DANILO LOURENÇO LOPES

ANÁLISE ESTATÍSTICA DE PROCESSOS PONTUAIS BIVARIADOS LIGADOS

Belo Horizonte
16 de junho de 2008

DANILO LOURENÇO LOPES

Orientador: Renato Martins Assunção

ANÁLISE ESTATÍSTICA DE PROCESSOS PONTUAIS BIVARIADOS LIGADOS

Proposta de dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Estatística.

Belo Horizonte
16 de junho de 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Análise Estatística de Processos Pontuais Bivariados Ligados

DANILO LOURENÇO LOPES

Proposta de dissertação defendida e aprovada pela banca examinadora constituída por:

Ph. D. Renato Martins Assunção – Orientador
Universidade Federal de Minas Gerais

Ph. D. Marcelo Azevedo Costa
Universidade Federal de Minas Gerais

Ph. D. Marcos Antônio da Cunha Santos
Universidade Federal de Minas Gerais

Ph. D. Gauss Moutinho Cordeiro
Universidade Federal de Pernambuco

Belo Horizonte, 16 de junho de 2008

Para meus pais e todos os grandes amigos, que, ao longo desta jornada acadêmica, sempre dedicaram apoio e confiança incondicionais em meu sucesso.

Agradecimentos

Este projeto de pesquisa foi financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, processo 301173/97-4. Os autores agradecem à Polícia Militar de Minas Gerais por fornecer o conjunto de dados analisados neste trabalho. Nossos agradecimentos a todos aqueles que, direta ou indiretamente, nos auxiliaram no progresso deste trabalho.

Resumo

Este trabalho considera uma nova classe especial de processos pontuais espaciais marcados, os processos pontuais bivariados ligados. Para cada evento de um padrão pontual N_1 existem um ou mais eventos correspondentes em outro padrão pontual N_2, observado na mesma região geográfica. Pares de eventos de origem-destino são os exemplos mais comuns para esse tipo de dados. Um teste de correlação espacial entre os dois padrões pontuais é desenvolvido e, a partir de uma proposta simplificadora, alguns resultados úteis foram obtidos. Sugerimos um modelo de Gibbs para estes processos pontuais bivariados ligados, derivamos algumas estatísticas descritivas das propriedades do modelo e, em seguida, apresentamos métodos para simular, testar e estimar tais processos.

Desenvolvemos uma ferramenta de análise exploratória gráfica - MaPPEA - para a visualização dos gráficos ligados dinâmicos apresentados. Esta ferramenta foi desenvolvida em C++ com uma interface gráfica para auxiliar o usuário na análise estatística dos dados. A partir do "brushing" de janelas linkadas, MaPPEA fornece uma ilustração da estrutura e relações entre marcas e coordenadas de padrões pontuais. As principais funções incluídas neste software são a função de distribuição das marcas e a superfície de intensidade condicional das posições de destino correspondente a uma posição de origem selecionada, ambas mudando dinamicamente. Os métodos são ilustrados com dados de furto de veículos e a eventuais recuperações dos mesmos, bem como com dados de posições de árvores e suas marcas correspondentes.

Abstract

We consider a subclass of spatial marked point processes, the bivariate linked point processes. That is, each spatial event of a point pattern N_1 has one or more corresponding events in another spatial point pattern N_2, observed in the same geographical region. Pairs of origin-destination events are the main example of this kind of data. Based on this model, a test for spatial correlation between the point processes is proposed. Due to the a simplifying proposal on modeling, a number of useful results can be established. We first propose a Gibbs model to the bivariate linked point processes and derive summary statistics for the properties of the model, and, second, we focus on how to simulate/test/estimate such processes.

We give an introduction to our exploratory graphical analysis tool - MaPPEA - for the visualization of the dynamic linked graphs presented. This tool was developed in C++ with a graphical interface to help the user with statistical data analysis. Using linked windows brushing, MaPPEA provides an illustration of the structure and relationships between marks and locations of point patterns. The main features included in this software are the dynamically-changing cumulative distribution function of the univariate marks and the dynamically-changing conditional intensity surface of the destination locations corresponding to a selected origin location. The methods are illustrated with data on car theft location and the eventual car retrieval location and on trees' locations and their associated marks.

Resumo Estendido

O presente trabalho está organizado em dois capítulos, apresentados a seguir:

O Capítulo 1 constitui-se de um trabalho apresentado no VIII Brazilian Symposium on GeoInformatics, GEOINFO 2006, e publicado como capítulo do livro "Advances in Geoinformatics" pela Springer em 2007 (Assunção and Lopes, 2007). Ele trata de um tipo de dados pouco analisado em estudos espaciais: processos pontuais bivariados ligados; esses processos são compostos por dois padrões pontuais para os quais cada evento pontual em um deles (denominado padrão pontual de origem) possui um ou mais eventos pontuais correspondentes no outro padrão pontual (denominado padrão pontual de destino). Dentre alguns exemplos desse processo incluem-se posições de furto e recuperação de veículos, migração humana entre diferentes regiões e dados de origem-destino de estudos de transporte. Um modelo estocástico para padrões pontuais bivariados linkados é apresentado e, a partir de uma proposta simplificadora, deriva-se um teste de associação entre os dois padrões pontuais baseado em uma estatística de escore. Essa estatística possui distribuição assintótica normal, e os seus dois primeiros momentos são apresentados. Apresentamos também uma aplicação do teste da estatística de escore para dados de furto e localização de veículos em Belo Horizonte, uma das maiores cidades brasileiras.

O Capítulo 2 apresenta uma continuação do trabalho desenvolvido no capítulo anterior. Trata-se de um artigo independente em fase de submissão ao Journal of Graphical and Computational Statistics. Uma vez que um teste de associação entre os padrões pontuais de origem e destino foi desenvolvido, o foco agora torna-se descrever essa associação quando evidenciada pelo teste. Para tanto um estimador de densidade condicional baseado em estimadores de Kernel foi desenvolvido. Esse estimador foi implementado em uma ferramenta gráfica - MaPPEA - que utiliza "brushing" de janelas "linkadas" para a visualização da intensidade condicional correspondente a uma posição selecionada pelo usuário. Esta ferramenta foi desenvolvida em C++ como um programa "stand-alone" e possui uma interface gráfica para auxiliar o usuário na análise estatística dos dados. Apesar desta motivação inicial, o Capítulo 2 possui uma abordagem muito mais generalizada: foram desenvolvidos e implementados métodos para visualização por meio de gráficos "linkados" dinâmicos da relação entre marcas e coordenadas pontuais para diversos processos pontuais espaciais marcados. Para o caso de marcas univariadas discretas, o software MaPPEA possui uma ferramenta que desenha o "barplot" correspondente aos eventos selecionados na janela da região geográfica. Para o caso de marcas univariadas contínuas, o software apresenta tanto opções de visualizar uma esti-

mativa de intensidade condicional das marcas correspondentes a uma determinada posição na região de estudo quanto de obter a função de distribuição acumulada empírica para os eventos selecionados.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Testing association between origin-destination spatial locations

(published in “Advances in Geoinformatics”, Springer Berlin Heidelberg, 2007)

1.1 Introduction

Point processes are probabilistic frameworks to analyze spatial patterns composed by random point features, called events, stored in GIS database. In a multivariate point process, the events are of two or more different types such as the locations of disease cases and a set of locations labeled as control individuals or as the positions of plants in a planar region labeled according to their species (Daley and Vere-Jones, 2003). Usually, the spatial analysis of multivariate point processes is concerned with two questions. The first one concentrates on the comparison between the individual patterns of the component processes. Typically, the interest is to decide if one spatial pattern (such as the disease cases) has some degree of spatial clustering with respect to another spatial pattern (such as the controls' pattern) (see Kelsall and Diggle, 1995), perhaps identifying some putative sources of increased relative intensity (Diggle, 1990; Diggle and Rowlingson, 1994). The second question concentrates on testing the independence of two (or more) point patterns and therefore attention is directed to the *joint* distribution of the processes (Lotwick and Silverman, 1982). It is common, for example, to test if the presence of an event of a certain type in a location either inhibits or stimulates the nearby presence of events of other types.

In this paper, we are interested in another type of situation that is analyzed less often in spatial studies: bivariate point processes structured as origin-destination locations. To illustrate, consider the locations of $m = 6339$ car thefts occurred in he years 2000 and 2001 in Belo Horizonte, a 2 million inhabitants Brazilian city (see left hand side map in Figure 2.1). Some of the stolen cars are never retrieved by the police or they are retrieved outside the city boundary. The right hand side map of Figure 2.1 shows the locations of the $n = 5257$ cars eventually found within the city limits. Conditioning on the car retrieval, the interest is to know if there is some type of spatial dependence between the two locations of each stolen car.

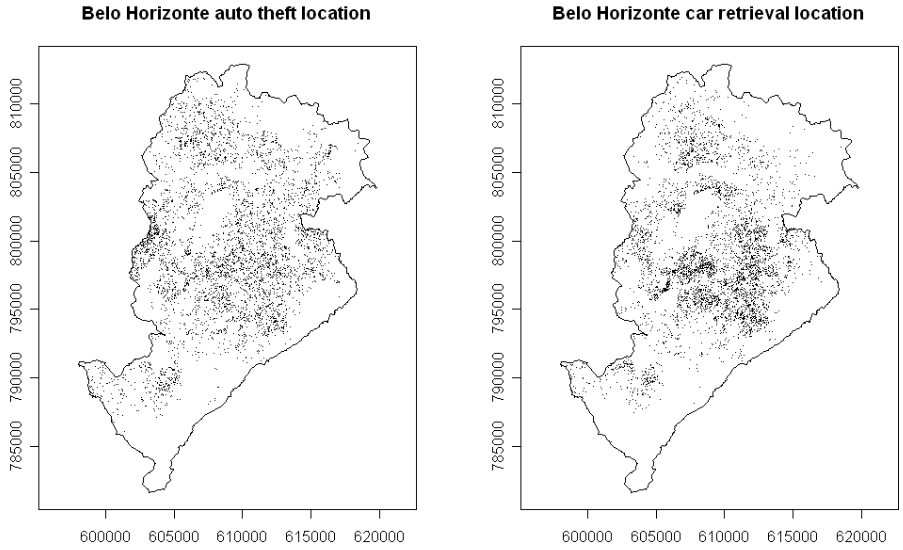

Belo Horizonte auto theft location
Belo Horizonte car retrieval location
785000
790000
795000
800000
805000
810000
600000
605000
610000
615000
620000

Figura 1.1: Maps of car theft locations and their subsequent retrieval locations in Belo Horizonte from Aug 2000 to July 2001

Other possible applied settings that generate bivariate linked point processes are: migration between different regions, murder location and the victim's residence address, origin-destination survey data from transportation studies.

These situations have in common that two (or more) spatial point patterns have linked events. That is, for each event in one point pattern, there is one (or more) corresponding events in the other point pattern. In this paper, we concentrate on bivariate patterns, one of them called origin process, and the other called destination process. We denote this kind of data as origin-destination point patterns or bivariate linked point processes.

We introduce a new correlation measure for origin-destination point patterns and we point its similarities with Knox's statistic, a common tool in spatial-temporal studies. Our measure of correlation tests if pairs of events that are close in origin tend to be close in destination too. Our proposal has some attractive features: it is simple to understand, it is easy to be calculated, and it has an asymptotic distribution that does not depend on the spatial-temporal pattern of the data. In particular, it can be applied to any type of spatial point pattern. Our measure derives its theoretical properties from the fact that it is based on a score score test statistic, the locally most powerful test within a certain probabilistic model.

We proceed as follows. Section 1.2 proposes a model to bivariate linked spatial point data. Section 1.3 develops the locally most powerful test for independence between origin-destination hypothesis. Section 1.4 describes an application to car theft-retrieval dataset in Belo Horizonte and Section 1.5 presents our conclusions.

1.2 A stochastic model for bivariate linked point processes

The main origin-destination point pattern motivation for this paper is composed by the locations of car thefts and the locations of their eventual retrieval in Belo Horizonte within a year. In Figure 1.2, the map in the right hand side shows all the linked locations of the 5257 vehicles stolen between August 2000 and August 2001. In the left hand side, only a random sample of 50 stolen vehicles in Belo Horizonte is shown with each arrow going from the theft position to its corresponding retrieval position.

