CHALMERS
UNIVERSITY OF TECHNOLOGY

Local inhomogeneous weighted summary statistics for marked point processes

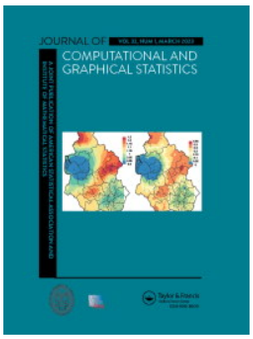
Downloaded from: https://research.chalmers.se, 2024-04-27 18:03 UTC

Citation for the original published paper (version of record):
D'Angelo, N., Adelfio, G., Mateu, J. et al (2023). Local inhomogeneous weighted summary statistics for marked point processes. Journal of Computational and Graphical Statistics, In press. http://dx.doi.org/10.1080/10618600.2023.2206441

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)



Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Local inhomogeneous weighted summary statistics for marked point processes

Nicoletta D'Angelo, Giada Adelfio, Jorge Mateu & Ottmar Cronie

To cite this article: Nicoletta D'Angelo, Giada Adelfio, Jorge Mateu & Ottmar Cronie (2023): Local inhomogeneous weighted summary statistics for marked point processes, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2023.2206441

To link to this article: https://doi.org/10.1080/10618600.2023.2206441

View supplementary material

Accepted author version posted online: 25 Apr 2023.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ucgs20

Check for updates

Local inhomogeneous weighted summary statistics for marked point processes

Nicoletta D'Angelo[1], Giada Adelfio[1], Jorge Mateu[2], Ottmar Cronie[3,*]

[1]Department of Businnes, Economics and Statistics, University of Palermo, Palermo, Italy

[2]Department of Mathematics, University Jaume I, Castellon, Spain

[3]Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden

[*]Corresponding author; email: Ottmar Cronie, ottmar@chalmers.se

Abstract

We introduce a family of local inhomogeneous mark-weighted summary statistics, of order two and higher, for general marked point processes. Depending on how the involved weight function is specified, these summary statistics capture different kinds of local dependence structures. We first derive some basic properties and show how these new statistical tools can be used to construct most existing summary statistics for (marked) point processes. We then propose a local test of random labelling. This procedure allows us to identify points, and consequently regions, where the random labelling assumption does not hold, e.g. when the (functional) marks are spatially dependent. Through a simulation study we show that the test is able to detect local deviations from random labelling. We also provide an application to an earthquake point pattern with functional marks given by seismic waveforms.

Keywords: earthquakes; functional marked point process; local envelope test; mark correlation function; marked *K*-function; random labelling

Accepted Manuscript

1 Introduction

The analysis of a point pattern, given as a collection of points in a region, typically begins with computing an estimate of some summary statistic which may be used to find specific structures in the data and suggest suitable models (Chiu et al., 2013; Daley and Vere-Jones, 2008; Gelfand et al., 2010; Illian et al., 2008; Van Lieshout, 2000)

The choice of summary statistic depends both on the pattern at hand and on the feature or hypothesis of interest.

A widely used summary statistic for descriptive analyses and diagnostics, which is obtained as an instance of the so-called reduced second moment measure (Cressie and Collins, 2001; Chiu et al., 2013; Møller, 2003), is Ripley's *K*-function (Ripley, 1976), which is based on the assumption of a non-marked stationary and isotropic point process. In the marked case, assuming discrete marks and stationarity, cross versions of the *K*- or nearest neighbour distance distribution functions have been proposed (Diggle, 2013). For real-valued marks, the mark correlation type-functions in Penttinen and Stoyan (1989); Illian et al. (2008) are widely used and such second order statistics have been studied in more detail and reformulated by Schlather (2001), in order to obtain a more rigorous formulation. However, although the assumption of stationarity is mathematically appealing, it can rarely be justified in practice since the intensity tends to change over the study region. This is to say that the underlying point process is inhomogeneous and, in the unmarked case, Baddeley et al. (2000) proposed an inhomogeneous extension of the *K*-function for a class of point processes, which are referred to as second order intensity-reweighted stationary. Their ideas were extended to spatio-temporal point processes in Gabriel and Diggle (2009); Møller and Ghorbani (2012). Further, Møller and Waagepetersen (2003) proposed an extension of this *K*-function to second order intensity-reweighted stationary multivariate point processes. As indicated in Cronie and van Lieshout (2016) and Iftimi et al. (2019), this structure may be extended to *K*-functions for general

Accepted Manuscript

marked point processes. To analyse higher order interactions in general stationary marked point processes, Van Lieshout (2006) proposed marked versions of the nearest neighbour distance distribution functions, the empty space function and the *J*-function. These summary statistics, which allow us to study spatial interactions between different mark groupings of the points, were later extended to the inhomogeneous setting by Cronie and van Lieshout (2016) and Iftimi et al. (2019). In particular, to test for random labelling, Cronie and van Lieshout (2016) proposed inhomogeneous Lotwick-Silverman-type Monte Carlo tests based on their new summary statistics, while Iftimi et al. (2019) proposed second order Monte Carlo tests based on permuting the attached marks. Further details on the random shift-type testing considered in Lotwick-Silverman-type tests can be found in Mrkvička et al. (2021).

Despite the relatively long history of point process theory (see e.g. Diggle, 2013; Stoyan and Stoyan, 1994; Daley and Vere-Jones, 2008), few approaches have been proposed to analyse spatial point patterns where the features of interest are functions/curves instead of qualitative or quantitative variables. Examples of point patterns with associated functional data include forest patterns where for each tree we have a growth function, curves representing the incidence of an epidemic over a period of time, and the evolution of distinct economic parameters such as unemployment and price rates, all for distinct spatial locations. The study of such configurations allows analysing the effects of the spatial structure on individual functions. Illian et al. (2006) consider for each point a transformed Ripley (1976)'s *K*-function to characterise spatial point patterns of ecological plant communities, whilst Mateu et al. (2007) build new marked point processes formed by spatial locations and curves defined in terms of Local Indicators of Spatial Association (LISA) functions, which describe local characteristics of the points. They use this approach to classify and discriminate between points belonging to a clutter and those belonging to a feature. Finally, the idea of analysing point patterns with

Accepted Manuscript

attached functions has been presented coherently by Comas et al. (2011); Ghorbani et al. (2021).

Ghorbani et al. (2021) introduced a very broad framework for the analysis of Functional Marked Point Processes (FMPPs), indicating how they connect the point process framework with both Functional Data Analysis (FDA; Ramsay and Silverman (2002)) and geostatistics. In particular, they defined a new family of summary statistics, so-called *weighted n-th order marked inhomogeneous K-functions*, together with their non-parametric estimators, which they exploited to analyse Spanish population structures, such as demographic evolution and sex ratio over time. This summary statistic family can be used to run a Monte Carlo test of random labelling, e.g. by means of global envelopes test (GET; Myllymäki et al. (2017)), to assess whether the functional marks of the analysed pattern are spatially dependent. However, this procedure is essentially global, since it does not provide information on the points which mostly contributed to the rejection of the random labelling hypothesis. Therefore, motivated by the need of detecting such points, and thus the regions in which they are located, where the functional marks really do depend on the surrounding structure, in this paper we introduce a new class of summary statistics, *local t-weighted marked n-th order inhomogeneous K-functions*. These are used to propose a *local test of random labelling*. Here *t* refers to a function which governs how much weight we put on different aspects of the marked point process/pattern.

Further, we use the developed tools to analyse seismic data. Note that while the spatial (and temporal) locations of the epicenters of earthquakes are typically analysed within the framework of point processes, the associated seismic waveforms are commonly investigated in separate analyses through FDA. Applying the local test allows us to identify where one would expect waveforms (i.e. functional marks) to be similar to those of nearby points.

Accepted Manuscript

All the performed analyses are carried out through the R Core Team (2022) software, and the codes are available from the first author. Preliminary data manipulation is performed through the software Python (Van Rossum and Drake Jr, 1995).

The structure of the paper is as follows. In Section 2, the motivation of this paper is presented, showing the dataset and problem that will be further analysed along the paper. Section 3 contains some preliminaries on functional marked point processes. In Section 4, we present our proposed local t-weighted n-th order inhomogeneous K-functions and their main properties, also relating them to their global counterparts. Section 5 outlines the main steps to run a local test of random labelling. In Section 6, we present a motivating example to show the further advantages of a local test, compared to a global one. To have a comprehensive understanding of the performance of the proposed local test, we show simulation results under different scenarios. Section 7 provides an application to seismic data. Finally, conclusions are drawn in Section 8.

2 Data and motivation

Earthquakes' detection provides a whole set of data which are usually studied separately, i.e. spatial (and temporal) occurrence of points through point process theory (Siino et al. (2017); Iftimi et al. (2019); D'Angelo et al. (2022), to cite just a few recent works), and the analysis of waveforms through FDA (Adelfio et al., 2011, 2012; Chiodi et al., 2013).

A recently released set of data on Italian seismic activity encompasses both of these data types. *The Italian seismic dataset for machine learning (INSTANCE)* is a dataset of seismic waveforms data and associated metadata (Michelini et al., 2021), which includes 54008 earthquakes for a total of 1159249 3-channel waveforms. It also contains 132330 3-channel noise waveforms. For each of these waveforms, 115 metadata (i.e. statistical variables) are available, providing information on station, trace, source, path and quality. Overall, the data are

Accepted Manuscript

collected by 19 networks which consist of 620 seismic stations. The dataset is available on http://www.pi.ingv.it/instance/.

The earthquake list in the dataset is based on the Italian seismic bulletin (http://terremoti.ingv.it/bsi) of the "Istituto Nazionale di Geofisica e Vulcanologia", includes events which occurred between January 2005 and January 2020, and in the magnitude range between 0.0 and 6.5. The waveform data have been recorded primarily by the Italian National Seismic Network. Figure 1 (a) - (b) depict the earthquake locations and the seismic stations which recorded the events.

In Figure 1 (c), some waveforms contained in the dataset are represented. All the waveform traces have a length of 120 seconds, are sampled at 100 Hz, and are provided both in counts and ground motion physical units after deconvolution of the instrument transfer functions. The waveform dataset is accompanied by metadata consisting of more than 100 variables providing comprehensive information on the earthquake source, the recording stations, the trace features, and other derived quantities.

3 Preliminaries on marked point processes

Throughout the paper, we consider a marked point process $Y = \{(x_i, m_i)\}_{i=1}^{N}$ (Daley and Vere-Jones, 2008, Definition 6.4.1), with ground points x_i in the d-dimensional Euclidean space $\mathbb{R}^d, d \geq 1$, which is equipped with the Lebesgue measure $|A| = \int_A dz$ for Borel sets $A \in \mathcal{B}(\mathbb{R}^d)$; a closed Euclidean r-ball around $x \in \mathbb{R}^d$ will be denoted by $b[x, r]$. By definition, the ground process $Y_g = \{x_i\}_{i=1}^{N}$, obtained from Y by ignoring the marks, is a well-defined point process on $\mathbb{R}^d$ in its own right. Note that, formally, Y is a random element in the measurable space $(N_{lf}, \mathcal{N})$ of locally finite point configurations/patterns $\mathbf{x} = \{((x_1, m_1), \ldots, (x_n, m_n))\}, n \geq 0$ (Daley and Vere-Jones, 2008; Van Lieshout, 2000). We assume that the mark space $\mathcal{M}$ is Polish

Accepted Manuscript

and equipped with a finite reference measure ν on the Borel σ-algebra $\mathcal{B}(\mathcal{M})$. The Borel σ-algebra $\mathcal{B}(\mathbb{R}^d \times \mathcal{M}) = \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathcal{M})$ is endowed with the product measure $A \times E \mapsto |A|\, \nu(E),\ A \times E \in \mathcal{B}(\mathbb{R}^d \times \mathcal{M})$. We will let $Y(A \times E) = \sum_{(x,m) \in Y} \mathbf{1}\{(x,m) \in A \times E\}$, where $\mathbf{1}$ is the indicator function, denote the cardinality of the random set $Y \cap (A \times E)$. We assume that Y is simple, that is, it almost surely (a.s.) does not contain multiple points in the sense that $P(Y(\{(x,m)\}) = 0 \text{ or } 1) = 1$ for all $(x,m) \in \mathbb{R}^d \times \mathcal{M}$.

