Two step estimation for Neyman-Scott point process with inhomogeneous cluster centers

Tomáš Mrkvička, Milan Muška, Jan Kubečka

May 2012

Motivation

- Study of the influence of covariates on the occurrence of fish in the inland reservoir.
- Study the interaction of the fish on the small scale.

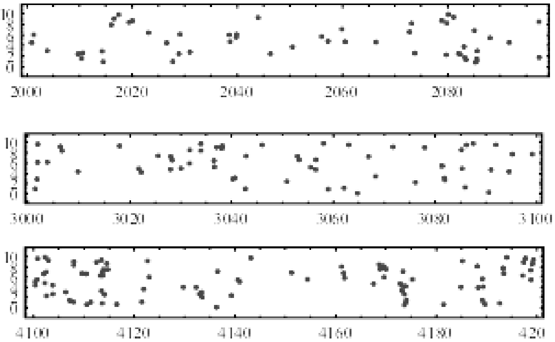


FIGURE: Three small parts of the fish positions.

Model

- Homogeneous Neyman-Scott process

κ - The intensity of the Poisson point process which forms the cluster centers.

α - The mean number of point per cluster.

ω - The size of the clusters. $k(\cdot, \omega)$ is a probability density function parameterized by ω which determines the spread of daughter points around cluster center.

If $k(\cdot, \omega)$ is symmetric normal distribution then the process is called modified Thomas process.

Inhomogeneity

- Inhomogeneous Neyman-Scott processes.

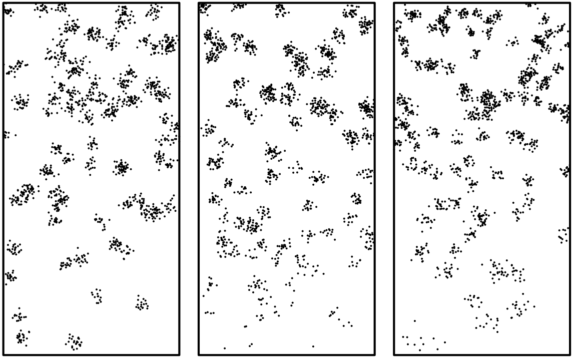


Figure: Three different types of inhomogeneities. Left : the cluster centers are thinned, center : the daughter points are thinned, right : the scale depends on the location

Inhomogeneity

- The clusters correspond to fish families or shoals which keep together and which are assumed to be homogeneous under similar environmental conditions.

- Therefore the inhomogenity is modeled by inhomogeneous cluster centers.

- Thus C the process of cluster centers is an inhomogeneous Poisson process with intenzity function

$$\rho_\beta(u) = \kappa \exp(z(u)\beta^T), \ u \in \mathbb{R}^2, \tag{1}$$

where $z = (z_1, \dots, z_k)$ is the covariate vector and $\beta = (\beta_1, \dots, \beta_k)$ is a regression parameter.

- The intenzity of the Neyman-Scott point process with inhomogeneous cluster centers is then

$$\lambda(u) = \alpha \mathbb{E} \sum_{c \in C} k(u-c, \omega) = \alpha \int k(u-c, \omega) \rho_\beta(c) dc, \ u \in \mathbb{R}^2. \tag{2}$$

Covariates

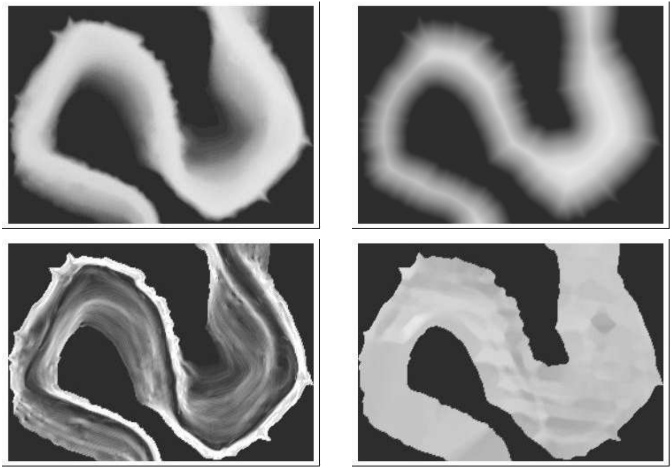


Figure: Four covariates, depth of the reservoir, distance from the bank, steepness of the bottom and light radiation. (Lighter colors correspond to the higher values.)

