CS-BIGS 1(2): 139-152 http://www.bentley.edu/csbigs/vol1-2/bonneu.pdf

Exploring and Modeling Fire Department Emergencies with a Spatio-Temporal Marked Point Process

Florent Bonneu
Université Toulouse I, France

We examine a database of fire department emergencies in the surroundings of the city of Toulouse during the year 2004, using methods of statistical analysis for spatial point patterns. Firemen emergencies are characterized by their positions and different features (time, duration, type,…) that one can model as a spatio-temporal marked point process. For our study, we consider the following characteristics of firemen emergencies: positions, time of occurrences and marks which take into account the duration and the number of firemen involved. We use graphical methods to explore the structure of the underlying spatial point process with a final objective of choosing a suitable model for future work. We first review the basic concepts and methods used in the paper. Considering the marginal spatio-temporal point pattern, we propose to evaluate the importance of the variation of intensity over time in comparison with spatial variation and to test the dependence between positions and time. Afterwards, we conduct an exploratory analysis of the marks to test their dependence with positions as well as their dependence with time. Our resulting framework of independence allows us to explore the dependence between categories in order to test the random labeling hypothesis. Then, under the hypothesis of random labeling and invariance in time which have been established in the first two parts, we fit a spatial point process model to the unmarked spatial point pattern aggregated over the whole year. Finally, we analyze the goodness-of-fit of our models by exploring the first and second order characteristics of simulations from the fitted models. Throughout this article, the exploratory analysis is made using mainly the R package ***spatstat****. An exposure to point processes is useful but not indispensable for understanding the exposition.*

1. Introduction

The database of firemen emergencies, provided to us by the fire department SDIS 31, contains the locations and characteristics (time, duration, number of firemen,) of emergencies which have required an intervention of firemen in the surroundings of the city of Toulouse, the largest town of the Midi-Pyrénées region in France, during the year 2004. Examples of emergencies include fires but also car accidents, assistance to injured people … After removing outliers and emergencies with missing values, we have the locations of 20,820 emergencies with 5,433 distinct points in an area of 620 km ^{2} . The important number of duplications for this spatial point pattern is caused by a positional error. The location of emergencies has not been recorded exactly but approximated by a nearby location which can be the centroid of the street for example. No information is available for this positional error even if we can think that it is closely linked with the level of urbanization.

This problem arises quite frequently in practice, for instance in econometrics or epidemiology (Benes *et al.*, 2005). For each emergency, we have in this dataset the location in the Lambert II extended coordinate system, the time of occurrence (in seconds since 1970) and the corresponding month. The mark we consider is the product of the duration of emergency by the number of firemen allocated to each emergency (number of man-hours) which thus represents a measure of total workload. Therefore, for us, an emergency with a low duration time and a high number of firemen will be considered as important as an emergency with a high duration time and a low number of firemen.

The exploratory analysis of this dataset is a preliminary step in the study of the following problem: find the optimal position of a new fire station in this area. The aim of the present paper is to analyze the point pattern in order to guide the choice of a well-founded spatio-temporal marked point process model to be used in the follow-up paper by Bonneu & Thomas-Agnan (2007).

The exploratory analysis of spatial point patterns often uses nonparametric estimates of various summary statistics based on first and second order properties of point processes. First of all, these characteristics are useful to test the hypothesis of Complete Spatial Randomness (CSR), which consists in determining whether the point pattern derives from a homogeneous Poisson process. Indeed, Poisson point processes model the absence of interaction between points. The intensity function is an important first order property which can be interpreted for homogeneous point processes as the mean number of points per unit area. Various functional summary statistics measure aggregation/clustering or regularity at distances less than different thresholds. In particular, we will introduce the L function derived from the so-called K function introduced by Ripley (1976) for stationary processes and extended to a more general class by Baddeley *et al.* (2000).

The generalization of these statistics to spatio-temporal marked point processes is theoretically feasible but is not yet implemented in software due to large dimensions which does not allow straightforward graphics. However, for spatial point processes with categorical marks (multitype point processes), there is a generalization of these statistics which allows one to judge whether the point patterns corresponding to the different categories are generated by the same point process model (Stoyan & Stoyan, 1994 and Schlather, 2001). The hypothesis of Complete Spatiotemporal Randomness (CSTR), which corresponds to a spatio-temporal point process where there is an absence of structure in time as well as in space, can be tested by generalizing the summary statistics to the temporal case (Cressie, 1993). In practice one often ignores the variation in time and the dependence between marks and positions in order to analyze the point pattern aggregated over time and to separately fit a spatial point process model for positions.

We suggest two different methods illustrated by graphics to evaluate the importance of the variation in time of the intensity in comparison with the variation in space. The first one consists in computing the intensity function for the point patterns associated with each month, for example, and comparing the graphs of their estimates. In the second one, we introduce estimates of a measure of the variation in time and variation in space and we compute the resulting ratio. This ratio enables us to understand if the temporal variation can be viewed as negligible compared to the spatial variation. For testing the dependence between positions and time, we present the results of separability tests of the marginal spatio-temporal point process introduced in Schoenberg (2004) which involve a comparison of the intensity and the product of marginal conditional intensities.

We then test the hypothesis of random labeling, i.e. whether the marks are i.i.d. and independent of the positions and time. The dependence between marks and positions can be explained by several aspects: intrinsic heterogeneity of the domain space, concurrence effects, etc. (Schlather *et al.*, 2004). We can use geo-statistical methods to test this dependence if the hypothesis that the point pattern is a realization of a stationary and isotropic spatial point process is reasonable. In our case, we have a high heterogeneity in the population density. Consequently, we discretize the marks into different categories and graphically compare the intensity estimates. We also use the method in Schoenberg (2004) for testing the dependence between marks and positions, and settle with the same separability tests the matter of the dependence between marks and time. Our framework of independence between marks and positions, and also between marks and time, allows us to test the dependence between the workload categories by computing a function denoted by L_{cross}. The absence of correlation between the workload categories suggests the random labeling of the marks. This leads us to the search for an adequate model for the marginal distribution of the marks.

The results thus obtained suggest that it is reasonable to aggregate the spatial point pattern over time. But, because of the high number of emergencies and duplicated locations, the modelling of the whole point pattern is very difficult. Consequently, we choose to analyze the emergencies of a particular month, for example, June. In this analysis, the major difficulty in choosing a suitable model consists primarily in adjusting the intensity function as well as possible. We present three methods of estimation of the intensity: parametric, nonparametric and semi parametric. For each different estimate of the background intensity, we test the absence of interaction for this point pattern by plotting an estimate of the L summary statistic and the pointwise envelope from simulations of an inhomogeneous Poisson process. Finally, we choose a fitted model which presents approximately the same first and second order properties as those of the spatial point pattern.

In this paper, we mainly use the R package **spatstat** for analyzing the spatial point process (Baddeley & Turner, 2005 and Baddeley & Turner, 2006).