Given this general setup, one may obtain various forms of marked point processes, most notably multivariate/multitype point processes with $\mathcal{M} = \{1, \ldots, k\}$ (Diggle, 2013) and functional marked point processes with $\mathcal{M}$ given by a suitable function space (Ghorbani et al., 2021).

3.1 Functional Marked Point Processes

In this section, we provide the definition of functional marked point processes following Ghorbani et al. (2021).

In classical FDA, one analyses a collection of functions $\{f_1(t), \ldots, f_n(t)\}, t \in \mathcal{T} \subset [0, \infty), n \geq 1$, which take values in some Euclidean space $\mathbb{R}^k, k \geq 1$, and belong to some suitable function space, typically an L_2-space. Although t usually represents time, it could also represent some other quantity, for example, spatial distance. Classically, one would assume that such a collection of functions constitute realisations or samples of some collection of independent and identically distributed (iid) random functions or stochastic processes $\{F_1(t), \ldots, F_n(t)\}, t \in \mathcal{T}$. Such an assumption may, however, be questioned in certain settings. For example, two functions f_i and f_j, which are spatially close to each other in $\mathbb{R}^k$, could gain (or lose) from being close to each other. Accordingly, it seems natural to relax the iid assumption for $F_1, \ldots, F_N$. A natural way to handle such a scenario is to generate $F_1, \ldots, F_N$ conditionally on

Accepted Manuscript

some collection of (dependent) random spatial locations. Note that the conditional distribution of $F_1,\ldots,F_N$ could render them either independent or dependent.

To facilitate such a setting, we consider a functional marked point process (Ghorbani et al., 2021), which is defined as a marked point process where the marks are random elements in some (Polish) function space, $\mathcal{M}$, most notably the space of L_2-functions $f:\mathcal{T}\to\mathbb{R}^k$. Realisations of FMPPs are called functional marked point patterns. It is noteworthy that the original formal construction of functional marked point processes by Ghorbani et al. (2021) also included an additional non-functional mark, so that each ground process point would be marked by a pair which consists of a function and a non-functional variable. We here do not consider such auxiliary non-functional marks.

3.2 Product densities

Provided that it exists, the *n-th order intensity/product density* function $\rho^{(n)}, n\geq 1$, which is the density of the *n*-th order factorial moment measure $\alpha^{(n)}$, may be specified through the *n*-th order *Campbell formula*. It states that, for any non-negative measurable function *h* on $(\mathbb{R}^d\times\mathcal{M})^n$, the expectation of the random sum of *h* satisfies

$$\mathbb{E}\left[\sum_{(x_1,m_1),\ldots,(x_n,m_n)\in Y}^{\neq} h((x_1,m_1),\ldots,(x_n,m_n))\right]= \\ =\int\cdots\int h((x_1,m_1),\ldots,(x_n,m_n))\rho^{(n)}((x_1,m_1),\ldots,(x_n,m_n))\prod_{i=1}^{n}\mathrm{d}x_i\nu(dm_i), \tag{1}$$

where $\neq$ indicates that the sum is over *n*-tuples of distinct points of *Y*. Heuristically, $\rho^{(n)}((x_1,m_1),\ldots,(x_n,m_n))\mathrm{d}x_1\nu(dm_1)\cdots\mathrm{d}x_n\nu(dm_n)$ gives the probability that *Y* has points in infinitesimal neighbourhoods $d(x_i,m_i)\ni(x_i,m_i)\in\mathbb{R}^d\times\mathcal{M}$ with measures $\mathrm{d}x_i\nu(dm_i), i=1,\ldots,n$. Moreover, we retrieve $\alpha^{(n)}((A_1\times E_1)\times\cdots\times(A_n\times E_n)), (A_i\times E_i)\in\mathcal{B}(\mathbb{R}^d\times\mathcal{M}), i=1,\ldots,n$, by letting *h* be given

Accepted Manuscript

by the indicator function $\mathbf{1}\{(x_1,m_1)\in(A_1\times E_1),\ldots,(x_n,m_n)\in(A_n\times E_n)\}$. It further follows that

$$\rho^{(n)}((x_1,m_1),\ldots,(x_n,m_n))=f_{x_1,\ldots,x_n}(m_1,\ldots,m_n)\rho_g^{(n)}(x_1,\ldots,x_n),$$

where $\rho_g^{(n)}$ is the *n*-th order product density of Y_g and $f_{x_1,\ldots,x_n}(\cdot)$ is a conditional density function on $\mathcal{M}^n$ which governs the joint distribution of *n* marks, given that their associated ground process points are given by $x_1,\ldots,x_n\in\mathbb{R}^d$. These, in turn, yield the corresponding mark distributions

$$M^{x_1,\ldots,x_n}(E_1,\ldots,E_n)=\int_{E_1}\cdots\int_{E_n}f_{x_1,\ldots,x_n}(m_1,\ldots,m_n)\prod_{i=1}^{n}\nu(dm_i),$$

which govern the joint distribution on *n* marks, given the associated ground process locations.

The intensity measure of *Y*, which coincides with the first order factorial moment measure, here satisfies

$$\begin{aligned}\alpha(A\times E)&=\mathbb{E}[Y(A\times E)]=\int_A\int_E\rho(x,m)\mathrm{d}x\nu(dm)\\&=\int_A\int_E f_x(m)\rho_g(x)\mathrm{d}x\nu(dm)=\int_A M^x(E)\rho_g(x)\mathrm{d}x,\end{aligned}\qquad(2)$$

where the first order intensity functions $\rho=\rho^{(1)}$ and $\rho_g=\rho_g^{(1)}$ are typically referred to as *the intensity functions* of *Y* and Y_g. Note that ρ may be viewed as a "heat map" which reflects the infinitesimal chance of having a point of *Y* at/around an arbitrary location in $\mathbb{R}^d\times\mathcal{M}$. When the intensity function (of the ground process) is constant, we say that the (ground) process is homogeneous, otherwise it is called inhomogeneous.

When, conditional on the ground process, all marks have the same marginal univariate distribution, so that $M^z(E)=\int_E f_z(m)\mathrm{d}\nu(dm)=\int_E f(m)\mathrm{d}\nu(m)=M(E)$, we say that *X* has a common (marginal) mark distribution. This holds e.g. when *Y* is

Accepted Manuscript

stationary, i.e. when its distribution is invariant under translations of the ground points; here $\alpha(A\times E)=\rho_g M(E)|A|$ and $\rho_g>0$ is the constant intensity of the ground process. We will see that, at times, it is particularly convenient to have here that the reference measure ν coincides with the common mark distribution M, which implies that the common mark density f is set to 1 and $\rho(x,m)=\rho_g(x)$.

When Y is independently marked, i.e. when the marks are independent conditional on the ground process, $f_{x_1,\ldots,x_n}(m_1,\ldots,m_n)=f_{x_1}(m_1)\cdots f_{x_n}(m_n)$ for any $n\geq 1$ and if, in addition, there is a common mark distribution, whereby the marks are iid conditional on the ground process, we say that Y is randomly labelled and note that $f_{x_1,\ldots,x_n}(m_1,\ldots,m_n)=f(m_1)\cdots f(m_n)$.

3.2.1 Intensity reweighted stationarity

We next turn to the notion of a k-th order marked intensity reweighted stationary (k-MIRS) marked point process Y (Ghorbani et al., 2021). We say that Y is k-MIRS, $k\in\{1,2,\ldots\}$, if ρ is bounded away from 0 and the correlation functions

$$g^{(n)}((x_1,m_1),\ldots,(x_n,m_n))=$$
$$=\frac{\rho^{(n)}((x_1,m_1),\ldots,(x_n,m_n))}{\rho(x_1,m_1)\cdots\rho(x_n,m_n)}=\frac{f_{x_1,\ldots,x_n}(m_1,\ldots,m_n)}{f_{x_1}(m_1)\cdots f_{x_n}(m_n)}\frac{\rho_g^{(n)}(x_1,\ldots,x_n)}{\rho_g(x_1)\cdots\rho_g(x_n)},\quad n\geq 1,$$

satisfy $g^{(n)}((x_1,m_1),\ldots,(x_n,m_n))=g^{(n)}((z+x_1,m_1),\ldots,(z+x_n,m_n))$ for any $z\in\mathbb{R}^d$ and any $n\leq k$. Note that $g^{(1)}(\cdot)\equiv 1$ and that the second ratio on the right hand side is the n-th order correlation function, $g_g^{(n)}$, of the ground process. Provided that the product densities of all orders exist, stationarity implies k-MIRS for all orders $k\geq 1$. Note further that $g^{(n)}(\cdot)\equiv 1, n\geq 1$, for a Poisson process and when $g^{(n)}((x_1,m_1),\ldots,(x_n,m_n))>1$ points of Y_g in infinitesimal neighbourhoods of $x_1,\ldots,x_n$ with marks in infinitesimal neighbourhoods of $m_1,\ldots,m_n$ tend to cluster/aggregate. Similarly, $g^{(n)}((x_1,m_1),\ldots,(x_n,m_n))<1$ indicates inhibition/regularity.

Accepted Manuscript

3.3 Palm distributions

Let Y be a simple marked point process whose intensity function exists. Many of the summary statistics we will consider can be expressed in terms of *reduced Palm distributions*. These satisfy the *reduced Campbell–Mecke* formula which states that, for any non-negative measurable function h on the product space $(\mathbb{R}^d \times \mathcal{M}) \times N_{lf}$,

$$\mathbb{E}\left[\sum_{(z,m)\in Y} h((z,m), Y \setminus \{(z,m)\})\right] = \int \mathbb{E}[h((x,m), Y^{!(x,m)})]\rho(x,m)\mathrm{d}x\nu(dm) \\ = \int \mathbb{E}^{!(x,m)}[h((x,m), Y)]\rho(x,m)\mathrm{d}x\nu(dm). \quad (3)$$

Here $Y^{!(x,m)}$ is the reduced Palm process at $(x,m) \in \mathbb{R}^d \times \mathcal{M}$, which we interpret as Y conditioned on the null event that there is a point in (x, m), which is removed upon realisation. The probability distribution $P^{!(x,m)}(\cdot) = \mathbb{P}^{!(x,m)}(Y \in \cdot) = \mathbb{P}(Y^{!(x,m)} \in \cdot)$ on $(N_{lf}, \mathcal{N})$, which corresponds to $\mathbb{E}^{!(x,m)}$, is called the reduced Palm distribution at (x, m).

4 Local weighted marked summary statistics

Global summary statistics have had a prominent role in the statistical analysis of point processes. More precisely, their non-parametric estimators are typically used to characterise the degree of spatial interaction present in the underlying data-generating point process. In Section 1, we have reviewed a few such examples, for instance K-functions.