Suppose that a car is stolen at position $\mathbf{x}$, corresponding to the origin of the event. Consider the probability distribution of a stolen car retrieval location $\mathbf{y}$, given that it has been stolen at location $\mathbf{x}$. Let $f(\mathbf{y}|\mathbf{x})$ be the density of the destination $\mathbf{y}$ given that the origin is $\mathbf{x}$. For each possible origin $\mathbf{x}$ we have a surface $f(\mathbf{y}|\mathbf{x})$ showing the most likely destinations of events originating at $\mathbf{x}$.

Before embarking on estimation of a stochastic model for $f(\mathbf{y}|\mathbf{x})$, it is worth to verify if the hypothesis of independence between origin-destination point patterns holds. This is a minimum requirement for such modeling. If there is no evidence for dependence, there is no point on estimating $f(\mathbf{y}|\mathbf{x})$. Hence, a first step in the analysis is to test if $f(\mathbf{y}|\mathbf{x})$ is the same, whatever the location $\mathbf{x}$. If this is so, there is origin and destination locations are independent. Otherwise, we say that there is dependence or association between origin and destination locations.

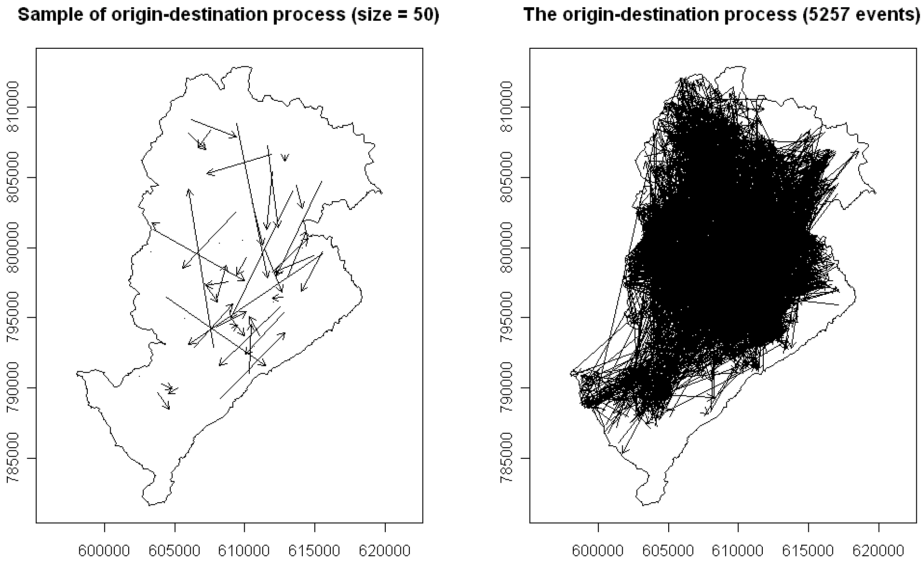

Sample of origin-destination process (size = 50)
The origin-destination process (5257 events)
785000
790000
795000
800000
805000
810000
600000
605000
610000
615000
620000

Figura 1.2: Sample (left) and entire dataset (right) of thefts and retrievals in Belo Horizonte. Each arrow goes from a theft location to the correspondent retrieval location

Let (N_1, N_2) be a bivariate linked point process observed in a finite polygon $A \subset R^2$. The available data are a set of n events $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \ldots, n$, composed by pairs of spatial locations in A. The first location $\mathbf{x}_i = (x_{1i}, x_{2i})$ is called the origin event and the second location $\mathbf{y}_i =$

(y_{1i}, y_{2i}) is the destination event. We denote by φ the unordered set $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ of pairs of locations.

Conditioning on the total number n of observed events, we can work with the probability density distribution $p(\varphi)$ of the n events in A^{2n}, which must be invariant with respect to the ordering of the locations (Daley and Vere-Jones, 2003, page 123-125). Under the assumption of independence between origin and destination locations, this probability density can be written as

$$p(\varphi) = C \exp\left(g(\mathbf{x}_1, \ldots, \mathbf{x}_n) + h(\mathbf{y}_1, \ldots, \mathbf{y}_n)\right) \tag{1.1}$$

where C is a normalizing constant. The function g models the interaction among the origin events as well as any spatial variation in the first order intensity of this marginal processes. The function h has the same role with respect to the destination events. The functions g and h can be chosen arbitrarily as long as the normalizing constant in (1.1) is finite and as long as each one of them is invariant with respect to all permutations of the events labels.

To introduce some kind of interaction between origin and destination, we propose a model similar to Gibbs processes with pairwise interaction functions. Namely, we assume that

$$p(\varphi) = C \exp\left(g(\mathbf{x}_1, \ldots, \mathbf{x}_n) + h(\mathbf{y}_1, \ldots, \mathbf{y}_n) - \sum_{i<j} \phi\left((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)\,; \boldsymbol{\theta}\right)\right) \tag{1.2}$$

where ϕ is a potential function depending on a set of parameters $\boldsymbol{\theta}$. One simple special case is to define two positive thresholds r_x and r_y and then (1.2) becomes

$$\begin{aligned} p(\varphi) &= C \exp\left(g(\mathbf{x}_1, \ldots, \mathbf{x}_n) + h(\mathbf{y}_1, \ldots, \mathbf{y}_n) - \theta \sum_{i<j} I[|\mathbf{x}_i - \mathbf{x}_j| < r_x]\, I[|\mathbf{y}_i - \mathbf{y}_j| < r_y]\right) \\ &= C \exp\left(g(\mathbf{x}_1, \ldots, \mathbf{x}_n) + h(\mathbf{y}_1, \ldots, \mathbf{y}_n) - \theta T(\varphi)\right) \end{aligned} \tag{1.3}$$

where $I[\cdot]$ is the indicator function and $T(\varphi)$ is the number of pair of events that are within the threshold limits at the origin and at the destination.

The case $\theta = 0$ corresponds to the independence between origin and destination processes. If $\theta < 0$, the result is an inhibition process: pairs of nearby events at the origin will tend to be farther apart in the destination. Values of $\theta > 0$ correspond to the spatial clustering of the pair of origin-destination events.

In principle, the functions g and h in (1.2) can be quite general. Isolated analysis of each one of the patterns, origin events and destination events, can provide parametric or non-parametric estimates of these functions under the null hypothesis. However, when the main interest is to test for the presence of spatial correlation between the two types of events, it is useful to condition on the observed marginal locations. This eliminates the nuisance functions g and h.

Let $\pi_1, \ldots, \pi_n$ be a permutation of the indexes $1, \ldots, n$. The probability density distribution for the unlinked and unordered locations is denoted by $p(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \{\mathbf{y}_1, \ldots, \mathbf{y}_n\})$ and given by

$$C \exp\left(g(\mathbf{x}_1,\ldots,\mathbf{x}_n)+h(\mathbf{y}_1,\ldots,\mathbf{y}_n)\right)\sum_{\pi}\exp\left(-\sum_{i<j}\phi\left((\mathbf{x}_i,\mathbf{y}_{\pi_i}),(\mathbf{x}_j,\mathbf{y}_{\pi_j})\,;\boldsymbol{\theta}\right)\right) \tag{1.4}$$

where the sum is taken over all $n!$ possible permutations.

Therefore, the distribution of φ conditional on the unordered set of origin locations $\mathbf{x}_1,\ldots,\mathbf{x}_n$ and on the unordered set of destinations $\mathbf{y}_1,\ldots,\mathbf{y}_n$ is given by

$$p(\varphi|\{\mathbf{x}_1,\mathbf{x}_n\},\{\mathbf{y}_1,\mathbf{y}_n\})=\frac{\exp\left(-\sum_{i<j}\phi\left((\mathbf{x}_i,\mathbf{y}_i),(\mathbf{x}_j,\mathbf{y}_j)\,;\boldsymbol{\theta}\right)\right)}{\sum_{\pi}\exp\left(-\sum_{i<j}\phi\left((\mathbf{x}_i,\mathbf{y}_{\pi_i}),(\mathbf{x}_j,\mathbf{y}_{\pi_j})\,;\boldsymbol{\theta}\right)\right)} \tag{1.5}$$

The hard to calculate normalizing constant C is substituted by a simpler normalizing constant although its sum over $n!$ terms is computationally demanding unless n is unrealistically small.

To simplify notation, we let $\pi\varphi$ to denote the unordered set of events with the destinations shuffled by some arbitrary permutation $(\pi_1,\ldots,\pi_n)$:

$$\pi\varphi=\{(\mathbf{x}_1,\mathbf{y}_{\pi_1}),\ldots,(\mathbf{x}_n,\mathbf{y}_{\pi_n})\}$$

1.3 Testing for spatial correlation

In the remaining of the paper, we focus our attention in the model (1.3). For the conditional distribution (1.5), the log-likelihood of the interaction parameter θ is given by

$$l(\theta)=-\theta T(\varphi)-\log\left(\sum_{\pi}e^{\theta T(\pi\varphi)}\right)$$

It is clear that $T(\varphi)$ is a natural sufficient statistic for the parameter θ.

The score statistic is then given by

$$\begin{aligned}\frac{\partial l}{\partial\theta} &= -T(\varphi)+\frac{\sum_{\pi}T(\pi\varphi)e^{\theta T(\pi\varphi)}}{\sum_{\pi}e^{\theta T(\pi\varphi)}}\\ &= E_{\pi,\theta}\left[T(\pi\varphi)\right]-T(\varphi)\,,\end{aligned} \tag{1.6}$$

a contrast between observed and expected values of the sufficient statistic, where the expectation of $T(\pi\varphi)$ is taken with probabilities given by all $n!$ normalized values

$$e^{\theta T(\pi\varphi)}\Big/\sum_{\pi}e^{\theta T(\pi\varphi)}\,.$$

The score test statistic is given by

$$\left.\frac{\partial l}{\partial\theta}\right|_{\theta=0}=\frac{1}{n!}\sum_{\pi}T(\pi\varphi)-T(\varphi)\,. \tag{1.7}$$

The moments of $T(\varphi)$ under the null hypothesis are easily obtained. With the origin and the destinations fixed and taking the expectation under all permutations of the destination indexes, we have the following for the expected value:

$$\begin{aligned} E_\pi\left[T(\pi\varphi)\right] &= E_\pi\left(\sum_{i<j} I[|\mathbf{x}_i - \mathbf{x}_j| < r_x]\, I[|\mathbf{y}_{\pi_i} - \mathbf{y}_{\pi_j}| < r_y]\right) \\ &= \sum_{i<j} I[|\mathbf{x}_i - \mathbf{x}_j| < r_x]\, E_\pi\left(I[|\mathbf{y}_{\pi_i} - \mathbf{y}_{\pi_j}| < r_y]\right) \end{aligned}$$

because of the independence between origin and destination and the conditioning on the marginal patterns. There are $n(n-1)/2$ pairs of events and n_y of them are close to each other at the destination. Hence, the probability that the pair (π_i, π_j) of a random permutation is one of them is $2n_y/n(n-1)$ and therefore

$$E_\pi\left[T(\pi\varphi)\right] = \frac{n_y}{\begin{pmatrix} n \\ 2 \end{pmatrix}} \sum_{i<j} I[|\mathbf{x}_i - \mathbf{x}_j| < r_x] = \frac{n_y n_x}{\begin{pmatrix} n \\ 2 \end{pmatrix}}$$

where n_x is the number of pairs of origin events that are close to each other.

The second moment is given by

$$\begin{aligned} E_\pi\{[T(\pi\varphi)]^2\} &= E_\pi\left(\sum_{i<j}\sum_{k<l} I[|x_i - x_j| < r_x] I[|x_k - x_l| < r_x] I[|y_{\pi_i} - y_{\pi_j}| < r_y] I[|y_{\pi_k} - y_{\pi_l}| < r_y]\right) \\ &= \sum_{i<j}\sum_{k<l} I[|x_i - x_j| < r_x] I[|x_k - x_l| < r_x] E_\pi\left(I[|y_{\pi_i} - y_{\pi_j}| < r_y] I[|y_{\pi_k} - y_{\pi_l}| < r_y]\right) \end{aligned}$$

also conditioning on the marginal patterns. Following, we present the development of such expectation for each of the cases can arise. First, for $i = k, j = l$ we have:

$$\sum_{i<j} (I[|x_i - x_j| < r_x])^2 E_\pi\left\{(I[|y_{\pi_i} - y_{\pi_j}| < r_y])^2\right\} \quad = \quad \frac{n_y n_x}{\begin{pmatrix} n \\ 2 \end{pmatrix}}$$

as demonstrated previously. For the case where only one of (i, j) is equal to one of (k, l): there are

$$\frac{n}{2}\begin{pmatrix} n-1 \\ 2 \end{pmatrix}$$

pairs of pair cases with one case in common. n_{2y} of them are composed by pairs of 'case pairs that are close to each other in destination' that have one case in common. Hence the probability that the the quadruplet $(\pi_i, \pi_j, \pi_i, \pi_l)$ is one of them is $4n_{2y}/n(n-1)(n-2)$ and therefore:

$$\frac{4n_{2y}}{n(n-1)(n-2)}\sum_{i<j}\sum_{i<l} I[|x_i - x_j| < r_x]I[|x_k - x_l| < r_x] \quad = \quad \frac{4n_{2y}n_{2x}}{n(n-1)(n-2)}$$

where n_{2x} is the number of pairs of origin events pairs that are close to each other and have one event in common.