2. Background on summary statistics

By definition, a spatial point process X is a random countable subset of a space S. As in our example, we focus on point processes X whose realizations are finite subsets of a compact set $W \subset S$. A spatio-temporal marked point process $Y = \{(\mathbf{x}, m_\mathbf{x}, t_\mathbf{x}) : \mathbf{x} \in X\}$ with points $\mathbf{x} \in S$, marks $m_\mathbf{x} \in M$ and times $t_\mathbf{x} \in T$ is defined to be a spatial point process on the product space $S \times M \times T$. In the sequel, the definitions are given for a spatial point process X in $W \subset \mathbb{R}^2$ but can be generalized to higher dimensions. For convenience, we number the points of a realization $\mathbf{x} = \{x_1, \cdots, x_n\}$ even if we must keep in mind that a point pattern is unordered. For a spatial point pattern with duplicated points, we often plot the distinct points with their number of duplications. However, due to the high proportion of duplicated points in our case, we choose to plot perturbed locations for a better readability. For the perturbation of locations, we use Gaussian noise with zero mean and standard deviation equal to 50 in each coordinate. Our choice for the standard deviation follows from the empirical distribution study of the inter-events distances. Figure 1 (Left) plots the perturbed locations of 2007 emergencies in June, suggesting a high inhomogeneity in the distribution of emergencies due to the density of population. In the sequel, we always consider the perturbed locations of emergencies obtained from the same Gaussian noise. This choice is justified later by the difficulty in using methods based on the K function for a point pattern with a high number of duplicated points (envelope, minimum contrast estimation…).

2.1. Estimation of the intensity λ

The process first-order characteristic is its intensity function λ defined as

$$\lambda(s) = \lim_{d\delta \to 0} \frac{\mathbb{E}[N(d\delta)]}{d\delta}$$

where $d\delta$ is the elementary area around s and $N(d\delta)$ the number of events in this area.

If λ is constant, then X is said to be homogeneous with intensity λ, otherwise it is inhomogeneous. The estimation of λ is the first step in any exploratory analysis of a point pattern and aims to evaluate the homogeneity of the process. In our case, it is inappropriate to assume homogeneity because of the spatial correlation of emergencies with the human settlement pattern which is not stationary. Due to the presence of inhomogeneity, we use a kernel method to estimate the intensity function. Our absence of information about the positional error of emergencies does not allow us to use the new kernel estimators introduced in Cucala (2006). Consequently, we choose to estimate the intensity function on the locations perturbed by the Gaussian noise defined before. The chosen estimate is presented in its anisotropic form with a border effects correction (Diggle, 1985) :

$$\hat{\lambda}(s) = \frac{\sum_{i=1}^{n} K_H(s - x_i)}{\hat{c}_{W,H}(s)}$$

where K_H is the kernel with covariance matrix H defined by $K_H(s) = |H|^{-1} k_2(H^{-\frac{1}{2}} s)$, k_2 is the density function of a standard bi-dimensional Gaussian variable and $\hat{c}_{W,H}(s)$ is an estimate of the edge correction factor $c_{W,H}(s) = \int_W K_H(s-u)du$. In its isotropic form K_h is the kernel with standard deviation h defined by $K_h(s) = h^{-2} k(h^{-1} s)$.

The choice of a good bandwidth h is difficult in practice, notably with very wide variations in λ as mentioned in Diggle *et al.* (2006). Indeed, the method proposed in Berman *et al.* (1989) of minimizing an estimation of the mean square error of $\hat{\lambda}$ produces in our case a value of h close to zero. For selecting an optimal diagonal matrix H we can use a plug-in method implemented in the R package **ks** with binned pilot estimation (Wand & Jones, 1994). The diagonal terms of the bandwidth matrix obtained are sufficiently small to capture changes in population density between urban and rural environments and to avoid problems of under-smoothing. For the point pattern of emergencies in June, the smoothing parameter is between 700 and 900 meters for both coordinates. We subsequently use the isotropic form with bandwidth $h = 800$. Figure 1 (Right) represents the logarithm transformation of the intensity estimate of emergencies in June. This transformation achieves an enhancement of variations of intensity around cities smaller than Toulouse.

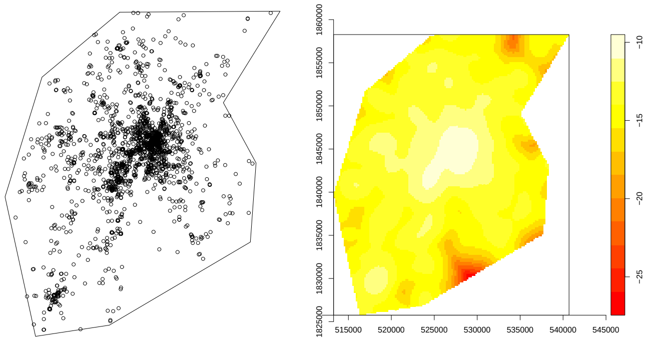

Figure 1 : Left : Perturbed locations of the 2007 emergencies in June. Right : Logarithm transformation of the estimated intensity of emergencies in June.

2.2. Estimation of the K-function

To study the spatial dependence over a wide range of scales, we can use summary statistics based on a number of known second order properties. Here, we only consider the L function derived from the K function introduced by Ripley (1976) for stationary processes and extended to the class of second order intensity-reweighted stationary processes by Baddeley *et al.* (2000). The theoretical K function for a stationary spatial point process is the expectation of the number of extra events within distance $r \geq 0$ of a randomly chosen event, divided by the intensity λ . The L function is then defined by $L(r) = \sqrt{K(r)/\pi}$ for all $r \geq 0$. At least for small values of r , $L(r)-r > 0$ indicates aggregation/clustering at distances less than r , and $L(r)-r < 0$ indicates regularity. More precisely, because K is a cumulative function, a significant peak of L above 0 shows the maximum range of aggregation and should be interpreted with care beyond this point. For second order intensity-reweighted stationary point processes X , we use the following estimate of the inhomogeneous K function introduced in Baddeley *et al.* (2000) :

$$\hat{K}_{inhom}(r)=\frac{1}{|W|}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\hat{w}_{x_i,x_j,r}1(\| x_i-x_j \|\leq r)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)},$$

$r \geq 0$

where $\hat{w}_{x_i,x_j,r}$ is a boundary correction factor and $|W|$ the area of W. The more common boundary correction factor is the translation correction factor $w_{x_i,x_j,r} = |W \cap W_{x_i-x_j}|^{-1}$, where $W_{x_i-x_j} = \{\xi + x_i - x_j : \xi \in W\}$, but it is computationally expensive for large point patterns. So, in our case, we prefer the border correction factor implemented in **spatstat**,

$$\hat{w}_{x_i,x_j,r} = \frac{1(d(x_i,\partial W) > r)}{\sum_{k=1}^{n}\left(1(d(x_k,\partial W) > r)/\hat{\lambda}(x_k)\right)}$$

where ∂W is the boundary of the observation window.

Afterwards, in order to study the correlation structure in multivariate point processes such as $X = (Y, Z)$, cross summary statistics $K_{inhom}^{cross}(Y,Z)$ can be introduced for the non-stationary case in the same way as was done with the K function. The definition of $K_{inhom}^{cross}(Y, Z)$ concerns cross-second-order-intensity-reweighted stationary processes; an estimate is given by

$$\hat{K}_{inhom}^{cross}(Y,Z)(r)=\frac{1}{|W|}\sum_{i=1}^{n_Y}\sum_{j=1}^{n_Z}\frac{\hat{w}_{y_j,z_j,r}1(\| y_i-z_j \|\leq r)}{\hat{\lambda}_Y(y_i)\hat{\lambda}_Z(z_j)},$$

$r \geq 0$

(1)

The L_{inhom}^{cross} function is the extension to the multivariate case of the L_{inhom} function for the univariate case. For convenience and where there is no possible confusion, we use the notations K , L and L_{cross} respectively for K_{inhom} , L_{inhom} and L_{inhom}^{cross} .

