Model based Bayesian spatio-temporal survey design for species distribution modelling

Jia Liu

joint work with Jarno Vanhatalo, University of Helsinki

University of Helsinki

Bayesian Statistics in the Big Data Era
in CIRM, Marseille

29 November 2018

Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

Motivation

- A central question of geostatistics is the prediction of spatial patterns over a ROI using data measured at finite set of locations. —A hierarchical Gaussian process model
- When the data are not fully observed, with a suitable model, the goodness of the spatial prediction and estimation depend on the spatial allocation of the measurement locations [Müller, 2007], i.e. *observational/experimental design*.
- The design in spatial data analysis– the spatial/spatiotemporal allocation of the data.
- Gaussian *v.s. non-Gaussian observation processes in spatial analysis.*
- *We study observational designs for spatiotemporal log-Gaussian Cox processes (LGCPs).*

LGCPs

- Why LGCPs?
- A LGCP arises from an inhomogeneous Poisson process with intensity λ whose logarithm has a Gaussian process.
- In terms of the spatialtemporal observation design, the key question is when and where we should do the survey in order to learn most of the essentials of $\lambda(\boldsymbol{s}, t)$.

  The interest is the saptiotemporal varies over the intensity surface.

Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

- Denote the study region by $\mathcal{D}$, and a vector of spatiotemporal covariates by $\boldsymbol{x} = [\boldsymbol{s}^T, t] \in \mathcal{D}$.
- The (approximate) likelihood can be written

$$\begin{aligned} L(y_1, \ldots, y_n | \lambda(\cdot)) &= L(y_1, \ldots, y_n | \lambda(\boldsymbol{x}_i)) \\ &= \prod_{i=1}^{n} \text{Poisson}(y_i | \lambda(\boldsymbol{x}_i)) \end{aligned} \tag{1}$$

where n is the number of observed discretized locations and y_i is the count observation at i'th location $\boldsymbol{x}_i$ and $\log(\lambda(\boldsymbol{x})) = \boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]^T$ is a vector of latent variables at those locations, and has a multivariate Gaussian distribution.

The additive model for spatiotemporal Gaussian process prior

- *Additive model*

$$\log \lambda(\boldsymbol{x}) = f(\boldsymbol{s}, t) \sim GP(\mu(\boldsymbol{s}, t), k(\boldsymbol{s}, \boldsymbol{s}') + k(t, t')). \tag{2}$$

$$f(\boldsymbol{s}, t) = \mu(\boldsymbol{s}, t) + g(\boldsymbol{s}) + h(t),$$

  where the additive terms are mutually independent Gaussian processes. $g(\boldsymbol{s}) \sim GP(0, k(\boldsymbol{s}, \boldsymbol{s}'))$ and $h(t) \sim GP(0, k(t, t'))$.

- Choices of covariance functions (e.g., Martèn, square exponential, etc.).
- Laplace approximation for posterior inference.
- GPstuff [Vanhatalo et al., 2013] software *v.s. other alternatives.*

Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

- What is the design?
- What is the problem that arises from the design?
- How to evaluate the design?

Expected utilities

We will denote by $D_n = \{d_n\}$ the set of all possible designs of size n in domain $\mathcal{D}$.

- The expected utility is then defined as

$$U(d_n) = \int_Y \int_f U(d_n, f, y) p(f|d_n, y_*) p(y|d_n) df \; dy_*, \tag{3}$$

  where $y_* \in Y$ the future data.

- The MC simulation.

Spatial balance designs

- The model-based optimal experimental design, simulated annealing algorithms [Müller, 1999, Müller et al., 2004], interactive MCMC methods [Amzal et al., 2006].
- Spatial balance design sampling methods to increase expected utilities and obtain good designs by means of good coverage rates of the survey region.

- Halton, Sobol designs, the Fibonacci lattice designs, distance based designs (simple inhibitory, inhibitory plus close pairs *lattice* designs [Chipeta et al., 2016], and the space-filling design [Nychka and Saltzman, 1998].

Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

Inclusion probability

- An even probability.
- Some locations are more informative than others.
- Rejection sampling scheme, more weights to certain covariates which are a priori expected to be more informative.

Rejection design

The general algorithm of the rejection sampling design proceeds as following:

1. Randomly generate a location $\boldsymbol{x}^*$ within the study domain (here any of the above random or quasi-random sequence can be used);
2. Calculate an inclusion probability $0 \leq p(\boldsymbol{x}^*) \leq 1$
3. Accept the location with probability $p(\boldsymbol{x}^*)$. If accepted, set $\boldsymbol{x}_j = \boldsymbol{x}^*$ and increase $j = j + 1$. If rejected, keep $j = j$ and return to step 1;
4. Repeat steps 1-3 until the size of design reaches to n.

Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

Two common utility functions in geostatistics

We will consider two commonly used utilities in geostatistics.