Finally, for the case where i, j, k, l are all different among them, there are

$$\binom{n}{2}\binom{n-2}{2}$$

possible quadruplets. The number of them that are close to destination in both pairs is the number of crossing products of the n_y pairs, minus the number of repeated pairs, minus those ones that have one case in common, which gives to us the quantity of $n_y^2 - n_y - n_{2y}$. So, the expectation for this case is given by:

$$\begin{aligned} &= \frac{4[n_y(n_y-1) - n_{2y}]}{n(n-1)(n-2)(n-3)}\sum_{i<j}\sum_{\substack{k<l \\ k\neq i,j \\ k\neq i,j}} I[|x_i - x_j| < r_x]I[|x_k - x_l| < r_x] \\ &= \frac{4[n_y(n_y-1) - n_{2y}][n_x(n_x-1) - n_{2x}]}{n(n-1)(n-2)(n-3)} \end{aligned}$$

Uniting all possible cases, we have:

$$E_\pi\{[T(\pi\varphi)]^2\} \quad = \quad \frac{n_y n_x}{\binom{n}{2}} + \frac{4n_{2y}n_{2x}}{n(n-1)(n-2)} + \frac{4[n_y(n_y-1) - n_{2y}][n_x(n_x-1) - n_{2x}]}{n(n-1)(n-2)(n-3)}$$

Hence, a asymptotic test based on the normal approximation to the score test statistic can be carried out. Alternatively, we can use a Monte Carlo test which do not require asymptotic arguments and which provides exact p-values (Besag and Clifford, 1989). By sampling a large number $B-1$ of permutations independently and with equal probability, the p-value is given by $(1+k)/B$ where k is the number of sampled values $|T(\pi\varphi)|$ greater or equal than the observed value of $|T(\varphi)|$.

$T(\pi\varphi)$, the score test statistic for origin-destination independence test, is similar to another well-known test statistic, the Knox test statistic (see Knox, 1964), proposed for testing space-time interactions. Specifying a spatial and a temporal critical distance, it is possible to indicate when a pair of events is close in space or close in space and in time. Knox test statistic X is defined as the number of pairs of events that are simultaneously close in space and in time. A large number X would be an indication that cases that are close in space tend also to be close in time leading to space-time interaction. In fact, the score test statistic $T(\varphi)$, that was derived from a simplifying proposal on the Gibbs process with pairwise interaction functions, may be also seen as an extension of Knox test statistic, used for spatial-temporal data, to origin-destination data.

1.4 Application

For illustration, we use the auto theft and car retrieval locations data from a large Brazilian city, Belo Horizonte, during the period from August 2000 to August 2001 collected by the Military Police of Minas Gerais based on their police records of crime events. From the 6339 vehicles listed in the database, 5257 of them are georeferenced by both of its theft and retrieval locations, 653 of them are georeferenced in destination but not in origin, 268 of them were stolen and not retrieved and 161 of them are not georeferenced on neither one of theft and retrieval locations.

Table 1 presents the results of applying the score test in our data set with six distance thresholds, the same for both, origin and destination. We used 3999 Monte Carlo simulations to carry out the test. For each distance threshold, we present the values of the observed statistic, as calculated with the real data set, the expectation of the test statistic under independence and the Monte-Carlo p-value. As seen in Table 1, the hypothesis of independence between origin and destination is rejected at a significance level of 5% in all threshold levels.

$r_x = r_y$	$T(\varphi)$	$E[T(\varphi)]$	P-value
750 m	13,884	1,920	2.5E-4
1,500 m	93,262	21,966.32	2.5E-4
2,500 m	371,307	129,216.81	2.5E-4
3,250 m	734,011	314,244.47	2.5E-4
4,000 m	1,241,616	622,437.05	2.5E-4
5,000 m	2,115,013	1,240,667.53	2.5E-4

Tabela 1.1: Score test for different distances

As there is evidence of dependence between origins and destinations, one further step in the analysis is to fit a stochastic model for the bivariate point patterns. A previous and superficial analysis of our dataset revealed that in Belo Horizonte, if a car is stolen in somewhere, the most probable regions for its retrieval locations are the proximities of the theft location or the Center-South region of the city. So, for Belo Horizonte data, it seems that a useful model is:

- Given that a car is stolen at the origin $\mathbf{x}$, its destination tends to be a mixture of two densities.
- With probability $p(\mathbf{x})$, it stays around $\mathbf{x}$.
- With probability $1 - p(\mathbf{x})$, it tends to be attracted towards the Center-South region.

1.5 Conclusion

In this paper, we introduce a new type of method to analyze bivariate linked point patterns. The interest is on two geographical patterns, the first containing the origin locations, while the second contains the destination pattern of point events in a map. Before any modeling of the association between the patterns, one must first to test if there is any evidence for this association. The main objective of this paper is the introduction of such a test. We showed that a simple statistic, the number of pair of events that are close in both, origin and destination maps, has some optimality properties to test the independence between origin and destination. It is also very simple to calculate and can be applied in virtually any specific bivariate point patterns. One additional advantage of our proposed test statistic is that it does not require risk population information. Hence, we do not need to know where all the cars are located or are more liked to be stolen to test the origin-destination hypothesis. It has an asymptotic normal distribution with known moments that can be used when simulation is too costly.

1.6 Bibliography

Besag, J. and Clifford, P. (1989) Generalized Monte Carlo significance tests *Biometrika*, **76**, 633-642.

Daley, D.J., and Vere-Jones, D. (2003) *An introduction to the theory of point processes*, 2nd edition. New York: Springer-Verlag.

Diggle, P.J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *Journal of the Royal Statistical Society A*, **153**, 349-362.

Diggle, P.J. and Rowlingson, B. S. (1994). A conditional approach to point process modelling of raised incidence. *Journal of the Royal Statistical Society A*, **157**, 433-440.

Kelsall, J. and Diggle, P.J. (1995). Kernel estimation of relative risk. *Bernoulli*, **1**, 3-16.

Knox, G. (1964). Epidemiology of childhood leukemia in Northumberland and Durham. *Brit. J. Prevent. & Social Med.*, **18**, 17-24.

Lotwick, H. W. and Silverman, B. W. (1982) Methods for analyzing spatial processes of several types of points. *Journal of the Royal Statistical Society* B, **44**, 406-413.

Capítulo 2

Visualizing marked spatial and origin-destination point patterns with dynamically linked windows

Danilo Lopes and Renato Assunção

2.1 Introduction

Marked point processes are important when the researcher is interested in modeling dependence between spatial locations and some other variables. For example, ecologists may map the spatial distribution of individual trees in an ecosystem and record their species, ages and sizes (Stoyan and Penttinen, 2000). The mark of each event is a three-dimensional random vector with these three characteristics. Epidemiologists commonly analyze a random spatial pattern of events labeled according to two categories, the disease cases and a set of control individuals (English et al., 1999). In this case, the mark is a binomial indicator. Earthquakes and forest fires are usually recorded as a four-dimensional marked point process with the spatial location, the occurrence time and some measure of the event intensity comprising the data (Ogata, 1998; Schoenberg, 2004).

Another interesting example of spatial marked point processes are the so-called *bivariate linked point processes* (Assunção and Lopes, 2007). These processes consist of events given by pairs of locations in the same geographical region: an origin and a destination positions. For example, we can record the coordinates of car thefts cases and the eventual car retrieval locations. Conditioning on a car being retrieved, the interest is to know if there is some type of spatial dependence between the two locations associated with each car. This dataset can be considered as a four-dimensional structure, where the marks associated with spatial locations are the car retrieval positions and hence, they are also spatial locations. The two-dimensional mark in this case is special because there is little interest in carrying on statistical analysis on the marginal distributions of each coordinate of the car retrieval locations. Rather, the two coordinates must be considered simultaneously to analyze the spatial pattern of the events.

The literature on marked point process has been concentrated on parameter estimation and on testing hypotheses, with little concern for exploratory tools. Several tests have been proposed in the literature recently for the hypothesis that marks and locations are independent (Stoyan, 2000; Schlather et al., 2004; Guan, 2006). We think that these tests should be complemented by exploratory data analysis based on visual displays. The reason is that graphical methods are capable of evoking immediate intuitive understanding of the data generating process. They can potentially suggest models for the marks-location dependence after a statistical test rejects the independence hypothesis. At the same time, they make easier to detect unanticipated features such as the presence of localized clusters of marks and locations dependence, and the mixing of different stochastic processes, as well as trends across regions. The fact that these exploratory methods tend to make few assumptions on the stochastic structure of the data is another advantage.

In this paper, we focus on graphical displays for marked point processes in general. Our visualization tools involve dynamic displays, linked windows or views, interactive controls, and animation. In Section 2.2, we revise the main definitions and notation for marked point process and introduce some spatial statistics implemented in our software. In Section 2.3, we revise some concepts associated with dynamic graphics and introduce the software we developed. In Section 2.4, we illustrate the exploratory techniques using three datasets: two ecological datasets of trees with associated species and stem diameters and the car theft example in a large Brazilian town. Section 2.5 presents our conclusions.

Our tools are implemented as a plug-in of TerraView, a free GIS visualization software with many other spatial statistics capabilities (TerraView, 2008). In the spirit of the open source philosophy, the authors make available their source code in the TerraLib platform (Câmara et al., 2008) as well as a stand alone software, available at `http://www.est.ufmg.br/leste/softwares.htm`.

2.2 Model and Estimation

2.2.1 Spatial Marked Point Processes

Spatial point processes are stochastic processes defined in a bounded polygon $A \subset \mathbb{R}^2$, with realizations that are finite random subsets $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subset A$. The points $\mathbf{x}_i = (u_{1i}, u_{2i})$ are called *events* to distinguish them from other arbitrary points of the plane $\mathbb{R}^2$. When we also have a random structure $\mathbf{m}_i$ associated to each event $\mathbf{x}_i$, we have a spatial marked point process. The marks can be one or more random variables such as trees' heights or trees' species when events are the locations of trees in A. For a recent review of spatial point process theory, see Møller and Waagepetersen (2007).

Many times, the marks $\mathbf{m}_i$ are simply a scalar. In the case of a multivariate point process (multivariate referring to more than one type of event), the marks are discrete (or qualitative) and represent labels, such as species of plants, types of crimes or any other categories. The marks characterizing the point process may also be continuous (or quantitative) expressing, for example, stem diameters of trees or estimated stolen-car values. Apart from space-time models (see, e.g. Schoenberg 2004, Assunção and Maia 2007), analysis of quantitative marks, however, are not yet frequently observed, possibly due to the laborious task of collecting geographical information at the same time that another variable is measured or more likely due to the lack of theoretical models and data analysis tools.

Some recent examples of marked point processes with multivariate marks include the space-time magnitude models (Ogata, 1998; Schoenberg, 2004), where events are collected in the form $\{t_i; \mathbf{x}_i; m_i\}$, representing time, spatial locations and magnitudes, respectively, and the bivariate linked point processes (see Section 2.2.2 for details), where the events can be represented as a pair $\{\mathbf{x}, \mathbf{y}\}$ of linked positions. In this latter case, the mark $\mathbf{y}$ is another location in the same polygon as $\mathbf{x}$.

2.2.2 Bivariate Linked Point Processes

In the case of bivariate linked point process, the events are of the form $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are spatial locations, and the mark is simply a two-dimensional vector $\mathbf{y}$, the destination event location. However, rather than considering this as a special case of marked point processes, it is more useful to treat it differently recognizing the special geographical character of the mark.

As an illustration, consider the dataset of car thefts occurred in Belo Horizonte, a Brazilian city with 331 km^2 of extension and approximately 2.4 million inhabitants. The data were provided by the State Police Force. In the period from August, 01 2000 to July, 31 2001 there were $m = 6339$ stolen cars in the city and their locations can be seen in the left hand side map of Figure 2.1. The right hand side map of Figure 2.1 shows the locations of the $n = 5257$ cars eventually found within the city limits. Conditioning on a car being retrieved, the interest is to know if there is some type of spatial dependence between the two locations of the stolen cars.