Methods of parametr estimation

1. likelihood-based inference - computationally very demanding and it is not straightforward to implement.
2. Two-step estimation methods
   - 2.1 First step : inhomogenity parameters are estimated by Poisson log likelihood function.
   - 2.2 Second step : clustering parameters are estimated.
     - 2.2.1 Minimum contrast method, where the contrast is measured on the K-function which is modified to be homogeneous under our model.
     - 2.2.2 Composite likelihood method.
     - 2.2.3 Bayesian method.

First step

- We approximate the intensity of X by

$$\rho_{\overline{\beta}}(u) = \exp(\overline{z}(u)\overline{\beta}^T), \; u \in \mathbb{R}^2, \tag{3}$$

where $\overline{z}(u) = (1, z_1, \ldots, z_k)$ and $\overline{\beta} = (\log(\alpha\kappa), \beta_1, \ldots, \beta_k)$.

- This approximation is intuitively justified if the range of interaction among the points is small with respect to range of changes of spatial covariates $z(u)$.

- The Poisson log likelihood function is used to estimate $\overline{\beta}$.

- It means, that we maximize the score function

$$l(\overline{\beta}) = \sum_{u \in X \cap W} \overline{z}(u)\overline{\beta}^T - \int_W \exp(\overline{z}(u)\overline{\beta}^T) du \tag{4}$$

Here W is the observation window.

Minimum contrast method

- The second order product density of the Neyman-Scott point process with inhomogeneous cluster centers is

$$\rho^{(2)}(u,v) = \lambda(u)\lambda(v)+\alpha^2 \int k(u-c,\omega)k(v-c,\omega)\rho_\beta(c)dc,\ u,v \in \mathbb{R}^2, \tag{5}$$

- The pair correlation function is

$$g(u,v) = 1+\frac{\int k(u-c,\omega)k(v-c,\omega)\rho_\beta(c)dc}{\int k(u-c,\omega)\rho_\beta(c)dc \int k(v-c,\omega)\rho_\beta(c)dc},\ u,v \in \mathbb{R}^2, \tag{6}$$

Minimum contrast method

The $g(u, v)$ can be approximated by

$$g(u,v) \sim 1 + \frac{\rho_\beta(\frac{u+v}{2})}{\rho_\beta(u)\rho_\beta(v)} \int k(u-c,\omega)k(v-c,\omega)dc, \ \ u,v \in \mathbb{R}^2. \tag{7}$$

The function $h(u,v,\omega) = \int k(u-c,\omega)k(v-c,\omega)dc$ depends only on the difference $u - v$ and it will be our homogeneous characteristic $(h(u,v,\omega) = h(v-u,\omega))$.

Integrate the $h(t,\omega)$ similarly like in the definition of K function

$$H(r,\omega) = \int_{\|t\| \leq r} h(t,\omega)dt, r \geq 0. \tag{8}$$

Minimum contrast method

The $H(r,\omega)$ can be computed, for example for Thomas process $H(r,\omega) = 1 - \exp(\frac{-r^2}{4\omega^2})$.

On the base of approximation 7 we have

$$\int_{\|u-v\|\leq r} (g(u,v)-1)\frac{\rho_\beta(u)\rho_\beta(v)}{\rho_\beta(\frac{u+v}{2})} dudv \sim H(r,\omega). \tag{9}$$

Since $\rho_\beta(u) = \rho_{\overline{\beta}}(u)/\alpha$ and $\rho_{\overline{\beta}}(u)$ is estimated in the first step, the left hand side of 9 can be estimated by

$$\sum_{x,y\in X}^{\neq} \frac{I_{\|x-y\|\leq r}}{\alpha\rho_{\overline{\beta}}(\frac{x+y}{2})|W \cap W_{x-y}|} - \int_{\|u-v\|\leq r} \frac{\rho_{\overline{\beta}}(u)\rho_{\overline{\beta}}(v)}{\alpha\rho_{\overline{\beta}}(\frac{u+v}{2})}. \tag{10}$$

Minimum contrast method

The unknown parameter α can be given out and we get that the homogeneous characteristic $\alpha H(r, \omega)$ can be estimated by

$$\alpha \widehat{H(r,\omega)} = \sum_{x,y \in X}^{\neq} \frac{I_{\|x-y\| \leq r}}{\rho_{\overline{\beta}}(\frac{x+y}{2})|W \cap W_{x-y}|} - \int_{\|u-v\| \leq r} \frac{\rho_{\overline{\beta}}(u)\rho_{\overline{\beta}}(v)}{\rho_{\overline{\beta}}(\frac{u+v}{2})}. \tag{11}$$

Note here that the second term is not estimated from the points of X, but it can be numerically integrated from estimated $\rho_{\overline{\beta}}(u)$.