2.3. Envelope

In general, let us consider a statistic $L(r)$ and a given hypothesis H_0. Typically, the null hypothesis can be the absence of interaction, the random labeling hypothesis or the goodness-of-fit of a given model. Critical intervals are necessary to judge the deviance from the null hypothesis of a nonparametric estimate of a summary statistic. Let $\hat{L}(r)$ be the estimate computed from the observed point process X in W, and $\hat{L}_1(r),\cdots,\hat{L}_m(r)$ those obtained from i.i.d. simulations $X_1,\cdots,X_m$ under H_0. For each value of r, we can estimate any quantile of the distribution of $\hat{L}(r)$ under H_0 from the empirical distribution of $\hat{L}_1(r),\cdots,\hat{L}_m(r)$, if m is large enough. The quantiles $L_l(r)$ and $L_u(r)$ used to construct the critical interval are called respectively the lower and the upper envelope. We obtain a pointwise envelope because we have a critical interval for each value of r. Throughout this paper, the envelope is computed from 39 simulations with pointwise minima and maxima in order to have for each r a 5% probability that the estimate of $L(r)$ falls outside the interval.

3. Time variation and spatial variation

In this section, we focus on comparing the relative importance of time variation and spatial variation in the intensity of the spatio-temporal point process. In practical situations, unless the importance of time is well-known and predominant (earthquakes…), the dependence on time is often ignored by aggregating the spatial point process over time. Even so, there does exist some literature on spatio-temporal point processes models (Diggle (2006)), so we would like to make sure that this aggregation is well justified in our case. Because it will be impractical to simultaneously model space, time and marks, we choose to ignore the possible dependence between marks and time. We thus perform this investigation by aggregating the marked process into a single unmarked one. Diggle *et al.* (2005) propose a Monte Carlo test to investigate temporal changes. We propose to decompose the spatio-temporal process into 12 monthly realizations. A first approach is to compare the logarithm transformation of estimates of the intensities for each month (Figure 2). At first sight, the estimates of intensities do not show important modifications in shape and seem to present a temporal trend with a low rate of change. Indeed, we only note a slight trend in the total number of emergencies through the year.

A second approach consists in computing the ratio between an estimate of the time variation and an estimate of the spatial variation of intensity. To measure the time variation at a location s, we introduce

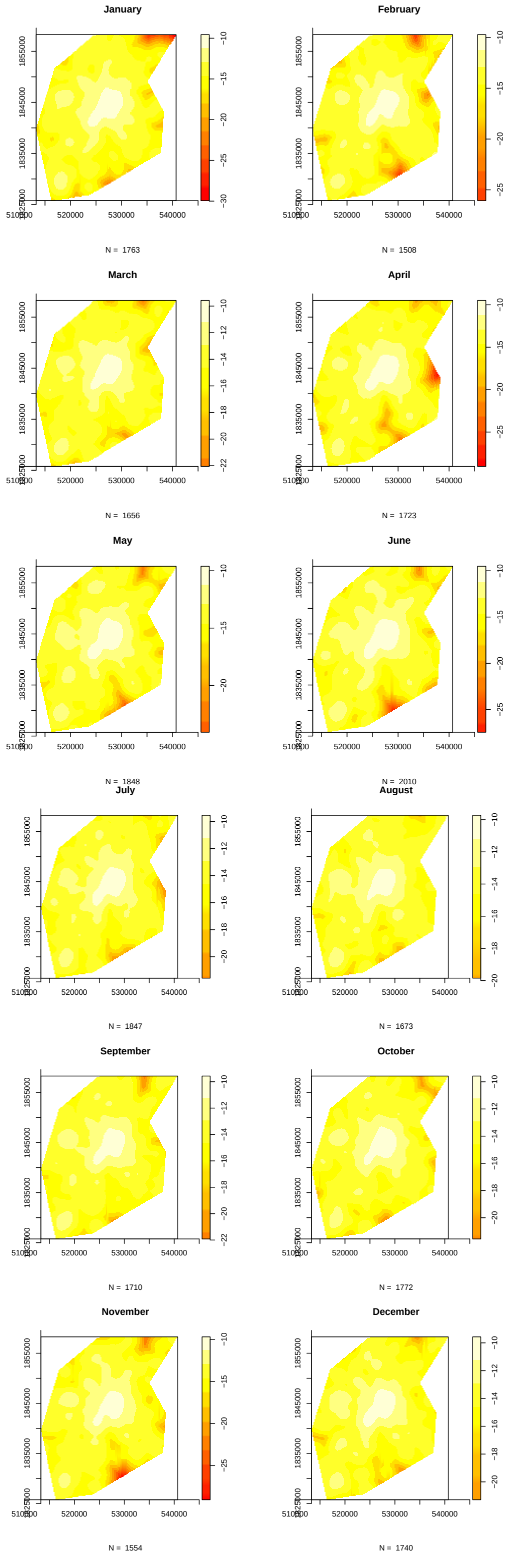

Figure 2 : Logarithmic transformation of intensity by month.

$$TMSE(s) = \frac{1}{12}\sum_{k=1}^{12}(\hat{\lambda}_k(s) - \bar{\lambda}(s))^2$$

where $\hat{\lambda}_k$ is the intensity estimate of the month k and $\bar{\lambda} = \frac{1}{12}\sum_{k=1}^{12}\hat{\lambda}_k$ is the mean intensity. This measure is computed on a regular grid of m points $\mathbf{s} = \{s_1, \cdots, s_m\}$ in the domain space. We denote by $TMSE$ the pixel image giving the value at each point of the grid.
To measure the spatial variation, we introduce

$$SMSE = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{\lambda}(s_i) - \frac{n}{|W|}\right)^2$$

The image ratio *TMSE*/*SMSE* indicates that the time variation is negligible in comparison with the spatial variation. Indeed, the time variation represents 0.5 percent of the spatial variation at most. Figure 3 shows the logarithm of this ratio which reflects well the high spatial inhomogeneity in this domain.

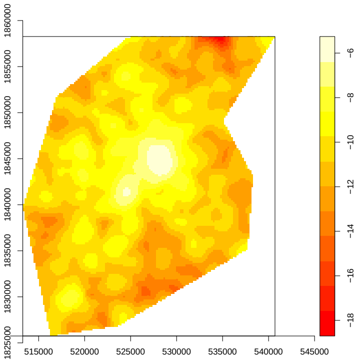

Figure 3 : Logarithm of the image ratio between the time variation and the spatial variation of intensity.

Finally, we investigate the separability of the intensity function of the spatio-temporal point process (X, D) as in Schoenberg (2004), i.e. we test whether we have

$$\lambda_{X,D}(s,t) = \lambda_X(s) f_D(t), \quad s \in W \text{ and } t \in T.$$

where $\lambda_{X,D}$ and λ_X are respectively the intensities functions of (X, D) and X , and f_D the density function of D. We denote by $\hat{\lambda}$ the estimate of $\lambda_{X,D}$ and by $\tilde{\lambda}$ that of $\lambda_X f_D$. We want to judge the difference between the two. The fact that $s \in W \subset \mathbb{R}^2$ and $t \in T \subset \mathbb{R}$ do not allow us to present straigthforward graphics as in section 4.2 where we discuss dependence between marks and time. To compare the two estimates, we compute four statistics defined in Schoenberg's article on a regular grid of m points $\mathbf{(s,t)} = \{(s_i, t_j) \in W \times T : i = 1, \cdots, m_\mathbf{s}; j = 1, \cdots, m_\mathbf{t}\}$.