There are other situations generating multivariate linked point processes. Migration, with its origin and destination places, is a major case. Another example is homicide information where we can have up to three linked positions, the murder location, and the victim and offender addresses. These situations have in common that two or more spatial point patterns have linked events. That is, for each event in one point pattern, there is one or more corresponding events in the other point pattern. In this paper, we concentrate on the bivariate situation, calling the patterns as origin and destination processes.

One exception to the lack of linked point processes statistical analysis are the papers by Zeman (1997), Krejčíř (2000), and Beneš et al. (2005). They consider 866 cases of Lyme borreliosis disease occurring during 1987-1991 in the Central Bohemian region of the Czech Republic. Associated with each single case, they have the coordinates of the individual home and of the reported infection location. The main interest was the derivation of a spatially constant function $g(d)$ that would describe the movement pattern of the risk population around the residences. For this purpose, only the distance between the two linked locations

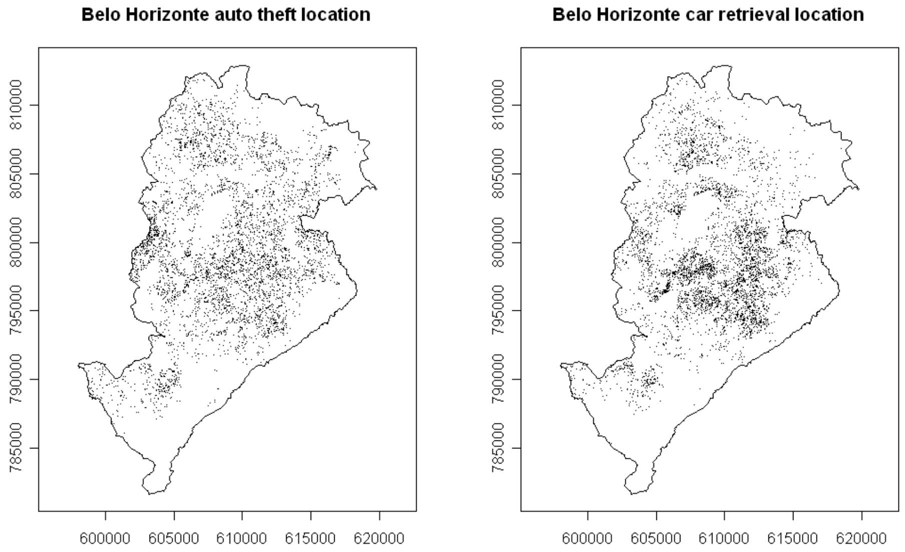

Belo Horizonte auto theft location
Belo Horizonte car retrieval location
785000
790000
795000
800000
805000
810000
600000
605000
610000
615000
620000

Figura 2.1: Maps of car theft locations and their subsequent retrieval locations in Belo Horizonte from Aug 2000 to July 2001

was used to obtain $g(d)$, both parametrically and non-parametrically. Therefore, although the data set had a linked point process structure, it was not explicitly used in these previous works.

A similar situation was studied by Banerjee, Gelfand and Polasek (2000). They analyzed the postal service performance with regard to delivery of priority mail using a designed experiment involving 79,557 letters. Flows between pairs of origin and destination were modeled by a random function of both locations. Although this structure is similar to ours in the sense that each data item has a linked origin and a destination, the stochastic structure is not of the point process type.

Assunção and Lopes (2007) propose a Gibbs model with pairwise interaction functions for the bivariate linked point processes and derive summary statistics for the main characteristics of the model. Since the interest is to model the association between origin and destination locations, they assume a distribution conditional on the unordered set of origin locations $\mathbf{x}_1, \dots, \mathbf{x}_n$ and on the unordered set of destinations $\mathbf{y}_1, \dots, \mathbf{y}_n$:

$$p(\varphi|\{\mathbf{x}_1, \mathbf{x}_n\}, \{\mathbf{y}_1, \mathbf{y}_n\}) = \frac{\exp\left(-\sum_{i<j} \phi\left((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)\, ; \boldsymbol{\theta}\right)\right)}{\sum_\pi \exp\left(-\sum_{i<j} \phi\left((\mathbf{x}_i, \mathbf{y}_{\pi_i}), (\mathbf{x}_j, \mathbf{y}_{\pi_j})\, ; \boldsymbol{\theta}\right)\right)} \tag{2.1}$$

where ϕ is called the *potential function*, φ denotes the unordered set $\{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ of pairs of locations and $\sum_\pi$ means the sum over all possible permutations of destinations with respect to the data. In a simple special case, we have:

$$\phi\left((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)\, ; \boldsymbol{\theta}\right) = \theta\, I[|\mathbf{x}_i - \mathbf{x}_j| < r_x]\, I[|\mathbf{y}_i - \mathbf{y}_j| < r_y] \tag{2.2}$$

where $I[\cdot]$ is the indicator function and r_x and r_y are threshold limits for distances between pairs of origin and destination events, respectively. For this special case, one can show that:

$$\sum_{i<j} \phi\left((\mathbf{x}_i, \mathbf{y}_{\pi_i}), (\mathbf{x}_j, \mathbf{y}_{\pi_j})\, ; \boldsymbol{\theta}\right) = \theta T(\pi\varphi)$$

where $\pi\varphi$ represents the unordered set of events with the destinations shuffled by some arbitrary permutation of indexes $(\pi_1, \cdots, \pi_n)$ and $T(\pi\varphi)$ corresponds to the number of pairs of events that are within the threshold limits at the origin and at the destination for this set. Therefore the statistic $T(\varphi)$ is sufficient for the distribution of the process conditional on the marginal point patterns.

If $\theta > 0$ the density function (2.1) puts more probability mass on configurations in which pair of events that are near each other in the origin tend to be near each other also in the destination. If $\theta < 0$, we have the opposite effect: a small number of pairs that are near each other in origin and in destination is more probable. Assunção and Lopes (2007) also show that, for this simple case, the score statistic only depends on the number of pair of events that are within the threshold limits at the origin and at the destination, and derive the first two moments for this statistic.

The score statistic for origin-destination independence test is similar to another well-known test statistic, the Knox test statistic (Knox, 1964), proposed for testing space-time interactions. Specifying a spatial and a temporal critical distance, it is possible to indicate when a pair of events is close in space or close in space and in time. Knox test statistic X is defined as the number of pairs of events that are simultaneously close in space and in time. Similarly to the Knox test statistic, the score statistics for our simple pairwise interaction model is the total number of pair of events that are simultaneously close in space and in time. We can approximate the distribution of the score statistic by a normal distribution and run a statistical test. Another alternative is to run a Monte based test. We will further discuss some details of both approaches in Section 2.4.3.

2.2.2.1 Simulation of the process

In this section we explore methods for simulating (2.1) using spatial birth-and-death processes (Ripley, 1977). This technique uses Metropolis-Hastings algorithm to propose transitions rates for a birth or a death from an initial state and converges in distribution to a Markov point process.

Let $\sim$ be a symmetric, reflexive relation on the study region A. We say that two events η and ξ are related under $\sim$ (and write $\xi \sim \eta$) if their origin locations are r_x-close and their destination locations are r_y-close. We define the neighborhood of ξ as the set of all $\eta \in \varphi \setminus \{\xi\}$ such that $\eta \sim \xi$ and denote it by $N_\varphi(\xi)$. We can easily see from the proposition in (2.2) that $\phi\left((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)\, ; \boldsymbol{\theta}\right) \neq 0$ implies $(\mathbf{x}_i, \mathbf{y}_i) \sim (\mathbf{x}_j, \mathbf{y}_j)$ and, therefore, $\exp(\phi(\cdot))$ is an *interaction function* (Ripley and Kelly, 1976). Hence the conditional density given in (2.1) is a Markov function, as defined by Ripley and Kelly (1976), and this special case of bivariate point process is a Markov point process with respect to the neighbor relation $\sim$. In special, the conditional intensity

$$\lambda(\xi|\varphi \setminus \{\xi\}, \{\mathbf{x}_1, \mathbf{x}_n\}, \{\mathbf{y}_1, \mathbf{y}_n\}) = \frac{p(\varphi|\{\mathbf{x}_1, \mathbf{x}_n\}, \{\mathbf{y}_1, \mathbf{y}_n\})}{p(\varphi \setminus \{\xi\}|\{\mathbf{x}_1, \mathbf{x}_n\}, \{\mathbf{y}_1, \mathbf{y}_n\})}$$

depends only on the neighborhood $N_\varphi(\xi)$.

For origin-destination data, given an origin process, a destination process, the parameter $\boldsymbol{\theta}$, and thresholds r_x and r_y, we start from any initial linking between the two processes. At each iteration, we propose a random permutation of the event's destination and accept or reject it according to a transition rate given by:

$$\begin{aligned} p(\pi\varphi_k, \pi\varphi_{k-1}|\{\mathbf{x}_1, \mathbf{x}_n\}, \{\mathbf{y}_1, \mathbf{y}_n\}) &= \min\left\{\frac{p(\pi\varphi_k|\{\mathbf{x}_1, \mathbf{x}_n\}, \{\mathbf{y}_1, \mathbf{y}_n\})}{p(\pi\varphi_{k-1}|\{\mathbf{x}_1, \mathbf{x}_n\}, \{\mathbf{y}_1, \mathbf{y}_n\})}; 1\right\} \\ &= \min\left\{\exp\left(-\theta\left[T(\pi\varphi_k) - T(\pi\varphi_{k-1})\right]\right); 1\right\} \end{aligned}$$

where $\pi\varphi_k$ to denote the unordered set of events with the destinations shuffled by some arbitrary permutation $(\pi_1, \ldots, \pi_n)$ at the k-th iteration where and $T(\pi\varphi_k)$ is the corresponding number of pair of events that are within the threshold limits at the origin and at the destination.

This process converges to a bivariate linked point process with probability density function given by:

$$C \exp\left(g(\mathbf{x}_1, \ldots, \mathbf{x}_n) + h(\mathbf{y}_1, \ldots, \mathbf{y}_n)\right) \sum_{\pi} \exp\left(-\sum_{i<j} \phi\left((\mathbf{x}_i, \mathbf{y}_{\pi_i}), (\mathbf{x}_j, \mathbf{y}_{\pi_j}); \boldsymbol{\theta}\right)\right)$$

We used this algorithm to obtain simulated point patterns in the unit square (Figure 2.3). Each simulated pattern consists of the same marginal point patterns for origin and destination, shown in Figure 2.2. These marginal patterns are fixed and simulations generate only the links between a origin event and a destination event. We used the simple potential function in 2.2 and threshold limits of 0.05 for r_x and r_y. On the left side of Figure 2.3, we have the generated point pattern under the hypothesis of independence between origin and destination locations ($\theta = 0$). Any linkage between the two point patterns is equally likely. In the middle of the figure, we have the generated point pattern with $\theta < 0$. Lower values of the statistic $T(\varphi)$ are more likely and the simulation tends to generate inhibition between origin and destination (events with origin locations close to each other have, in general, destination locations within a distance greater than the destination threshold). Observe that there is a strong restriction in the form of the fixed marginal point patterns. Finally, for the bivariate linked point pattern on the right side, we have $\theta > 0$, indicating clustering, i.e. the events tend to be close to each other in both origin and destination locations.

The assumption of a Gibbs model for the bivariate linked point processes allows the use of several methods for parameter estimation of point processes. Geyer (1999) gives a description

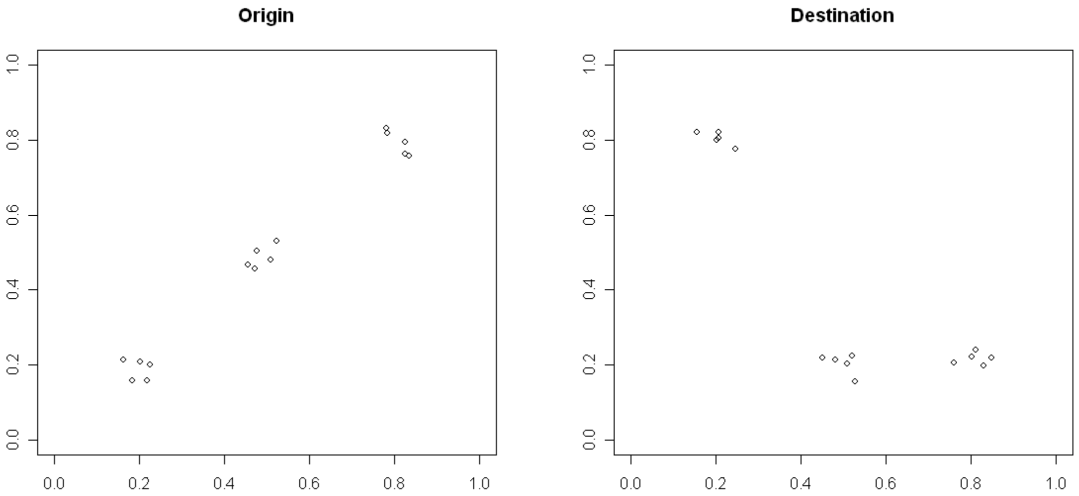


Figura 2.2: Origin-destination point patterns that were used in simulation for patterns in Figure 2.3

of simulation and estimation methods for spatial point processes. The application of these methods for origin-destination processes is, however, in its initial stage of implementation in our software.