The individual contributions to a global statistic, which are commonly called Local Indicators of Spatial Association (LISA) functions, can be used to identify outlying components measuring the influence of each contribution to the global statistic (Anselin, 1995). This is the case of the scatter plot based on the local Moran index (Anselin, 1996). On the other hand, the individual contributions can be used to test for specific local structures, such as spatial association and hot spot

Accepted Manuscript

detection in areal data (Getis and Ord, 1992). Basically, the local statistics mentioned so far are often used to analyse areal data but Getis and Franklin (1987) introduced a local version of the *K*-function for spatial point processes to show that trees exhibit different kinds of heterogeneity when examined at different scales of analysis. The notion of individual functions for certain statistics has also been studied in Stoyan and Stoyan (1994) and Mateu et al. (2010) showed that the local product density function (Cressie and Collins, 2001) is more sensitive to identifying different local structures and unusual points than the local *K*-function. Applications of LISA functions range from detection of features in images with noise (Mateu et al., 2007) to detection of disease clusters (Moraga and Montes, 2011). In Siino et al. (2018) the authors extend local indicators of spatial association to the spatio-temporal context (LISTA functions) based on the second order product density, and these local functions have been used to define a proper statistical test for clustering detection. Recently, LISTA functions have been used both for diagnostic (Adelfio et al., 2020) and fitting purposes (D'Angelo et al., 2023). Finally, D'Angelo et al. (2021) extended LISTA functions to spatio-temporal point processes living on linear networks.

As we have clearly indicated, an alternative to studying the aforementioned global summary statistics for marked point processes is considering local summary statistics which describe the spatial interaction in the vicinity of a given marked point. In order to do so here in the marked context, we introduce the function

$$L((x,m),\mathbf{x}) = L_n((x,m),\mathbf{x};\tilde{t},\tilde{\rho}) = \sum_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in\mathbf{x}}^{\neq} \frac{\tilde{t}((x,m),(x_1,m_1),\ldots,(x_{n-1},m_{n-1}))}{\tilde{\rho}(x,m)\tilde{\rho}(x_1,m_1)\cdots\tilde{\rho}(x_{n-1},m_{n-1})}, \quad (4)$$

for $(x,m)\in\mathbb{R}^d\times\mathcal{M}$, point pattern $\mathbf{x}\in N_{lf}$, and measurable $\tilde{t}:(\mathbb{R}^d\times\mathcal{M})^n\to\mathbb{R}$, $n\geq 2$. Note that, formally, the argument $\tilde{\rho}$ does not need to be the true intensity function ρ of Y, it could e.g. be a plug-in estimator. We will exploit Definition 1,

Accepted Manuscript

and thereby (4), to define proper notions of (mark-weighted *n*-th order inhomogeneous) local summary statistics.

Definition 1. Given a marked point process *Y*, we refer to $L((x,m),Y \setminus \{(x,m)\};\tilde{t},\tilde{\rho}),(x,m) \in Y$, as the family of *n*-th order local marked cumulative summary statistics of *Y* associated with $\tilde{t}$ and $\tilde{\rho}$.

The construction of a specific local statistic is obtained by identifying when, for some function family $\{\tilde{t}_r\}$,

$$G(r,Y) = \sum_{(x,m)\in Y} L_n((x,m),Y \setminus \{(x,m)\};\tilde{t}_r,\tilde{\rho}) \quad (5)$$

forms an estimator of an existing global summary statistic.

Using *n*-th order local marked cumulative summary statistics to quantify local spatial interactions for a point pattern **x** entails inserting an estimate $\hat{\rho}(x,m) = \hat{f}_z(m)\hat{\rho}_g(x)$ for the unknown intensity $\rho(x,m) = f_z(m)\rho_g(x)$, i.e. setting $\tilde{\rho} = \hat{\rho}$. When we assume that there is a common mark distribution which coincides with the mark reference measure *v*, we obtain that $\hat{\rho}(x,m) = \hat{\rho}_g(x)$, i.e. the intensity estimate does not depend on the mark values. Imposing this assumption is particularly convenient when dealing with functional marks since estimation of the mark density, which here is a density on a function space, is rather challenging and beyond the scope of this paper. Note that when *Y* is randomly labelled, it has a common mark distribution and in this setting the assumption $\hat{\rho}(x,m) = \hat{\rho}_g(x)$ thus makes sense.

Turning to the distributional properties of the *n*-th order local marked cumulative summary statistics, we next derive their expectations under the assumption of *k*-MIRS. Note, in particular, that the choice of $\tilde{t}$ plays a significant role here.

Accepted Manuscript

Theorem 1. *When Y is k-MIRS and* $\tilde{\rho}=\rho$*, for any* $W\in\mathcal{B}(\mathbb{R}^d)$ *the expectation of* $L((x,m),Y\setminus\{(x,m)\}\cap W\times\mathcal{M};\tilde{t},\rho),(x,m)\in Y\cap W\times\mathcal{M}$*, is almost everywhere given by*

$$\int_{W-x}\cdots\int_{W-x}\Big(\int_{\mathcal{M}}\cdots\int_{\mathcal{M}}\tilde{t}((x,m),(x_1+x,m_1),\ldots,(x_{n-1}+x,m_{n-1}))\times$$
$$\times\frac{f_{0,x_1,\ldots,x_{n-1}}(m,m_1,\ldots,m_{n-1})}{f_0(m)\,f_{x_1}(m_1)\cdots f_{x_{n-1}}(m_{n-1})}\,\nu(dm_1)\cdots\nu(dm_{n-1})\Big)g_g^{(n)}(0,x_1,\ldots,x_{n-1})\mathrm{d}x_1\cdots\mathrm{d}x_{n-1},$$

when $2\le n\le k$*. Moreover, the expectation of* $G(r,Y\cap W\times\mathcal{M})$ *is obtained by replacing* $\tilde{t}$ *by* $\tilde{t}_r$ *in the expression above and integrating it over* $W\times\mathcal{M}$ *with respect to the reference measure on* $\mathbb{R}^d\times\mathcal{M}$*.*

Proof. Note first that the expectation coincides with

$$\mathbb{E}^{!(x,m)}[L_n((x,m),Y\cap W\times\mathcal{M};\tilde{t},\rho)]=$$
$$=\mathbb{E}^{!(x,m)}\left[\sum^{\neq}_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in Y\cap W\times\mathcal{M}}\frac{\tilde{t}((x,m),(x_1,m_1),\ldots,(x_{n-1},m_{n-1}))}{\rho(x,m)\rho(x_1,m_1)\cdots\rho(x_{n-1},m_{n-1})}\right].$$

Hence, our starting point will be the reduced Campbell-Mecke formula. Consider an arbitrary bounded $A\times E\in\mathcal{B}(\mathbb{R}^d\times\mathcal{M})$. It follows that

$$\mathbb{E}\left[\sum_{(x,m)\in Y\cap A\times E}\ \sum^{\neq}_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in Y\setminus\{(x,m)\}\cap W\times\mathcal{M}}\frac{\tilde{t}((x,m),(x_1,m_1),\ldots,(x_{n-1},m_{n-1}))}{\rho(x,m)\rho(x_1,m_1)\cdots\rho(x_{n-1},m_{n-1})}\right]=$$
$$=\int_{A\times E}\mathbb{E}^{!(x,m)}\left[\sum^{\neq}_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in Y\cap W\times\mathcal{M}}\frac{\tilde{t}((x,m),(x_1,m_1),\ldots,(x_{n-1},m_{n-1}))}{\rho(x_1,m_1)\cdots\rho(x_{n-1},m_{n-1})}\right]\mathrm{d}x\,\nu(dm).$$

On the other hand, by the Campbell formula we have that

Accepted Manuscript

$$\mathbb{E}\left[\sum_{(x,m)\in Y\cap A\times E}\ \sum_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in Y\setminus\{(x,m)\}\cap W\times\mathcal{M}}^{\neq}\frac{\tilde{t}((x,m),(x_1,m_1),\ldots,(x_{n-1},m_{n-1}))}{\rho(x,m)\rho(x_1,m_1)\cdots\rho(x_{n-1},m_{n-1})}\right]=$$

$$=\int_{A\times E}\int_{\mathbb{R}^d\times\mathcal{M}}\cdots\int_{\mathbb{R}^d\times\mathcal{M}}\mathbf{1}\{x_1,\ldots,x_{n-1}\in W\}\tilde{t}((x,m),(x_1,m_1),\ldots,(x_{n-1},m_{n-1}))\times$$

$$\times g^{(n)}((x,m),(x_1,m_1),\ldots,(x_{n-1},m_{n-1}))\mathrm{d}x_1\nu(dm_1)\cdots\mathrm{d}x_{n-1}\nu(dm_{n-1})\mathrm{d}x\nu(dm)$$

$$=\int_{A\times E}\int_{\mathbb{R}^d\times\mathcal{M}}\cdots\int_{\mathbb{R}^d\times\mathcal{M}}\prod_{i=1}^{n-1}\mathbf{1}\{u_i\in W-x\}\tilde{t}((x,m),(u_1+x,m_1),\ldots,(u_{n-1}+x,m_{n-1}))\times$$

$$\times g^{(n)}((0,m),(u_1,m_1),\ldots,(u_{n-1},m_{n-1}))\mathrm{d}u_1\nu(dm_1)\cdots\mathrm{d}u_{n-1}\nu(dm_{n-1})\mathrm{d}x\nu(dm)$$

by the imposed *k*-MIRS and a change of variables, $u_i+x=x_i$. Hence, since $A\times E\in\mathcal{B}(\mathbb{R}^d\times\mathcal{M})$ was arbitrary, for almost every (*x*, *m*) we have that

$$\mathbb{E}^{!(x,m)}[L_n((x,m),Y;\tilde{t},\rho)]=$$

$$=\int_{(W-x)\times\mathcal{M}}\cdots\int_{(W-x)\times\mathcal{M}}\tilde{t}((x,m),(u_1+x,m_1),\ldots,(u_{n-1}+x,m_{n-1}))\times$$

$$\times g^{(n)}((0,m),(u_1,m_1),\ldots,(u_{n-1},m_{n-1}))\mathrm{d}u_1\nu(dm_1)\cdots\mathrm{d}u_{n-1}\nu(dm_{n-1})$$

$$=\int_{W-x}\cdots\int_{W-x}\Big(\int_{\mathcal{M}}\cdots\int_{\mathcal{M}}\tilde{t}((x,m),(u_1+x,m_1),\ldots,(u_{n-1}+x,m_{n-1}))\times$$

$$\times\frac{f_{0,u_1,\ldots,u_{n-1}}(m,m_1,\ldots,m_{n-1})}{f_0(m)\,f_{u_1}(m_1)\cdots f_{u_{n-1}}(m_{n-1})}\nu(dm_1)\cdots\nu(dm_{n-1})\Big)g_g^{(n)}(0,u_1,\cdots,u_{n-1})\mathrm{d}u_1\cdots\mathrm{d}u_{n-1},$$

by Fubini's theorem. □

The first thing we note is that when Y is independently marked then the density ratio in the expression for the expectation vanishes. In addition, if Y is a Poisson process on $\mathbb{R}^d\times\mathcal{M}$ which satisfies being a marked point process with mark space $\mathcal{M}$, then the expectation reduces to an integral with $\tilde{t}$ as integrand. These observations may be used as benchmarks for when Y exhibits mark (in)dependence and spatial interaction locally.