- *(1) The average predictive variance (APV)*

$$\Rightarrow \hat{L}_{\text{APV}}(d_n) = \frac{1}{|\mathcal{D}|} \sum_{y \in \mathcal{N}^n} p(y|d_n) \int_{\boldsymbol{x}_* \in \mathcal{D}} \text{Var}\{\lambda(\boldsymbol{x}_*)|d_n, y\} d\,\boldsymbol{x}_* \,. \tag{4}$$

- *The MC approximation of (APV)*

$$\hat{L}_{\text{APV}}(d_n) \approx \frac{1}{M} \sum_{j=1}^{M} \left[\frac{1}{N} \sum_{\boldsymbol{x}_* \in X_*} \text{Var}\{\lambda_j(\boldsymbol{x}_*)|d_n, Y_j\} d\,\boldsymbol{x}_* \right],$$

- The intensity function

$$\mu(\lambda(\boldsymbol{x}_*)) = \exp\Big(\mu(f(\boldsymbol{x}_*)) + \text{Var}(f(\boldsymbol{x}_*))/2 \Big),$$

$$\text{Var}[\lambda(\boldsymbol{x}_*)] = \Big[\exp(\text{Var}(f(\boldsymbol{x}_*)) - 1 \Big] \exp\Big(2\mu(f(\boldsymbol{x}_*) + \text{Var}(f(\boldsymbol{x}_*) \Big).$$

KL divergence

- *The mutual information*

$$U_{\mathrm{KL}}(d_n, Y) = \mathrm{KL}\Big(d\,\mathcal{P}(f(\cdot)|X, Y)||d\,\mathcal{P}(f(\cdot))\Big)$$

$$\Rightarrow \hat{U}_{\mathrm{KL}}(d_n) = \sum_{y \in \mathcal{N}^n} p(y|d_n)\mathrm{KL}\Big(d\,\mathcal{P}(f(\cdot)|X, y)||d\,\mathcal{P}(f(\cdot))\Big). \tag{5}$$

- *The Kullback-Leibler divergence (KL) [Kullback, 1987]*

$$U_{\mathrm{KL}}(d_n, y) = \frac{1}{2}\Big(\log|K_* K_{*|y}^{-1}| + tr(K_*^{-1} K_{*|y})$$

$$+ (\mu_* - \mu_{*|y})^T K_*^{-1}(\mu_* - \mu_{*|y})) - c\Big), \tag{6}$$

where c is the dimension of the covariance matrices $K_{*|y}$ and K_*.

The KL divergence

- *The KL-divergence from the prior to the posterior*

$$\mathrm{KL}\Big(d\mathcal{P}(f(\cdot)|y)||d\mathcal{P}(f(\cdot))\Big) = \int \log \frac{p(y|f(\cdot))d\mathcal{P}(f(\cdot))}{d\mathcal{P}(f(\cdot))\int p(y|f(\cdot))d\mathcal{P}(f(\cdot))} d\mathcal{P}(f(\cdot)|y)$$

$$= \int \log p(y|f(\cdot))d\mathcal{P}(f(\cdot)|y) - \log p(y)$$

$$= \int \log p(y|\boldsymbol{f})\; d\mathcal{P}(\boldsymbol{f}|y) - \log p(y), \tag{7}$$

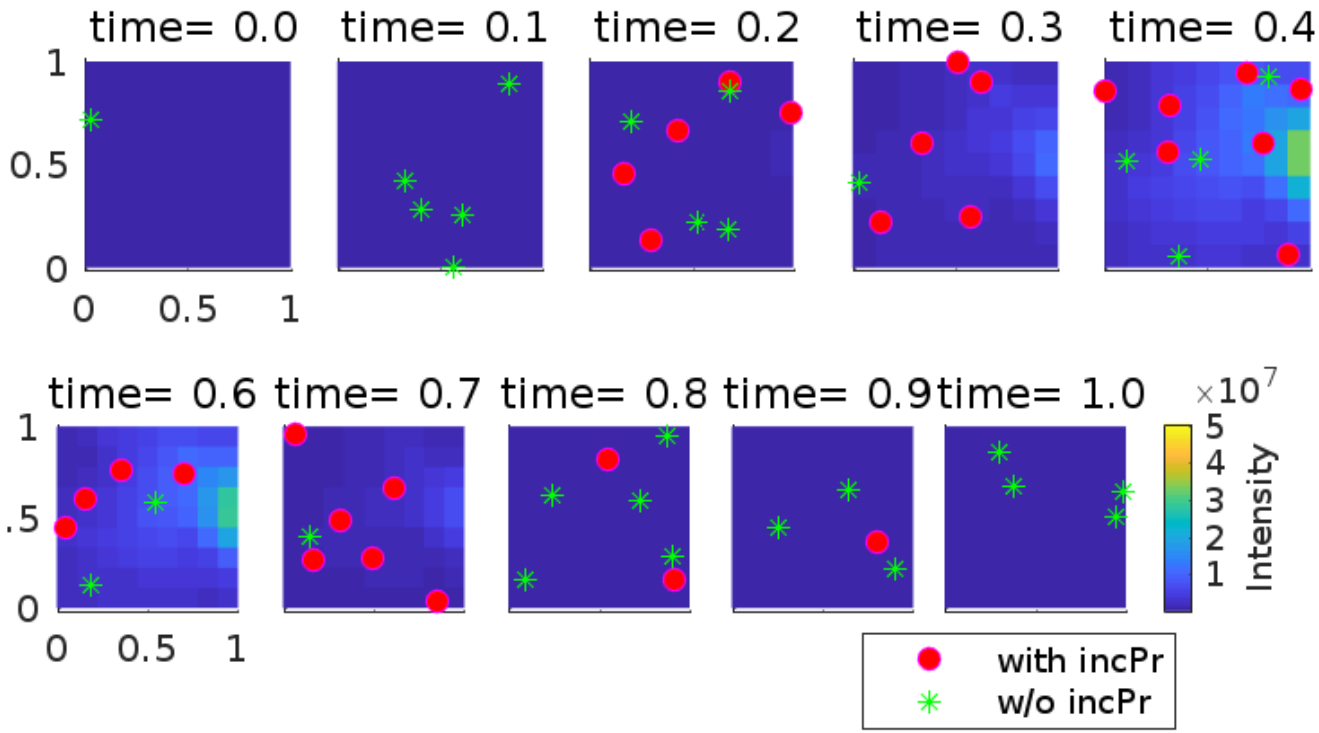
where $p(y) = \int p(y|f(\cdot))d\mathcal{P}(f(\cdot)) = \int p(y|\boldsymbol{f})p(\boldsymbol{f})\; d\boldsymbol{f}$. When $X \subset \mathcal{D}$, we get the last equality.

Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

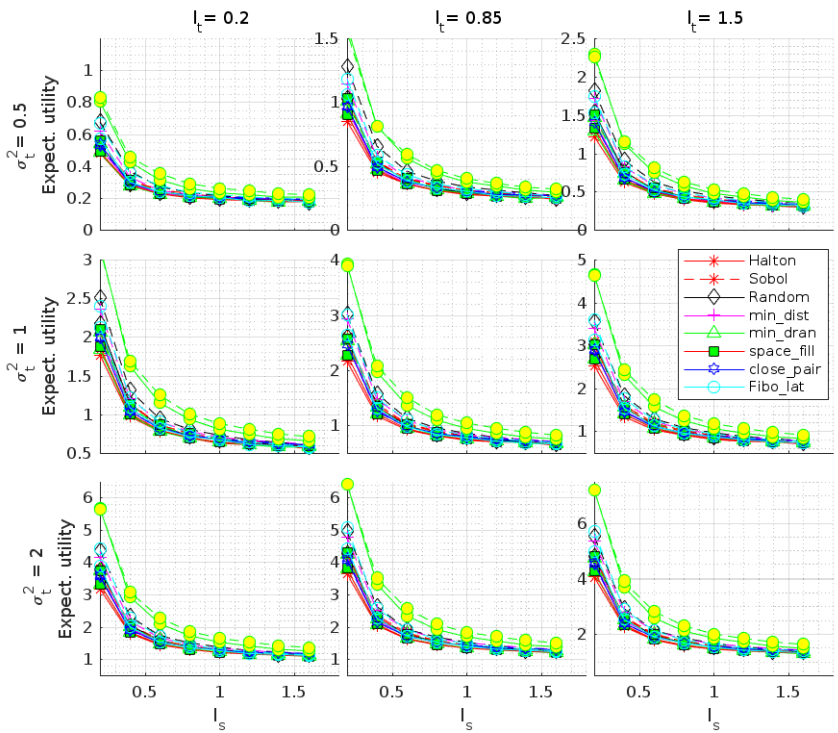
Examples of spatiotemporal designs

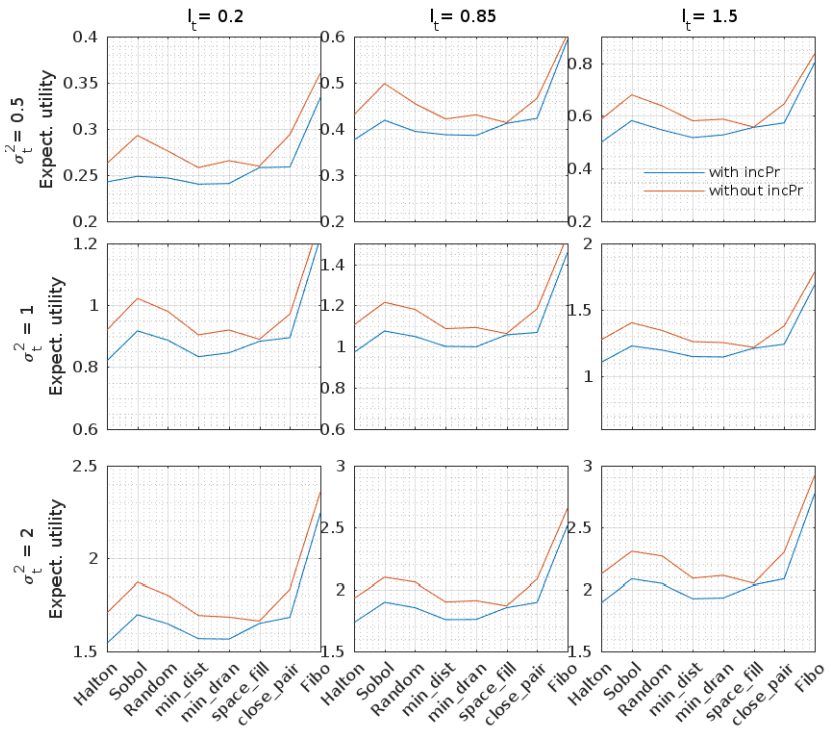
A random draw from an additive GP with unimodal mean function along time (color surface) and samples from Sobol design ($n = 30$).



