Kybernetika

Tomáš Mrkvička; François Goreaud; Joël Chadoeuf

Spatial prediction of the mark of a location-dependent marked point process: How the use of a parametric model may improve prediction

Kybernetika, Vol. 47 (2011), No. 5, 696--714

Persistent URL: `http://dml.cz/dmlcz/141685`

KYBERNETIKA — VOLUME 47 (2011), NUMBER 5, PAGES 696–714

SPATIAL PREDICTION OF THE MARK OF A LOCATION-DEPENDENT MARKED POINT PROCESS: HOW THE USE OF A PARAMETRIC MODEL MAY IMPROVE PREDICTION

Tomáš Mrkvička, François Goreaud and Joël Chadœuf

We discuss the prediction of a spatial variable of a multivariate mark composed of both dependent and explanatory variables. The marks are location-dependent and they are attached to a point process. We assume that the marks are assigned independently, conditionally on an unknown underlying parametric field. We compare (i) the classical non-parametric Nadaraya–Watson kernel estimator based on the dependent variable (ii) estimators obtained under an assumption of local parametric model where explanatory variables of the local model are estimated through kernel estimation and (iii) a kernel estimator of the result of the parametric model, supposed here to be a Uniformly Minimum Variance Unbiased Estimator derived under the local parametric model when complete and sufficient statistics are available. The comparison is done asymptotically and by simulations in special cases. The procedure for better estimator selection is then illustrated on a real-life data set.

Keywords: kernel estimation, marked Poisson process, mean mark estimation, location-dependent mark distribution, segment process

Classification: 62M30, 62G05

1. INTRODUCTION

Marked point processes (X, L) are used to describe the position and characteristics of objects randomly spread in space. They can be used, for example, in forestry to describe the position $X = (x_1, \dots x_n)$ and characteristics (height, diameter, ...) denoted as $L = (l_1, \dots l_n)$ of trees, where $l_i = (l_i^{(1)}, \dots, l_i^{(k)})$, k being the number of characteristics observed. At a given point, these characteristics can be dependent and may also depend on unobserved characteristics (soil properties for example). For such point processes the marks are then location-dependent – depend on the position of the point. Furthermore we assume that the marks are independent of each other and of other points. Most of the literature deal with the stationary marked point processes (e. g. [8, 18]) and study the dependence structure among marks. Here we are interested in with the location-dependent varying of the marks which are independent of other locations and other marks.

Suppose one characteristic $l^{(1)}$ is of interest (e. g., the tree volume or tree height). This characteristic can be analyzed by (i) looking at the univariate marked point process $(X, L^{(1)})$, or (ii) by using local models describing the dependence within the vector of a mark if available, so as to use all information. Local models can be available from prior biological or physical knowledge. (iii) In the case of local statistical models, (ii) should be improved using Uniformly Minimum Variance Unbiased estimators.

In the case (i), non-parametric estimation is widely used in order to take into account the spatial non-stationarity, as in spatial epidemiology ([9, 10, 11]). Main estimation procedures, focusing on local mean value estimation, are completely non-parametric. The most classical example is the estimation of the mean value of marks of a marked point process, where non-parametric Nadaraya–Watson kernel estimators are used. In case (ii) explanatory variables at each location will be estimated by using a kernel estimation of the observed explanatory variables $(L^{(2)}, \dots, L^{(k)})$ and will be plugged into the local model. In case (iii), we propose to plug these kernel estimates in the Uniformly Minimum Variance Unbiased estimators derived from the local model.

The question arises then to know whether it is better (i) to estimate directly the mean $l^{(1)}$ non-parametrically, or (ii) to estimate the explanatory variables $l^{(2)}, \dots, l^{(k)}$, first non-parametrically and then estimate $l^{(1)}$ by pluging-in the estimated values in the expectation formula. This question has been already formulated (see for example [22] where it was reformulated recently in an environmental and agricultural context), but the answer to this question is not known. At the moment, scientists invest mainly on the local model by introducing as much physical components as possible. For example, in [4] and [17] the physiology and bioclimatology in plan models in order to predict potential yield or carbon release in atmosphere is introduced. Such models are then used locally everywhere in the zone of interest, by interpolating the explanatory variables, as can be seen for example in [2]. A better solution may also be (iii) to estimate the complete and sufficient statistics of the local model obtained under location-independent assumption by applying it locally, and then estimate $l^{(1)}$ by plugging these statistics in the Uniformly Minimum Variance Unbiased (UMVU) estimators for which better statistical properties are expected ([12, 13, 14, 15]).

The answer will depend on the mean value estimator which is used instead of the sample mean and on how much are the estimator properties degraded when it is used in a location-dependent case. Intuitively, it will depend on the estimator characteristics and the characteristics of the spatial variability. In the location-independent case, both the UMVU estimator and the sample mean being unbiased, the obvious choice is the UMVU estimator. In the location-dependent case however, the more spatial variability occurs, the more bias is expected from each estimator. Choosing one or the other estimator in the location-dependent case will then depend, in particular, on how much bias each estimator generates. We compare the approaches asymptotically in Section 2 for marked stationary Poisson point process. Progress is made in the case where the parameters of the local model are expectations of functions under this local model, $\theta = \mathbb{E}(M(l))$. This case is encountered, for example, with distributions which have complete and sufficient statistic generated by the distribution moments. In this case, the marks $(l_i^{(2)}, \ldots, l_i^{(k)}) = M(l_i^{(1)})$ will be such functions of the mark $l_i^{(1)}$. Thus we will focus throughout the rest of the paper on one dimensional marks l for the simplicity. But additional explanatory variables can be also added. An example of this case is given in Section 3. In Section 4, we study a less classical case, where the mark is uniformly distributed between two unknown bounds and compare UMVU estimators to the direct non-parametric mean mark estimator. At last, in the Section 5, we illustrate our method on a real-life data from forestry.

2. GENERAL CASES

2.1. Statistical framework

We consider a point process X in $\mathbb{R}^2$ with intensity λ and a marked point process (X, L) obtained by attaching to each point $x \in X$ a univariate random mark $l(x) \in R$, the marks being independent to each other and to the position of the other points of the process ([21]). Let $P_x(l) = P(l(x) \leq l)$ denote the distribution of a mark attached to point x.

We consider that this distribution admits a density that can be expressed as a parametric function $\mathrm{d}P_x(l) = f(l, \theta(x))\,\mathrm{d}l$ where $\theta(x)$ is the set of parameters of the distribution at point x. For a given parameter $\theta(x)$, we can calculate the expectation and the variance of the distribution of the mark at point x : the expectation is $g(\theta) = \int_l l f(l, \theta)\,\mathrm{d}l$ and the variance is $v(\theta) = \int_l (l - g(\theta))^2 f(l, \theta)\,\mathrm{d}l$. Assume that $g(\theta)$ is second order differentiable.