4.1 Special cases

We next illustrate how (5), through Definition 1 and (4), reduces to several existing summary statistic estimators by varying $\tilde{t}$ and $\tilde{\rho}$.

Accepted Manuscript

4.1.1 Ground *K*-functions

First, set $n = 2$ and $\tilde{t}$ to $\tilde{t}_r((x,m),(x_1,m_1)) = w(x,x_1)\mathbf{1}\{x_1 \in x + C\}/|W|, r \geq 0$, where $W \subseteq \mathbb{R}^d, |W| > 0$, and $w(\cdot)$ is an edge correction term. If the ground process is stationary with intensity $\rho_g > 0$ and $\tilde{\rho}(x,m) \equiv \rho_g$, then (5) with Y set to $Y \cap W \times \mathcal{M}$ reduces to an estimator of Ripley's *K*-function when $x + C = x + b[0,r] = b[x,r]$ whereas if the ground process is inhomogeneous and we set $\tilde{\rho}(x,m) = \rho_g(x)$, it follows that (5) reduces to an estimator of the inhomogeneous *K*-function (Baddeley et al., 2000) for Y_g. The extension to space-time is straightforward; replace the Euclidean ball $b[0,r]$ by $C = \{(x,s) : \|x\| \leq r, |s| \leq t\} \in \mathcal{B}(\mathbb{R}^{d+1})$, where $\|\cdot\|$ denotes the Euclidean norm (Cronie and Van Lieshout, 2015; Gabriel and Diggle, 2009; Iftimi et al., 2019).

4.1.2 Marked *K*-functions

When $n = 2$, by instead letting $\tilde{t}_r((x,m),(x_1,m_1)) = w(x,x_1)\mathbf{1}\{x_1 \in x + C\}\mathbf{1}\{m \in E, m_1 \in E_1\}/(|W|\nu(E)\nu(E_1))$ and $\tilde{\rho} = \rho$ in (4), using a suitable edge correction function $w(\cdot)$, then $G(r, Y \cap W \times \mathcal{M})$ in (5) reduces to an estimator of the *marked second order reduced moment measure* $\mathcal{K}^{EE_1}(C)$ of Iftimi et al. (2019), which measures the intensity reweighted interactions between points with marks in E and points with marks in E_1, when their separation vectors belong to $C \in \mathcal{B}(\mathbb{R}^d)$. We note that measures of this kind are in general not symmetric, i.e. $\mathcal{K}^{EE_1}(\cdot) \neq \mathcal{K}^{E_1E}(\cdot)$ (Iftimi et al., 2019). Furthermore, choosing C to be the closed origin-centred ball $b[0,r]$ of radius $r \geq 0$, we consider the marked inhomogeneous *K*-function $K_{inhom}^{EE_1}(r)$ of Cronie and van Lieshout (2016), which measures pairwise intensity reweighted spatial dependence within distance r between points with marks in E and points with marks in E_1.

By additionally letting $n > 2$, we obtain a definition of a *marked n-th order reduced moment measure*, $\mathcal{K}^{E \times \prod_{i=1}^{n-1} E_i}(C_1 \times \cdots \times C_{n-1})$, which measures the intensity

Accepted Manuscript

reweighted spatial interaction between an arbitrary point with mark in E and distinct $(n-1)$-tuples of other points, where the separation vectors between the E-marked point and these $n-1$ points, which have marks in $E_1,\ldots,E_{n-1}$, belong to $C_1,\ldots,C_{n-1}$. We note that $C_i = b[0,r], i=1,\ldots,n-1, r\geq 0$, yields an n-point version of the marked inhomogeneous K-function $K_{inhom}^{E\times\prod_{i=1}^{n-1}E_i}(r)$ of Cronie and van Lieshout (2016), which may be used to analyse intensity reweighted interactions between a point with mark in E and $n-1$ of its r-close neighbours, which have marks belonging to the respective sets $E_1,\ldots,E_{n-1}$.

4.1.3 Weighted marked reduced moment measures and *K*-functions

Finally, by letting $\tilde{\rho}=\rho$ and $\tilde{t}((x,m),(x_1,m_1),\ldots,(x_{n-1},m_{n-1}))$ be given by the product of

$$\begin{aligned}\tilde{t}(m,m_1,\ldots,m_{n-1}) &= t(m,m_1,\ldots,m_{n-1})\frac{\mathbf{1}\{m\in E\}}{\nu(E)}\prod_{i=1}^{n-1}\frac{\mathbf{1}\{m_i\in E_i\}}{\nu(E_i)},\\ \tilde{w}(x,x_1,\ldots,x_{n-1}) &= w(x,x_1,\ldots,x_{n-1})\prod_{i=1}^{n-1}\mathbf{1}\{x_i\in(x+C_i)\},\end{aligned} \tag{6}$$

for $E\in\mathcal{B}(\mathcal{M}),\ \nu(E)>0$, and $C_i\times E_i\in\mathcal{B}(\mathbb{R}^d)\times\mathcal{B}(\mathcal{M})=\mathcal{B}(\mathbb{R}^d\times\mathcal{M}),\ \nu(E_i)>0, i=1,\ldots,n-1$, we obtain an unbiased estimator $\mathcal{K}_t^{E\times\prod_{i=1}^{n-1}E_i}(C_1\times\cdots\times C_{n-1}) = G(r, Y\cap W\times\mathcal{M})$ of the t-weighted marked n-th order reduced moment measure of Ghorbani et al. (2021),

$$\begin{aligned}&\mathcal{K}_t^{E\times\prod_{i=1}^{n-1}E_i}(C_1\times\cdots\times C_{n-1}) =\\ &=\frac{1}{|W|\,\nu(E)\prod_{i=1}^{n-1}\nu(E_i)}\mathbb{E}\Bigg[\sum_{(x,m)\in Y\cap W\times E}\ \sum_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in Y\setminus\{(x,m)\}}^{\neq}\frac{t(m,m_1,\ldots,m_{n-1})}{\rho(x,m)}\times\\ &\times\prod_{i=1}^{n-1}\frac{\mathbf{1}\{x_i-x\in C_i\}\mathbf{1}\{m_i\in E_i\}}{\rho(x_i,m_i)}\Bigg]\end{aligned} \tag{7}$$

assuming that the edge correction function w is such that unbiasedness holds. Examples of such w include the minus sampling edge correction and the

Accepted Manuscript

translational edge correction (Ghorbani et al., 2021). Note here that one just as well could have merged the scaled indicators in the expression for $\tilde{t}$ with t so that $\tilde{t} = t$; Ghorbani et al. (2021) included this mark set filtering to highlight that their summary statistic generalises previously proposed ones.

4.2 Local t-weighted marked n-th order inhomogeneous K-function

In this section, we provide the estimator corresponding to the local contributions of (7) and discuss its properties.

Definition 2. Let $\tilde{t}$ be (up to indicator-scaling) as in (6) and consider

$$\mathcal{K}_t^{(x,m)\times\prod_{i=1}^{n-1}E_i}(C_1\times\cdots\times C_{n-1}) = L_n((x,m), Y\setminus\{(x,m)\}\cap W\times\mathcal{M}; \tilde{t}, \tilde{\rho}) =$$
$$= \frac{1}{\tilde{\rho}(x,m)\nu(E)\prod_{i=1}^{n-1}\nu(E_i)} \sum_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in Y\setminus\{(x,m)\}\cap W\times\mathcal{M}}^{\neq} w(x,x_1,\ldots,x_{n-1})\times \qquad (8)$$
$$\times t(m,m_1,\ldots,m_{n-1})\prod_{i=1}^{n-1}\frac{\mathbf{1}\{x_i - x\in C_i\}\mathbf{1}\{m_i\in E_i\}}{\tilde{\rho}(x_i,m_i)}, \quad (x,m)\in Y\cap W\times\mathcal{M},$$

for some suitable edge correction w in (6), $W\in\mathcal{B}(\mathbb{R}^d)$, $E\in\mathcal{B}(\mathcal{M})$, $\nu(E)>0$, and $C_i\times E_i\in\mathcal{B}(\mathbb{R}^d\times\mathcal{M})$, $\nu(E_i)>0$, $i=1,\ldots,n-1$. We refer to $\mathcal{K}_t^{(x,m)\times\prod_{i=1}^{n-1}E_i}(r) = \mathcal{K}_t^{(x,m)\times\prod_{i=1}^{n-1}E_i}(b[0,r]^{n-1})$, $r\geq 0$, as a *local t-weighted marked n-th order inhomogeneous K-function*. In particular, $\mathcal{K}_{t,n}^{(x,m)}(r) = \mathcal{K}_t^{(x,m)\times\mathcal{M}^{n-1}}(r)$ does not perform any explicit mark set filtering.

Note first that when there is a common mark distribution which coincides with the reference measure on $\mathcal{M}$, setting $\tilde{\rho}=\rho$ we, for instance, obtain

$$\mathcal{K}_{t,n}^{(x,m)}(r) = \sum_{(x_1,m_1),\ldots,(x_{n-1},m_{n-1})\in Y\setminus\{(x,m)\}\cap(b[x,r]\cap W)\times\mathcal{M}}^{\neq} \frac{t(m,m_1,\ldots,m_{n-1})w(x,x_1,\ldots,x_{n-1})}{\rho_g(x)\rho_g(x_1)\cdots\rho_g(x_{n-1})}$$

since ν must be a probability measure here.

Accepted Manuscript

Regarding the distributional properties of (8), when Y is k-MIRS, Theorem 1 tells us that the expectation is given by

$$\frac{1}{\nu(E)}\prod_{i=1}^{n-1}\frac{1}{\nu(E_i)}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} w(x,x_1+x,\ldots,x_{n-1}+x)\prod_{i=1}^{n-1}\mathbf{1}\{x_i\in(x+C_i)\cap(W-x)\}\times$$
$$\times\left(\int_{E_1}\cdots\int_{E_{n-1}} t(m,m_1,\ldots,m_{n-1})\frac{f_{0,x_1,\ldots,x_{n-1}}(m,m_1,\ldots,m_{n-1})}{f_0(m)\,f_{x_1}(m_1)\cdots f_{x_{n-1}}(m_{n-1})}\,\nu(dm_1)\cdots\nu(dm_{n-1})\right)\times$$
$$\times g_g^{(n)}(0,x_1,\ldots,x_{n-1})\mathrm{d}x_1\cdots\mathrm{d}x_{n-1}.$$

In particular, under independent marking the mark related integral within brackets reduces to $\int_{E_1}\cdots\int_{E_{n-1}} t(m,m_1,\ldots,m_{n-1})\nu(dm_1)\cdots\nu(dm_{n-1})$, whereby (8) is given by the product of this term and a term measuring intensity reweighted spatial interaction.

4.2.1 Test functions for FMPPs

Turning to the FMPP case, by choosing different test functions $t(\cdot)$ for the functional marks, we may extract different features. We here focus on pairwise interactions, i.e. $n = 2$.

The test function t is intended to reflect similarities between functions. Hence, a natural starting point would be a metric $t(f_1,f_2)=d(f_1,f_2)$ on the function space $\mathcal{M}$, which does not necessarily need to be the underlying assumed metric on $\mathcal{M}$. The first candidate that comes to mind is an L_p-distance:

$$t(f_1,f_2)=\left(\int_a^b |f_1(t)-f_2(t)|^p\,\mathrm{d}t\right)^{1/p},\quad 1\le p\le\infty, \qquad (9)$$

where $p=\infty$ represents the supremum metric. For any choice of p in (9), similarity between functions implies a small value of the test function. Other tentative functions are semi-metrics based on the L_p distance between the s-th derivatives of the functions, for different combinations of p and s, with the L_1 and

Accepted Manuscript

L_2 distances being particular cases, and semi-metrics based on functional principal component analysis.

