Spatial Statistics 35 (2020) 100392

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Bayesian model based spatiotemporal survey designs and partially observed log Gaussian Cox process

Jia Liu [a,b,*], Jarno Vanhatalo [b,c,**]

[a] Finnish Meteorological Institute, P.O. Box 503, FI00101 Helsinki, Finland
[b] Department of Mathematics and Statistics, Faculty of Science, University of Helsinki, P.O. Box 68, FI00014, Finland
[c] Organismal and Evolutionary Biology Research Program, Faculty of Bio- and Environmental Sciences, University of Helsinki, P.O. Box 68, FI00014, Finland

ARTICLE INFO

Article history:
Received 15 September 2018
Received in revised form 18 August 2019
Accepted 9 October 2019
Available online 18 October 2019

Keywords:
Experimental design
Bayesian inference
Kullback–Leibler information
Log Gaussian Cox process
Rejection sampling design
Species distribution

ABSTRACT

In geostatistics, the spatiotemporal design for data collection is central for accurate prediction and parameter inference. An important class of geostatistical models is log-Gaussian Cox process (LGCP) but there are no formal analyses on spatial or spatiotemporal survey designs for them. In this work, we study traditional balanced and uniform random designs in situations where analyst has prior information on intensity function of LGCP and show that the traditional balanced and random designs are not efficient in such situations. We also propose a new design sampling method, a *rejection sampling design*, which extends the traditional balanced and random designs by directing survey sites to locations that are *a priori* expected to provide most information. We compare our proposal to the traditional balanced and uniform random designs using the expected average predictive variance (APV) loss and the expected Kullback–Leibler (KL) divergence between the prior and the posterior for the LGCP intensity function in simulation experiments and in a real world case study. The APV informs about expected accuracy of a survey design in point-wise predictions and the KL-divergence measures the expected gain in information about the joint distribution of the intensity field. The case study concerns planning a survey design for analyzing larval areas of two commercially important

* Corresponding author at: Department of Mathematics and Statistics, Faculty of Science, University of Helsinki, P.O. Box 68, FI00014, Finland.
** Corresponding author.
E-mail addresses: jia.liu@fmi.fi (J. Liu), jarno.vanhatalo@helsinki.fi (J. Vanhatalo).

https://doi.org/10.1016/j.spasta.2019.100392

fish stocks on Finnish coastal region. Our experiments show that the designs generated by the proposed rejection sampling method clearly outperform the traditional balanced and uniform random survey designs. Moreover, the method is easily applicable to other models in general.

1. Introduction

A central question of geostatistics is the prediction of a spatial pattern over a region using data measured at a finite set of locations. A widely used stochastic model for such tasks is a hierarchical Gaussian process model (Cressie, 1993; Gelfand et al., 2010). It is well known that the goodness of the spatial prediction with such models depends on the spatial allocation of the data locations (Müller, 2001; Diggle and Lophaven, 2006) — that is on the *survey/experimental design*. The problem of finding a good survey design for spatial prediction when the observation process is Gaussian has received much interest in spatial statistics (e.g., Müller, 1999; Müller, 2001; Diggle and Lophaven, 2006). What has received less attention are spatiotemporal survey designs and survey designs for models with non-Gaussian observation processes (however, see Chipeta et al., 2016, 2017, for few examples). In this work, we study survey designs in particular for partially observed spatiotemporal log-Gaussian Cox processes (LGCPs).

In classical examples of LGCPs, the spatial study region is observed fully and the statistical analysis reduces to inference concerning the underlying intensity function and hyperparameters (Møller et al., 1998; Møller and Waagepetersen, 2007; Illian et al., 2008). However, recently LGCPs have gained increasing interest in applications where the study region is not fully observed in which case in addition to inference we want to predict the intensity field over unobserved regions. For example in ecology, LGCPs are used in species distribution modeling where observations comprise of animal counts in survey plots or transects that cover only small proportion of the whole study region (Yuan et al., 2017; Vanhatalo et al., 2017; Mäkinen and Vanhatalo, 2018). In these applications, the LGCP describes the process generating locations of individual specimen and the observations are a thinned version of the underlying LGCP. The thinning process describes the survey design by assigning zero probability of observing individual points (e.g., animals) at regions outside the survey sites. Inside the survey sites the observation probability can be either constant, corresponding to constant survey effort and observation probability (Kallasvuo et al., 2017), or it may vary between zero and one as, for example, in distance sampling (Yuan et al., 2017). The statistical inference includes then predictions for intensity in regions where observations have not been made (e.g. Yuan et al., 2017; Kallasvuo et al., 2017; Vanhatalo et al., 2017) resulting in a spatial prediction problem where the survey design plays critical role.