2.2. Problem description in the i.i.d. case

In this subsection, we describe the problem in the simpler case of the stationary marked point process where the marks are not position-dependent (i. e. when marks are i.i.d.). In the next subsection, we look at the same problem, but for marked point process with position-dependent marks, as it was said in the Introduction.

Consider a sample of marks for a given location x, $l_{(n)} = (l_1, \ldots, l_n)$ from the distribution $f(l, \theta(x))$. (i) A first simple estimator of the expectation of $l(x)$ is the classical mean $m_1(l_{(n)}) = \frac{1}{n} \sum_i l_i$. Let consider another estimator taking into account the knowledge of the distribution of marks, i. e., its distribution function $f(l, \theta(x))$. The parameter is not known but we may assume that we have a consistent estimator $\hat{\theta}(l_{(n)})$ of this parameter. (ii) We can then define a second estimator of the expectation of $l(x)$: $m_2(l_{(n)}) = g(\hat{\theta}(l_{(n)}))$. (iii) Finally, these two estimators can be compared to the UMVU estimator $m_3(l_{(n)})$, when available, for which by definition $E(m_3(l_{(n)})) = g(\theta)$ and its variance is minimal among unbiased estimators.

Furthermore, assume, in the rest of this Section, that $\hat{\theta}(l_{(n)})$ can be expressed as the mean of functions of the l_i, $\hat{\theta}(l_{(n)}) = \frac{1}{n} \sum_i M(l_i)$ so that $\mathrm{var}(\hat{\theta}(l_{(n)})) = \frac{1}{n} u(\theta)$, for a function u. A case encountered, for example, if the parameter is expressed as moments of functions of l_i.

When n tends to infinity, mean and variance of these two estimators satisfy:

- $E(m_1(l_{(n)})) = g(\theta)$ and $n{\cdot}\mathrm{var}(m_1(l_{(n)})) = v(\theta)$,
- $E(m_2(l_{(n)})) = E(g(\hat{\theta}(l_{(n)}))) \to g(\theta)$ and $n \cdot \mathrm{var}(m_2(l_{(n)})) \to g'(\theta)^t u(\theta) g'(\theta)$ where $g'(u)$ is the gradient vector of $g(x)$ at point x and v^t is the transpose of vector v and
- $E(m_3(l_{(n)})) = g(\theta)$ and $\mathrm{var}(m_3(l_{(n)})) \leq \mathrm{var}(m_1(l_{(n)}))$.

The ratio of the mean squared error of the two first estimators tends to $\frac{g'(\theta)^t u(\theta) g'(\theta)}{v(\theta)}$ and using $m_2(l_{(n)})$ instead of $m_1(l_{(n)})$ can be of interest as soon as this ratio is less than 1. But when dealing with non i.i.d. versions as will be the case henceforth, there remains the question, whether the same estimators remain interesting as in the i.i.d. case.

2.3. Problem description for position-dependent marks

The versions of these three estimators in the case of position-dependent marks are:
$m_1(x,h) = \frac{\sum_i \omega_h(x-x_i) l_i}{\sum_i \omega_h(x-x_i)}$ ([20]) and
$m_2(x,h) = g(\hat{\theta}(x,h)) = g\left(\frac{\sum_i \omega_h(x-x_i) M(l_i)}{\sum_i \omega_h(x-x_i)}\right)$
where $\omega_h(x) = \frac{1}{h^2}\omega(x/h)$ is a two dimensional kernel with bandwidth h.

Suppose that there exists a function e of the complete and sufficient statistic $S(l_1,\ldots,l_n)$ of the local model which constructs the UMVU estimator of the expectation of l: $e(S(l_1,\ldots,l_n))$. If $S(l_1,\ldots,l_n)$ can be written as $\sum_i M(l_i)$ in the local model (e.g. when marks are normally distributed or exponentially or lognormally, ...), then for position-dependent marks the estimator can be defined as
$m_3(x,h) = e\left(\frac{\sum_i \omega_h(x-x_i) M(l_i)}{\sum_i \omega_h(x-x_i)}\right)$
and this version of the UMVU $m_3(x,h)$ can be handled in the same way as $m_2(x,h)$.

To study how these three estimators behave, we look at its asymptotic behaviour in the stationary Poisson case.

2.4. Asymptotic behaviour of $m_1(x,h)$ and $m_2(x,h)$ in the stationary Poisson case

Let us suppose, throughout this subsection, that (X,L) is the stationary marked Poisson point process with location-dependent marks.

At each point x and each h, these two estimators can be written as $F(\frac{1}{n}\sum_i (U_i,V_i))$ where (U_i,V_i) are independent random vectors and F is two times differentiable. Their asymptotic behaviour is then classically obtained by using first a central limit theorem, then applying a Delta method ([5, Section 5.2.4]).

Point convergence of $m_1(x,h)$.

Lemma 2.1. Let us suppose that the number of points n tends to infinity in the fixed window $W \subset \mathbb{R}^2$ and h tends to 0, then the asymptotic behavior of $m_1(x,h)$ is

- $m_1(x,h) - g(\theta(x)) = \frac{h^2}{2}\int \omega(y) y^t A_x y \,\mathrm{d}y + o(h^2)$,
- $\mathrm{var}(m_1(x,h)) = \frac{1}{nh^2} v(\theta(x)) \int_y \omega^2(y)\,\mathrm{d}y$

where A_x is the matrix of the second derivatives of $g(\theta(x))$ at x.

Proof. Denote $(U_i, V_i) = (\omega_h(x - x_i) l_i, \omega_h(x - x_i))$. These vectors being i.i.d,

$$\frac{1}{n}\sum_i (U_i, V_i) \to \left(\int_W \omega_h(x-y) g(\theta(y))\,\mathrm{d}y, \int_W \omega_h(x-y)\,\mathrm{d}y\right) = (U_\infty, V_\infty)$$

and $\sqrt{n}(\frac{1}{n}\sum_i (U_i,V_i) - (U_\infty, V_\infty))$ is asymptotically Gaussian with mean $(0,0)^t$ and variance matrix Σ defined by:

- $\Sigma_{11} = \int_y \omega_h^2(x-y)(v(\theta(y)) + g^2(\theta(y)))\,\mathrm{d}y - (\int_y \omega_h(x-y) g(\theta(y))\,\mathrm{d}y)^2$
- $\Sigma_{12} = \Sigma_{21} = \int_y \omega_h^2(x-y) g(\theta(y))\,\mathrm{d}y - \int_y \omega_h(x-y)\,\mathrm{d}y \int_y \omega_h(x-y) g(\theta(y))\,\mathrm{d}y$
- $\Sigma_{22} = \int_y \omega_h^2(x-y)\,\mathrm{d}y - (\int_y \omega_h(x-y)\,\mathrm{d}y)^2$.