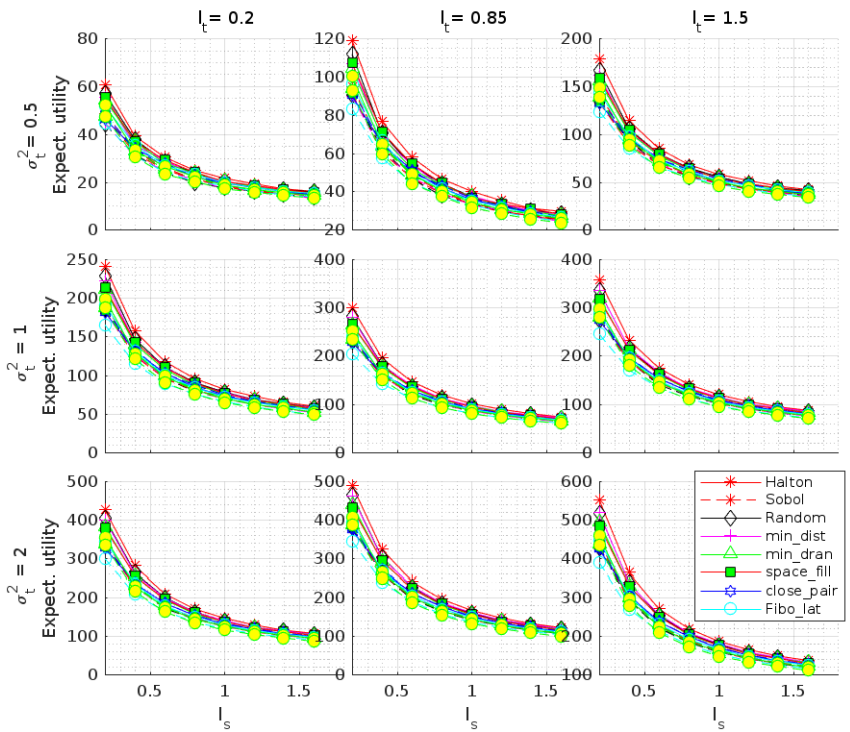
Poisson additive model for latent function, *EAPV*

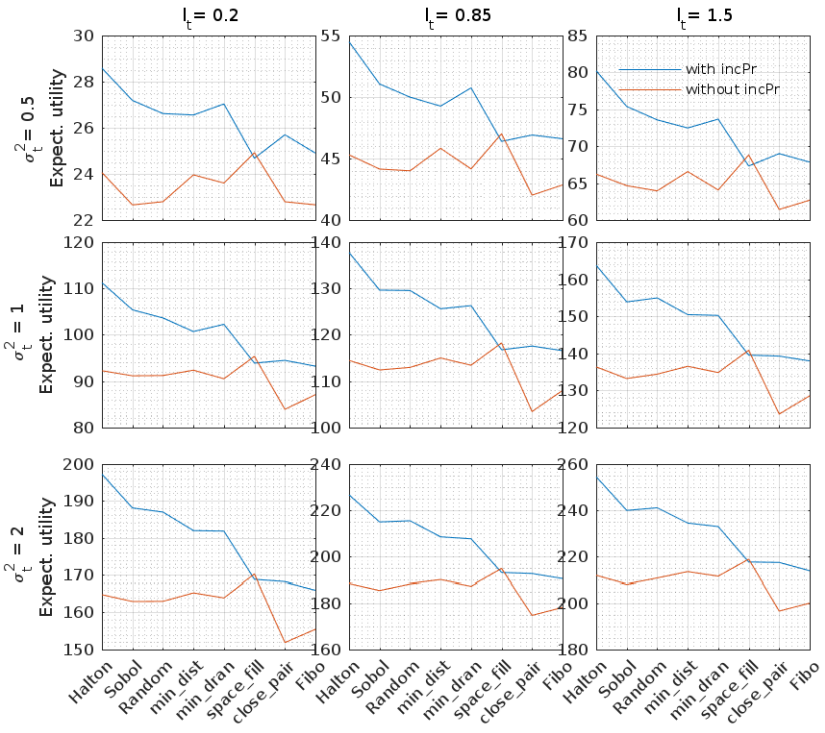
The dimension of the designs, $n = 100$.