Model-based optimal survey design concerns a problem of maximizing (minimizing) the expected utility (loss) of future data over the design space. Much of the literature on optimal design focus on the development of computational algorithms for optimizing the expected utility over the design space (Stein and Handcock, 1989; Robert, 2004; Vlachos and Gelfand, 1996; Van Groenigen and Stein, 1998; Müller, 1999). Alternatively, many authors have aimed at developing sampling methods that can generate designs with better, if not optimal, expected utility for a given class of spatial models than random allocation of survey sites (Müller, 2001, 2007; Ryan et al., 2016; Chipeta et al., 2017). These approaches are computationally easier since the expected utilities of the candidate designs need to be calculated only once and there is no need to build optimization algorithm. In this work, we studied different classes of the designs: Spatially balanced designs provide more uniform coverage over the study area than random sampling (Robertson et al., 2013), which decreases uncertainty in spatial interpolation. A balanced design maintains spatial regularity of sampling locations by spreading observation points as evenly as possible in the design space by

means of specific sampling methods or criteria (Müller, 2007; Stevens and Olsen, 2004; Grafström et al., 2012). Quasi-random methods use quasi-random number generators such as the Sobol and Halton sequence to generate balanced and well-distributed designs. Distance-based designs are either deterministic algorithms, such as space filling design (see e.g., Müller, 2007 or random, such as the recently proposed simple inhibitory and the inhibitory plus close pairs designs (Chipeta et al., 2017). This class of sampling methods generates designs by considering the distance between any two sampling locations in order to achieve good allocations of the observations (Russo, 1984; Royle and Nychka, 1998).

Sampling locations in the above mentioned designs are commonly selected with even or roughly equal probabilities. This may be inappropriate in applications where some locations are *a priori* expected to be more informative than others. As we demonstrate in this work, this happens with LGCPs when there is prior knowledge on its intensity function. Such prior information may arise, for example, from earlier surveys and can be used to more efficiently plan new studies. In this work, we are specifically interested in spatiotemporal design problems and extend traditional spatially balanced and random designs with a novel rejection sampling scheme that gives more weights to times and spatial regions which are *a priori* expected to be more informative about the intensity function of the point process.

The structure of the paper is as the follows. In Section 2 we present a motivating case study and in Section 3 we review partially observed LGCP models. We then motivate and formulate one utility and one loss function that are used in this work for evaluating the alternative designs and discuss some of their properties in LGCPs (Section 4). Then we review the widely used balanced and random designs, and propose the novel rejection sampling design method (Section 5). In Section 6, we study the designs with simulation experiments and in Section 7 we present a case study of survey design concerning fish reproduction areas in fisheries management. We close the work with discussion and conclusions in Section 8.

2. Motivating case study: species distribution modeling

The main motivation for our work comes from species distribution modeling where LGCPs have received increasing interest in recent years (e.g., Yuan et al., 2017; Mäkinen and Vanhatalo, 2018; Vanhatalo et al., 2019). Species distribution models are key tools in the ecologists' toolbox. They are widely used, for example, to improve identification and management of conservation areas and natural resources (Kallasvuo et al., 2017) and to evaluate species responses to environmental filtering under climate change scenarios (Clark et al., 2016; Kotta et al., 2019). One of the main goals of statistical inference in species distribution modeling is to use species observations to predict over regions of unsampled locations to build thematic species distribution maps (Gelfand et al., 2006; Elith and Leathwich, 2009). The LGCP is routinely used to model count abundance data (Yuan et al., 2017; Mäkinen and Vanhatalo, 2018) and many species distribution models that are built using Poisson observation model can be seen as special cases of LGCP even if they were not explicitly written as such (see e.g. Kallasvuo et al., 2017; Vanhatalo et al., 2017, and discussion in Section 7). In resent years, LGCPs have gained a lot of interest also in modeling presence only data (Warton and Shepherd, 2010; Chakraborty et al., 2011; Renner and Warton, 2013) The LGCP model arises in other applications as well (Illian et al., 2012; Lombardo et al., 2018) and the methods presented here are applicable beyond species distribution modeling.