Then $m_1(x,h) = F(\frac{1}{n}\sum_i (U_i, V_i))$ with $F((u,v)) = u/v$. F is two times differentiable, bounded in a neighborhood of (u,v) as soon as $v \neq 0$ and the Delta method ([5, Section 5.2.4]) gives

- $m_1(x,h) \to \frac{\int_y \omega_h(x-y) g(\theta(y))\,\mathrm{d}y}{\int_y \omega_h(x-y)\,\mathrm{d}y}$
- $\sqrt(n)\left(m_1(x,h) - \frac{\int_y \omega_h(x-y) g(\theta(y))\,\mathrm{d}y}{\int_y \omega_h(x-y)\,\mathrm{d}y}\right)$ is asymptotically Gaussian with variance

$$\frac{1}{\left(\int_y \omega_h(x-y)\,\mathrm{d}y\right)^4}\Big(\begin{array}{l} \Sigma_{11}(\int_y \omega_h(x-y)\,\mathrm{d}y)^2 \\ -2\Sigma_{12}\int_y \omega_h(x-y)\,\mathrm{d}y \int_y \omega_h(x-y) g(\theta(y))\,\mathrm{d}y \\ +\Sigma_{22}(\int_y \omega_h(x-y) g(\theta(y))\,\mathrm{d}y)^2 \Big). \end{array}$$

Thus letting $h \to 0$ leads to the result. □

Point convergence of $m_2(x,h)$.

Lemma 2.2. Let us suppose that the number of points n tends to infinity in the fixed window W and h tends to 0, then the asymptotic behavior of $m_2(x,h)$ is

- $m_2(x,h) - g(\theta(x)) = \frac{h^2}{2} g'(\theta(x))^t \int \omega(y) y^t B_x y\,\mathrm{d}y + o(h^2)$,
- $\mathrm{var}(m_2(x,h)) = \frac{1}{nh^2} g'(\theta(x))^t C_x g'(\theta(x)) \int_y \omega^2(y)\,\mathrm{d}y$,

where $g'(\theta)$ is the vector of first derivatives of $g(\theta)$ at x, $B_x = (\frac{\partial^2 \theta_k(x)}{\partial x_i \partial x_j})$ $1 \le i \le 2$, $1 \le j \le 2$, $1 \le k \le K$ the array of second derivatives of $\theta(x)$ with respect to x and C_x is the covariance matrix of $M(l_x)$.

Proof. In the same way as in the previous proof one gets that for $n \to \infty$, if W_i denotes the vector $W_i = (\omega_h(x-x_i), \omega_h(x-x_i)M_1(l_i), \dots, \omega_h(x-x_i)M_K(l_i))^t$, then

$$\frac{1}{n}\sum_i W_i \to$$

$$\left(\int_y \omega_h(x-y)\,\mathrm{d}y, \int_y \omega_h(x-y)E(M_1(l(y))\,\mathrm{d}y, \dots, \int_y \omega_h(x-y)E(M_K(l(y)))\,\mathrm{d}y\right)^t = W_\infty$$

and $\sqrt{n}\left(\frac{1}{n}\sum_i W_i - W_\infty\right) \to \mathcal{N}(0, \Sigma')$ where

$$\Sigma'_{j,j'} = \int_y \omega_h^2(x-y)(\operatorname{cov}(M_j(l_y), M_{j'}(l_y)) + E(M_j(l_y))E(M_{j'}(l_y)))\,\mathrm{d}y$$
$$-\int_y \omega_h(x-y)E(M_j(l_y))\,\mathrm{d}y \int_y \omega_h(x-y)E(M_{j'}(l_y))\,\mathrm{d}y, \text{ where } 0 \le j, j' \le K$$

and define $M_0 = 1$.

Then using the variational theorem for $m_2(x,h) = g(W)$ and letting $h \to 0$, one gets the result.

Note that, in the i.i.d. case, the variance of the second estimator verifies $\operatorname{var}(\theta(x)) \simeq \frac{1}{n} g'(\theta(x))^t C_x g'(\theta(x))$ when $h \to 0$. □

Point convergence of $m_3(x,h)$.

Lemma 2.3. Let us suppose furthermore that $m_3(x,h) = e(\frac{\sum_i \omega_h(x-x_i)M(l_i)}{\sum_i \omega_h(x-x_i)})$ (see Section 2.3), then the asymptotic behavior of $m_3(x,h)$ is

- $m_3(x,h) - e(\theta(x)) = \frac{h^2}{2} e'(\theta(x))^t \int \omega(y) y^t B_x y\,\mathrm{d}y + o(h^2)$,
- $\operatorname{var}(m_3(x,h)) = \frac{1}{nh^2} e'(\theta(x))^t C_x e'(\theta(x)) \int_y \omega^2(y)\,\mathrm{d}y$,

where $e'(\theta)$ is the vector of first derivatives of $e(\theta)$ at x.

Proof. The proof is same as for $m_2(x,h)$ with changing g to e. □

Global comparison. As the UMVU derived estimator $m_3(x,h)$ is similar to $m_2(x,h)$, we concentrate on the comparison between $m_1(x,h)$ and $m_2(x,h)$.

The integrated mean squared errors (IMSE) become:

$$\mathrm{IMSE}_1(h) = \frac{h^4}{4} \int_x (\int \omega(y) y^t A_x y\,\mathrm{d}y)^2\,\mathrm{d}x + \frac{1}{nh^2} \int_x v(\theta(x)) \int_y \omega^2(y)\,\mathrm{d}y\mathrm{d}x$$
$$\mathrm{IMSE}_2(h) = \frac{h^4}{4} \int_x (g'(\theta(x))^t \int \omega(y) y^t B_x y\,\mathrm{d}y)^2\,\mathrm{d}x + \frac{1}{nh^2}{}^t g'(\theta(x)) C_x g'(\theta(x)) \int_y \omega^2(y)\,\mathrm{d}y$$

for which the minimum values are:

- $\min_1 = \frac{2^{-4/3}+2^{-1/3}}{n^{2/3}} \left(\int_x (\int \omega(y) y^t A_x y\,\mathrm{d}y)^2\,\mathrm{d}x\right)^{1/3} \left(\int_x v(\theta(x)) \int_y \omega^2(y)\,\mathrm{d}y\mathrm{d}x\right)^{2/3}$
- $\min_2 = \frac{2^{-4/3}+2^{-1/3}}{n^{2/3}}$

$$\left(\int_x \left(g'(\theta(x))^t \int \omega(y) y^t B_x y\,\mathrm{d}y\right)^2 \mathrm{d}x\right)^{1/3} \left(\int_x g'(\theta(x))^t C_x g'(\theta(x)) \int_y \omega^2(y)\,\mathrm{d}y\,\mathrm{d}x\right)^{2/3}.$$

The ratio of these two values depends on the ratio of integrated variances,

$$\frac{\int_x g'(\theta(x))^t C_x g'(\theta(x))}{\int_x v(\theta(x))}$$

whose value measures how much the smaller variance of the second estimator improves the IMSE and the ratio of integrated squared bias

$$\frac{\int_x \left(g'(\theta(x))^t \int \omega(y) y^t B_x y \, \mathrm{d}y\right)^2 \, \mathrm{d}x}{\int_x (\int \omega(y) y^t A_x y \, \mathrm{d}y)^2 \, \mathrm{d}x}$$

which measures the effect of the mark location-dependence on the bias of the two estimators.