A further alternative is the functional marked counterpart of the test function for the classical variogram, given by

$$t(f_1, f_2) = \int_a^b (f_1(t) - \bar{F}(t))(f_2(t) - \bar{F}(t))\mathrm{d}t, \quad (10)$$

with $\bar{F}(t) = (1/n)\sum_{i=1}^{n} f_i(t)$ being the average functional mark at time t for the observed functional part of the point pattern; such averaging is motivated by the assumption of a common mark distribution.

5 Local test for random labelling

Simple hypotheses for spatial point patterns, such as Complete Spatial Randomness, are commonly tested using an estimator of a global summary statistic, e.g., Ripley's K-function. In this context, one typically resorts to Monte Carlo testing. The first step is then to generate Q simulations under the null hypothesis, and to estimate the chosen summary statistic for both the observed pattern and the simulations. In order to study whether there is *random labelling* in a (functional) marked point process, the simulations are obtained by permuting the (functional) marks, that is, randomly assigning them to the spatial points of the ground pattern, which are kept fixed. Then, the chosen summary statistic is estimated for each of these permutations and global envelopes at a given nominal level are generated based on them. The result of the test can be assessed graphically: if the summary statistic estimate for the observed pattern exits the envelopes, we proceed with the assumption that the underlying FMPP is not randomly labelled. Furthermore, it is possible to calculate a p-value based on the position of the observed summary statistic within the qth envelopes, following Myllymäki et al. (2017). We know, however, that the conclusion drawn from the application of the above-mentioned global test pertains to the whole analysed

Accepted Manuscript

process, indicating whether all the functional marks are randomly labelled or not. Motivated by the will to further detect the specific points, and regions, where the functional marks really do depend on the other marked points, we propose a *local test for random labelling*. The main idea is to run a global envelope test on each point of the analysed pattern by means of the previously proposed *local t-weighted marked inhomogeneous K-functions*, to draw different conclusions about the individual points, based on the obtained *p*-values. In Algorithm 1 we outline the proposed local test. Note that we alternatively may use sampling without replacement in step 5 of Algorithm 1. Moreover, if convinced that multiple testing issues are present here, one may adjust the type I error probability α by using e.g. the Holm-Bonferroni method.

Algorithm 1 Local test of random labelling

1: Set a fixed nominal value α for type I error;

2: Consider a (functional) marked point pattern $\mathbf{x} = \{(x_j, m_j)\}_{j=1}^{k}, k \geq 1$;

3: Set a number of simulations, $Q \geq 1$;

4: **for** each $q = 1, \ldots, Q$: **do**

5: Randomly sample k (functional) marks, with replacement, from the original k ones;

6: Denote the resulting point pattern by $\mathbf{x}_q = \{(x_j, m_j^q)\}_{j=1}^{k}$;

7: **end for**

8: **for** each $j = 1, \ldots, k$, **do**

9: Compute $L_n^{(j,q)} = \{\mathcal{K}_t^{(x_j, m_j^q) \times_{i=1}^{n-1} E_i}(r; \mathbf{x}_q)\}_{r \in [0, r_{max}]}$ for all $q = 1, \ldots, Q$;

Accepted Manuscript

10: Apply global envelope testing, using the functions $L_n^{(j,q)}, q = 1,\ldots,Q$, to generate the envelopes;

11: Obtain a *p*-value p_j from the test;

12: Reject the null hypothesis for the j^{th} point if $p_j \leq \alpha$.

13: **end for**

6 Motivating example and simulation study

This section is dedicated to simulation studies to assess the performance of our proposed local test. First, section 6.1 provides a motivating example of the use of such a test, by means of simulated data resembling seismic events, which in turn have motivated this work. In particular, this means simulating the functional marks as seismic waveforms, following the typical abrupt change in variance of the signal in correspondence with the arrivals of the first P- and S-waves. Then, section 6.2 presents an extensive simulation study, showing diverse and more general settings. Specifically, we assess the performance of the test by summarising the results in terms of classification rates.

6.1 The need for a local test

We simulate a homogeneous spatial point pattern with 250 points on the unit square, $W = [0,1] \times [0,1]$, which represents the ground pattern. For each ground point x_i, we simulate a functional mark of the from

$$
\begin{aligned}
&f_i(t) = y(t) = \mu(t) + \epsilon(t), \qquad t \in \mathcal{T} = [0,1], \\
&\epsilon(t) \sim N(0, \sigma(t)^2), \\
&\sigma(t)^2 = 0.2 + 7.5\mathbf{1}\{t > 0.4\} - 5\mathbf{1}\{t > 0.6\},
\end{aligned}
$$

where the mean signal $\mu(t)$ is taken to be zero. The spatial ground point pattern and the corresponding waveform for a given point are shown in Figure 2(a)-(b). Since the marks/waveforms are simulated from the same model, and

Accepted Manuscript

independently of each other and the spatial locations of the points, we see that such a process is indeed randomly labelled.

Having generated the data, we first run a *global envelope test for random labelling*, by randomly permuting the simulated waveforms, i.e. the functional marks, keeping the location of the points fixed. We run the test by means of the t-weighted marked *n*-th order inhomogeneous *K*-function of Ghorbani et al. (2021), with $n = 2$, making it a second order summary statistic, and t given by the test function (10), i.e. the functional marked counterpart of the test function for the classical variogram. As previously mentioned, we assume that there is a common mark distribution which coincides with the reference measure on the mark space so that the intensity function is estimated by the ground process intensity estimate. To be as objective as possible, we do not use the homogeneous intensity estimator $\hat{\rho}_g(\cdot) = Y_g(W)/|W|$ here but instead we use a kernel intensity estimator, as in practice it would be unknown to us whether the actual ground process is (in)homogeneous. We use a Gaussian kernel intensity estimator $\hat{\rho}_g(\cdot)$, where we select the bandwidth, h, according to Cronie and Van Lieshout (2018). More specifically, we minimise the discrepancy between the area of the observation window and the sum of reciprocal estimated intensity values at the points of the point pattern, i.e. we minimise $CvL(h) = (|W| - \sum_i 1/\hat{\rho}_g(x_i;h))^2$, where the sum is taken over all the data points x_i and $\hat{\rho}_g(x_i;h)$ is the kernel intensity estimate with bandwidth h, evaluated in x_i. Then, once the bandwidth has been selected, the intensity estimate is corrected for edge effects through global edge correction (the option `diggle=FALSE` in the `spatstat` function `density.ppp`), i.e. dividing the estimate by the convolution of the Gaussian kernel with the window of observation (Diggle, 1985). Finally, for w we use Ripley's isotropic edge correction in the summary statistic to correct for edge effects. We repeated the procedure 39 times, obtaining the result depicted in Figure 2(c). We stress that our approach seems to be robust with respect to

Accepted Manuscript

the bandwidth specification in this particular scenario setting, i.e. the choice of bandwidth selection approach plays a minor role in the final result.

As evident from Figure 2(c), the observed summary statistic completely lies within the envelopes, and this confirms the expected result of lack of spatial dependence/structure of the functional marks. This result is further corroborated by the non-significant *p*-value, equal to 0.25.

6.1.1 Simulating spatially dependent functional marks

To make the functional marks spatially dependent, we then superimpose a homogeneous spatial point pattern with 50 points, generated in the $[0,0.5]\times[0,0.5]$ square, i.e. the bottom left region of the entire study region *W*. For these additional points, we generate different functional marks than before, namely with the underlying trend $\mu(t)=10+6\sin(3\pi z_t)$. Consequently, we have simulated a FMPP with spatially varying functional marks, i.e. not random labelled. We therefore expect a global test of random labelling to confirm this.

We first run the same global test of random labelling as before. Here, the *K*-function is based on a kernel intensity estimate whose bandwidth is selected by Diggle (2013)'s rule. It represents a good alternative to Cronie and Van Lieshout (2018)'s one, being slightly faster to compute. We use $Q = 39$ and obtain a global *p*-value of 0.025. This, together with the observed *K*-function lying outside the envelopes (Figure 3(a)), indicates the ability of the global test to correctly detect the spatial dependence of the functional marks.

We know, however, that this conclusion should not be drawn for each point of the pattern, if we consider local restrictions of it, but specifically for those in the vicinity of the $[0,0.5]\times[0,0.5]$ square. We therefore proceed by running our proposed local test, based on the proposed second order local *K*-function $\mathcal{K}_{t,2}^{(x,m)}(r), r\in[0,r_{max}]$, in Definition (2), with the same choice of test function $t(\cdot)$ and the same intensity estimation scheme as for the global one. Figure 3(b)

Accepted Manuscript

depicts the points of the simulated point pattern, and it displays as black triangles those points for which the local test came out significant. Hence, this illustrates that the proposed local test is able to correctly identify some of the points, and consequently some parts of the region, where the hypothesis of random labelling does not hold locally. Note that a universally preferable option for r_{max} does not exist. In this paper, it is set to $\min(x_W, y_W)/4$, where x_W and y_W represent the maximum width and height of the observation region W, respectively; note that this rule of thumb is supported by Diggle (2013). Indeed, changing the value of r_{max} has an impact on the final results, and we found that our choice provided the best compromise among the options.

6.2 Extended simulation study

This section aims to study the proposed method's performance in terms of classification rates considering different scenarios, concerning both the ground processes and the functional marks' structures. To this end, we simulate under different such scenarios, to obtain a comprehensive understanding of the results of the local test in different settings.

In detail, we consider three types of ground process structures, all with an expected point count of 200: (1) a homogeneous Poisson process; (2) an inhomogeneous Poisson process with intensity function $\rho_g(x) = \rho_g(x_1, x_2) = \exp(3.5 + 3x_2),\ x \in W$; (3) a Thomas process, with intensity of the Poisson process of cluster centres equal to 25, standard deviation of random displacement of a point from its cluster centre equal to 0.05, and mean number of points per cluster equal to 7. They are all generated in W, i.e. the unit square, and will be referred to as the *base patterns*. Then, we superimpose additional simulated patterns in the $[0,0.5]\times[0,0.5]$ square, coming from the same generating processes, but with an expected number of points of 50; hereby the expected total number of points on $[0,0.5]\times[0,0.5]$ is $50 + 200/4 = 100$ and on its complement it is 150. These additional patterns will be referred to as *feature*

Accepted Manuscript

patterns. A graphical representation of these three ground patterns comes in Figure 4 (a) - (c).

As for the functional marks, we consider the time domain $\mathcal{T} = [0,10]$ and, practically, we sample each simulated mark function in 100 equally spaced time points in $\mathcal{T}$. We assume that each functional mark satisfies $f_i(t) = Z(x_i, t)$, where x_i is the ith ground point and

$$Z(x,t) = \mu + \xi(x,t), \quad (x,t) \in W \times \mathcal{T}, \quad (11)$$

for a zero-mean stationary Gaussian random field ξ with covariance function $C(h, u)$; here h and u denote the spatial and the temporal lags, respectively. For the base patterns, we consider $\mu = 5$ and a pure nugget effect model with covariance function $C(h,u) = \sigma^2 \mathbf{1}\{h = 0\}, \sigma^2 = 0.01$. In other words, each f_i is random noise with mean 5 and variance 0.01 and all f_is are iid; see the grey curves in Figure 4 (d) - (f). For the feature patterns, we consider three different marking models:

1. Shifted base model: We here let ξ have the same form as in the base model but let $\mu = 5.5$.
2. Decreased variance base model: We here let ξ have the same form as in the base model but let $\sigma^2 = 0.001$.
3. Non-separable space-time model: We here let $\mu = 5$ and consider a space isotropic covariance function given by $C(h,u) = (\psi(u)+1)^{-\delta/2} \phi(h/\sqrt{\psi(u)+1})$. Here, ϕ is a normal mixture and the corresponding covariance function only depends on the distance between two points, while ψ is a variogram model, which we choose according to a fractal Brownian motion with fractal dimension $\alpha = 1$; this is an intrinsically stationary isotropic variogram model.