Poisson additive model, *EKL*





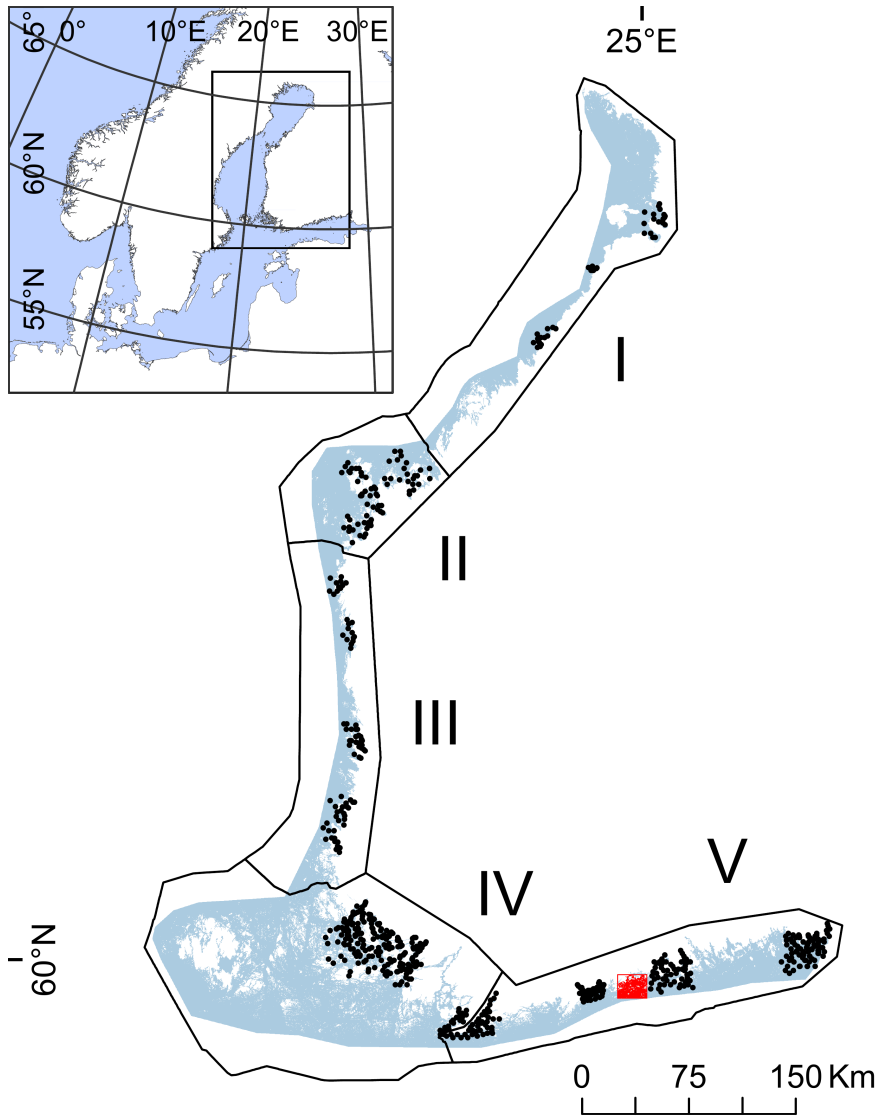
Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

Case study

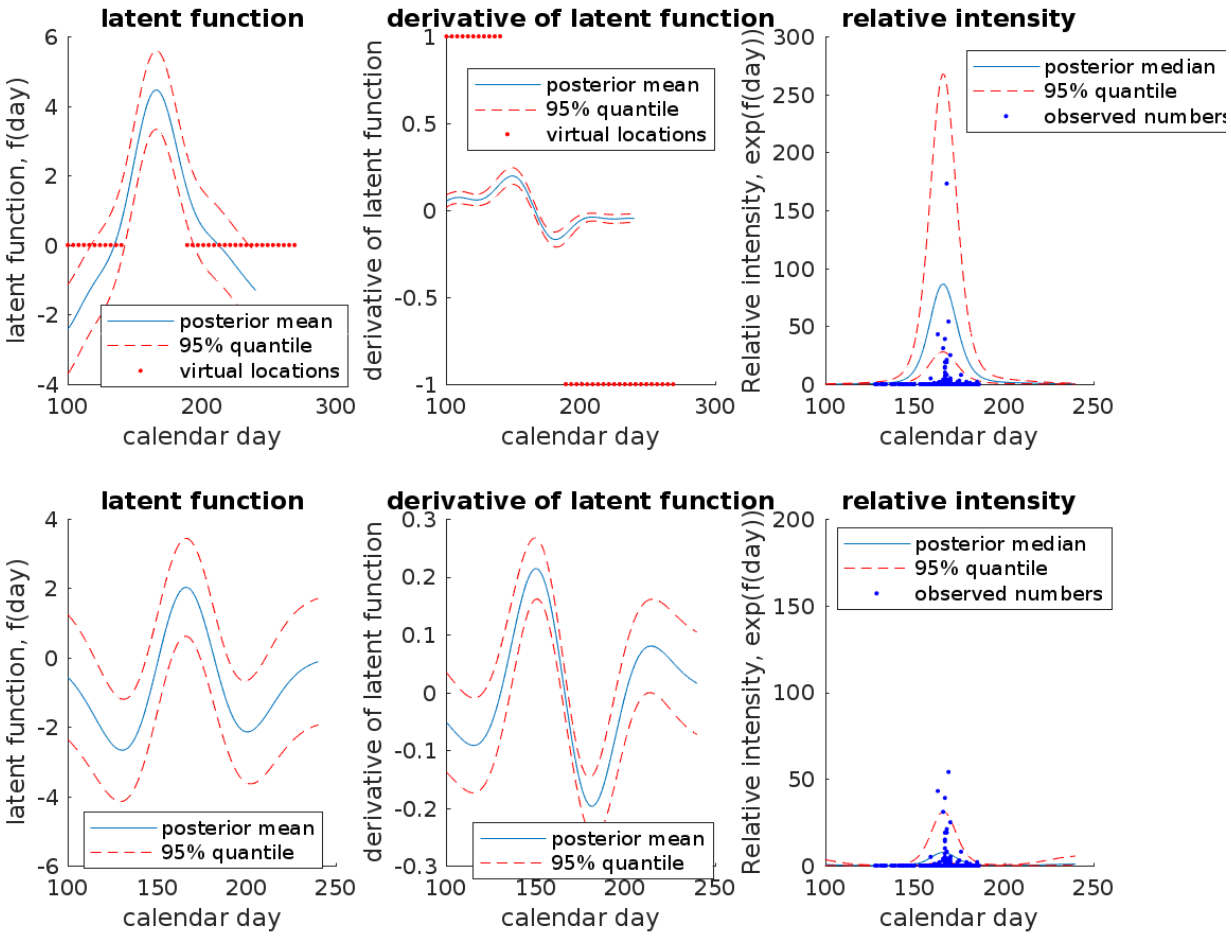
- We design a survey to inform spatial distribution of fish larval areas on Finnish coastal region in the northern Baltic Sea. The data contain several different species, count data between year 2007-2014, and from early May and early July (the calender days 128 -188. Ten different covariates that include times and spatial regions.