In particular, the ratio of minima of IMSE does not depend on n for large n. Therefore, increasing n, although leading to a smaller bandwidth window, thereby to less bias, does not lead to a preference for the second estimator systematically. Choosing the first or the second estimator depends on the variance improvement if one chooses the second estimator, but also to the characteristics of the mark location-dependence which must remain small enough in order not to lose this gain through the bias.

3. EXAMPLE WHEN $M_3(X, H)$ AND $M_2(X, H)$ ARE IN THE SAME FORM

In this section we look at one concrete case where $\hat{\theta}(l_{(n)}) = \frac{1}{n} \sum_i M(l_i)$ and the complete and sufficient statistic has also the same form $S(l_{(n)}) = \frac{1}{n} \sum_i M(l_i)$. This is the case of theoretical result showed in Section 2.4 for stationary Poisson point process.

We look at the case of the log-normal distribution and consider the case of known σ^2 for the sake of simplicity. We then have a marked point process with independent marks whose mark distribution follows a lognormal distribution with parameters $\theta(x) = [\log(m(x)) - \sigma^2/2, \sigma^2]$. This choice gives the mean mark $g(\theta(x)) = m(x)$. We focused on the estimation of $g(\theta(x))$ either by $m_1(x, h)$ or by UMVU derived estimator $m_3(x, h)$.

I.i.d. case. For an i.i.d. sample $l_{(n)} = (l_1, .., l_n)$ issued from the lognormal distribution with parameters $[\mu, \sigma^2]$, the empirical mean is unbiased with the variance $\mathrm{var}(m_1(l_{(n)}))=\frac{1}{n}\exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)$.

The UMVU estimator of the mean m expressed as

$$m_3(l_{(n)}) = \exp(\frac{\sum_i \log l_i}{n} + \frac{n-1}{n}\frac{\sigma^2}{2})$$

has the variance $\mathrm{var}(m_3(l_{(n)})) = \exp(2\mu + \sigma^2)(\exp(\sigma^2/n) - 1)$.

The variance ratio between these two estimators is $\frac{n(\exp(\sigma^2/n)-1)}{\exp(\sigma^2)-1} \to \frac{\sigma^2}{\exp(\sigma^2)-1}$. For $\sigma^2 = 0.75$, this limit is equal to 0.67.

Location-dependent case. In the location-dependent case, the UMVU derived estimator becomes

$$m_3(x, h) = \exp\left(\frac{\sum_i \omega_h(x - x_i)\log(l_i)}{\sum_i \omega_h(x - x_i)} + \frac{n-1}{n}\frac{\sigma^2}{2}\right),$$

where n is the number of points for which $\omega_h(x - x_i) > 0$.

In practice, minimum of IMSE cannot be explicitly computed so we can not compare the estimators directly. Thus we have to rely on simulations. When the above estimators are applied the bandwidth has to be selected. Several methods are used to choose an optimal bandwidth, one of the most popular being the cross-validation method. Consider now an estimator $m(x,h)$. Let $\hat{m}(x,h) = \frac{1}{n}\sum_i \tilde{m}(i,x,h)$, where $\tilde{m}(i,x,h)$ is the estimator same as $m(x,h)$ but which omits the ith mark and n is the number of all points visible in the observation window. For example: $\tilde{m}(i,x,h) = \frac{\sum_{j\neq i}\omega_h(x-x_j)l_j}{\sum_{j\neq i}\omega_h(x-x_j)}$ for classical non-parametric kernel estimator. Then the mean integrated square error (MISE) of the estimator $\hat{m}(x,h)$, is $E(\int_W (m(x)-\hat{m}(x,h))^2\,\mathrm{d}x) = \int_W m^2(x)\,\mathrm{d}x + E(C(h))$, where $C(h) = \int_W \hat{m}^2(x,h)\,\mathrm{d}x - 2\frac{1}{n}\sum_i l_i\tilde{m}(i,x_i,h)$. Thus, it is necessary to minimize $C(h)$ for finding an optimal bandwidth of the estimator $\hat{m}(x,h)$ by cross-validation method. We will use then $\hat{m}(x,h)$ with selected bandwidth instead of $m(x,h)$.

To illustrate this, we considered the following simulation example. We simulated stationary marked Poisson point process with intenzity $\lambda = 200$ in the unit square window. The mark location-dependence was given by $m(x) = \sin(4\pi(x_1+x_2))a/80+1/20$ with $a = 1$ for large variation, $a = 2$ for smaller variations, and $m(x) = (x_1 + x_2 + 1)/40$ for small variations. σ^2 was chosen to be equal to 0.75 and the Epanatchnikov kernel ω_h was used in the estimation. The corresponding surfaces of the location-dependence are shown in Table 1 (first row), examples of realizations in Table 1 (second row). On the pictures of realizations the points correspond to the point process X and the segments attached to the points show the size of the mark $l(x)$. Orientation is random and brings no information relative to the marked point process. As usual with such distributions, one can observe some large mark values which can perturb direct mean estimation.

Table 1 (fourth row) shows the estimated means with the classical non-parametric Nadaraya–Watson kernel estimator for a particular realization. Corresponding integrated square errors are shown in the third row. Table 1 (sixth row) shows the estimated means with the proposed estimator for same realization as above. Corresponding integrated square errors $\int_W (m(x)-\hat{m}(x,h))^2\,\mathrm{d}x$ are shown in the fifth row. For each model we have generated 100 simulations and we computed the integrated square error for every realization. Means, standard deviations and histograms for the integrated square error are displayed in Table 2. It is displayed for the classical non-parametric Nadaraya–Watson kernel estimator in the first, second and third row respectively and for the proposed estimator in the fourth, fifth and sixth row. In the last row in Table 2 there are the ratios of means of integrated square errors for the two tested estimators $(MISE(m_3)/MISE(m_1))$. These numbers show the efficiencies of the estimator m_3 with respect to the classical non-parametric Nadaraya–Watson kernel estimator m_1.