The estimates of α and ω are then obtained by minimizing

$$\int_{R_l}^{R_u} (\alpha \widehat{H(r,\omega)} - \alpha H(r,\omega))^2 dr,$$

where R_l and R_u are user specified constants.

Composite likelihood method

The estimate of the interaction parameters is obtained by maximizing the composite likelihood, which is defined by :

$$CL(\alpha, \omega) = \sum_{x \neq y \in X \cap W, \|x-y\| < R} [\log \rho^{(2)}(x-y) -$$

$$- \log \left(\int_W \int_W \rho^{(2)}(u-v) I(\|u-v\| < R) dudv \right)],$$

here R is the user specified constant. And the intensity function estimated in the first step is plug in the second order product density $\rho^{(2)}$ computed for our model in Formula 5.

Similarly like composite likelihood it possible to use Palm likelihood. Since those two method seems to get similar results, we worked only with composite likelihood (Prokešová & Jensen 2011) .

Bayesian approach

- C is the inhomogeneous point process of cluster centers with the intensity $\rho_{\overline{\beta}}/\alpha$,
- $p(C|\alpha)$ is the probability density of the point process C under the knowledge of α with respect to homogeneous Poisson point process
- and $p(X|C,\alpha,\omega)$ is the probability density of the point process X with respect to homogeneous Poisson point process under the knowledge of C and all parameters.

$$p(X|C,\alpha,\omega) = \exp(|W| - \int_W \widetilde{\lambda}(u)du) \prod_{x \in X} \widetilde{\lambda}(x), \qquad (12)$$

here $\widetilde{\lambda}(u) = \alpha \sum_{c \in C} k(u-c,\omega)$.

Bayesian approach

The joint posterior distribution of the of the process X and the parameters is then

$$p(C, \alpha, \omega | X) \propto p(X | C, \alpha, \omega) p(C | \alpha) p(\alpha) p(\omega). \tag{13}$$

Here $p(\alpha)$ and $p(\omega)$ denote the probability densities of priors.

- Two different updates of MCMC are needed.

1) Update for centers C - Birth-Death-Move algorithm.

2) Update for parameters of interest α, ω - Metropolis-Hastings algorithm.

The Bayesian point estimates of α and ω are then the expected values of the posterior distribution.

Simulation study

inhomogeneous intensities - smooth and wavy intensity. Both intensities are given as a combination of two covariates.

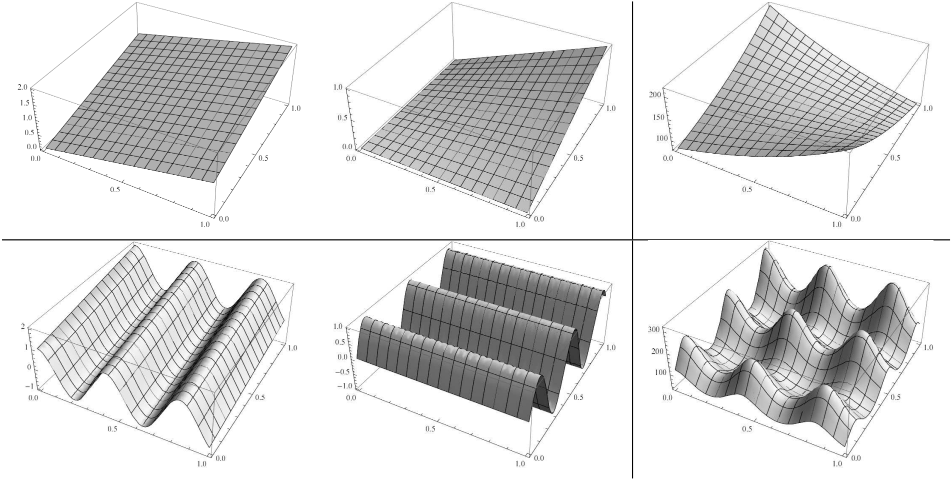


FIGURE: The covariates (first and second column), the intensity (third column).

Simulation study

Parameters :

$\kappa = 80, \alpha = 2.5, \omega = 0.02$	$\kappa = 80, \alpha = 2.5, \omega = 0.04$
$\kappa = 26.66, \alpha = 7.5, \omega = 0.02$	$\kappa = 26.66, \alpha = 7.5, \omega = 0.04$

This gives us in mean 334 points for first intensity and 304 points for second intensity.

