Bayesian model-based spatiotemporal survey design for Log Gaussian Cox process

Jarno Vanhatalo

Assistant Professor of statistics
Department of Mathematics and Statistics, Faculty of Science, and Organismal and Evolutionary Biology Research Programme, Faculty of Bio- and Env. Sciences University of Helsinki, Finland
www.helsinki.fi/~jpvanhat
Outline of the talk

• Bayesian hierarchical models
• Species distribution modeling with Log Gaussian Cox processes
• Model based survey design
• New sampling method for species distribution modeling
Hierarchical Bayesian modeling

• **A process model:** Let $f(x, \theta)$ denote a process that is a function of variables / covariates / drivers $x$ and parameters $\theta$
  - "Statistical" linear model
    
    $f(x, \theta) = \theta_1 + \theta_2 x$
  - Stochastic time series
    
    $f_t = L(x, \theta) f_t + \epsilon_t$
  - Systems of (stochastic) partial differential equations, such as a diffusion model
    
    $$\frac{\partial f(x, t)}{\partial t} = \frac{\partial}{\partial x_1} \left( \delta(x) \frac{\partial f(x, t)}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left( \delta(x) \frac{\partial f(x, t)}{\partial x_2} \right) + \theta f(x, t)$$

• **Observation model:** Let $p(y|f(x, \theta), \sigma)$ denote a model describing how measured data $y = [y_1, \ldots, y_n]$ is related to the process model
  - A regular Gaussian noise $p(y|f(x, \theta), \sigma) = \prod N(y_t|f_t, \sigma)$
  - In principle any kind of relation is possible
  - The observation model gives us the likelihood function

• **Parameter model:** Let $p(\theta)$ denote a prior probability distribution for the parameters
  - Describes the *a priori* relative plausibility of different parameter values
Bayes rule

• Deterministic process model

\[ p(\theta, \sigma|y, x) = \frac{p(y|f(\theta, x), \sigma)p(\theta)}{p(y)} \]

• Stochastic process model

\[ p(\theta, \sigma, f|y, x) = \frac{p(y|f(\theta, x), \sigma)p(f|\theta)p(\theta)}{p(y)} \]
Sea surface salinity and temperature in Kara Sea, statistical analysis

- **Data**
  - Point wise insitu observations on temperature and salinity
  - River inflow
  - Ice cover
  - Arctic oscillation index

- **Model**
  - \( p(y|f, \theta, \sigma) = \prod N(y_i|f_i, \sigma) \)
  - \( f(x, s, t\theta) = \theta_1(s, t) + \theta_2(s, t)x(s, t) \)
Spatially varying coefficients and temporal changes
Point process model for species distribution

• denote by s a spatial location (e.g. longitude-latitude coordinates – and possibly depth)

• $x_s$ the vector of environmental covariates associated with that location (e.g. depth, surface roughness, temperature etc.)

• A key component of point process models is the intensity function $\lambda(s, x_s)$ and observation process (sampling effort) $z(s, x_s)$
  • the “probability” that one individual is present at location s”

• [#animals in region A] $\sim$ Poisson $\left( \int_A \lambda(s, x_s)ds \right)$

• [#observed animals in region A] $\sim$ Poisson $\left( \int_A \lambda(s, x_s)z(s, x_s)ds \right)$
**SDM with point process models**

**Count data from survey transect**

<table>
<thead>
<tr>
<th>Id</th>
<th>Long</th>
<th>Lat</th>
<th>Effort, z</th>
<th>Count observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Survey transect**

**Individual observations along transect**

**Likelihood for data on surveyed grid cells**

\[
L(y_1, \ldots, y_n | \lambda, z) = \prod_{i=1}^{5} \text{Poisson}(\lambda(s_i, x_i)z(s_i, x_i)A_i)
\]

**Prediction for unsurveyed grid cell**

\[
\log \lambda(s, x_s) = f(s, t, x_{s,t}) = \alpha + x_s^T \beta + \varphi(s) + \rho(s, t)
\]

**Model for intensity**

- **Intercept**
- **Environmental covariates**
- **Spatial random effect**
- **Spatiotemporal process**
Arctic Marine mammals: Accounting for unequal sampling effort (Mäkinen and Vanhatalo, 2018)

• If we don’t know the sampling effort \( z(s, x_s) \), we can model it

• Allows also analysing presence only data
  • However, requires prior assumptions on sampling effort to be formulated in \( z(s, x_s) \)
Arctic Marine mammals: Distributional changes under climate change (Mäkinen and Vanhatalo, in press)

- Columns 1-4: seasonal average relative densities in 1997 – 2013
- Column 5: Change in spring relative densities between 1997 – 2004 (high ice concentration) and 2005 – 2013 (low ice concentration).
Spatio-temporal fluctuations of Crown of thorns star fish (COTS) in the Great Barrier Reef (Vanhatalo et al., 2017)

Long term spatiotemporal data

COTS eat coral

Observation: # number of COTS on an area covered by diver

Manta tow sampling
Spatio-temporal fluctuations of COTS in the Great Barrier Reef

Predictions on a raster covering the reefs and along the reef line.