Map of the case study area on the Finnish coastal region. The study region includes 229 429 very dense spatial grid cells, 20 weeks, in total 4 588 580 spatiotemporal grid cells.

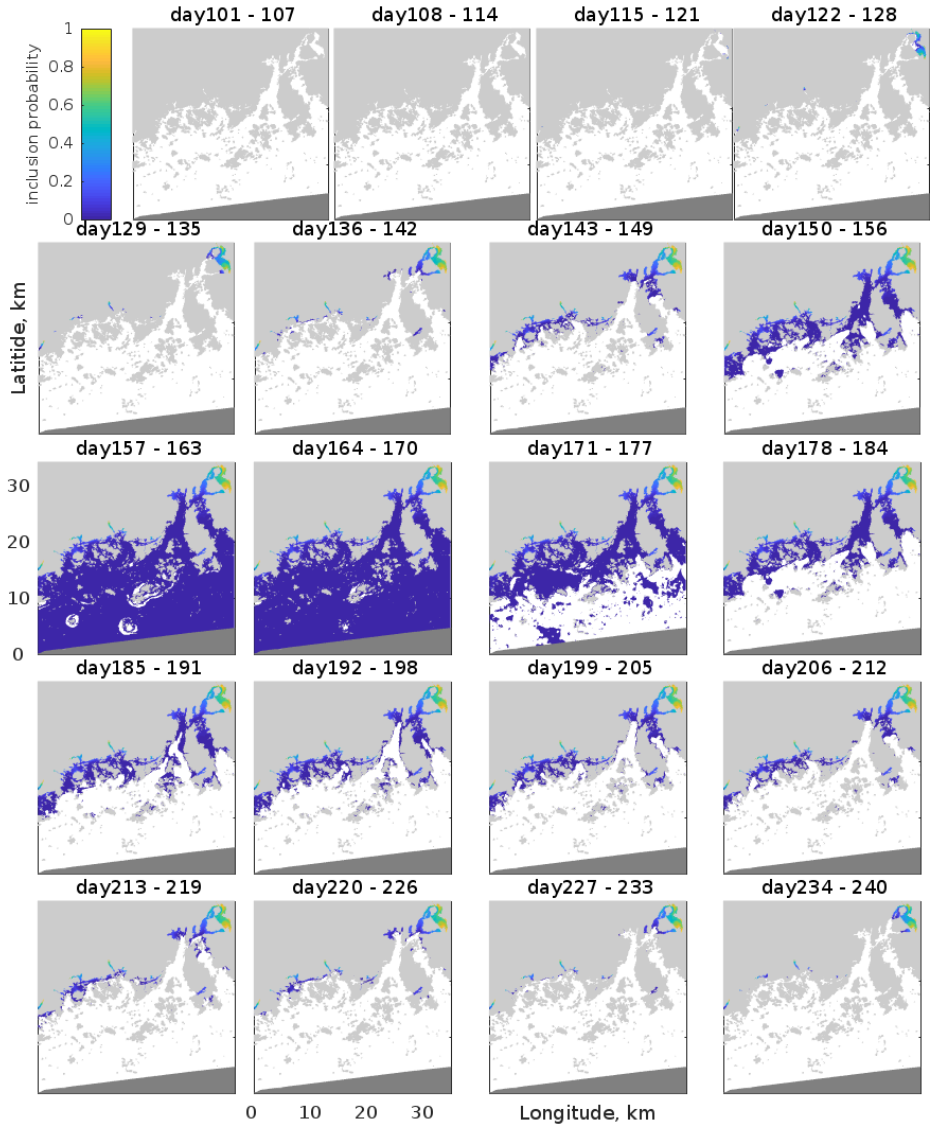