We performed 100 simulations for all 8 combinations of parameters

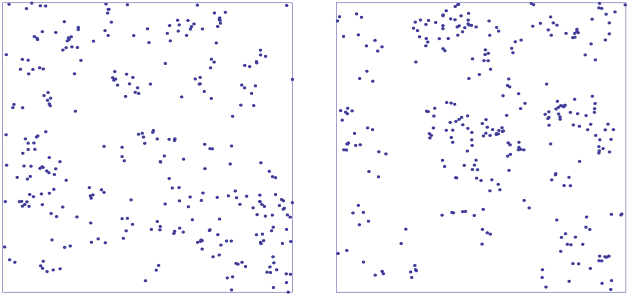


FIGURE: Realizations for two considered inhomogeneities.

Intensity	smooth	smooth	smooth	smooth	wavy	wavy	wavy	wavy
$\kappa\alpha$	200	200	200	200	200	200	200	200
β_1	1	1	1	1	0.4	0.4	0.4	0.4
β_2	-2	-2	-2	-2	0.6	0.6	0.6	0.6
κ	80	80	26.66	26.66	80	80	26.66	26.66
α	2.5	2.5	7.5	7.5	2.5	2.5	7.5	7.5
ω	0.02	0.04	0.02	0.04	0.02	0.04	0.02	0.04
First step								
Mean $\widehat{\kappa\alpha}$	216.7	198.7	224.6	221.7	202.7	209.1	203.1	211.5
SD $\widehat{\kappa\alpha}$	93.34	70.81	137.0	135.2	31.61	32.83	48.83	46.75
MSE $\widehat{\kappa\alpha}$	8911	4961	19214	18608	997.5	1149	2372	2300
Mean $\hat{\beta}_1$	0.984	1.110	1.064	1.067	0.391	0.344	0.368	0.350
SD $\hat{\beta}_1$	0.572	0.612	1.028	0.962	0.141	0.139	0.224	0.182
MSE $\hat{\beta}_1$	0.324	0.383	1.053	0.922	0.020	0.022	0.051	0.035
Mean $\hat{\beta}_2$	-2.022	-2.271	-2.065	-2.115	0.545	0.504	0.582	0.493
SD $\hat{\beta}_2$	0.997	1.216	1.886	1.774	0.162	0.141	0.257	0.217
MSE $\hat{\beta}_2$	0.986	1.537	3.529	3.132	0.029	0.029	0.050	0.058
Min. Contrast								
Mean $\hat{\alpha}$	2.497	3.763	6.949	5.845	2.230	3.702	7.304	5.745
SD $\hat{\alpha}$	1.125	2.227	2.123	3.857	0.830	2.378	2.320	3.928
MSE $\hat{\alpha}$	1.253	6.503	4.770	17.48	0.756	7.045	5.374	18.37
Mean $\hat{\omega}$	0.180	0.161	0.058	0.184	0.170	0.238	0.054	0.183
SD $\hat{\omega}$	0.333	0.297	0.189	0.312	0.330	0.355	0.185	0.296
1000 × MSE $\hat{\omega}$	135.5	102.2	37.1	117.3	130.2	163.7	34.98	107.2
Composite Lik.								
Mean $\hat{\alpha}$	3.090	6.783	8.613	7.132	3.350	5.306	8.220	7.500
SD $\hat{\alpha}$	1.839	4.261	2.831	4.463	2.418	3.901	4.223	4.320
MSE $\hat{\alpha}$	3.695	36.34	9.173	19.86	6.396	22.95	17.08	18.00
Mean $\hat{\omega}$	0.0213	0.0643	0.0194	0.0398	0.0215	0.066	0.0187	0.0417
SD $\hat{\omega}$	0.0057	0.0189	0.0024	0.0126	0.0062	0.019	0.0026	0.011
1000 × MSE $\hat{\omega}$	0.035	0.948	0.006	0.159	0.040	1.064	0.008	0.133
Bayesian								
Mean $\hat{\alpha}$	2.724	4.168	7.769	7.679	2.697	3.556	7.815	7.578
SD $\hat{\alpha}$	0.513	2.175	0.815	1.786	0.387	1.234	0.930	1.505
MSE $\hat{\alpha}$	0.310	7.471	0.730	3.191	0.185	2.622	0.903	2.190
Mean $\hat{\omega}$	0.0207	0.0494	0.0201	0.398	0.0200	0.0447	0.0204	0.0401
SD $\hat{\omega}$	0.0023	0.0146	0.0010	0.0045	0.0016	0.0078	0.0018	0.0029
1000 × MSE $\hat{\omega}$	0.005	0.300	0.001	0.021	0.0025	0.0829	0.003	0.008

The results of the simulation study First step

1. The estimation of inhomogeneity parameters perform well in all cases.
2. The results of the first step of the estimation procedure is better for less clustered processes.
3. The wavy inhomogeneity structure does not bring (with respect to smooth inhomogeneity structure) deterioration of the performance of estimation of inhomogeneity parameters neither the interaction parameters.
4. Thus the assumption, that the range of interaction is small with respect to the range of changes of covariates, is not completely crucial.