**Effect of fishing closures**

**Spatial changes**

- **Source of outbreak**

**Spatial-temporal changes**

- Median of spatio-temporal relative change in COTS intensity
- Coefficient of variation of spatio-temporal relative change in COTS intensity
- Mean relative change in COTS intensity across GBR
- Probability of COTS outbreak
Case study: Coastal fish reproduction areas
(Kallasvuuo et al., 2017)

- Fish usually have the most specific habitat demands during spawning and early-life stages
- Commercial and recreational fisheries are strongly linked to the ecological state of these habitats.
- The reproduction areas are also heavily exploited and threatened by anthropogenic pressures
→ there is a growing need for better understanding on
  - current reproduction habitats
  - predictions concerning changes in these habitats

Data
- years 2007-2014
- 655 distinct sampling sites
  - stratified according to bottom depth and exposure
- Sampled with tow net traveling in the surface water (~1.5 m dept)
Coastal fish: abundance vs. occurrence

• What extra does intensity tell us compared to probability of occurrence?

• Let’s divide the study area into three classes
  • **Unsuitable** for reproduction:
    • $P(\text{catch larvae}) < 0.5$
  • **Suitable** for reproduction
    • $P(\text{catch larvae}) > 0.5$
  • **Important** (and suitable) for reproduction
    • the smallest area whose expected cumulative intensity is 80% of the total expected intensity over study area
Coastal fish: abundance vs. occurrence

Perch

Baltic herring

Smelt

60°N

20°E

65°N

I

II

III

IV

V

Not suitable

Suitable

Important
Coastal fish: abundance vs. occurrence

<table>
<thead>
<tr>
<th></th>
<th>Total number of larvae in 1.5 m deep water column, $\times 10^9$ mean (95% credible interval)</th>
<th>Percent of water area suitable for larvae</th>
<th>Percent of water area producing 80% of larvae</th>
</tr>
</thead>
<tbody>
<tr>
<td>perch</td>
<td>1.56 (0.89, 2.55)</td>
<td>13.66</td>
<td>3.03</td>
</tr>
<tr>
<td>pikeperch</td>
<td>0.54 (0.12, 1.56)</td>
<td>3.68</td>
<td>1.37</td>
</tr>
<tr>
<td>Baltic herring</td>
<td>8.72 (5.65, 12.86)</td>
<td>99.79</td>
<td>52.89</td>
</tr>
<tr>
<td>smelt</td>
<td>5.91 (2.88, 10.81)</td>
<td>22.50</td>
<td>4.44</td>
</tr>
</tbody>
</table>

Modelling abundance matters:
- very limited areas can be suitable for fish production (*occurrence model*)
- The area that contributes most to the fish production can be even more restricted (*abundance model*)

Threshold for important areas should depend on the problem
- Here we used 80% but
  - the value could be higher for endangered species, lower for common species
  - In some cases we should account also for connectivity between sites
Survey design

\[ p(\theta, \sigma, f|y) = \frac{p(y|f, \theta, \sigma)p(f|\theta)p(\theta)}{p(y)} \]

- The inference (estimation) on model parameters (and the process) is directly influenced by
  - the observation model and data
    \[ p(y|f, \theta, \sigma) \]
  - But also of the prior knowledge
    \[ p(f|\theta)p(\theta) \]
- If the prior information on parameters or process components is strong there is not much to learn from the data
  - we should plan surveys so that they inform most the uncertain parameters and components.
  - Informative prior is typically based on earlier data
  - More generally, all parameters might not be equally important
Survey design for log Gaussian Cox processes

• Where (and when) to sample?
• I will denote by $d_n = \{s_1, ..., s_n\}$ a design with $n$ sampling locations and by $y(d_n)$ data to be sampled at these locations
• We compare the goodness of different designs with a utility function:

$$U(d_n, y(d_n), \theta, \sigma, f)$$

• Here, I will consider two typical utilities in spatial statistics

  • Average point-wise variance: $U(d_n, y(d_n)) = - \int_{s \in A} Var[f(s) | y(d_n)] \, ds$

  • Increase in information $U(d_n, y(d_n)) = KL[p(f_A(s) | y(d_n)) || p(f_A(s))]$
Survey design for log Gaussian Cox processes