We note that the first two of these scenarios represent independent but not identically distributed marks, whereas in the third scenario we additionally have

Accepted Manuscript

that the marks are also dependent. In Figure 4 (d) - (f), the functional marks corresponding to the marking models in item 1, 2, and 3 are depicted.

We show the results of the local test in terms of true-positive rate (TPR), false-positive rate (FPR), and accuracy (ACC), averaging over 100 simulated point patterns in Table 1. The rates are defined as

$$TPR = \frac{\text{true positives}}{\text{positives}}, \quad FPR = \frac{\text{false negatives}}{\text{negatives}}, \quad ACC = \frac{\text{true positives and negatives}}{\text{positives and negatives}}.$$

We of course wish to have TPR and ACC close to 1 and FPR close to 0.

As shown in Table 1, the performance of the local test in terms of classification rates strongly depends on the difference in the functional marks. Specifically, changing only the mean of the underlying random field is not enough for properly identifying the points of the feature patterns. This sufficiently improves when changing the variance only, but the best result is obtained when the whole model is changed, that is, changing the correlation structure. The effect of the type of ground pattern is less evident but still present. The inhomogeneous Poisson scenario reports the best classification rates, followed by the Thomas and homogeneous Poisson ones.

Finally, we found that the test function $t(\cdot)$ based on the L_2 distance in Equation (9) gave the better results overall. To further explore how the choice of test function influences the test, we also compared to a test function incorporating a derivative function accounting for the shape of the functional marks. This yielded similar results but turned out to be more computationally demanding.

7 Real seismic data analysis

We analyse data coming from the *ISTANCE* dataset, presented in Section 2. More specifically, we analyse a sample dataset provided at http://www.pi.ingv.it/instance/. The observed point pattern consists of 300 seismic

Accepted Manuscript

events which occurred in a period ranging from 21st July 2012 to the 9th December 2016. As shown in Figure 5, the observation area is $[6.729, 18.002] \times [36.64, 46.46]$, including also seismic events occurring around Italy. They tend to gather into two main clusters. The northernmost originated in May 2012, when two major earthquakes struck Northern Italy, causing 27 deaths and widespread damage. The events are known in Italy as the 2012 Emilia earthquakes, because they mainly affected the Emilia region. Then, Central Italy seismic sequence began in August 2016, and it is now defined by the INGV as the Amatrice-Norcia-Visso seismic sequence. The analysed events' magnitudes vary between 0.5 to 4.8.

We first compute the proposed local *K*-function. Figure 6 depicts the estimated local summary statistics. In particular, the steady black lines represent the global statistics, while the grey ones represent the individual contributions. In dashed lines we also represent the theoretical value. In panel (a), the *K*-function is based on a kernel intensity estimate whose bandwidth is selected by Diggle (2013)'s rule, while in panel (b) the bandwidth is chosen as in Cronie and Van Lieshout (2018). We observe some relevant differences: while with Cronie and Van Lieshout (2018)'s rule we depict different local *K*-functions deviating from the global one, following Diggle (2013), we find a unique outlying local *K*-function. This may be explained by the fact that Cronie and Van Lieshout (2018)' s approach tends to yield a bit too large bandwidths when large parts of the study region contain no points, while Diggle (2013)'s approach tends to yield too small bandwidths in general; see Cronie and Van Lieshout (2018) for details. Note that by increasing the bandwidth we decrease the intensity estimate and, as a consequence, the summand denominators in (8) are decreased. Therefore, we run the proposed local test of random labelling with both options for the bandwidth selection and, as expected, the differences observed in the computation of the local *K*-functions are reflected in the results of the test.

Accepted Manuscript

Figure 7 displays the significant points (black triangles) and the non-significant ones (grey points). Panel (a) shows the results with Diggle (2013)'s bandwidth while the ones in panel (b) are obtained with Cronie and Van Lieshout (2018)'s bandwidth. For both choices, we selected a significance level of 0.1. We observe that the significant points tend to be similar in both cases, therefore the choice of bandwidth (selection method) does not seem to be crucial. We note that such bandwidth-induced differences were missing in the previously run simulation study. We attribute this sensitivity of the procedure to the shapes of the functional marks, that are obviously more variable, if compared to the simulated ones.

Nevertheless, both bandwidths lead to significant events belonging to important well known Italian seismic sequences. Of course, these sequences are likely generated by different underlying processes, giving rise to long-term and highly correlated aftershocks. The implication of this result is twofold. On one hand, we have been able to correctly identify seismic events belonging to important well-known Italian seismic sequences. On the other hand, we have found that the shocks related to these sequences exhibit different local dependence structure and therefore, these events are likely generated by different underlying processes, corresponding to different seismic sources.

8 Conclusions

In this work, we have proposed a general form for local summary statistics for marked point processes, which has been exploited to define the family of local inhomogeneous mark-weighted summary statistics for spatial point processes with functional marks, i.e. Functional Marked Point Processes (FMPP). We have employed such local summary statistics to construct a local test for random labelling, that is, to identify points, as well as regions, where this hypothesis does not hold.

More specifically, we first introduce a general local function for marked point patterns. With this specification, we are able to show that this function may be

Accepted Manuscript

exploited to generate most summary statistics established in the literature. With particular reference to the functional marked context, we define the family of local *t*-weighted marked *n*-th order inhomogeneous summary statistics based on the *K*-function, which is a local contribution to a global summary statistic estimator. We obtain a result for the expectation of the general local summary statistic and exploit it to derive an expression for the expectation of our *t*-weighted local statistics.

Having access to these tools, we have proposed a local test of random labelling, resorting to the second order version of our proposed local estimator, obtaining a local test useful for identifying specific regions where a global test would not detect atypical behaviour of the points.

To study the performance of the test in terms of classification rates, we have conducted a simulation study, considering a number of scenarios with different ground processes and structures for the functional marks. Such simulations have shown that in many settings, the local test performs well in identifying points of a pattern where the hypothesis of random labelling is not verified.

We can draw a number of future work paths. Nevertheless, the local functions proposed in this paper can be considered as a very informative synthesis of the local second order behavior, useful for characterising the study area by an extended marked model, based on the FMPP theory. Incorporating local characteristics as functional marks would become part of the so called *Constructed functional marks* (CFMs), which are marks reflecting the geometries of point configurations in neighbourhoods of the individual points.

Concerning the application to seismic data, we aim at including also auxiliary (non-functional) marks into the analysis. These could contain synthetic information about the waveforms, such as the arrival times of the seismic event, or the inter-time between the two. The achievement of the unification of earthquake data and the FMPP theory would result in building a framework

Accepted Manuscript

where it would be possible to exploit the available information of the seismic point process altogether.

A final comment concerns the possible extension of this paper's tools to spatio-temporal ground processes, which of course are of importance for processes which typically exhibit spatio-temporal interactions, such as the seismic one. Undoubtedly, such extensions would be crucial for accounting for the temporal dimension of the seismic events, whose realization depends on their past history, as proved by the existence of aftershocks. This would mean to consider a spatio-temporal marked point process $Y = \{(x_i, m_i)\}_{i=1}^{N}$, with ground points x_i in the 3-dimensional space $\mathbb{R}^2 \times \mathbb{R}^+$ and exploit the methodological framework introduced in this paper. Moreover, local summary statistics in space and time are well established, both theoretically (Siino et al., 2018; Adelfio et al., 2020) and computationally (Gabriel et al., 2021). Although such an extension could be straightforwardly achieved by essentially having our summary statistic functions incorporate an additional argument, t, which controls the temporal lags (cf. Iftimi et al. (2019)), this adds another level of complexity which we believe is out of the scopes of this paper, but it surely represents an interesting path to cover in future.

Supplementary material

Supplementary material contains the source codes to reproduce experimental results.

Disclosure Statement

The authors have no potential conflict of interest to report.

Accepted Manuscript

Funding

This work was supported by “FFR 2023 - Giada Adelfio”, “FFR 2023 - Nicoletta D ’Angelo”, and by the PNRR project “Growing Resilient, INclusive and Sustainable - GRINS” Spoke 06: UNIPD “Low Carbon Policies”.

Open access

The authors would like publish open access, making use of the Bibsam deal which Swedish universities have with Taylor & Francis.

Accepted Manuscript

References

Adelfio, G., Chiodi, M., D'Alessandro, A., and Luzio, D. (2011). Fpca algorithm for waveform clustering. *Journal of Communication and Computer*, 8(6):494–502.

Adelfio, G., Chiodi, M., D'Alessandro, A., Luzio, D., D'Anna, G., and Mangano, G. (2012). Simultaneous seismic wave clustering and registration. *Computers & geosciences*, 44:60–69.

Adelfio, G., Siino, M., Mateu, J., and Rodríguez-Cortés, F. J. (2020). Some properties of local weighted second-order statistics for spatio-temporal point processes. *Stochastic Environmental Research and Risk Assessment*, 34(1):149–168.

Anselin, L. (1995). Local indicators of spatial association-lisa. *Geographical analysis*, 27(2):93–115.

Anselin, L. (1996). Chapter eight the moran scatterplot as an esda tool to assess local instability in spatial association. *Spatial Analytical*, 4:121.

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.

Chiodi, M., Adelfio, G., D'Alessandro, A., and Luzio, D. (2013). Clustering and registration of multidimensional functional data. In *Statistical Models for Data Analysis*, pages 89–97. Springer.

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic Geometry and Its Applications*. John Wiley & Sons, third edition.

Comas, C., Delicado, P., and Mateu, J. (2011). A second order approach to analyse spatial point patterns with functional marks. *Test*, 20(3):503–523.

Accepted Manuscript

Cressie, N. and Collins, L. B. (2001). Analysis of spatial point patterns using bundles of product density lisa functions. *Journal of agricultural, biological, and environmental statistics*, 6(1):118–135.

Cronie, O. and Van Lieshout, M. (2015). A J-function for inhomogeneous spatio-temporal point processes. *Scandinavian Journal of Statistics*, 42(2):562–579.

Cronie, O. and van Lieshout, M. N. M. (2016). Summary statistics for inhomogeneous marked point processes. *Annals of the Institute of Statistical Mathematics*, 68(4):905–928.

Cronie, O. and Van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105(2):455–462.

Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*. Springer-Verlag, New York, second edition.

D'Angelo, N., Adelfio, G., and Mateu, J. (2021). Assessing local differences between the spatio-temporal second-order structure of two point patterns occurring on the same linear network. *Spatial Statistics*, 45:100534.

D'Angelo, N., Adelfio, G., and Mateu, J. (2023). Locally weighted minimum contrast estimation for spatio-temporal log-gaussian cox processes. *Computational Statistics & Data Analysis*, 180:107679.