We can see from Table 2 that the tested estimator has lower MISE in all our cases of non-stationarity structure for the case of log-normal distribution.

To see, whether the non-stationarity of the points influence the performances of the tested estimators, we simulate also the non-stationary Poisson marked point processes with intensity function $r(x) = \sin(4\pi(x_1 + x_2))/4 + 1$ (the multiple of the function $m(x)$ which gives the mark location-dependence in the large variation case). The expected number of points in the observation window is equal to 200. All other parameters of the simulation were left same as for the stationary case. Table 3 shows the results of this simulations. The results do not reveal any significant difference with respect to the stationary case, except a slight increase of the inaccuracy of both estimators.

4. EXAMPLE WHEN $M_3(X,H)$ AND $M_2(X,H)$ ARE NOT THE SAME

We look at the case of the uniform distribution of the marks of a marked point process. In such a case the complete and sufficient statistics consist of min and max functions which are not of the shape $\frac{1}{n}\sum_i M(l_i)$ (for details about complete and sufficient statistics see [12]. We focus again on comparison of the mean estimators $m_1(x,h)$ and $m_3(x,h)$.

I.i.d. case. For an i.i.d. sample $l_{(n)} = (l_1,\dots,l_n)$ issued from the uniform distribution with parameters $[A,B]$, the empirical mean is unbiased with the variance $\mathrm{var}(m_1(l_{(n)})) = \frac{1}{n}\frac{1}{12}(B-A)^2$.

The UMVU estimator of the mean m, expressed as $m_3(l_{(n)}) = \dfrac{\max l_i + \min l_i}{2}$ has the variance $\mathrm{var}(m_3(l_{(n)})) = \frac{1}{2(n+1)(n+2)}(B-A)^2$. The variance ratio between these two estimators is $\frac{6n}{(n+1)(n+2)}$.

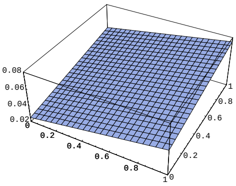
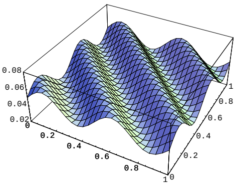
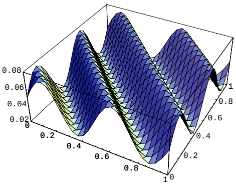
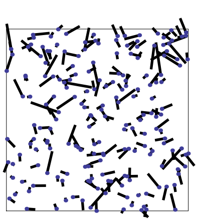
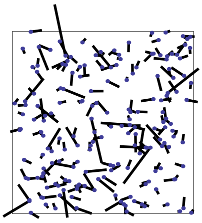
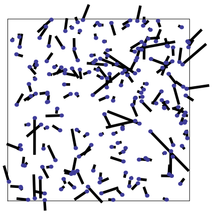
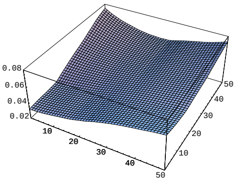
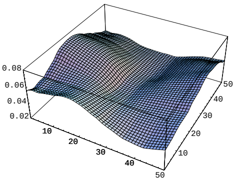
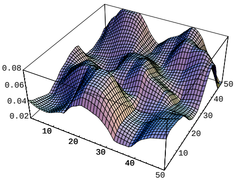
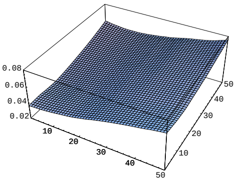
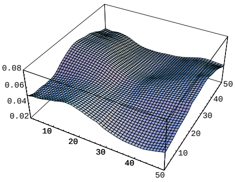
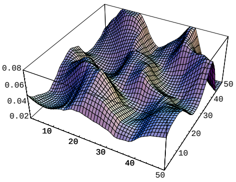
Location-dependent case. In the location-dependent case consider the following example. Assume a stationary Poisson marked point process with the intensity 200 and with independent marks whose mark distribution follows a uniform distribution with parameters $\theta(x) = [0.2m(x), 1.8m(x)]$ on the unit square. This choice gives again $g(\theta(x)) = m(x)$. Location-dependence was given in the same way as in the lognormal example. Examples of realizations are shown on Figure 4 (first row).

For the middle variation location-dependence structure, the average number of points in the window $[0,2b]^2$ is 14. Here b is the optimal bandwidth obtained by cross-validation. The above ratio is equal to 0.35 for $n = 14$. Thus one could expect that the efficiency computed as in the lognormal case will be a bit above 0.35. In fact, preliminary studies showed that the estimator $m_3(x,h) = (\max_{\{i:x_i - x\in[-h,h]^2\}} l_i + \min_{\{i:x_i-x\in[-h,h]^2\}} l_i)/2$ is not robust with respect to the location-dependence.

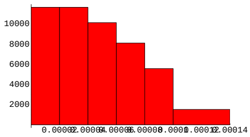
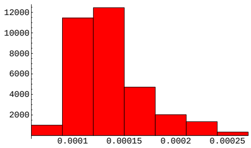
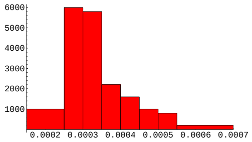
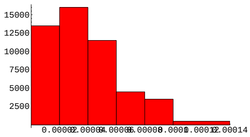
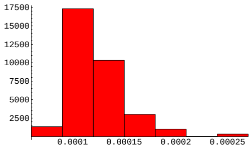
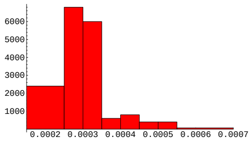
Modification in the location-dependent case. Since the $m_3(x,h)$ is not robust with the location-dependence, we propose a modification which should admit the properties of the complete and sufficient statistic but which is more robust with respect to location-dependence.

One possibility is to approximate the complete and sufficient statistic $(\min l_i, \max l_i)$ by a kernel estimator

$$(\widehat{\min}(x), \widehat{\max}(x)) = \left(\frac{\sum_i \omega_h(x-x_i)N_1(x,l_i)}{\omega_h(x-x_1)}, \frac{\sum_i \omega_h(x-x_i)N_2(x,l_i)}{\omega_h(x-x_1)}\right),$$

	Slow gradient	middle variations	large variations
expected mean			
simulated example			
ISE	0.000033	0.00015	0.00025
estimated mean by classical estimator			
ISE	0.000027	0.00011	0.00020
estimated mean by tested estimator			

Tab. 1. Comparison of the estimators m_1 and m_3 for one realization.

	Slow gradient	middle variations	large variations
mean for m_1	0.000049	0.000138	0.000341
SD for m_1	0.0000339	0.0000412	0.0000913
histogram for classical estimator			
mean for m_3	0.000038	0.000122	0.000297
SD for m_3	0.0000241	0.0000284	0.0000734
histogram for tested estimator			
efficiency	0.77	0.88	0.87

Tab. 2. Comparison of the estimators m_1 and m_3 for 100 realizations.