• The future observations are uncertain so we want to find a design that has high expected utility

\[
\bar{U}(d_n) = E[U(d_n, y(d_n))]
\]

• Two main strategies
  • Choose the best design among ”intuitively good” designs
  • Try to optimize the design

\[
d_n = \arg \max_{d_n \in D} \bar{U}(d_n)
\]
  • Difficult. For example in case of spatial design \( D \subset R^{2^n} \)
Intuitively good survey design for LGCP

• Typical spatial sampling design is a space filling (quasi) random design
  • The aim is to locate the design points so that they fill the domain ”evenly”
  • Theoretically optimal under Gaussian model with a priori uniform variance throughout the region

• With LGCP this does not apply if we have a priori information on the shape of the intensity function.
  • This happens when we have earlier data
SDM with point process models

\[ \log \lambda(s, x_s) = f(s, t, x_{s,t}) = \alpha + x_s^T \beta + \varphi(s) + \rho(s, t) \]

- We give Gaussian prior for the linear weights
  - \( \alpha \sim N(0, \sigma^2) \)
  - \( \beta \sim N(0, \sigma^2 I) \)
- and Gaussian process prior for the random effects
  - \( \varphi(s) \sim GP(0, k(s, s')) \)
  - \( \rho(s, t) \sim GP(0, k((s, t), (s', t')) ) \)
- This implies that the joint prior of log intensities at any collection of survey locations is multivariate Gaussian
  \[
  \log \lambda = f \sim N(0, K)
  \]
  - Where \( K = K_\alpha + K_\beta + K_\varphi + K_\rho \)
Intuitively good survey design for LGCP

• Laplace approximation for the posterior

\[ p(f \mid y_1, \ldots, y_n) \approx N(f \mid m_p, K_p) \]

\[ \Rightarrow p(\lambda \mid y_1, \ldots, y_n) = \log - N(m_p, K_p) \]

\[ \Rightarrow Pr(y(d_n) = 0 \mid y_1, \ldots, y_n) \approx 0 \text{ if } m_p \ll 0 \]

• And in this case we do not expect to learn anything from \(y(d_n)\)

• \(\Rightarrow\) When allocating sampling locations give more weight to regions with high a priori expected intensity
Spatially balanced design with rejection sampling

1. Randomly generate a location \((s, t)\) within the study domain using some traditional spatially balanced design.

2. Accept the location with probability
   \[
   p(s, t) \propto \min(p_{\text{max}}, e^{m_p(s,t) + \sigma_p^2(s,t)}).
   \]
   If accepted, set \((s, t)\) into design. If rejected return to step 1.

3. Repeat steps 1-2 until the size of design reaches to \(n\).
Case study: New sampling design to improve estimates on fish reproduction areas in Porvoo region

Results based on current data:
Calendar day effect on intensity and the spatial intensity at peak larval week

Black dots are current data sampling locations
Red represents the new sampling region
Comparing spatially balanced designs with rejection sampling and regular spatially balanced designs

The inclusion probability in the rejection sampling phase

The best balanced design with rejection sampling and pure balanced design

Table 1: Weekly distribution of the design points (number of sampling locations) for pike perch (Random design) and herring (Halton design) sampling.

<table>
<thead>
<tr>
<th>Calendar Day</th>
<th>With rejection sampling (pike perch / herring)</th>
<th>Without rejection sampling (pike perch / herring)</th>
</tr>
</thead>
<tbody>
<tr>
<td>101-128</td>
<td>0 / 0</td>
<td>23 / 23</td>
</tr>
<tr>
<td>129-156</td>
<td>11 / 9</td>
<td>22 / 18</td>
</tr>
<tr>
<td>157-184</td>
<td>65 / 78</td>
<td>21 / 19</td>
</tr>
<tr>
<td>185-212</td>
<td>19 / 13</td>
<td>20 / 20</td>
</tr>
<tr>
<td>213-240</td>
<td>4 / 0</td>
<td>14 / 20</td>
</tr>
</tbody>
</table>
Expected utility

- We did also extensive simulation study where the balanced designs with rejection sampling worked clearly better than the corresponding traditional balanced designs with both
  - the average predictive variance and
  - KL-divergence utility
Discussion

• The proposed design algorithm is easy to implement even without statistical skills
  • The sampling is conducted by local administrative people for whom we need to give simple instructions
  • The rejection sampling can be implemented by instructing people to allocate sampling locations based on probability maps.

• Currently we work on methods to optimize the design
  \[ \bar{d}_n = \arg \max_{d_n \in D} \overline{U}(d_n) \]
  • More on this sometime later
Thank you!

References


Funding and collaboration