D'Angelo, N., Siino, M., D'Alessandro, A., and Adelfio, G. (2022). Local spatial log-gaussian cox processes for seismic data. *Advances in Statistical Analysis. https://doi.org/10.1007/s10182-022-00444-w*.

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.

Accepted Manuscript

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press.

Gabriel, E. and Diggle, P. J. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, 63(1):43–51.

Gabriel, E., Diggle, P. J., Rowlingson, B., and Rodriguez-Cortes, F. J. (2021). *stpp: Space-Time Point Pattern Simulation, Visualisation and Analysis*. R package version 2.0-5.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Getis, A. and Franklin, J. (1987). Second-order neighborhood analysis of mapped point patterns. *Ecology*, 68(3):473–477.

Getis, A. and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189–206.

Ghorbani, M., Cronie, O., Mateu, J., and Yu, J. (2021). Functional marked point processes: a natural structure to unify spatio-temporal frameworks and to analyse dependent functional data. *Test*, 30(3):529–568.

Iftimi, A., Cronie, O., and Montes, F. (2019). Second-order analysis of marked inhomogeneous spatiotemporal point processes: Applications to earthquake data. *Scandinavian Journal of Statistics*, 46(3):661–685.

Illian, J., Benson, E., Crawford, J., and Staines, H. (2006). Principal component analysis for spatial point processes–assessing the appropriateness of the approach in an ecological context. In *Case studies in spatial point process modeling*, pages 135–150. Springer.

Accepted Manuscript

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, volume 70. John Wiley & Sons.

Mateu, J., Lorenzo, G., and Porcu, E. (2007). Detecting features in spatial point processes with clutter via local indicators of spatial association. *Journal of Computational and Graphical Statistics*, 16(4):968–990.

Mateu, J., Lorenzo, G., and Porcu, E. (2010). Features detection in spatial point processes via multivariate techniques. *Environmetrics*, 21(3-4):400–414.

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V. (2021). Instance–the italian seismic dataset for machine learning. *Earth System Science Data*, 13(12):5509–5544.

Møller, J. (2003). Shot noise cox processes. *Advances in Applied Probability*, pages 614–640.

Møller, J. and Ghorbani, M. (2012). Aspects of second-order analysis of structured inhomogeneous spatio-temporal point processes. *Statistica Neerlandica*, 66(4):472–491.

Møller, J. and Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton.

Moraga, P. and Montes, F. (2011). Detection of spatial disease clusters with lisa functions. *Statistics in Medicine*, 30(10):1057–1071.

Mrkvička, T., Dvořák, J., González, J. A., and Mateu, J. (2021). Revisiting the random shift approach for testing in spatial statistics. *Spatial Statistics*, 42:100430.

Accepted Manuscript

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):381–404.

Penttinen, A. and Stoyan, D. (1989). Statistical analysis for a class of line segment processes. *Scandinavian Journal of Statistics*, pages 153–168.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. Springer.

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

Schlather, M. (2001). On the second-order characteristics of marked point processes. *Bernoulli*, pages 99–117.

Siino, M., Adelfio, G., Mateu, J., Chiodi, M., and D'alessandro, A. (2017). Spatial pattern analysis using hybrid models: an application to the hellenic seismicity. *Stochastic Environmental Research and Risk Assessment*, 31(7):1633–1648.

Siino, M., Rodríguez-Cortés, F. J., Mateu, J., and Adelfio, G. (2018). Testing for local structure in spatiotemporal point pattern data. *Environmetrics*, 29(5-6):e2463.

Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields: methods of geometrical statistics*. Wiley, Chichester.

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

Accepted Manuscript

Van Lieshout, M. (2006). A j-function for marked point patterns. *Annals of the Institute of Statistical Mathematics*, 58(2):235–259.

Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Accepted Manuscript

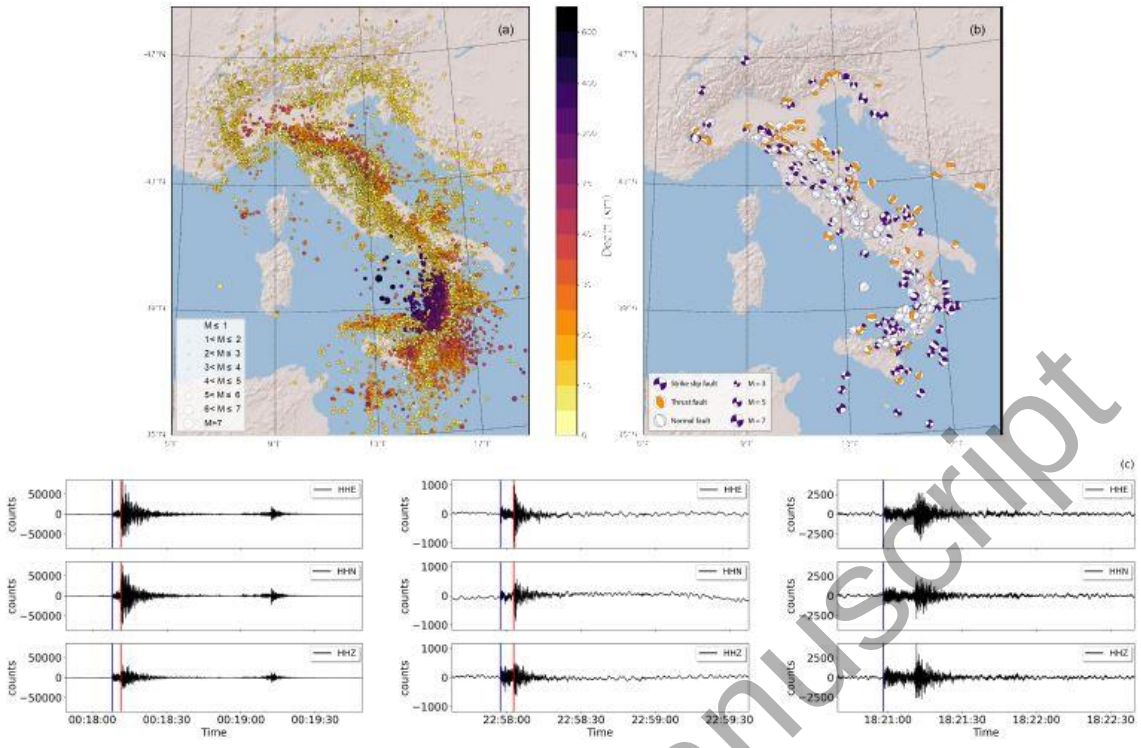


Fig. 1 The Italian seismic dataset for machine learning (INSTANCE). (a) Earthquake locations; (b) Seismic stations used for waveforms extraction. The symbol sizes are proportional to earthquake magnitude and number of arrival phases recorded by stations, respectively; (c) Seismic waveforms of some events with magnitude in the range [2, 4]. Vertical lines indicate the seismic waves' arrival times. *Source: Michelini et al. (2021)*.

Accepted Manuscript

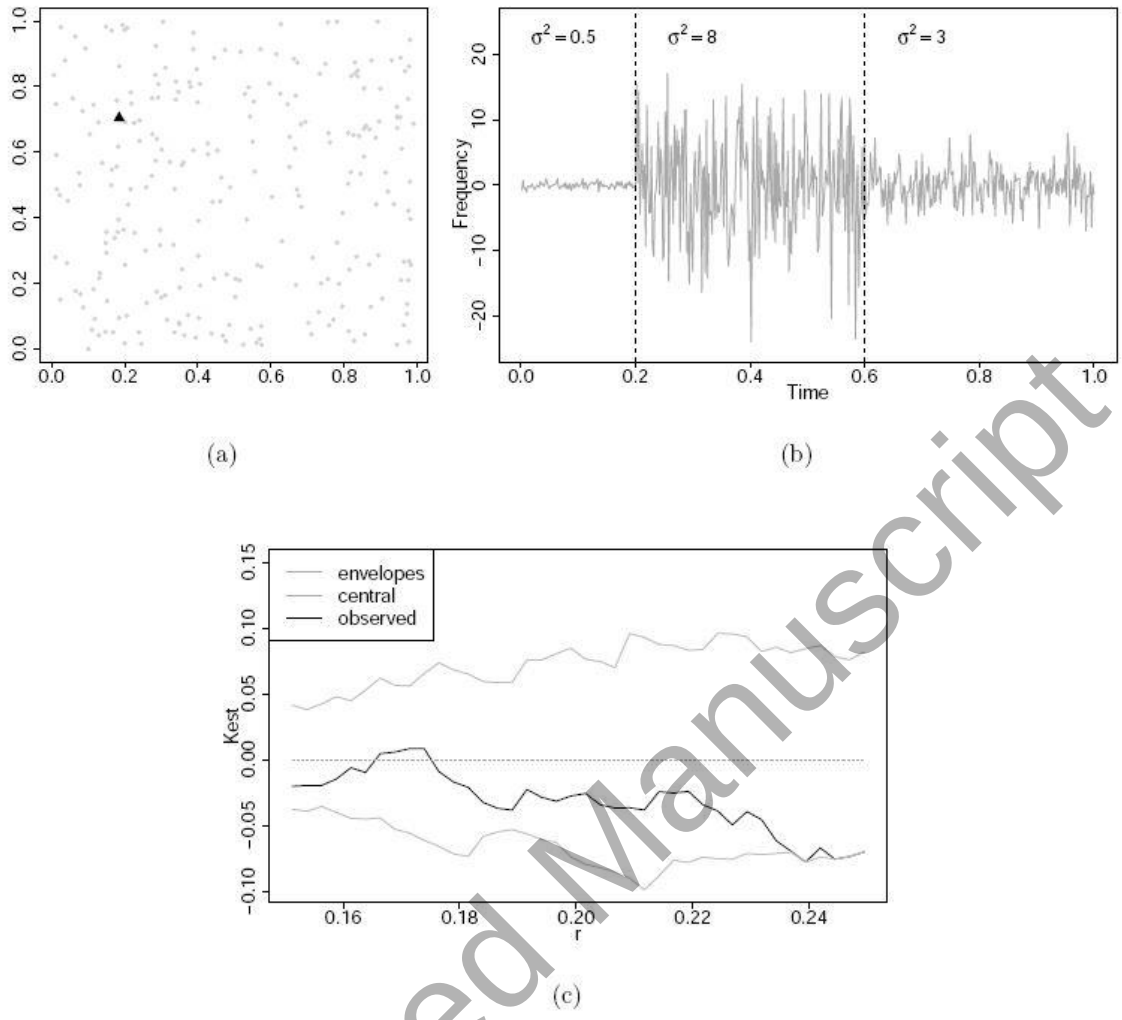


Fig. 2 (a) Simulated earthquake locations. (b) Simulated waveform marking the highlighted point on panel (a). (c) Result of the global test.