2.2.3 Conditional Density Estimation

Of special interest in the analysis of marked point processes is the exploratory analysis of the relationship between marks and spatial locations. One helpful method for this exploration is the estimation of conditional densities. Hyndman, Bashtannyk and Grunwald (1996) extended the kernel conditional density estimator proposed by Rosenblatt (1969) in the following way: let $\{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \cdots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$ be a sample of n observations from the distribution of $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{X}$ is defined on $\mathbb{R}^m$ and $\boldsymbol{Y}$ is defined on $\mathbb{R}^d$. Then the estimator of the conditional density of $\boldsymbol{Y}|\boldsymbol{X}$ has the form:

$$\hat{f}(\mathbf{y}|\mathbf{x}) = \frac{(1/nab)\sum_{i=1}^{n} K\left((||\mathbf{x}-\boldsymbol{X}_i||_x)/a\right) K\left((||\mathbf{y}-\boldsymbol{Y}_i||_y)/b\right)}{(1/na)\sum_{i=1}^{n} K\left((||\mathbf{x}-\boldsymbol{X}_i||_x)/a\right)} \tag{2.3}$$

where $||\cdot||_x$ and $||\cdot||_y$ are distance metrics on the spaces of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, $K(\cdot)$ is a traditional kernel function (symmetric, univariate probability density function) and the parameters a and b control the degree of smoothing applied to the density estimate. Hyndman et al. (1996) give the bias, variance, mean square error and convergence properties of $\hat{f}(y|x)$.

In this article we use the Euclidean distance for all the applications, and our choice of kernel function for continuous random variables is the quartic kernel:

$$K(x) = \begin{cases} \frac{3}{\pi}\left(1-x^2\right)^2 & \text{for } |x|<1; \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

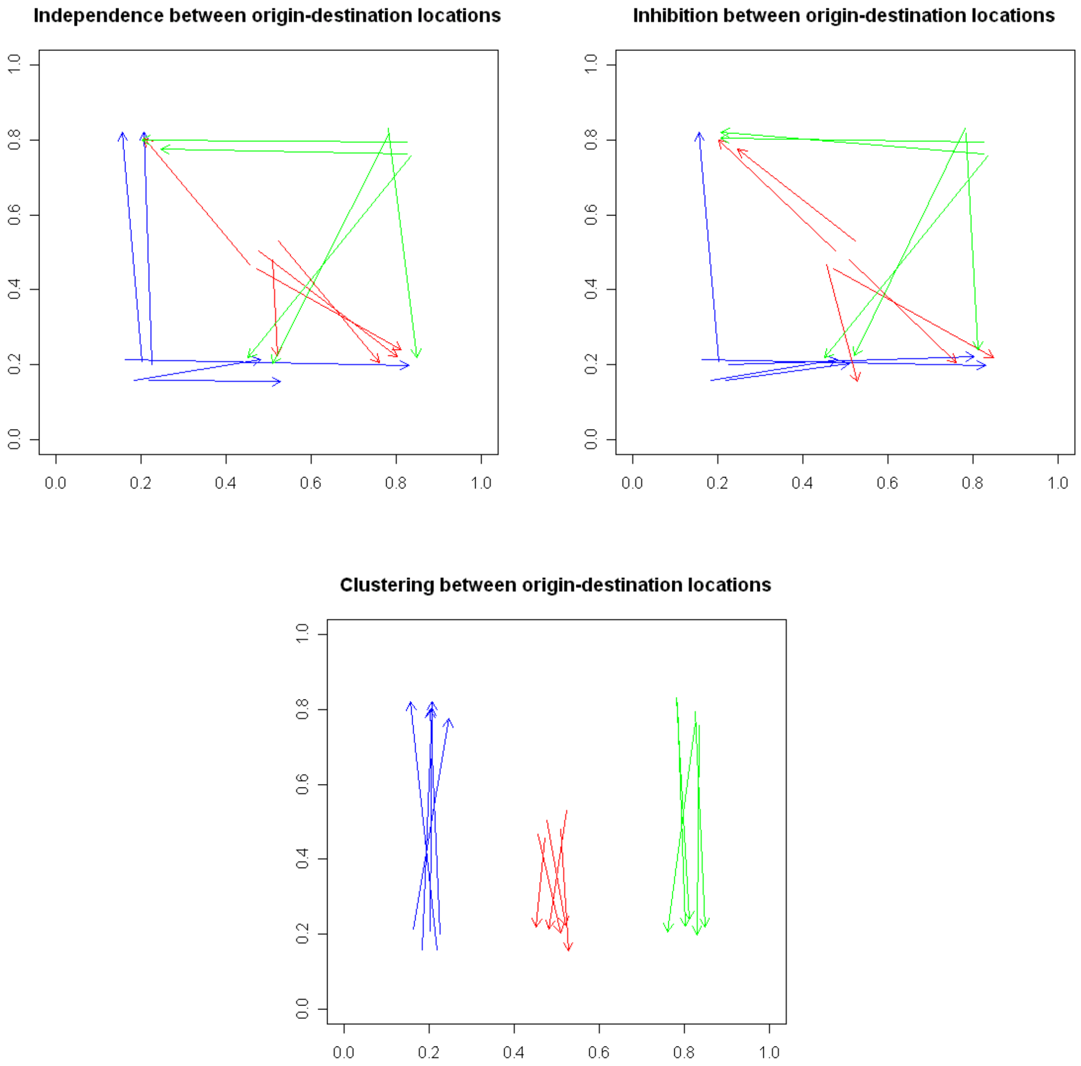


Figura 2.3: Simulated point patterns. Top left: independence between origin and destination ($\theta = 0$). Top right: inhibition between origin and destination ($\theta = -1$). Bottom: clustering between origin and destination ($\theta = 1$)

Hall, Racine and Li (2004) generalize the conditional density estimator defined in (2.3) to the case where $\boldsymbol{X}$ has components of many different types: continuous, unordered discrete, ordered discrete. They divide $\boldsymbol{X}_i$ into two components, $\boldsymbol{X}_i^c = (X_{i1}^c, \cdots, X_{iq}^c)$ and $\boldsymbol{X}_i^d = (X_{i1}^d, \cdots, X_{ip}^d)$, denoting the continuous and discrete parts, respectively. Assuming that $\boldsymbol{X}_{ij}^d$ takes the values $\mathbf{x}_0^d, \mathbf{x}_1^d, \cdots, \mathbf{x}_{r-1}^d$, the generalized kernel is defined as:

$$K(\mathbf{x}, \boldsymbol{X}) = K^c(\mathbf{x}^c, \boldsymbol{X}^c) K^d(\mathbf{x}^d, \boldsymbol{X}^d)$$

with

$$K^d(\mathbf{x}^d, \boldsymbol{X}^d) = \prod_{j=1}^{q} \{\lambda_j/(r_j - 1)\}^{1-N_{ij}} (1 - \lambda_j)^{N_{ij}} \tag{2.5}$$

where $K^c(\mathbf{x}^c, \boldsymbol{X}^c)$ corresponds to the same kernel function used in (2.3), $N_{ij} = I\left(\boldsymbol{X}_{ij}^d = \mathbf{x}_j^d\right)$, and $\lambda_1, \cdots, \lambda_q$, the smoothing parameters for the discrete component, are constrained by $0 \leq \lambda_j \leq (r_j - 1)/r_j$.

For simplicity, we here consider X as scalar. In this case, the kernel function in (2.5) is appropriate if X is a discrete variable with r unordered categories. If $\lambda = 0$, then $K^d(x_i, X) = I(X = x_i)$ and the kernel conditional density estimator in (2.3) is the naive cell-proportion (or maximum-likelihood) estimator. If $\lambda = (r-1)/r$ then $\widehat{\Pr}(X = x_i|\mathbf{x}) = K^d(x_i, X) = 1/r$ and the estimator resembles the discrete uniform distribution independently of the conditional location. Hall et al. (2004) also develop the mean square error for their estimator and introduce an algorithm that optimizes the smoothing parameters.

In the case of ordered discrete variables, we use the uniform weight function proposed by Wang and Ryzin (1981). Assuming that X takes values $\{\cdots, -1, 0, 1, \cdots\}$, the kernel weights are defined as:

$$K(j, X_i) = \begin{cases} \frac{1}{2}s/k & (|X_i - j| = 1, \cdots, k;\ k \text{ fixed integer } \geq 1), \\ 1 - s & (X_i = j) \\ 0 & (|X_i - j| \geq k + 1) \end{cases}$$

where k determines the number of symmetrical neighboring cells being smoothed and s, the weight function parameter, is constrained by $0 \leq s \leq 1$. Again, if $s = 0$, we have $K(j, X) = I(X = j)$. This function may assign weights to values outside the support set, so Wang and Ryzin (1981) propose the reflection of the data or the placement of the weights on the nearest valid value.

The application of this estimator to the context of marked point processes seems unexplored until this moment. Given the observed pattern $\{(\mathbf{x}_1, \mathbf{m}_1), \cdots, (\mathbf{x}_n, \mathbf{m}_n)\}$ we can estimate the density of the mark $\mathbf{m}$ conditional on the location $\mathbf{x}$ or vice versa by using Equation (2.3). In the case of marked point processes with scalar marks M, we can obtain the conditional probability $\Pr(M = m_i|\mathbf{x})$ that the mark takes the value m_i (if the mark is discrete) or the the conditional density $f(m|\mathbf{x})$, expressing the likelihood of getting the mark value m given that we have observed an event at location $\mathbf{x}$ (if the mark is continuous). In the special case of bivariate linked point processes, the conditional density $f(\mathbf{y}|\mathbf{x})$ gives the likelihood of ending at destination $\mathbf{y}$ given that the origin is $\mathbf{x}$. Using the same estimator is also possible to make the reverse analysis: given an observed value for the mark, we can use (2.3) to access the conditional distribution of the spatial locations. Our software, described in Section 2.3, produces this density in a dynamically linked window environment.

2.2.4 Spatial Cumulative Distribution Function

Another statistical summary that captures the empirical distribution of marks over a spatial domain of interest is the spatial cumulative distribution function (SCDF) (Kaiser et al. 1997, Lahiri et al. 1999). Let M be a univariate real-valued continuous mark and $R \subset A$ be a subregion for which we are interested in analyzing the mark distribution; then a basic estimator for the SCDF is:

$$\hat{F}_n(m; R) = n^{-1} \sum_{i=1}^{n} I[M_i \leq m]; \quad m \in \mathbb{R}. \tag{2.6}$$

Kaiser et al. (1997) and Lahiri et al. (1999) use this statistic in the context of continuous random fields (although the latter applies the SCDF to measurements in individual trees). Approximations of marked point processes to random field mark models are common in ecology, even tough this simplification ignores interaction among trees for very short distances (Stoyan, 2000).

2.2.4.1 Confidence Envelopes

An *independently marked point process* occurs when the marks are identically distributed and independent of locations (Schlather et al, 2004). Under the hypothesis that the point process is independently marked, we can use bootstrap to resample estimators for the SCDF given by Equation (2.6). This provides confidence envelopes as we will show in Section 2.4.3.

2.3 Graphical Display

Becker and Cleveland (1988) popularized the scatterplot brushing technique among statisticians. Since then, the concept of linked views of a dataset became common and it was implemented in several software tools. The idea is wonderfully simple: different windows show different views of the same data items. One view could be, for example, a scatterplot of X versus Y while another view could be a scatterplot of Z versus W, or it could be a box plot or a histogram of Z. The views are dynamically linked through a brush, which usually is a rectangular area on the computer screen that can be moved with the mouse. Clicking on one view with the brush marks a subset of items with a different symbol. The items so marked are also highlighted with the same symbol and at the same time in the other linked views. Move the mouse around brushing the first view and follows what patterns happen to occur in the other views.