$$S_1 = \sup_{i,j}\{|\hat{\lambda}(s_i,t_j) - \tilde{\lambda}(s_i,t_j)| / \sqrt{\tilde{\lambda}(s_i,t_j)}; (s_i,t_j) \in \mathbf{(s,t)}\}$$

$$S_2 = \inf_{i,j}\{|\hat{\lambda}(s_i,t_j) - \tilde{\lambda}(s_i,t_j)| / \sqrt{\tilde{\lambda}(s_i,t_j)}; (s_i,t_j) \in \mathbf{(s,t)}\}$$

$$S_5 = \frac{1}{m}\sum_{(s_i,t_j)\in\mathbf{(s,t)}}(\hat{\lambda}(s_i,t_j) - \tilde{\lambda}(s_i,t_j))^2$$

$$S_6 = \sup_{i,j}\{(\hat{\lambda}(s_i,t_j) - \tilde{\lambda}(s_i,t_j))^2; (s_i,t_j) \in \mathbf{(s,t)}\}$$

Abnormally large value of these test statistics indicate a departure from the separability hypothesis. The intensities and the probability density are computed with the **kde** function in the R package **ks**, initially programmed for density estimation. This function allows computing of the density/intensity for three dimensional point patterns. So, in order to obtain intensity estimates, we multiply the result by the number of points. These estimates are not adjusted by a correction factor for border effects. To judge the significance of these statistics, we construct one-sided Monte-Carlo tests from 19 simulations of a Poisson point process under the null hypothesis of separability. If the statistic test S is lower than the maximum value obtained from the simulations then we accept the separability assumption at level 5%. For computational reasons, we limit our study to a subsample of 2000 emergencies randomly chosen. Here, the four tests conclude to the separability assumption. Note that this Monte-Carlo inference is based on simulations from the Poisson model, an assumption that we will discuss later in the paper.

The different approaches show that the variation in time is negligible in comparison with the variation in space and that there is independence between positions and time. We therefore consider that we can aggregate the point pattern over time without losing important information.

4. Analysis of the workload mark

4.1. Dependence between marks and positions

Marks and positions are often assumed to be independent but this may not hold in practice. For instance, in forestry, the diameters of trees can be dependent on the nature of the soil and of the presence of others trees nearby. In the case of firemen emergencies, it is possible that the frequency of large workloads emergencies is higher in some areas. Another type of dependence arises from the fact that an occurrence may have an influence on future emergencies around it. The first type of dependence seems more likely here.

Summary statistics for marked point processes are introduced in Stoyan & Stoyan (1994) and Schlather (2001) to test the dependence between continuous marks and positions. However, these statistics are just defined for stationary and isotropic marked point processes. In these articles, the marks process is modeled as a random field and the authors can apply geostatistical methods. In Schlather (2004), the test of dependence is valid for any random field model where the marks are given by a strictly monotone transformation of a Gaussian random field. This last assumption on the marks is not necessary for the test based on the conditional expectation of marks developed in Guan (2006). Guan's method allows the treatment of examples with a bimodal distribution of marks. However, to our knowledge, tests of dependence between continuous marks and positions in the case of inhomogeneous point processes are not available.

We next use two empirical approaches to test the validity of the independence. We first use the same method developed for testing the temporal trend. We discretize the logarithm of workloads, to mitigate the influence of outliers, into three categories: Low, Medium, and High. This discretization is performed by applying the k-means method minimizing the within category variance. Figure 4 represents the logarithm transformation of the estimated intensity for the different categories of the multitype point patterns. The patterns of estimates are close together across categories but different in total mass. This suggests that the point patterns could be generated by the same point process model with a different expectation of the number of points.

To confirm the conclusion of independence between marks and positions given by the previous approach we now investigate Schoenberg's method. The statistic tests based on S_1 and S_2 accept the separability assumption whereas those based on S_5 and S_6 reject this hypothesis. This situation is not clear-cut and allows us to consider one of the two cases. However, taking into account this dependence implies looking for a more complicated model. One can find some reasons to believe in dependence between marks and positions for some categories of emergencies (fires, car accidents …) but we think that this dependence is not very important when considering all types of emergencies simultaneously.

Finally, taking into account the different methods used, the hypothesis of independence between the occurrences of emergencies and the workload marks is not as clearly established as in the case of time and location. Nevertheless, we maintain this hypothesis of independence in order to avoid an intractable model.

4.2. Dependence between marks and time

We investigate the dependence between marks and time through the separability method. In this case, the bidimensional framework allows to present straightforward plots of the estimates of $\hat{\lambda}$ and $\tilde{\lambda}$ for the marginal marked temporal process (Figure 5).

The graph supports the separability assumption. This conclusion is emphasized by the Monte-Carlo separability tests which do not reject the separability assumption.

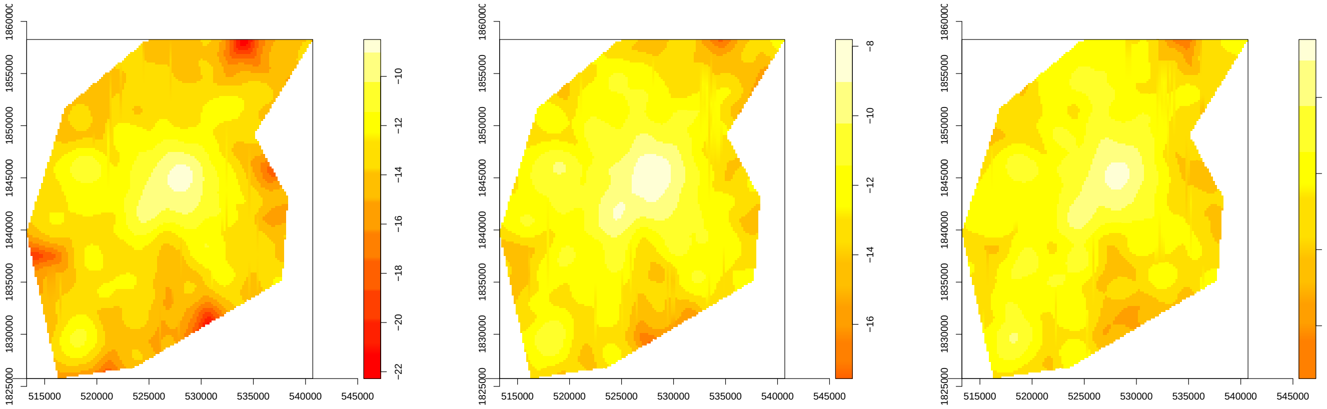

Figure 4 : Logarithm transformation of estimated intensity by category. Up : Small workloads. Middle : Medium workloads. Down : High workloads.

4.3. Dependence between mark categories

Next, we test the dependence between the emergencies of different categories of marks by estimating the functions L_{cross} for all pairs of categories. These functions measure the dependence between points of types i and j at distances $r \geq 0$. The calculation of L_{cross} requires a great deal of memory, so we restrict our study to the June emergencies. We estimate the overall intensity of all the emergencies by a semiparametric method using a model with a single covariate (population) as is done in section 5. As in Moller & Waagepetersen (2004), the estimated intensity for each category is chosen to be proportional to the overall intensity estimate in order to have an expectation of the number of points equal to the number of emergencies in each category. The 39 simulations for the envelope calculation are obtained by taking the same positions of the multitype point pattern but with a random permutation of categories. Figure 6 presents the estimates and envelopes of L_{cross} corresponding to the three pairs of categories: (Low,Medium), (Low,High) and (Medium,High). For the three pairs of categories the estimated $L_{cross}(r)-r$ lies within the envelope even though it appears to track the upper envelope boundary and sometimes exceed it in a few instances. So, we can consider that the emergencies of different categories of marks are independent. The independence is not a surprising hypothesis in this practical example and is verified under the assumption of independence between marks and locations (proportional intensities).