	Slow gradient	middle variations	large variations
mean for m_1	0.000047	0.000165	0.000464
SD for m_1	0.0000294	0.0000632	0.0001583
mean for m_3	0.000037	0.000138	0.000372
SD for m_3	0.0000201	0.0000450	0.0001088
efficiency	0.79	0.84	0.80

Tab. 3. Comparison of the estimators m_1 and m_3 for 100 realizations of non-stationary Poisson point process.

where $N_1(x, l_i) = \min\{l_i - \min_{\{j:||x-x_j||<||x-x_i||\}} l_j, 0\}$, $N_2(x, l_i) = \max\{l_i - \max_{\{j:||x-x_j||<||x-x_i||\}} l_j, 0\}$ and x_1 is the nearest point to X. Roughly speaking every mark adds to the maximum the difference between the mark and the maximum of marks closer to x (if it is positive), but with weight dependent on the distance from x. The proposed estimator of the mean is then

$$m_4(x, h) = \left(\frac{\sum_i \omega_h(x - x_i) N_1(x, l_i)}{\omega_h(x - x_1)} + \frac{\sum_i \omega_h(x - x_i) N_2(x, l_i)}{\omega_h(x - x_1)} \right) /2.$$

As the weighting function, we chose the Gaussian function with parameters $(0, h)$.

Tables 4 and 5 compare the classical non-parametric Nadaraya–Watson kernel estimator with the proposed estimator $m_4(x, h)$. The comparison is done in the same way as in Tables 1 and 2.

Here the tested estimator has slightly lower MISE in all cases.

The non-stationary case was performed here in the same way as in the log-normal case. Table 6 shows the results of this simulations. The results are varying here, therefore it is necessary to check which estimator is more accurate in practice for a real data. One possible check is described in the next Section.

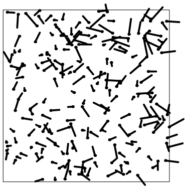
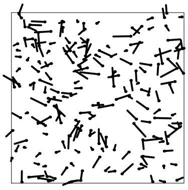
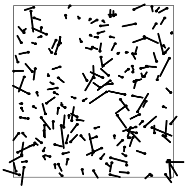
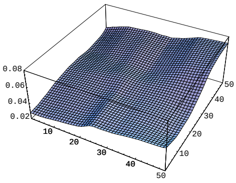
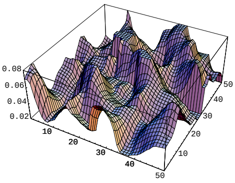
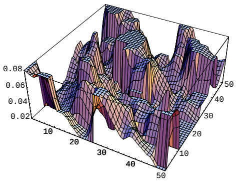
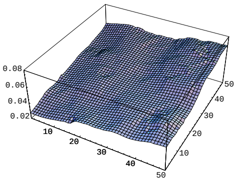
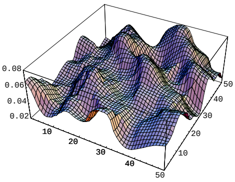
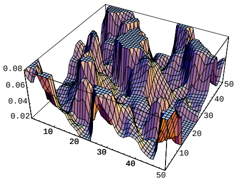
5. EXAMPLE DATA: THE LOCAL ESTIMATION OF TREE'S MEAN HEIGHT

In this section, we will show, how to use approach described above in practice. As an illustrative example, we study a data set composed of the position and height of 232 trees which fell during two large wind gusts (1967 and 1990) in the west of France ([19]). The studied area is a biological reserve preserved for at least four centuries, so that there has been hardly no human influence for a long time ([6]). The forest stand follows an almost natural dynamics, and corresponds to an uneven-aged Beech stand, with only a few old oaks.

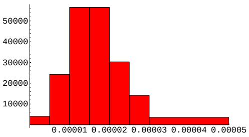
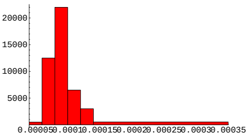
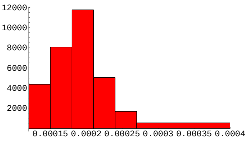
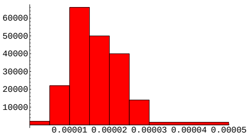
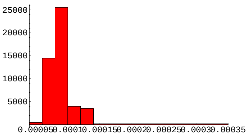
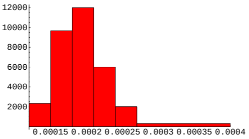
The trees that have fallen during these events usually lead to gap openings of variable size. Therefore all trees of different heights and status are concerned by these storm damages (from young and dominated trees in the understorey to dominant trees in the canopy). Therefore, the height of fallen trees can be considered as a good representation of the local variability of the height of the stand.

Soil conditions in this stand are known to present some heterogeneity ([1]). Indeed, the soil can be leached, and has a variable depth of 0.5 to 2m. These soil conditions highly influence the local fertility, and thus tree height growth. Therefore, mean height of fallen trees can be a good way to look at variation in soil characteristics.

The position of the fallen trees is indicated as a black dot on Figure 1, their height is corresponding to the length of the attached segment. As no theoretical distribution for local height was available, we chose the uniform distribution. It corresponds to the idea that these trees heights are randomly distributed between a minimum height, for trees in the understorey, and a maximum height, for dominant trees in the canopy. We consider that this maximum height depends on local soil potentiality, better soil conditions leading to locally higher trees. The assumption of

	Slow gradient	middle variations	large variations
simulated example			
ISE	0.000013	0.00016	0.00041
estimated mean by classical estimator			
ISE	0.0000067	0.000092	0.00034
estimated mean by tested estimator			

Tab. 4. Comparison of the estimators m_1 and m_4 for one realization.

	Slow gradient	middle variations	large variations
mean for m_1	0.0000177	0.000101	0.000202
SD for m_1	0.0000087	0.0000420	0.0000619
histogram for classical estimator			
mean for m_4	0.0000168	0.000090	0.000196
SD for m_4	0.0000062	0.0000201	0.0000334
histogram for tested estimator			
efficiency	0.95	0.89	0.97

Tab. 5. Comparison of the estimators m_1 and m_4 for 100 realizations.

	Slow gradient	middle variations	large variations
mean for m_1	0.000023	0.000146	0.000279
SD for m_1	0.0000134	0.0000692	0.0001107
mean for m_4	0.000024	0.000107	0.000246
SD for m_4	0.0000196	0.0000273	0.0000541
efficiency	1.07	0.74	0.88

Tab. 6. Comparison of the estimators m_1 and m_4 for 100 realizations of non-stationary Poisson point process.

the uniform length distribution of the trees was tested in [16] and it was not rejected by all tests studied there.