We show how this can be useful for understanding the structure and relationships between marks and locations of two linked point patterns. The visualization framework allows diverse complex models to be visually compared in a way that illuminates the similarities and differences in the underlying models, facilitates the choice of a particular model structure, and provides a useful check for implausible predictions of future observations in regions with little or no data.

The idea of applying interactive and dynamic graphics to data analysis opened an entire new avenue of research in statistics. In particular, the application of these ideas in a geographic context has been called ESDA, an acronym for exploratory spatial data analysis (Anselin, 1998). For a recent and thorough review, see Symanzik (2004).

Haining (2003) contrasts ESDA to confirmatory analysis: instead of exploring model fits and orienting decision makers, the aim is to describe and detect patterns in order to give insights into the data to the researchers. Visualization, he says, plays an important role in this context, because it provides analysis in a quickly and intuitive and more informative way, allowing non-statisticians to participate in the process of understanding the data structure.

2.3.1 Functionality

ESDA requires the ability of rapidly drawing maps and graphs according to parameters modified interactively by the user; for this purpose, dynamic links between geographical and statistical displays are necessary. The windows are disposed side by side and allow the analyst to instantly recognize patterns and make comparisons between subsets of the data using the appropriate selections.

We here define some notions introduced by Buja et al. (1996) that compose the heart of our graphical toolkit: *linking multiple views, dynamic brushing* and *dynamic projection mechanisms*. The first feature gives to the user the ability to pose the query graphically and to receive the response graphically as well. In the context of the spatial statistics presented in section 2.2, if a location in the map is selected in the geographical display, the application responds with the corresponding plot in the statistical window.

Dynamic brushing consists of moving the query and noting the changing pattern of highlighted cases in a graph. Finally, *Dynamic projection mechanisms* is characterized by small changes in parameters of the graph performed at rapid speed resulting in real-time animation controlled by user's input. This feature allows the analyst to keep track of observed patterns as the animation moves smoothly.

2.3.2 Software Architecture and Design

MaPPEA (**Ma**rked **P**oint **P**rocesses **E**xploratory **A**nalysis) is an ongoing toolkit for interactive graphical exploratory data analysis of several types of spatial marked point processes using linked dynamic graphics. MaPPEA components were developed in C++ as a part of the activities of the Spatial Statistics Laboratory (LESTE/UFMG) to facilitate spatial data analysis. MaPPEA's visualization framework allows diverse and complex models to be visually compared in a way that illuminates some similarities and differences in these models, facilitates the choice of a particular model structure, and provides a useful check for implausible predictions of future observations in regions with little or no data.

All the graphical summaries presented in Section 2.2 were built from scratch in a modular approach and displayed in linked dynamic graphics using the CImg Library (version 1.2.8) (Tschumperlé, 2008), an open-source C++ template image processing library with many utilities such as processing and displaying objects and accessing user actions. The software is distributed in a stand-alone version as well as a plug-in version integrated with the open-source GIS TerraLib classes and functions library (Câmara et al., 2008). The same developers of TerraLib also released TerraView, an application based on this library, for visualization and analysis of spatial databases. In the plug-in version, the MaPPEA application may be launched from within TerraView via a custom option on TerraView menu bar. It is important to point out that MaPPEA is a completely separate application, which runs in parallel to the GIS.

Once launched, both versions of MaPPEA display a small window constructed in Qt 3.2.0 with a set of tabs that organizes the collection of spatial analysis routines available to the user in two modules: *Visualization* and *Inference*. The algorithms included in the visualization module are the conditional density estimator, the spatial cumulative distribution function, the conditional probability estimator, a dynamic brushing barplot for quantitatively marked point processes and the spatial birth-and-death simulation of bivariate linked point processes. On the other hand, the inference module currently offers routines for the Knox test for space-time interaction and the score test for association between origin-destination spatial locations (Assunção and Lopes, 2007). Further versions of the software will include the tests described in Schlather et al. 2004, Guan 2006 and Assunção et al. 2007 in the inference module. MaPPEA interface with the visualization module enabled is shown in Fig. 2.4.

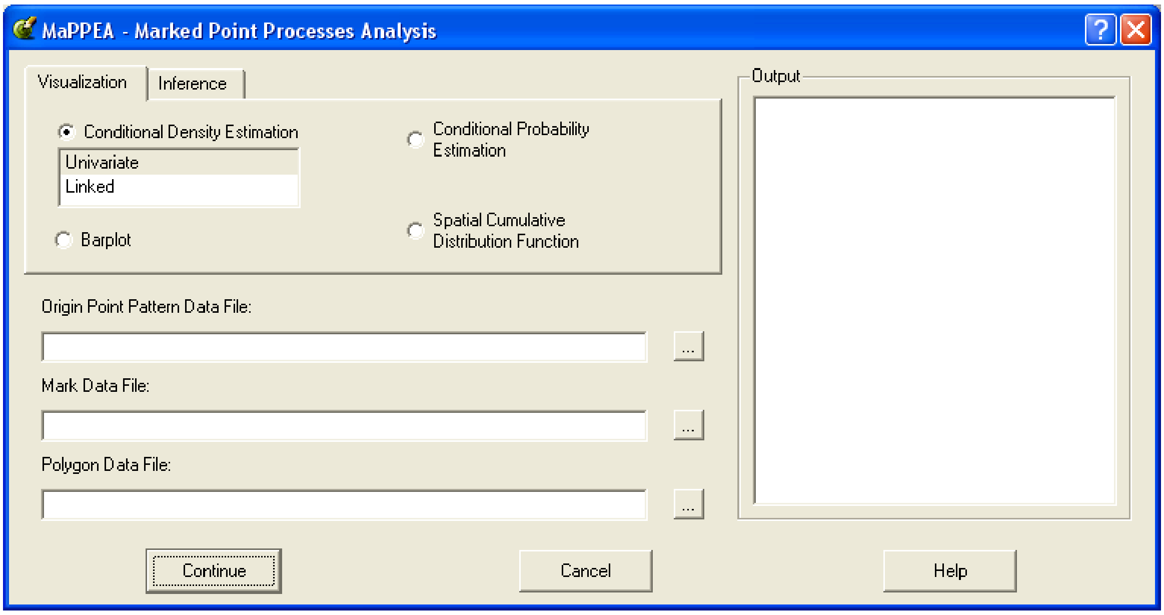

MaPPEA - Marked Point Processes Analysis
Visualization
Inference
Output
● Conditional Density Estimation
Univariate
Linked
○ Conditional Probability Estimation
○ Barplot
○ Spatial Cumulative Distribution Function
Origin Point Pattern Data File:
...
Mark Data File:
...
Polygon Data File:
...
Continue
Cancel
Help

Figura 2.4: MaPPEA displaying the visualization module

2.3.3 Graphics Input and Output

MaPPEA as a stand-alone software reads data from text files; the software needs separate files in table format for spatial coordinates, marks and polygon boundary. The two files of coordinates and marks must have the same number of lines, in which each line of the files corresponds to a single event; marks and locations are linked by order of appearance. Each line of the polygon data file corresponds to its vertices' coordinates. As a TerraView plug-in, MaPPEA interacts with the GIS to import coordinates information from geographic themes on the database and marks information from selected columns of static tables. In both cases, the graphical user interface text or combo box for data input.

When running a specific algorithm, MaPPEA also uses the pointing device for graphics input, in a linked multiple views fashion. The position of the cursor on the map is continuously captured and, after a single-click inside the polygon boundary, it calculates spatial statistics conditioned on the corresponding geographical coordinates.

The output is a map or graph corresponding to the visualization of the spatial statistic calculated. As we said before, the plug-in application runs in parallel to the GIS, but we plan to integrate it completely to the TerraView application in future versions, allowing the user to alternate visualizations and features from both softwares in the same database.

2.4 Applications

2.4.1 Lansing Wood data

Gerrard (1969) describes an investigation of a 924 ft × 924 ft (19.6 acre) plot in Lansing Woods, Clinton County, Michigan, USA. He has provided the location of 2,251 trees in the plot, with their corresponding botanical classification (as hickories, maples, red oaks, white oaks, black oaks and miscellaneous trees). The original plot size has been rescaled to the unit square. These data have been analyzed by several researchers. Holmes and Adams (2002) use nearest neighbor method to produce smooth posterior class conditional probability fields for trees in Lansing Woods. We here apply the dynamic brushing barplot tool to this example of multivariate point process.

The left side of Figure 2.5 shows the geographical window containing the square plot with all the tree's locations. We chose a radius of 0.07 units for the brushing circle drawn to the left of the plot; this circle is centered at the cursor in a way that any movement of the cursor over the plot results in the same movement for the circle. The right side of Figure 2.5 shows the statistical window. At the top, the barplot corresponds to the empirical frequencies of the species of all 2,251 trees. This graph does not change as the analysis proceeds. The x-axis shows all the categories (species) observed and the y-axis represents the number of trees in each category. When the user single-click inside the geographical plot, all the events inside the circle of radius 0.07 centered at the cursor are highlighted in red and the barplot of the empirical distribution of the corresponding marks is drawn at the bottom of the statistical window. If the mouse is dragged over the square plot, the barplot changes in a dynamic projection mechanism.

Future versions of MaPPEA software will also include estimated conditional probabilities (Eq. **??**) for analysis of qualitatively marked point processes.

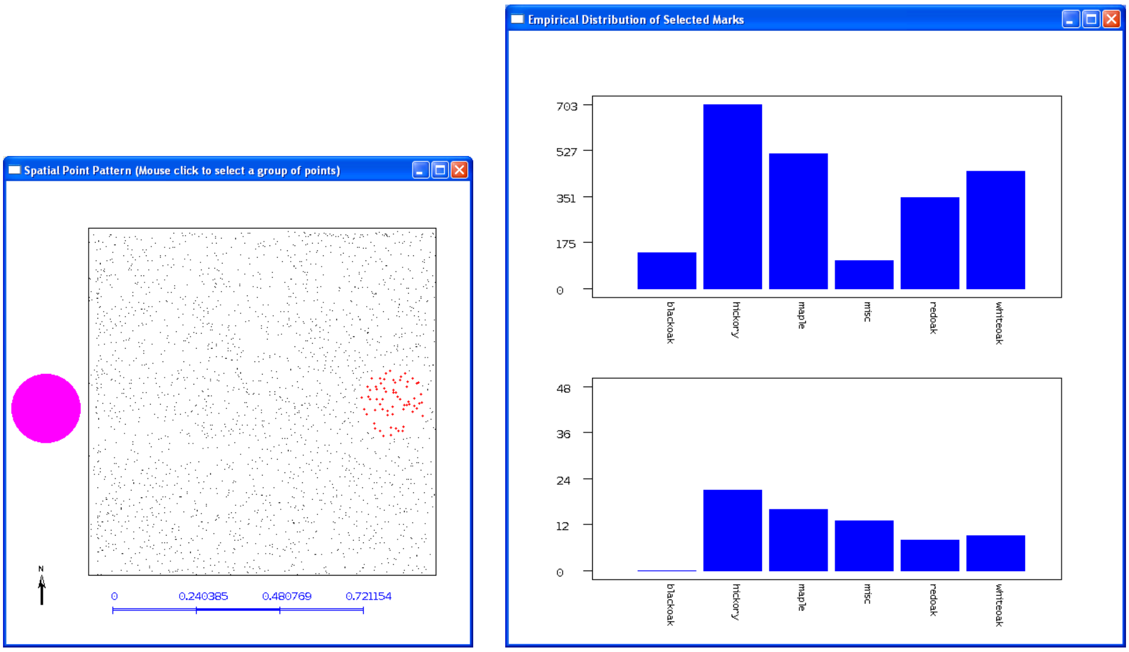


Figura 2.5: Dynamic brushing barplot feature for analysis of qualitatively marked point processes

2.4.2 Coulissenhieb data

We now illustrate our toolkit by applying it to a quantitatively marked point process. Data give the locations and the stem diameters of a stand of 476 spruces at *Coulissenhieb* in the Fichtelgebirge, Northern Bavaria, Germany. The data were collected by Pedro Gerstberger, BITOEK, University of Bayreuth, Germany, and published as part of R package "MarkedPointProcess" by Martin Schlather (Schlather et al, 2004).