The results of the simulation study Second step

1. The Bayesian method performs best.
2. But this method is rather computationally demanding with many implementation pitfalls.
3. For the two remaining simpler methods, the minimum contrast method performs better for the estimation of α and the composite likelihood method performs better for the estimation of ω.
4. But both simpler methods are quite sensitive for the choice of tuning parameters.

Fish spatial distribution

- 4351 fish recorded in the representative middle part of the reservoir.
- The fish were recorded along the trace of the boat, which was 12 km long.
- The fish were recorded in the distance 10 to 20 meters from the boat and in the depth 1 to 1.75 meters.

Covariates

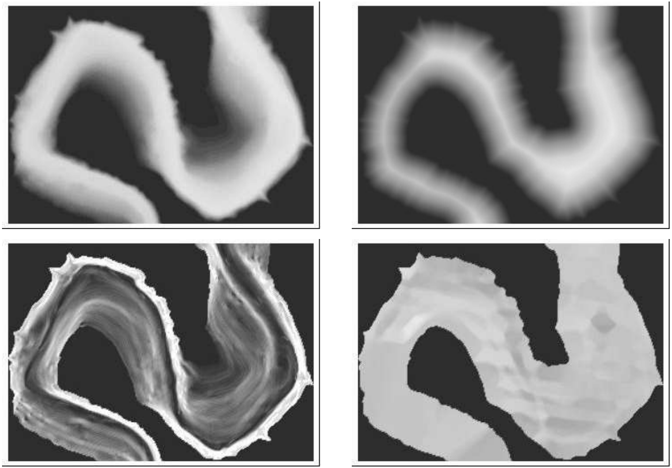
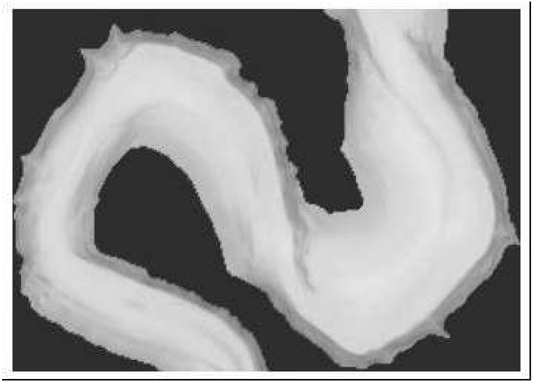


FIGURE: Four covariates, depth of the reservoir, distance from the bank, steepness of the bottom and light radiation. (Lighter colors correspond to the higher values.)

Estimated inhomogeneity intensity function



The estimated parameters with their 95% confidence intervals.

Parameters	$\kappa\alpha$	Depth	Distance to bank	Steepness	Radiation	α	ω
Estimates	0.0304	0.0039	0.0038	- 0.0147	0.0574	4.57	3.76
Standard dev.	0.0515	0.0108	0.0021	0.0163	0.0818	0.41	0.21
Lower bound	0.0181	- 0.0407	- 0.0033	- 0.0649	- 0.05		
Upper bound	0.2214	0.0127	0.0058	- 0.0056	0.015		

Testing complete spatial randomness

- The method of (Brix et. al. 2001) was chosen since it tests only the Poisson assumption and does not test the goodness of fit of the inhomogeneous function.
- The resulted p-value of this test is less than 10^{-6}.
- Thus we clearly reject the hypothesis of independent structuring of the fish in the reservoir.
- Since the shorter nearest-neighbor distances appear more often than it should be under the Poisson hypothesis, the clustering structure of the fish is evident.

Estimation of interaction parameters

- Since the Bayesian method is the most accurate method, we use it.

- Finally we performed the parametrical bootstrap to obtain the confidence intervals of the estimated inhomogeneous parameters. We simulated 250 inhomogeneous Thomas processes with estimated parameters.

Conclusions

- The properties of the two step estimation procedures for the Neymann-Scot process with inhomogeneous cluster centers were studied.
- Since we use some approximation of the first order intenzity function in the first step, we have to rely on the simulation study only.
- The first step, the estimation of inhomogeneity parameters performs reasonably well.
- For the second step we introduced 3 estimation procedures.
- The Bayesian method reveals the best and the most stable results in our simulation study.
- Therefore we chose this method and applied it to fisheries data set, which was the motivation of this study.

Conclusions for real data

- The clustering structure of fish were proven.
- The mean number of fish in the cluster was estimated to 4.57.
- The steepness of the ground is the only significant covariate.