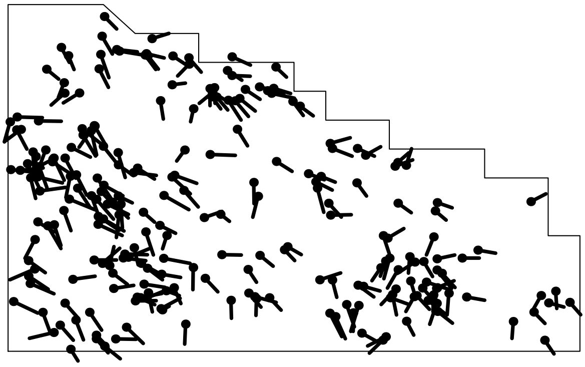


Fig. 1. The fallen trees data observed in the non-rectangular area with dimensions 720 × 480 meters.

The position of the fallen trees do not fit with stationary Poisson point process, thus the comparison of the described estimators can not be done asymptotically by Lemma 1 and 2. The position of the fallen trees fit with non-stationary Poisson point process ([16]), therefore the comparison of the described estimators is necessary to be done by further simulations.

Estimating the mean length. We assumed that the length distribution has locally uniform distribution with parameters $[A(x), B(x)]$ and we estimated the mean length using either a classical kernel method or our proposed method. The optimal bandwidth in this case is rather small thus there exist areas with no points within the Epanetchnikov kernel range and thus the mean length is not estimable for such areas. Therefore we choose a Gaussian kernel, which has infinite range, instead of Epanetchnikov. Thus we will compare the standard Gaussian kernel estimator $m_1(x, h)$ with the kernel estimator $m_4(x, h)$ described in the previous section. For the two estimates, the variances of the Gaussian kernel are estimated through the cross-validation procedure and are equal to $\sigma^2 = 182$ for the classical estimator and $\sigma^2 = 100$ for the proposed one. The two mean value estimates are mapped on Figure 2. The global variations are similar for the two estimates. Differences between the two surfaces are mainly local, the classical kernel estimate being smoother than the one we propose.

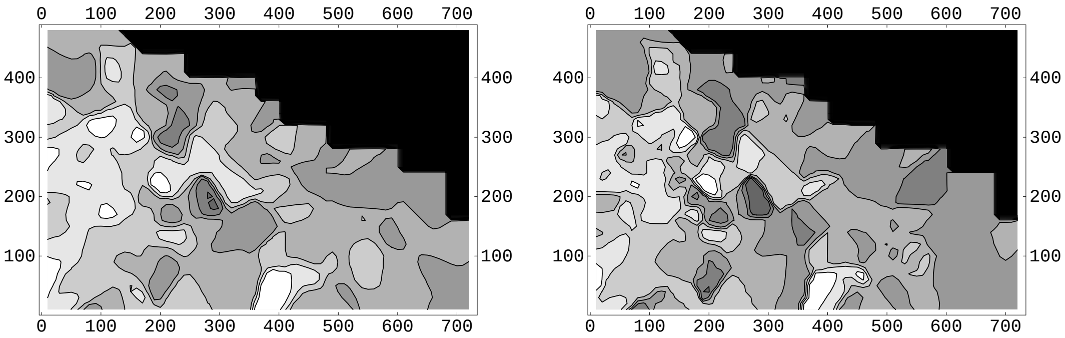


Fig. 2. Estimates of height density of the fallen trees. Left: by the standard Gaussian kernel estimator $m_1(x,h)$, right: by the tested estimator $m_4(x,h)$. The darker colours corresponds to the higher value of estimate.

Cross-validation leads to estimators presenting good asymptotic properties ([7]) and avoids the calculus of the IMSE. On the other hand, the MISE is a natural statistics to compare the two methods, but this supposes to know the true mean value. We propose to approximate it by a common mean value estimation and then compute the MISEs by simulation in the case of such chosen mean value. Common mean value estimation is obtained in the following way. First we estimate the parameter field $(A(x),B(x))$ by both approaches. We used the moment method in the Gaussian kernel approach

$$(\widehat{A}(x),\widehat{B}(x)) = \left(m_1(x) - \sqrt{3*(m_1^2(x)-(m_1(x))^2)},\, m_1(x) + \sqrt{3*(m_1^2(x)-(m_1(x))^2)}\right),$$

where $m_1^2(x) = \frac{\sum_i \omega_h(x-x_i) l_i^2}{\sum_i \omega_h(x-x_i)}$. And in our approach we estimated the parameters fields directly by an estimator

$$(\widetilde{A}(x),\widetilde{B}(x)) = \left(\widehat{\min}(x)\frac{n(x)-1}{n(x)}, \widehat{\max}(x)\frac{n(x)+1}{n(x)}\right),$$

where $n(x) = \frac{\sum_i \omega_h(x-x_i)}{\omega_h(x-x_1)}$. Then we had made the averages of the parameter fields over the two approaches $\big(\overline{A}(x),\overline{B}(x)\big) = \left(\frac{\widehat{A}(x)+\widetilde{A}(x)}{2}, \frac{\widehat{B}(x)+\widetilde{B}(x)}{2}\right)$. Then we simulated the lengths of trees from the averaged parameter fields $\big(\overline{A}(x),\overline{B}(x)\big)$ and estimated a new mean length by $m_1(x,h)$ and $m_4(x,h)$. And we computed the integrated square errors of these two estimates with respect to chosen mean value $\frac{\left(\overline{A}(x)+\overline{B}(x)\right)}{2}$. We performed 100 simulations and computed the mean integrated square errors (MISE) of the estimates $m_1(x,h)$ and $m_4(x,h)$ with respect to chosen mean value.

$$\mathrm{MISE}(m_1(x,h)) = 3.0389 \times 10^6, \quad \mathrm{MISE}(m_4(x,h)) = 2.7781 \times 10^6.$$

This leads to a mean square error at a random location of about 11.9 for m_1 and 10.8 for m_4. These MISE's are computed under the model of uniform height distribution with parameters $\big(\overline{A}(x), \overline{B}(x)\big)$. This is a model which is certainly close to the reality, thus we choose the estimator $m_4(x,h)$ in this example as a more precise estimator of the mean height.

6. CONCLUSIONS

Measuring the effect of interpolation strategies on non-linear predictors with explanatory variables, which are spatially location-dependent, is a recurrent question (see for example [22] for a recent reformulation). The problem we addressed is in fact only a small part of it, but the question remains the same: shall we (i) interpolate the model results obtained locally, in our case estimate directly the mean non-parametrically or (ii) interpolate the explanatory variables and then use as predictor the model output computed on the interpolated variables? Even in our simple case, the answer is not known beforehand, as it is dependent on both model predictor statistical properties and non-stationary properties of the random field.