4.4. Marginal distribution

The previous sections have concluded that we can model the marginal point pattern of positions and marks separately in order to avoid a more complicated model. This is the reason why now we analyze the marginal distribution of marks. Figure 7 (Left) presents the histogram of the logarithm of workloads. At first sight, one may think that a log-normal model is acceptable considering the empirical marginal distribution of marks. However, the Normal Q-Q plot of the logarithm of marks in Figure 7 (Right) shows that it is not a reasonable choice. The kurtosis value is far away from the kurtosis value of the adjusted normal model. Many others transformations with the aim to obtain a normal distribution as well as different models were attempted but none of these were satisfactory. The transformations considered include the Box-Cox transformation with an optimal parameter chosen by *boxcox.fit* (package **geoR**), inverse transformations, etc. Moreover, we have tried in vain to fit several models from the logarithm of marks (Cauchy,

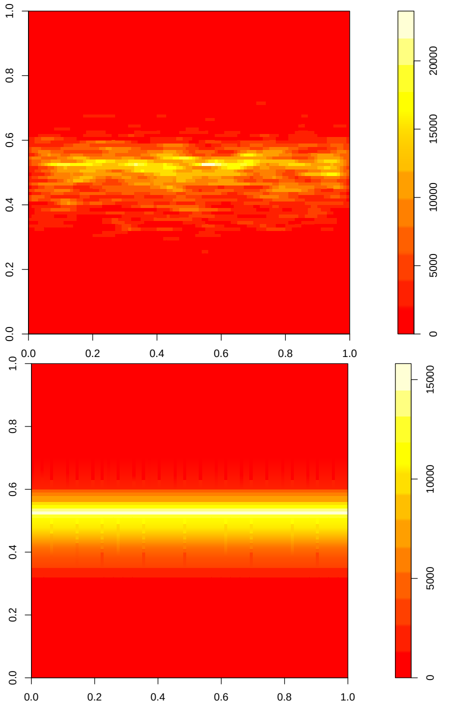

Figure 5 : Up : Intensity estimate $\hat{\lambda}$. Down : Intensity estimate $\tilde{\lambda}$.

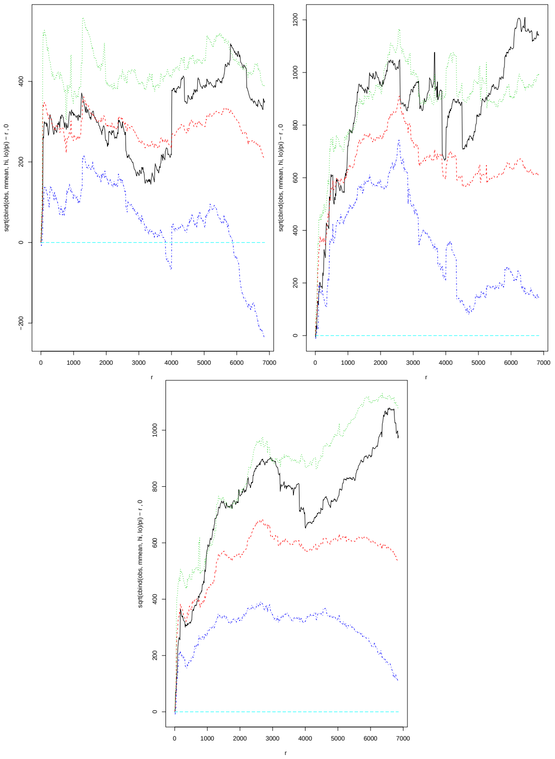

Figure 6 : Estimated $L_{cross}(r)-r$ for the three pairs of categories on emergencies in June (solid line), average and envelope from 39 multitype point patterns with same locations but categories given by a random permutation (dashed lines).

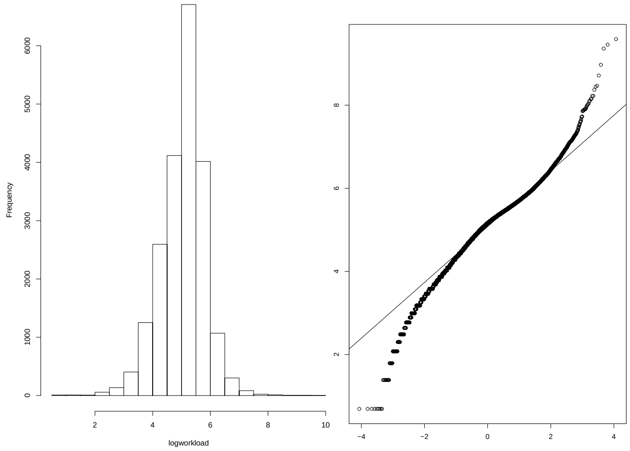

Figure 7 : Left : Histogram of the logarithm of workloads. Right : Normal Q-Q plot of the logarithm of workloads (xaxis : Theoretical Quantiles, yaxis : Sample Quantiles).

Gaussian Mixture …) or directly from the marks (Exponential, Generalized Pareto, Generalized Extreme Values …). This difficulty in obtaining a satisfactory model for the workload marks suggests that we should consider a bootstrap procedure for generating “simulated” samples.

5. Model

On the basis of the previous analysis, we decide to consider in this section the marginal spatial point pattern of positions aggregated over the year for fitting a spatial point process model ignoring the marks. But, due to the high number of locations, we take a subsample of this point pattern corresponding to emergencies of a particular month, for example, June. For testing the absence of interaction, we choose to apply a Monte-Carlo test by computing simulated envelopes of the inhomogeneous L function under an inhomogeneous Poisson process model. The first step in order to estimate the L function and to simulate realizations of an inhomogeneous Poisson process is to estimate the intensity function. We investigate three methods for estimating the intensity: parametric, nonparametric and semi parametric with one covariate.

5.1. Parametric and Nonparametric estimation

The parametric method consists in estimating the logarithm of the intensity with a polynomial in the coordinates. We estimate the polynomial coefficients by the method of maximum pseudo-likelihood (Moller & Waagepetersen (2004)). However, parametric models with a reasonable degree (< 5) are often unsatisfactory in the presence of high inhomogeneity of the locations of points in the domain. The resulting intensity is a rough estimate and the coefficients are difficult to compute for higher degrees.

An alternative is to use nonparametric methods that are more adaptable. A major problem is always to separate inhomogeneity explained by the intensity λ and interactions measured by the L function. Figure 8 (Left-Middle1) shows that our choice of $h = 800$ yields an intensity estimate and a simulated point pattern close to the point pattern of emergencies in June. So, from the point of view of the first order characteristic, an inhomogeneous Poisson point process seems to be an appropriate model. The choice of the bandwidth for the kernel estimation is of primary importance. As in Diggle (2003), the estimated $L(r) - r$ in Figure 8 (Middle2) shows that the selected bandwidth is too small and involves an over-fitting problem. Figure 8 (Right) also displays the estimated $L(r) - r$ and envelope when we use the leave-one-out estimate $\overline{\lambda}$ of the intensity function introduced in Baddeley *et al.* (2000) to correct the bias in the estimate of $L(r) - r$. Its formula is given by

$$\overline{\lambda}(s) = \frac{1}{\hat{c}_{W,h}(s)} \sum_{i=1}^{n} K_h(s - x_i) 1_{(x_i \neq s)}$$

If $\hat{\lambda}$ and $\overline{\lambda}$ are approximated by their values evaluated at a fixed grid of points, the two estimators of the intensity surface agree with probability 1. The difference with the usual estimator consists in not taking into account in the summation the point of the pattern at which we estimate the intensity. The use of this estimate in the estimation of K gives a better bias in the simulation example of Baddeley *et al.* than the classical one. The bandwidth $h = 800$ with the leave-one-out estimator gives here an envelope which is difficult to interpret due to the surprising form of the mean curve under the null hypothesis of a Poisson process (Figure 8 (Below)).

Moreover we think that the use of the same data to estimate non-parametrically both λ and L is problematic. Indeed, we obtain better results by using the emergencies in May for the estimation of λ , which is then used in the estimate of the L function for the point pattern in June. Figure 9 shows that this method allows to fit a Poisson process model with a similar first order characteristic, a good simulated process and a better graph for $L(r) - r$. Finally, the envelope of $L(r) - r$ implies that we conclude to neither aggregation nor repulsion in the point pattern; neither do we conclude to an over-fitted model. However, in this case, the envelope is highly dependent on the choice of the subsample for the estimation of λ . For instance, the choice of the month of

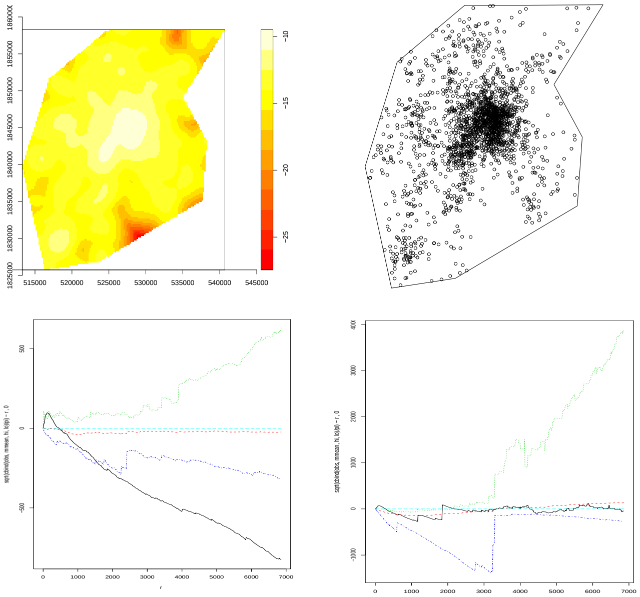

Figure 8 : Top left: Nonparametric density estimation with bandwidth *h*=800 (June emergencies). Top right: A simulation from a inhomogeneous Poisson process model. Bottom left: Estimated *L(r)-r* on emergencies in June (solid line), average and envelope from 39 simulations of a inhomogeneous Poisson process (dashed lines) with $\hat{\lambda}$. Bottom right: As previously with the leave-one-out estimate $\bar{\lambda}$ of the intensity.

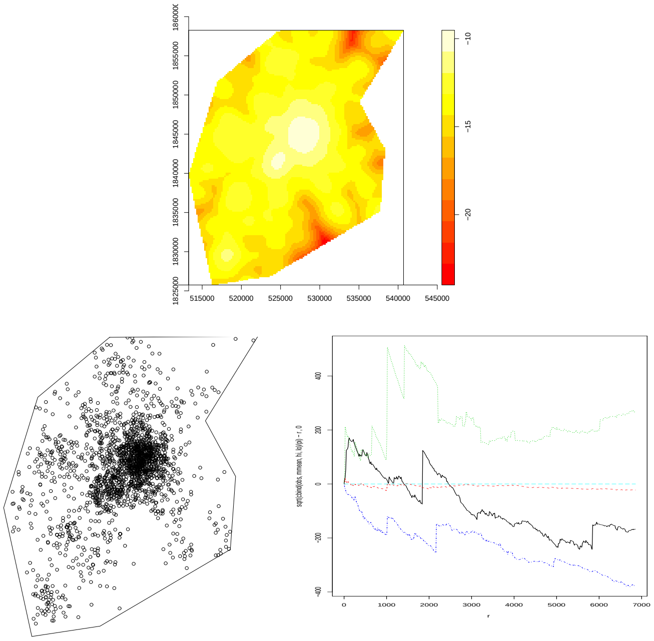

Figure 9 Top: Nonparametric intensity estimation with bandwidth *h*=800(May emergencies). Left: A simulation from a inhomogeneous Poisson process model. Down : Estimated *L(r) -r* on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Poisson process (dashed lines).

July for the estimation of λ would involve an "artifact" on the envelope estimate. This is a reason why we did not pursue this direction further.

5.2. Semi parametric estimation with one covariate

In many cases, the intensity of the spatial point pattern depends on covariates. For instance, our spatial point pattern is influenced by environmental and economic covariates: population, presence of woods … In our study, we have a population covariate which allows us to estimate the intensity of emergencies from an estimate of the population density. Our population covariate is the number of inhabitants in 296 INSEE[1] administrative units named IRIS[2]. We know the total population and the centroid of each IRIS. Figure 10 represents these units with a circle centered at those centroids with radius proportional to the number of inhabitants. We denote by $\xi_1,\cdots,\xi_{296}$ the centroids of the administrative units and by $N_1,\cdots,N_{296}$ their number of inhabitants. It is necessary to know the values of this covariate at every point in the window in order to estimate the background intensity. Consequently, we predict the covariate on a regular grid with a non-parametric predictor and then estimate the coefficients α and β in the expression $\lambda(s)=\exp(\alpha+\beta\log(\hat{C}(s)))$ by maximum pseudo-likelihood, where $\hat{C}(s)$ is the estimate of the covariate.

5.2.1. Two density estimates

We present two nonparametric methods to estimate the population density. The first one $\hat{C}_1(s)$ is a classical nonparametric kernel method with a selected global bandwidth of $h=900$ and a border correction factor. The second kernel method uses an adaptive choice of bandwidth based on k -nearest neighbors. At each point of a regular grid, we estimate the population density by applying an Epanechnikov kernel k_e with its support adapted in order to take into account only k centroids. We arbitrarily choose $k=5$. The expression for this estimator is

[1] Institut National de la Statistique et des Etudes Economiques; National Institute for Statistics and Economic Studies

[2] Ilôts Regroupés pour l'Information Statistique ; areas formed for purposes of statistical information

$$\hat{C}_2(s) = \frac{1}{N_{Pop} h_s} \sum_{i=1}^{296} N_i k_e\left(\frac{s-\xi_i}{h_s}\right)$$

where $N_{Pop} = \sum_{i=1}^{296} N_i$, $k_e(s) = \frac{3}{4}(1-\|s\|^2)$ and $h_s = \|s-\xi\|_{(5)}$ is the fifth order statistic of distances between s and the IRIS centroids. We do not use here any correction factor for border effects.

Figure 10 displays the logarithm transformation of these two density estimates. We note that the second approach has the advantage of clearly identifying the biggest cities in this region. Unfortunately, the intensity is under-estimated near the boundary of the observation region.