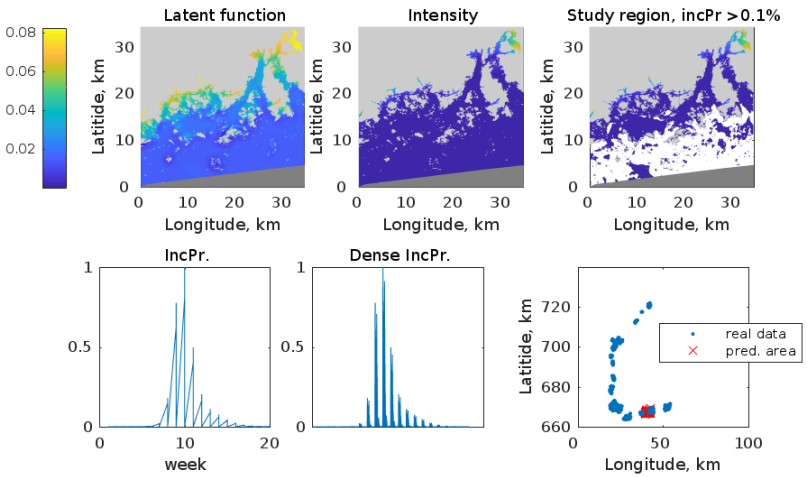
Results

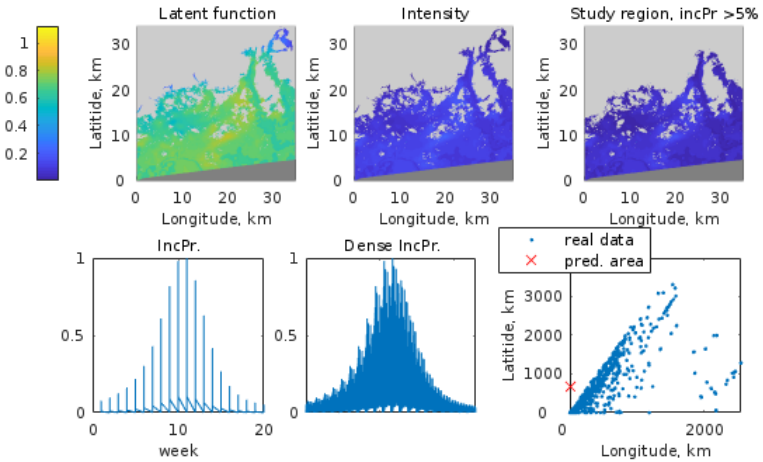
Posterior inference with monotonic constraints.



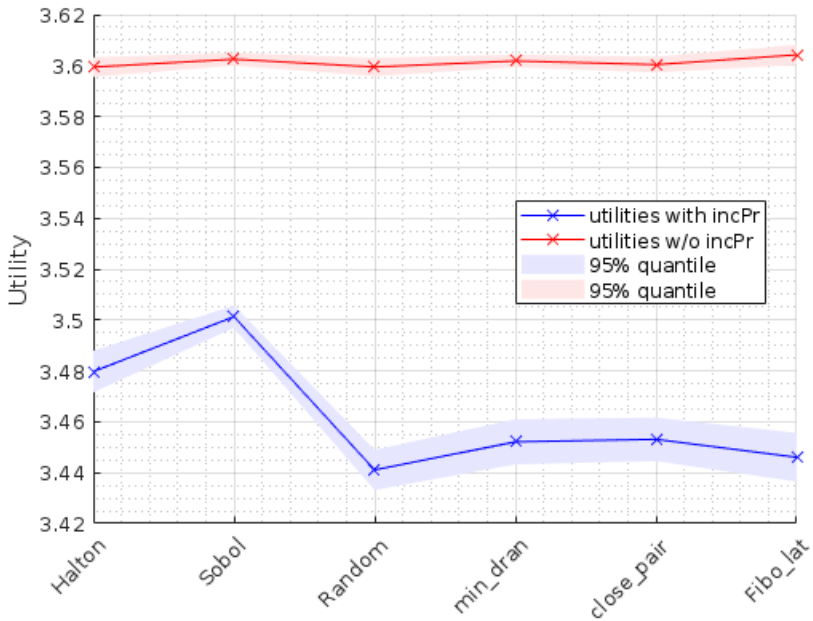
Shrink the survey region in the study.