We first apply the conditional density estimator to the data. The chosen smoothing parameters were 7.5 m for the tree's locations and 0.025 m for the stem diameters. The purple-circle radius in Figure 2.6 corresponds to the value of the smoothing parameter. The larger the radius of the circle, the more smoothing will be applied to the estimator (and the more events will be considered in the estimation, since the quartic kernel only consider events inside the circle). When the user single-click in point inside the provided polygon boundary, the function corresponding to the conditional density estimator is drawn on the statistical window and a red cross is left on the clicked position, letting the user know to which position the graph corresponds. Figure 2.6 shows both displays after the single-clicking. If the user drags the cursor on the geographical window, the conditional density estimator changes dynamically in a continuous motion.

Next, we apply the spatial cumulative distribution function to the Coulissenhieb data. We evaluate the SCDF for circles of radius 7.5 m in the study region. The left side of Figure 2.7 shows the circle aside the geographical area and some selected events in red. The right side of the figure shows the SCDF for the entire study region at the top. Once we select the

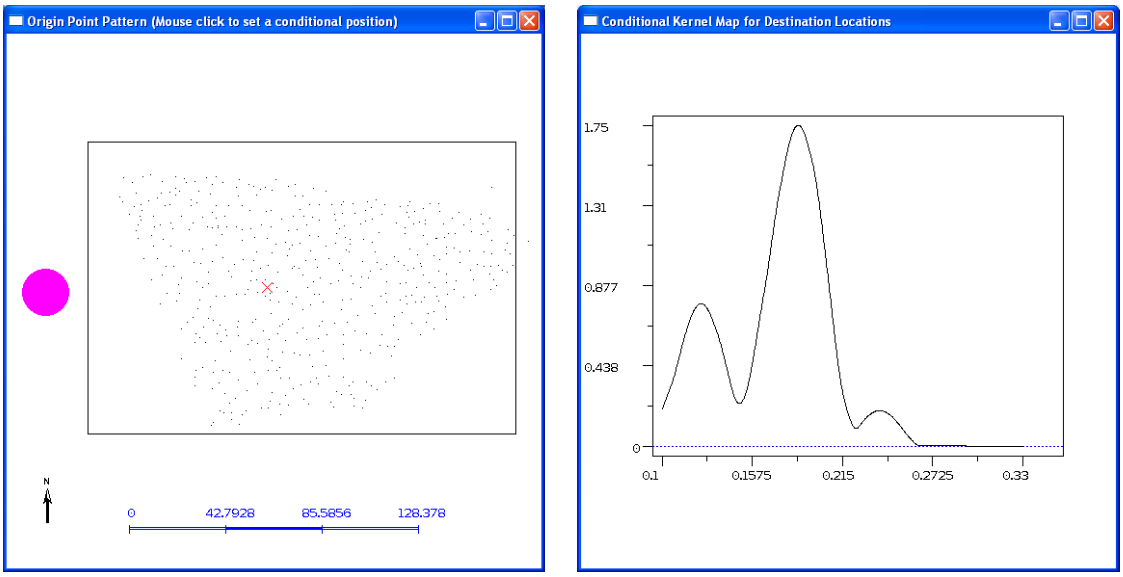

Origin Point Pattern (Mouse click to set a conditional position)
N
0
42.7928
85.5856
128.378
Conditional Kernel Map for Destination Locations
1.75
1.31
0.877
0.438
0
0.1
0.1575
0.215
0.2725
0.33

Figura 2.6: Conditional density estimator for analysis of univariate continuous marks

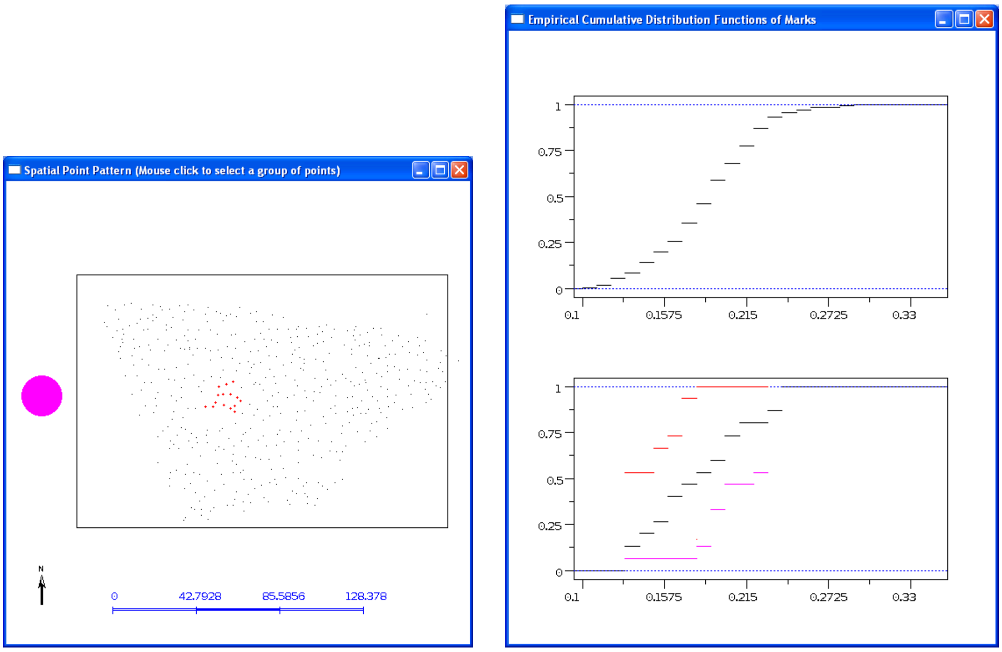

Spatial Point Pattern (Mouse click to select a group of points)
N
0
42.7928
85.5856
128.378
Empirical Cumulative Distribution Functions of Marks
1
0.75
0.5
0.25
0
0.1
0.1575
0.215
0.2725
0.33
1
0.75
0.5
0.25
0
0.1
0.1575
0.215
0.2725
0.33

Figura 2.7: Spatial Cumulative Distribution Function with 95% confidence envelopes

events by single-clicking, the algorithm calculates the SCDF corresponding to these selected events and, under the null hypothesis of independently marked point processes, it creates 95% confidence envelopes based on 10,000 simulations for the spatial statistic and draws the curves at the bottom. At this point, for each one of the 10,000 simulations, a sample of the same size as the number of selected events is taken and the corresponding cumulative distribution function is calculated; the algorithm then plot the upper and lower envelopes based on the lines connecting the 2.5th and 97.5th percentiles of the simulated CDF's for each value of the x-axis. Such envelopes offer diagnostic suggestion of departure from the null hypothesis but do not provide formal statistical inference in the form of a test statistic. SCDF is a function defined on the real line that increases from 0 to 1 and is right continuous; the common image of the functions for different regions facilitates the comparison of how geographical regions affect the distribution of marks. Dynamic projection mechanism is also available, with both envelopes and observed SCDF changing dynamically.

2.4.3 Belo Horizonte data

Finally, we use the MaPPEA software for the analysis of a bivariate linked point process. We use the auto theft and car retrieval locations data from a 3-million-inhabitant Brazilian city, Belo Horizonte, during the period from August 2000 to August 2001 collected by the Military Police of Minas Gerais recorded the crime events and geocoded their locations. Out of 6,339 vehicles listed in the database, 4,554 were georeferenced by its theft and retrieval locations, 703 were carjackings with vehicles being abandoned a few meters from the theft location, 653 were georeferenced in destination but not in origin, 268 were stolen and not retrieved, and 161 were not georeferenced on neither the theft or the retrieval locations. We focus on the 4,554 cars that have origin and destination information and do not correspond to carjackings.

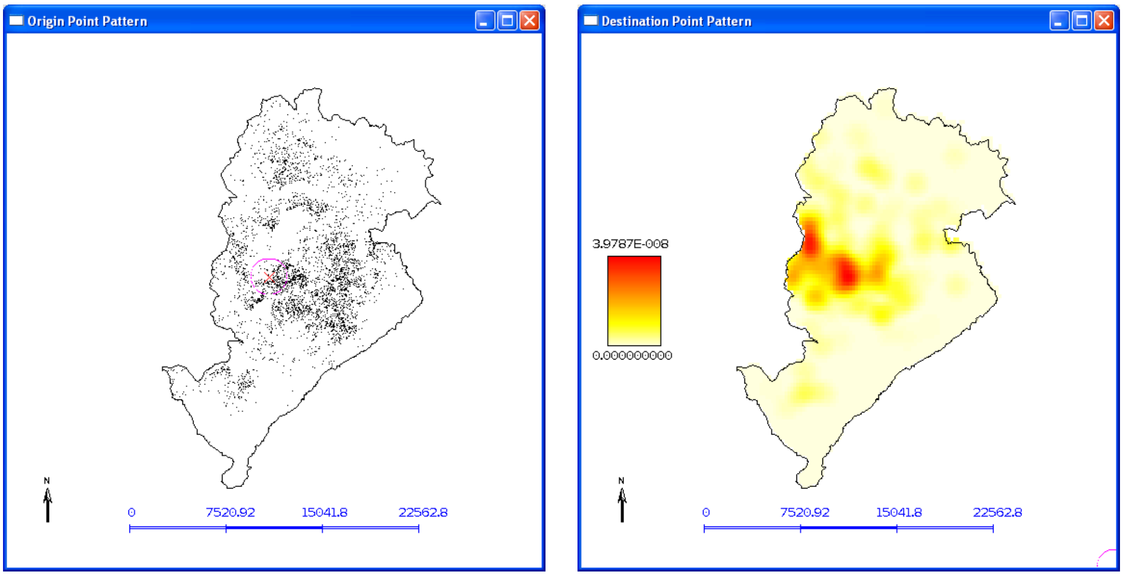


Figura 2.8: Conditional density of destination given the selected origin location

The kernel estimator of conditional density defined in Equation (2.3) generated a powerful tool for visualizing the relationship between origin and destination point patterns. This tool works in a way similar to the usual kernel intensity estimation: an origin location $\mathbf{x}$ is fixed and the estimator in Eq. (2.3) is evaluated for a grid of locations $\mathbf{m}$ that covers the entire study region, producing a conditional intensity surface estimate. Using dynamic graphics, this procedure becomes extremely intuitive for the user: after entering the data and selecting

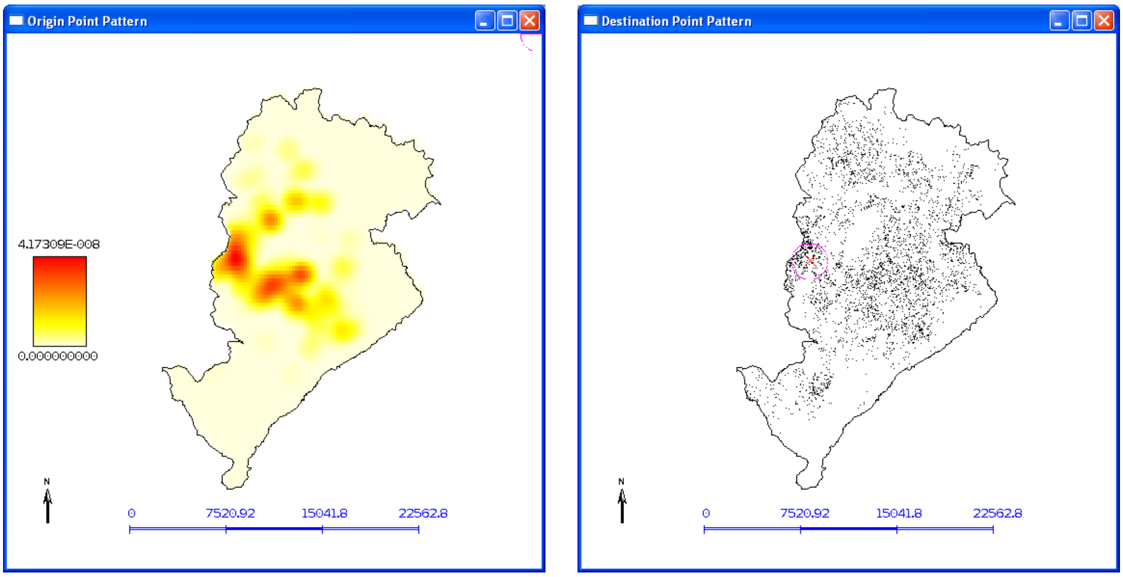
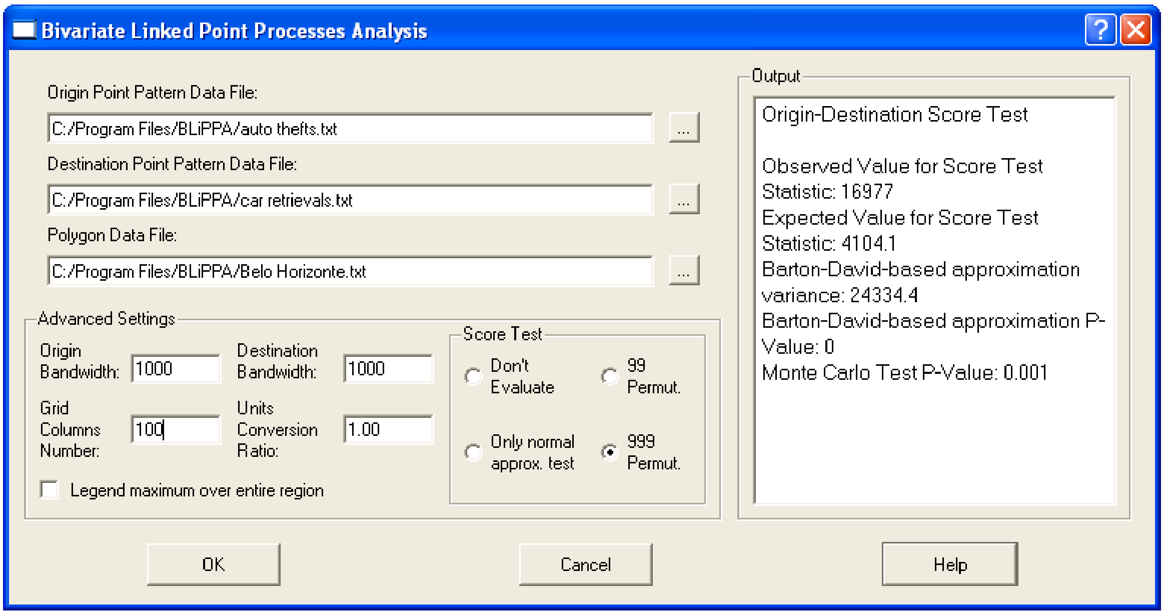


Figura 2.9: Conditional density of origin given the selected destination location

this tool, the software shows two different windows of the same data items, one representing the origin point pattern and the other representing the destination point pattern. If the user mouse-clicks on a location in the origin data window, the conditional surface of the destination locations corresponding to this origin location is drawn on the second window. In our example, we chose the value of 1000 m for the origin's and the destination's smoothing parameters. Figure 2.8 shows the response of the software to a click in the origin data window: on the the left side, we have the plot of the auto theft locations in Belo Horizonte with the selected location marked with a red cross; the radius of the purple circumference represents the origin's smoothing parameter of 1000 m. On the right side we have the corresponding conditional density surface for the selected location. This surface shows the most probable region to observe the destination location corresponding to the selected origin location. Additionally, if the user drags the mouse over the origin data window, the destination data window shows a surface changing smoothly according to the corresponding origin locations indicated by the arrow, as if the graphical device was animated. This feature allows the user to visualize how the relationship between the two point patterns changes when we focus on different regions of the map.

It is important to point out that this feature is bifunctional: using the same data input, it is possible to get not only conditional densities of destination locations given the origin, but also conditional densities of origin locations given the destination. If the user interacts with the destination data window in the same manner, the origin data window will display the corresponding conditional density surfaces. Using the same smoothing parameters of 1000 m for origin and destination point patterns, figure 2.9 illustrate the MaPPEA response to a single-click on the destination data window: on the right side, we have the plot of car retrieval locations with the selected position marked by a red cross and, on the left side, we have the conditional density of theft locations corresponding to the selected position.



Origin Point Pattern Data File: C:/Program Files/BLiPPA/auto thefts.txt

Destination Point Pattern Data File: C:/Program Files/BLiPPA/car retrievals.txt

Polygon Data File: C:/Program Files/BLiPPA/Belo Horizonte.txt

Advanced Settings

Origin Bandwidth: 1000

Destination Bandwidth: 1000

Grid Columns Number: 100

Units Conversion Ratio: 1.00

□ Legend maximum over entire region

Score Test

○ Don't Evaluate ○ 99 Permut.

○ Only normal approx. test ● 999 Permut.

Output

```
Origin-Destination Score Test

Observed Value for Score Test
Statistic: 16977
Expected Value for Score Test
Statistic: 4104.1
Barton-David-based approximation
variance: 24334.4
Barton-David-based approximation P-
Value: 0
Monte Carlo Test P-Value: 0.001
```

Figura 2.10: Test based on normal approximation and Monte Carlo test with 999 simulations for independence between auto thefts and corresponding retrieval locations in Belo Horizonte

Another tool present in the MaPPEA software is the score test for independence between the origin and the destination point patterns (Assunção and Lopes, 2007). On the same interface of the conditional density estimation tool, there is a check box where the user can choose to evaluate or not the score test; the options for the test are the normal approximation only (the fastest option) and Monte Carlo test with 99 or 999 simulations. The MaPPEA interface has an output field that presents the test statistic, the corresponding p-value and some summaries of the point patterns after the dynamic windows (origin data window and destination data window) are closed. Figure 2.4 shows the output of the score test for the Belo Horizonte data with a Monte Carlo test of 999 simulations.

2.5 Discussion

In this paper we introduce a software for the analysis of several types of marked point processes. We focused on tools for visualization of patterns in the relationship between marks and spatial locations, including features that facilitates the interaction of the analyst with software such as dynamic linked views. We also present some spatial analysis that will be included in future versions of the toolkit.

Gahegan (1999) identifies barriers to the development of visualization methods for data from continuous surfaces. One of Gahegan's barriers is the speed with which plots can be implemented within any system that has to handle large complex data sets since interactivity is deemed critical to success of visualization. A second is the complexity of the mapping from geographical data sets to the visual domain and the enormous number of views of spatial data that are possible.

Our tools have the advantage of working with events locally (for each time the user clicks on the geographical window, only the events inside the specified circle are considered in the the estimation), which facilitates a rapid response of the software to the user commands and permits the user-controlled animation impression. One of our preoccupations during the development of this toolkit was the use of graphical summaries that look intuitive to user; although surface visualizations, for example, represent a lot of information for the visual domain, analysts are usually interested in critical areas and significant differences between different areas. We believe that our software adds a useful contribution in these cases.

2.6 Bibliography

Anselin, L. (1998), "Exploratory Spatial Data Analysis in a Geocomputational Environment". In Longley, P.A., Brooks, S.M., McDonnell, R., and Macmillan, B. (Editors), *Geocomputation - A Primer*, pages 77-94. Wiley, Chichester.

Assunção, R., and Lopes, D. (2007), "Testing association between origin-destination spatial locations". In: Davis, C. A. Jr., and Monteiro, A. M. V. (Editors). *Advances in GeoInformatics*, pages 293-304, Springer Berlin Heidelberg.

Assunção R., and Maia, A. (2007), "A Note on Testing Separability in Spatial-Temporal Marked Point Processes," *Biometrics*, **63**, 290-293.

Assunção, R., Tavares, A., Correa, T., and, Kulldorf, M. (2007), "Space-time cluster identification in point processes", *The Canadian Journal of Statistics*, **9**, 9-25.

Banerjee, S., Gelfand, A.E., and Polasek, W. (2000), "Geostatistical Modelling of Spatial Interaction Data with Application to Postal Service Performance," *Journal of Statistical Planning and Inference*, **90**, 87-105.

Becker, R.A., and Cleveland, W.S. (1988), "Brushing Scatterplots". In: Cleveland, W.S., and McGill, M.E. (Editors), *Dynamic Graphics for Statistics*, pages 201-224. Wadsworth & Brooks/Cole, Belmont, CA.

Beneš, V., Bodlák, K., Møller, J., and Waagepetersen, R.P. (2005), "A case study on point process modelling in disease mapping," *Image Analysis and Stereology*, **24**, 159-168.

Buja, A., Cook, D., and Swayne, D.F. (1996) "Interactive high-dimensional data visualization," *Journal of Computational and Graphical Statistics*, **5**, 78-99.

Câmara, G., Vinhas, L., Ferreira, K, de Queiroz, G., de Souza, R., Monteiro, A., de Carvalho, M., Casanova, M., and de Freitas, U. (2008), "TerraLib: An Open Source GIS Library for Large-scale Environmental and Socio-economic Applications" (submitted). In: Hall, G. Brent; Leahy, Michael J. (Eds.), *Open Source Approaches to Spatial Data Handling.* Berlin, Springer-Verlag.

English, P., Neutra, R., Scalf, R., Sullivan, M., Waller, L., and Zhu, L. (1999), "Examining associations between childhood asthma and traffic flow using a geographic information system," *Environmental Health Perspectives*, **107**, 761-767.

Gahegan, M. (1999), "Four barriers to the development of effective exploratory visualization tools for geosciences," *International Journal of Geographical Information Sciences*, **13**, 289-309.

Gerrard, D.J. (1969), "Competition quotient: a new measure of the competition affecting individual forest trees," *Research Bulletin* **20**, Agricultural Experiment Station, Michigan State University.

Geyer, C. J. (1999), "Likelihood inference for spatial point processes," In: O. E. Barndorff-Nielsen, W. S. Kendall & M. N. M. van Lieshout (Editors) *Stochastic geometry: likelihood and computation*, 79-140. Chapman and Hall/CRC, London.

Guan, Y. (2006), "Tests for Independence between Marks and Points of a Marked Point Process," *Biometrics* **62**, 126-134.

Haining, R. (2003) *Spatial data analysis: theory and practice.* Cambridge, UK; New York: Cambridge University Press.

Hall, P., Racine, J., and Li, Q. (2004), "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, **99**, 1015-1026.

Holmes, C., and Adams, N. (2002),"A Probabilistic Nearest Neighbour Method for Statistical Pattern Recognition," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **64**, 295-306.

Hyndman, R.; Bashtannyk, D., and Grunwald, G. (1996), "Estimating and Visualizing Conditional Densities," *Journal of Computational and Graphical Statistics*, **5**, 315-336.

Kaiser, M.; Hsu, N.; Cressie, N., and Lahiri, S, (1997), "Inference for Spatial Processes Using Subsampling: A Simulation Study," *Environmetrics*, **8**, 485-502.

Knox, G. (1964), "Epidemiology of childhood leukemia in Northumberland and Durham," *Brit. J. Prevent. & Social Med.*, **18**, 17-24.

Krejčíř, P. (2000), "A maximum likelihood estimator of an inhomogeneous Poisson point process intensity using beta-splines", *Kybernetika*, **36**, 455-464.

Lahiri, S.; Kaiser, M.; Cressie, N., and Hsu, N. (1999), "Prediction of Spatial Cumulative Distribution Functions Using Subsampling" (with discussion), *Journal of the American Statistical Association*, **94**, 86-97.

Møller, J., and Waagepetersen, R. (2007), "Modern Statistics for Spatial Point Processes," *Scandinavian Journal of Statistics*, **34**, 643-684.

Ogata, Y. (1998), "Space-Time Point-Process Models for Earthquake Occurrences," *Annals of the Institute of Statistical Mathematics*, 50, 379-402.

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

Ripley, B. D., and Kelly, F. P. (1977), "Markov point processes", *J. London Math. Soc.*, **15**, 188-192.

Rosenblatt, M. (1969), "Conditional probability density and regression estimators", in: *Multivariate Analysis II*, Academic Press, New York, 25-31.

Schlather, M. (2004), "Detecting dependence between marks and locations of marked point processes," *Journal of the Royal Statistical Society* **B, 44**, 79.

Schoenberg, F. P. (2004), "Testing separability in spatial-temporal marked point processes," *Biometrics*, **60**, 471.

Stoyan, D. (2000), "On variograms in point process statistics, II: Models of markings and ecological interpretation," *Biometrical Journal*, **42**, 171.

Stoyan, D., and Penttinen, A. (2000), "Recent applications of point process methods in forestry statistics. *Statistical Science*, **15**, 61-78.

Symanzik, J. (2004), "Interactive and Dynamic Graphics". In Gentle J E, Härdle W, Mori Y (Editors). *Handbook of Computational Statistics - Concepts and Methods*, pages 293-336, Springer-Verlag New York.

TerraView home page (2008), `http://www.dpi.inpe.br/terraview`

Tschumperlé, D. (2008), *The CImg Library* : `http://cimg.sourceforge.net`. The C++ Template Image Processing Library.

Wang, M. C., and van Ryzin, J. A. (1981), "A class of smooth estimators for discrete distributions", *Biometrika*, **68**, 301-309.

Zeman, P. (1997), "Objective assessment of risk maps of tick-borne encephalitis and Lyme borreliosis based on spatial patterns of located cases," *International Journal of Epidemiology*, 26, 1121–1129.