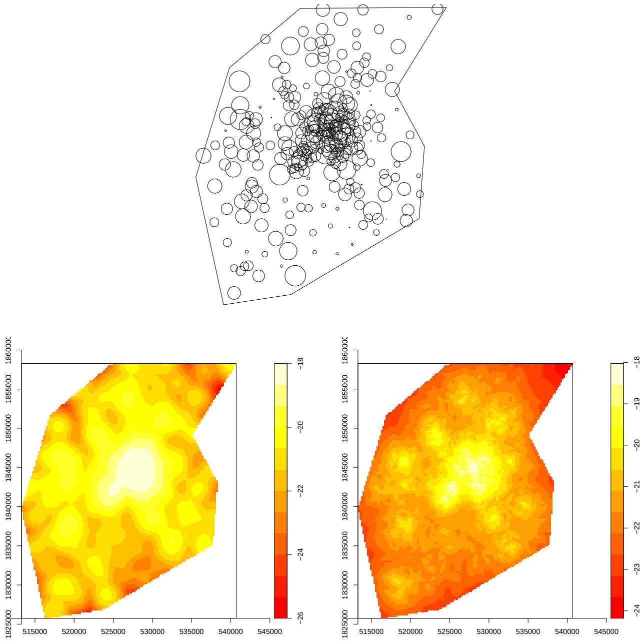

Figure 10 : Top: Proportional symbol map of the number of inhabitants per IRIS. Borrom row: Logarithmic transformation of the population density estimated by a nonparametric kernel method with a global bandwidth (Left) and a local bandwidth obtained by *k*-nearest neighbors (Right).

5.2.2. Models based on density estimate $\hat{C}_1$

First of all, we focus on models constructed from the density estimate $\hat{C}_1$ obtained by the first method. By maximum pseudo-likelihood, we obtain $\hat{\alpha}$ and $\hat{\beta}$ and write $\hat{\lambda}_1(s) = \exp(\hat{\alpha} + \hat{\beta}\log(\hat{C}_1(s)))$ for all s in the regular grid. Figure 11 shows a simulation of a Poisson point process with intensity $\hat{\lambda}_1$ and an envelope which lead us to reject the hypothesis of no interaction, and suggests aggregation for $r \leq 1500$.

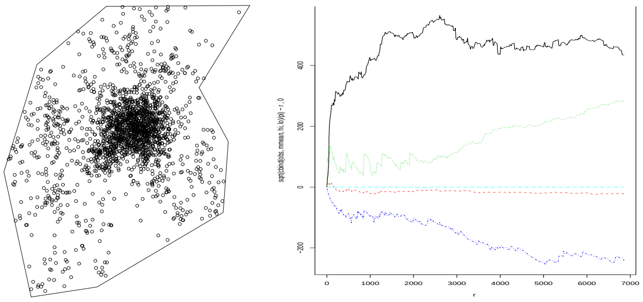

Figure 11 : Up : A simulation from a inhomogeneous Poisson process model with intensity $\hat{\lambda}_1$. Down : Estimated $L(r) - r$ on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Poisson process (dashed lines).

We tested three inhomogeneous point processes models of clustering : the Matern cluster process and the Thomas cluster process which belong to the class of Neyman-Scott processes and the Log Gaussian Cox Process. Neyman-Scott processes and Log Gaussian Cox processes are cluster processes in the class of Cox processes. A Cox process is obtained by considering the intensity function of the Poisson process as a realisation of a random field. Neyman-Scott processes are obtained by clustering points around a homogeneous Poisson point process with intensity κ ("mother" process). A realization of a Neyman-Scott process is obtained by all the realizations of independent Poisson processes at each "mother" point. This daughter point process has an intensity function which depends on a kernel function. The two point process models considered here are given by a specific kernel (Moller & Waagepetersen (2004)). For a Log Gaussian Cox process, the intensity function is the exponential transformation of a Gaussian field (Moller *et al.* (1998)).

The inhomogeneity can be incorporated by different methods (Jonsdottir (2004)). But, for the class of Neyman-Scott processes, it is necessary to incorporate this inhomogeneity by thinning as in Waagepetersen (2006). Indeed, this method is the only one that allows to get an inhomogeneous Neyman-Scott process which is second-order intensity reweighted and for which the inhomogeneous K function is well defined. Thinning is an easy method by which to simulate inhomogeneous point processes : we simulate a realization of a stationary point process X and afterwards apply an independent thinning method by the field defined from $\hat{\lambda}_1$ to obtain a realization of an inhomogeneous point process Y. The advantage is also that the inhomogeneous K function of Y coincides with the K function of X. This fact allows us to estimate the parameters κ and ω of the point process model by minimizing the contrast

$$\int_0^a (\hat{K}_{inhom}(r)^q - K(r;\kappa,\omega)^q)^2 \, dt$$

where $K(r;\kappa,\omega)$ is known for the class of point process models presented before. For the choice of a and q, Diggle (2003) recommends to choose a considerably smaller than the dimension of the observation plot and $q = 1/4$. We take $a = 4000$ meters.

The thinning method modifies the structure of the point process model. For example, Figure 12 (Left) displays a simulation of the fitted inhomogeneous Thomas cluster process which shows that the expected number of points per cluster is different. If we incorporate the inhomogeneity by considering an inhomogeneous Poisson point process for the "mother" points, then the usual structure is maintained in comparison with the stationary case. Indeed, the expected number of points per cluster is constant in this case. Figure 12 presents a simulation of the fitted Thomas point process which looks quite different from the emergencies in June. The minimum contrast estimation yields an expected number of "mother" points and a scale parameter which are too small. By construction, the second order characteristic of this fitted point process model is close to that of the point pattern (Figure 12, Right).

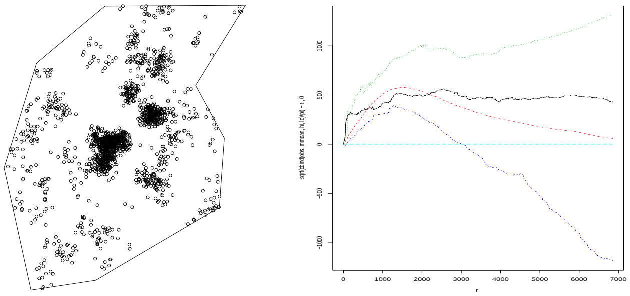

Figure 12 : Up: A simulation from an inhomogeneous Thomas point process model obtained by thinning. Down : Estimated *L(r)-r* on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Thomas point process model obtained by thinning (dashed lines).

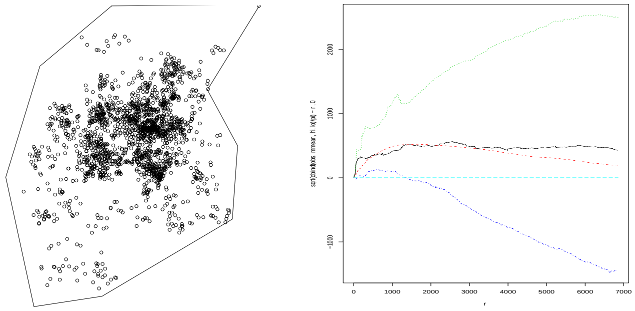

Figure 13 : Left: A simulation from an inhomogeneous LGCP model obtained by thinning. Right: Estimated *L(r)-r* on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous LGCP model obtained by thinning (dashed lines).

The simulation of the fitted LGCP in Figure 13 (Left) features aggregation areas as in our point pattern. However, these areas are wider and mainly concentrated around the city of Toulouse. Moreover, several areas exhibit no points or very few points in the simulation whereas they are important areas of emergencies in the point pattern. This is the case of the area in the bottom-left of the region which corresponds to the large city of Muret. We also note that the boundary of the region has generally few points in the simulation. The envelope of $L(r)-r$ is large due to the fact that the variability of the expected number of points in the simulations is relatively important. The envelope suggests that the goodness-of-fit of this model is satisfactory.

We now generalize to the case of Log Gaussian Cox processes (LGCP) the method proposed in Waagepetersen (2006) for Neyman-Scott processes.