The simulations showed significant increase of the accuracy when the UMVU estimators, derived from the local model, are used with respect to the classical non-parametric Nadaraya–Watson kernel estimator. The significant increase was shown for the case where the complete and sufficient statistic of the local model is expressed in the form $S(l_{(n)}) = \frac{1}{n}\sum_i M(l_i)$.

UMVU estimators are a natural choice to build non-parametric estimators in the proposed framework. Being unbiased and with minimum variance, one hopes that these properties will be transmitted to the non-parametric estimator. However, the UMVU estimator has also to be robust or the optimal properties of the UMVU may be lost in the location-dependent context. In such a case, one has to find some more robust version as the one we proposed for the case of the uniform distribution for which the complete and sufficient statistic is not expressed in the form $S(l_{(n)}) = \frac{1}{n}\sum_i M(l_i)$. In this case the simulations showed only a slight increase of the accuracy when the UMVU estimator is used with respect to the classical non-parametric Nadaraya–Watson kernel estimator.

The estimation quality, which can be approached through MISE values, did not increase much in the real data example, the MISE of the proposed estimator being only 10% less than the MISE of the classical non-parametric estimator. Judging if 10% justifies the use of specific estimators will depend on the question addressed by the data set. On the other hand, the proposed estimator, as it takes explicitly into account a local model, can lead locally to different prediction shapes than the classical estimator. In our case, the classical estimator leads to a smoother prediction. If sharp changes of the random field are expected, the proposed estimator may be better in the neighborhood of these changes.

If the interest is not on the expectation of the mark, as was the case in our examples, but on the expectation of a non-linear function of the mark, (as, for example, the square of the mark, which can be of interest in variance estimation) UMVU estimates are much more efficient ([3], [15]).

Note that, the asymptotic properties of the described estimators were shown under the assumption of the stationary Poisson point proces. It pointed out the major qualities which influence which estimator to use. But the final decision should be made after more concrete studies, as it was shown for the real data example. Such studies are not bounded by the Poisson assumption.

ACKNOWLEDGEMENT

We thank JY Pontailler for authorising us to use the data set and publish the corresponding results. The work was supported by the Grant Agency of Czech Republic, Projects No. P201/10/0472 and by the Grant Agency of Academy of Science, Project No. KJB101420801.

(Received December 8, 2009)

REFERENCES

[1] J. Bouchon, Faille, G. Lemée, A. M. Robin, and A. Schmitt: Cartes et notice des sols, du peuplement forestier et des groupements végétaux de la réserve biologique de la Tillaie en forêt de Fontainebleau. University of Orsay 1973.

[2] C. Coudun and J. C. Gegout: Quantitative prediction of the distribution and abundance of Vaccinium myrtillus (L.) with climatic and edaphic factors. J. Vegetation Sci. *18* (2007), 4, 517–524.

[3] D. J. Finney: On the distribution of a variable whose logarithm is normally distributed. J. Roy. Statist. Soc. Ser. B *7* (1941), 155–161.

[4] F. Flénet, P. Villon, and F. Ruget: Methodology of adaptation of the STICS model to a new crop: spring linseed (Linum usitatissimum, L.) Agronomie *24* (2004), 6–7, 367–381.

[5] W. H. Green: Econometric Analysis. Prentice Hall, New Jersey 2003.

[6] Ph. Guinier: Foresterie et protection de la nature. L'exemple de Fontainebleau. Rev. Forestière Française *II* (1950), 703–717.

[7] W. Härdle: Applied Non-parametric Regression. Cambridge University Press, Cambridge 1990.

[8] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan: Statistical Analysis and Modelling of Spatial Point Patterns. Wiley, New York 2008.

[9] J. Kelsall and P. J. Diggle: Kernel estimation of relative risk. Bernoulli *1* (1995), 3–16.

[10] J. Kelsall and P. J. Diggle: Non-parametric estimation of spatial variation in relative risk. Statist. Medicine *14* (1995), 2335–2342.

[11] A. B. Lawson: Statistical Methods in Spatial Epidemiology. Wiley, Chichester 2001.

[12] E. L. Lehmann: Theory of Point Estimation. Wadsworth &Brooks, California 1991.

[13] T. Mrkvička: Estimation variances for Poisson process of compact sets. Adv. Appl. Prob. (SGSA) *33* (2001), 765–772.

[14] T. Mrkvička: Estimation variances for parameterized marked point processes and for parameterized Poisson segment processes. Comment. Math. Univ. Carolin. *45,1* (2004), 109–117.

[15] T. Mrkvička: Estimation of intersection intensity in Poisson processes of segments. Comment. Math. Univ. Carolin. *48* (2007), 93–106.

[16] T. Mrkvička, S. Soubeyrand, and J. Chadœuf: Goodness-of-fit Test of the Mark Distribution in a Point Process with Non-stationary Marks. Research Report 36, Biostatistics and Spatial Processes Research Unit. INRA, Avignon 2009.

[17] N. de Noblet-Ducoudré, S. Gervois, P. Ciais, N. Viovy, N. Brisson, B. Seguin, and A. Perrier: Coupling the soil-vegetation-atmosphere-transfer scheme ORCHIDEE to the agronomy model STICS to study the influence of croplands on the european carbon and water budgets. Agronomie *24* (2004), 6–7, 397–407.

[18] A. Penttinen, D. Stoyan, and H. Hentonnen: Marked point processes in forests statistics. Forest Sci. *38* (1992), 4, 806–824.

[19] J. Y. Pontailler, A. Faille, and G. Lemee: Storms drive successiinal dynamics in natural forests: a case study in Fontainebleau forest (France). Forest Ecology and Management *98* (1997), 1–15.

[20] B. W. Silverman: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London 1986.

[21] D. Stoyan, W. S. Kendall, and J. Mecke: Stochastic Geometry and Its Applications. Second edition. John Wiley and Sons, New York 1995.

[22] P. Van Bodegom, P. H. Verburg, A. Stein, S. Adiningsih, H. A. C. Denier Van Der Gon: Effects of interpolation and data resolution on methane emission estimates from rice paddies. Environ. Ecol. Statist. *9* (2002), 5–26.

Tomáš Mrkvička, Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia, Studentská 13, 37005 České Budějovice. Czech Republic.
e-mail: mrkvicka@prf.jcu.cz

François Goreaud, Cemagref, Campus des Cézeaux, 2' Avenue de Landais, BP 50085, 63172 Aubiere, cedex. France.
e-mail: goreaud@avignon.inra.fr

Joël Chadœuf, INRA-Biométrie, Domaine St Paul,84914 Avignon, cedex 9. France.
e-mail: joel@avignon.inra.fr