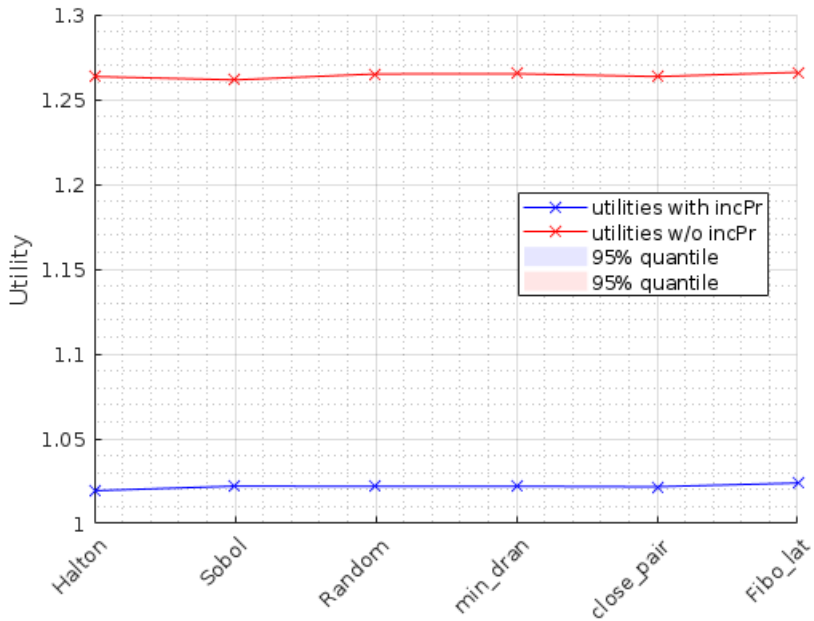


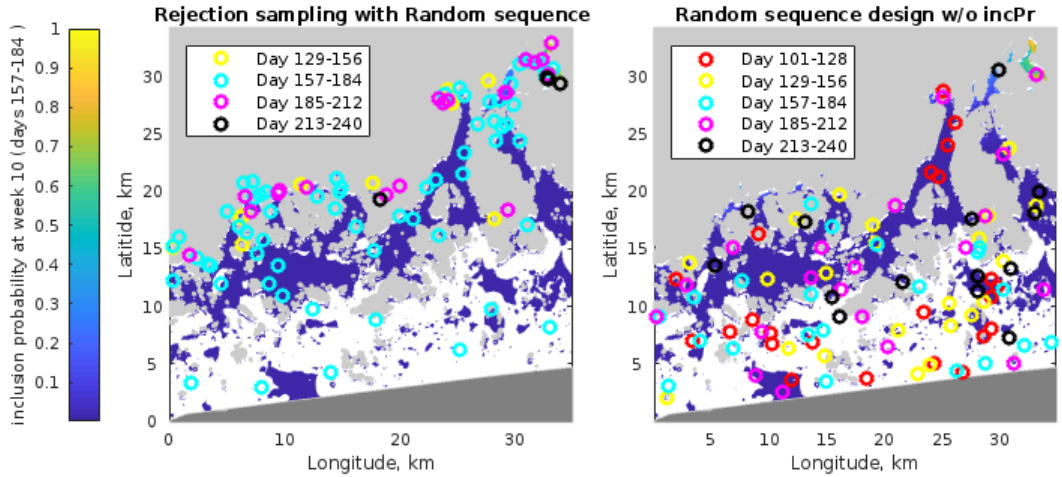

Latent function
Intensity
Study region, incPr >0.1%
Latitide, km
Longitude, km
IncPr.
Dense IncPr.
week
real data
pred. area


Latent function
Intensity
Study region, incPr >5%
Latitide, km
Longitude, km
IncPr.
Dense IncPr.
week
real data
pred. area

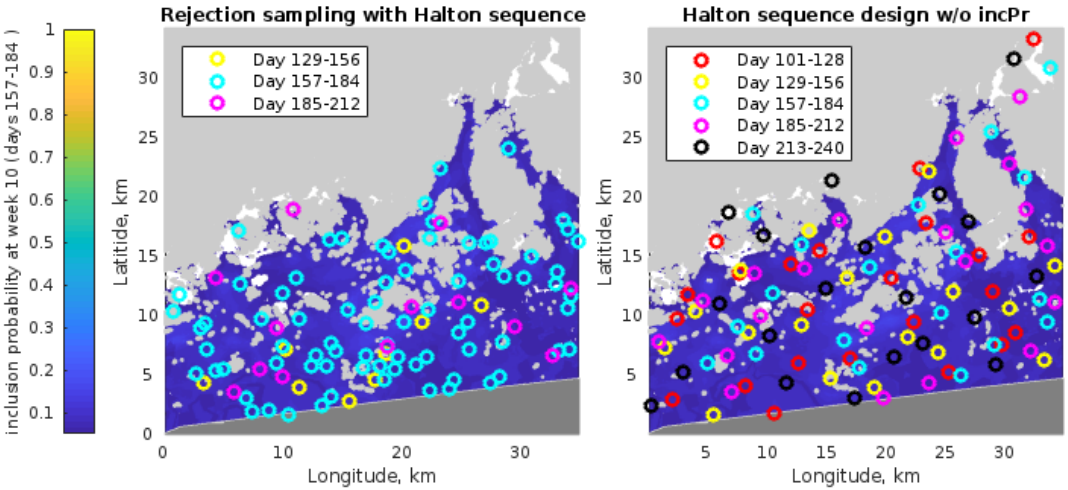
The crosses connected with solid lines show the Monte Carlo estimate and the highlighted regions show the 95% credible interval of this estimate.







a) pike perch sampling design



b) herring sampling design

Outline

1. Introduction
2. The Bayesian modeling and LGCP
3. The design
4. Rejection design
5. Utility functions and computational challenge
6. Simulation study
7. Case study
8. Conclusion

Conclusion

1. Realistic prior information can increase the expected utility of the designs with the observations that have LGCPs.
2. The design with inclusion probability keeps randomness and inherits the advantages from the spatial balance designs.
3. We need good/optimal designs: reduce the cost, good inference, etc.
4. This work has an arXiv version (arXiv:1808.09200).

On going works

1. New computational algorithms to make the computation of utilities and relatives (covariance matrix and inversion, the Cholesky decomposition) to be feasible and efficient with Big data.
2. Bayesian optimal design, new stochastic methods based annealing simulations to work with high dimensional cases. Study the discretized and continuous design spaces. Good proposals for fast mixing rates of the Markov chains.

References


Amzal, B., Bois, F. Y., Parent, E., and Robert, C. P. (2006).
Bayesian-optimal design via interacting particle systems.
Journal of the American Statistical association, 101(474):773–785.


Chipeta, M., Terlouw, D., Phiri, K., and Diggle, P. (2016).
Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure.
Environmetrics.


Kullback, S. (1987).
Letter to the editor: The Kullback-Leibler distance.
American Statistician, 41(4):340–341.


Müller, P. (1999).
Simulation based optimal design.
Bayesian statistics, 25:459–474.


Müller, P., Sansó, B., and De Iorio, M. (2004).
Optimal bayesian design by inhomogeneous markov chain simulation.
Journal of the American Statistical Association, 99(467):788–798.

References (continued)


Müller, W. G. (2007).
Collecting spatial data.


Nychka, D. and Saltzman, N. (1998).
Design of air-quality monitoring networks.
In *Case studies in environmental statistics*, pages 51–76. Springer.


Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013).
GPstuff: Bayesian modeling with Gaussian processes.
Journal of Machine Learning Research, 14(Apr):1175–1179.

Thank you very much!

Merci beaucoup!

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI