11} = \psi_{12} = \psi_{21} = \psi_{22} := \psi$, we get the final result of

$$\begin{aligned} Var\left[T\right] = & \left(B(m_1)\frac{1}{h_1} + B(m_2)\frac{1}{h_2}\right) 2R(K)\psi + A(m_1)A(m_2)\psi R(K)\left(\frac{1}{h_1} + \frac{1}{h_2}\right) \\ & + A(m_1)A(m_2)\psi\left(\frac{1}{h_1}\int K(u)K_{h_2/h_1}(u)du + \frac{1}{h_2}\int K(u)K_{h_1/h_2}(u)du\right) \\ & + O(B(m_1)) + O(B(m_2)) + O(A(m_1)A(m_2)). \end{aligned}$$

Asymptotic normality

Our test statistic can be expanded and written as a sum of non-duplicated points, where each of the addends is a U-statistic on a Poisson point process. Moreover, every of the addends is absolutely convergent in the sense defined by [30], hence following the Theorem 4.7 in that paper we can assure the normality of each term. Hence the normality of our test statistic with the mean and variance detailed in the main body of Theorem 2.1.