5.2.3. Models based on density estimate $\hat{C}_2$

We consider the case where the background intensity estimate $\hat{\lambda}_2$ is derived from the density estimate $\hat{C}_2$. Compared to the simulation of a Poisson point process with estimated intensity $\hat{\lambda}_1$, the simulation in Figure 14 (Left) now features more aggregated areas and reveals the area of the city of Muret in the bottom-left of the region. From the first order characteristic point of view, this point process is a good model of the emergencies. The Monte-Carlo test of no interaction in Figure 14 concludes that our point pattern is more aggregated for approximately $r \le 1500$ and more regular for large r than the Poisson model. We reject the hypothesis of no interaction and fit a LGCP model next.

The parameters of the LGCP are estimated as previously by the minimum contrast method and the inhomogeneity is incorporated by thinning. The obtained simulation shows that this point process model is interesting because the distribution of the aggregated areas is close to those of our point pattern (Figure 15 (Left)). There is no void large

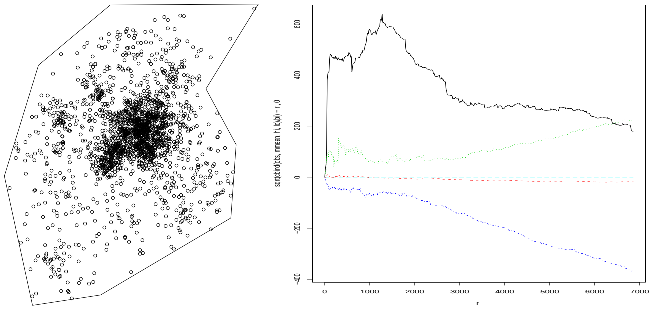

Figure 14 : Left: A simulation from an inhomogeneous Poisson process model with intensity $\hat{\lambda}_2$. Right: Estimated *L(r)-r* on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous Poisson process (dashed lines).

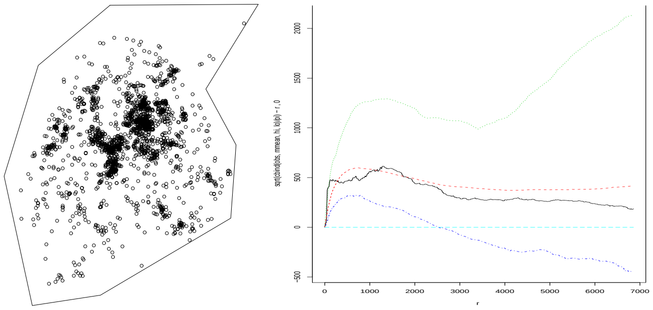

Figure 15 : Left: A simulation from an inhomogeneous LGCP model obtained by thinning. Right: Estimated *L(r)-r* on emergencies in June (solid line), average and envelope from 39 simulations of an inhomogeneous LGCP model obtained by thinning (dashed lines).

area except near the boundary of the region. Therefore, the inhomogeneous LGCP model obtained from thinning by the field $\hat{\lambda}_2$ yields a satisfactory model of the point pattern of emergencies in June. We only add that the estimate $\hat{\lambda}_2$ should be improved by considering an edge correction factor in the density estimate $\hat{C}_2$.

6. Conclusions

The study of this spatio-temporal marked point pattern of emergencies during one year underlines the numerous difficulties faced when analyzing complex and large data sets. First of all, the high inhomogeneity and the many duplicated points are a major problem in the estimation of the background intensity of the emergencies. These difficulties result in problems in finding a good bandwidth h which does not lead to over-fitting.

In the case of inhomogeneity of the positions, the global test of independence between positions, time and continuous marks is intricate. So, we have tested this dependence two by two on different subsamples for computational reasons. It seems hard to make a definite choice between the different point process models considered here. Indeed, the Poisson point process with intensity estimated nonparametrically yields a simulation with localizations of events close to that of our point pattern, but the estimate of the L function presents some over-fitting. It is difficult to decide whether this phenomenon is due or not to an "artifact" in the estimate of L.

With the semiparametric approach, we observe that the adaptive kernel estimation of the population density yields better point process models than the classical kernel estimation with a global bandwidth. So we retain as acceptable models for our data set the Poisson point process with intensity $\hat{\lambda}_2$ and the LGCP with inhomogeneity obtained with thinning by $\hat{\lambda}_2$. The goodness-of-fit is satisfactory for the first order characteristic for the Poisson model while it is good for the first and second order characteristics for the LGCP model. In spite of the boundary errors generated by the adaptive kernel estimation and the variability of the number of points per simulation, the LGCP appears to us to be a good enough model of the June emergencies.

The generalization to the other months is made by considering intensities proportional to $\hat{\lambda}_2$ according to the expected number of points for each month. The marks realizations are obtained by a bootstrap procedure and are affected independently to the point pattern of positions.

Correspondence: bonneu@cict.fr

REFERENCES

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

Baddeley, A. and R. Turner. 2005. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software 12*: 1–42.

Baddeley, A. and R. Turner. 2006. Modelling spatial point patterns in *R*. In *Case studies in spatial point process modeling*, vol. 185. Springer, New York, pp. 23–74.

Benes, V., K. Bodlak, J. Moller and R. Waagepetersen. 2005. A case study on point process modelling in disease mapping. *Image Analysis and Stereology 24*: 159–168.

P. Berman, M. and P. Diggle. 1989. Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society. Series B. 51*:81–92.

Bonneu, F. and C. Thomas-Agnan. 2007. Spatial point process models for capacitated location-allocation problems. *Working paper*.

Cressie, N. 1993. *Statistics for spatial data*. Wiley, New York. Second edition.

Cucala, L. 2006. Intensity estimation for spatial point processes observed with noise. *Preprint*.

Diggle, P. 1985. A kernel method for smoothing point process data. *Applied Statistics 34*: 138–147.

Diggle, P. 2003. *Statistical analysis of spatial point patterns*. Arnold.

Diggle, P. 2006. *Statistical analysis of spatio-temporal point process data*. In Finkenstadt, B., Held, L., and Isham V., editors, Semstat2004, 1-45. CRC Press, London.

Diggle, P., P. Zheng and P. Durr. 2005. Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society. Series C. Applied Statistics 54*: 645–658.

Diggle, P. J. , V. Gomez-Rubio, P. E. Brown, A. G. Chetwynd and S. Gooding. 2006. Second-order analysis of inhomogeneous spatial point processes using case-control data. *Biometrics (OnlineEarly Articles)*.

Guan, Y. 2006. Tests for independence between marks and points of a marked point process. *Biometrics 62*: 126–134.

Jonsdottir, K., U. Hahn and E. Jensen. 2004. Inhomogeneous spatial point processes, with a view to spatio-temporal modelling. In *Proceedings of the Conference on Spatial Point Process Modelling and its Applications* (Castellon), A. Baddeley, P. Gregori, M. J., S. R., and S. D., Eds., pp. 131–136.

Moller, J., A. R. Syversveen and R. P. Waagepetersen. 1998. Log Gaussian Cox processes. *Journal of Applied Probability 25*: 451–482.

Møller, J. and R. Waagepetersen. 2004. *Statistical inference and simulation for spatial point processes*, vol. 100. Chapman & Hall/CRC.

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 7*: 99–117.

Schlather, M., J. Ribeiro, J. Paulo and P. Diggle. 2004. Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 66: 79–93.

Schoenberg, F. P. 2004. Testing separability in spatial-temporal marked point processes. *Biometrics* 60: 471–481.

Stoyan, D. and H. Stoyan. 1994. *Fractals, random shapes and point fields*. Wiley, Chichester.

Waagepetersen, R. 2006. An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics (OnlineEarly Articles)*.

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