Accepted Manuscript

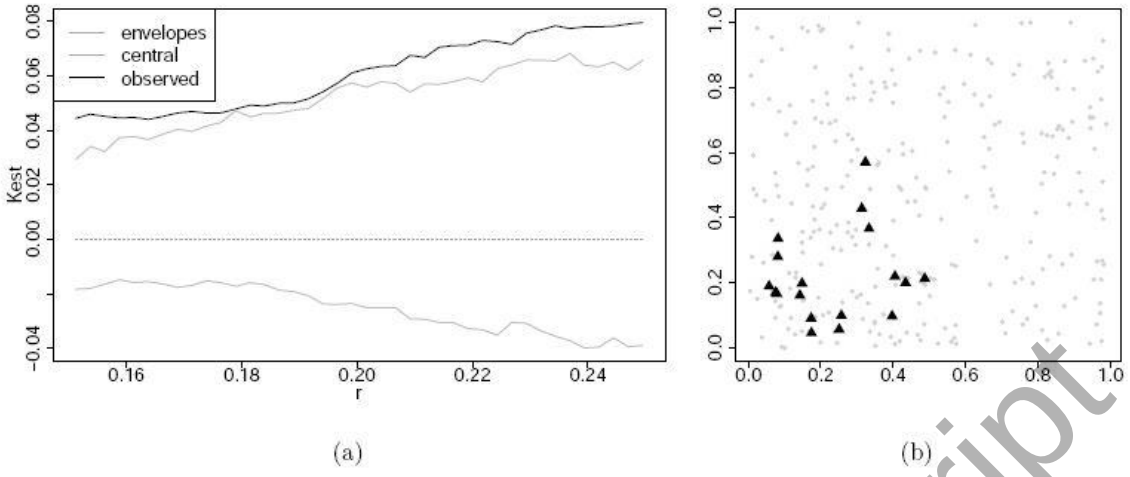


Fig. 3 (a) Result of global test for the spatially dependent simulated data. (b) Output of the local test: the black triangles are the significant points for which the hypothesis of random labelling is rejected.

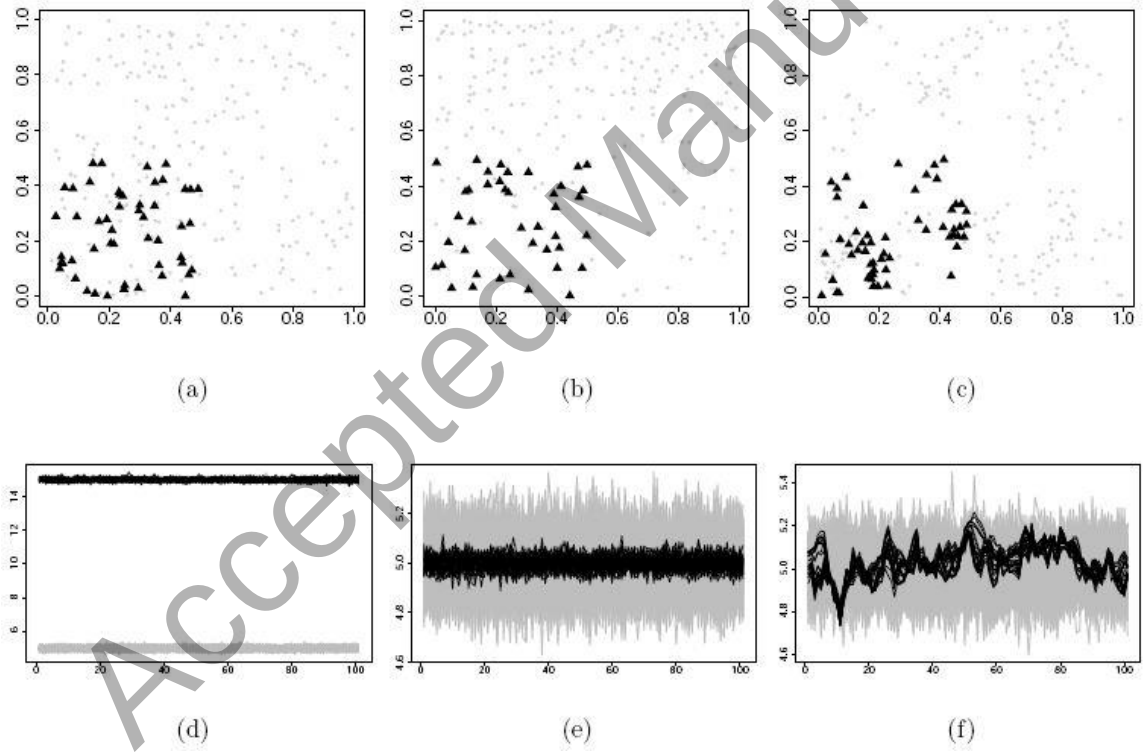


Fig. 4 Simulation scenarios. (a) - (c) Spatial ground patterns; (d) - (f) Functional marks of model (11) (in grey) and of the marking models in item 1, 2, and 3, respectively (in black).

Accepted Manuscript

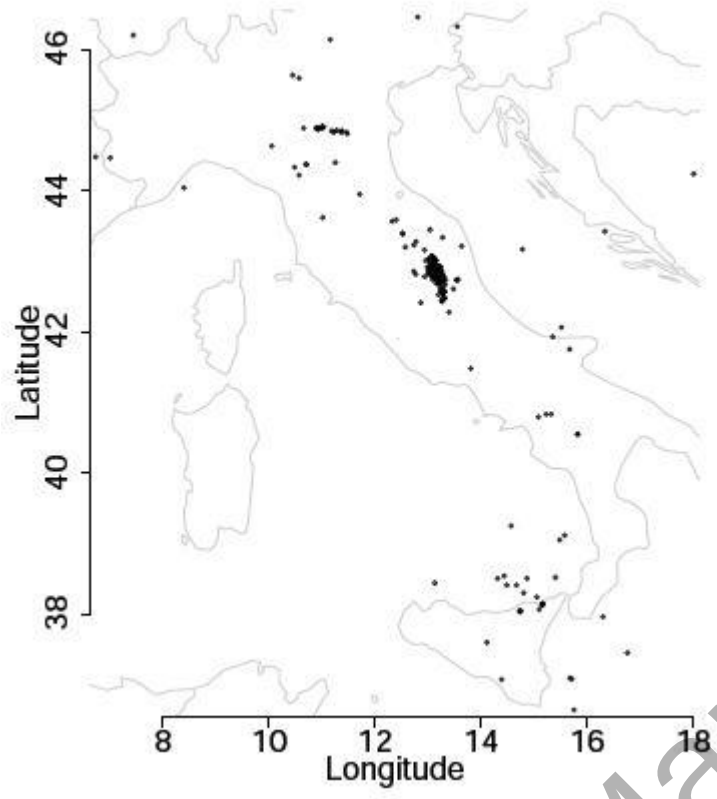


Fig. 5 Earthquake locations

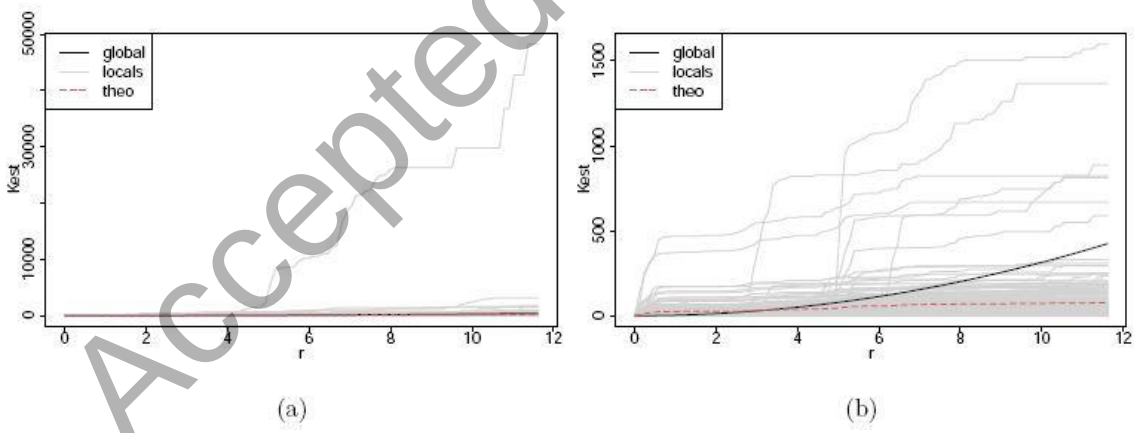


Fig. 6 Local *K*-functions. (a) The *K*-function is based on a kernel intensity estimate whose bandwidth is selected by Diggle (2013)'s rule. (b) The bandwidth is chosen as in Cronie and Van Lieshout (2018).

Accepted Manuscript

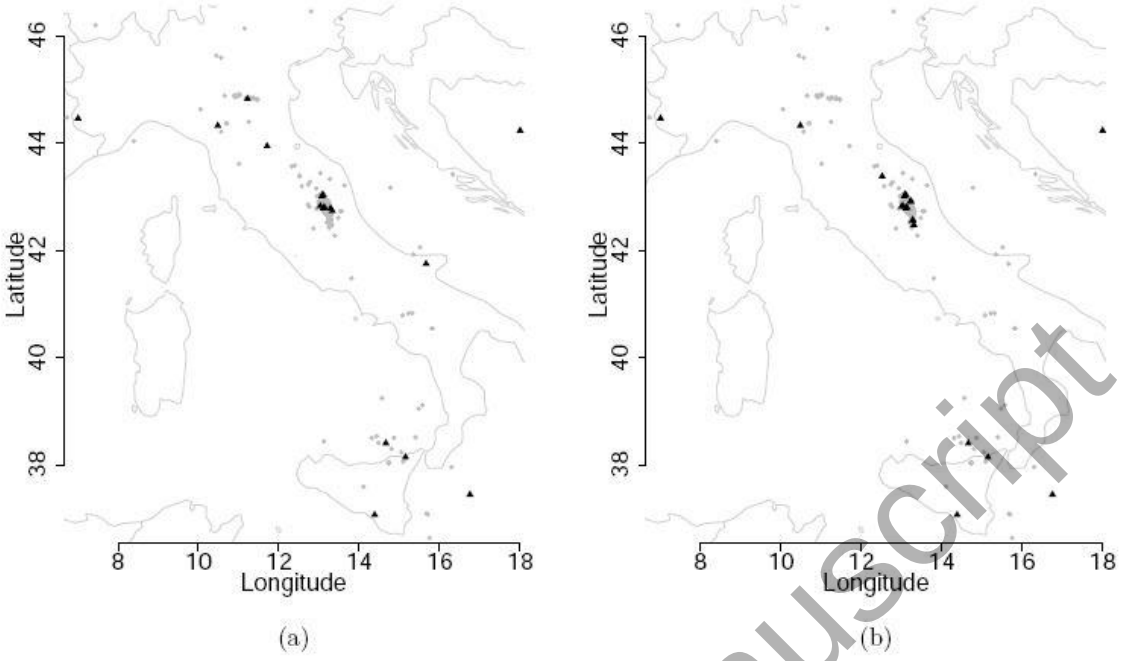


Fig. 7 Results of the local test at $\alpha = 0.1$. Non-significant events are displayed as grey points and significant events are the black triangles. (a) The *K*-function is based on a kernel intensity estimate whose bandwidth is selected by Diggle (2013)'s rule. (b) The bandwidth is chosen as in Cronie and Van Lieshout (2018).

Accepted Manuscript

Table 1 Results of the local test averaged over 100 simulated point patterns with an expected point count of 250 each.

Ground process	Marking model	TPR	FPR	ACC
Homogeneous Poisson	(1)	0.112	0.346	0.583
Homogeneous Poisson	(2)	0.583	0.066	0.820
Homogeneous Poisson	(3)	0.870	0.024	0.896
Inhomogeneous Poisson	(1)	0.032	0.585	0.449
Inhomogeneous Poisson	(2)	0.648	0.084	0.856
Inhomogeneous Poisson	(3)	0.895	0.023	0.932
Thomas	(1)	0.109	0.394	0.571
Thomas	(2)	0.637	0.088	0.846
Thomas	(3)	0.865	0.025	0.925

Accepted Manuscript