In our case study, we look for efficient survey designs for analyzing larval areas of pike perch (*Sander lucioperca*) and Baltic herring (*Clupea harengus membras*) in Finnish coastal region in the northern Baltic Sea (see Fig. 1). These are commercially important fish and information on their larval areas is needed for protecting the reproduction of these fish stocks. Previously, Kallasvuo et al. (2017) showed that in case of pike perch, the most important reproduction areas, which produce over 80% of new larvae, are extremely local whereas the most important herring reproduction areas are rather uniformly distributed throughout the region. Moreover, the larvae density varies significantly within a year since the spawning periods of pike perch and Baltic herring start after an ice break-up and last until late spring or early summer. The peak larval period is, however, rather short and its exact time poorly known so an efficient survey design should provide information

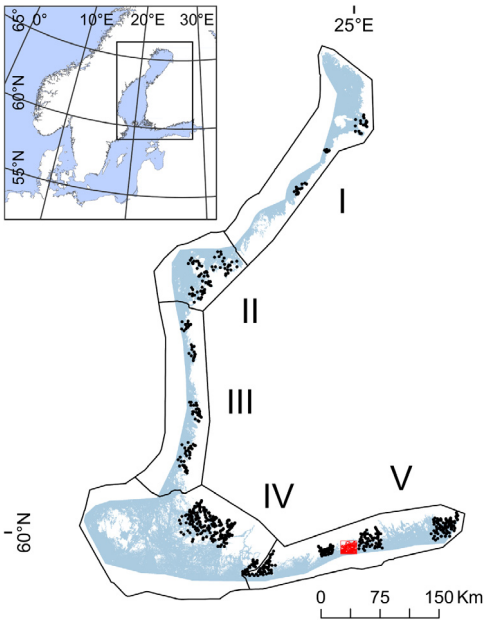


Fig. 1. Map of the case study area on the Finnish coastal region. The black dots show the survey sites of existing data and the red square shows the region over which we want to plan a new survey design.

on where the highest larval densities are located at and when they occur. Moreover, since the larval density information is used to estimate the total larvae biomass (Kallasvuo et al., 2017), in addition to accurate point-wise information the survey design should provide also information on dependencies between densities at different locations.

Our existing species distribution data (n = 1788) were collected during years 2007–2014. The locations of survey sites vary between years and, since the exact time of larval hatching is not known, each location was visited several times between the calender days 128 (early May) and 188 (early July). Each survey site is a transect of length 400–500 meters along which a net with 0.028 m^2 opening was towed behind a boat. The net sampled the surface water (depth 0.5–1.0 m) and the species observations consist of the number of larvae in the volume of sampled water. The survey sites were combined with seven abiotic environmental covariates that were available as raster maps with resolution of 50 m throughout the Finnish coastal region. See Kallasvuo et al., 2017 for detailed summary of the data.

In the case study, the aim is to construct a new survey design to improve the larval density estimates in a sub-region that was not included in earlier data collection. The new sampling region is located near Helsinki and is approximately 40 km wide and 40 km long (Fig. 1). We aim at a spatiotemporal design within the subregion between calendar days 100–240 (from early April to the end of August).

3. Partially observed spatiotemporal log Gaussian Cox process

We denote the study domain by $\mathscr{D} = \mathscr{A} \times [t_o, t_1]$ where $\mathscr{A} \subset \Re^2$ is the spatial region and $[t_o, t_1]$ is the time interval of interest. A spatiotemporal LGCP arises from an inhomogeneous Poisson process with a spatially and temporally varying intensity whose logarithm is given a Gaussian process (GP) prior (Chakraborty et al., 2011; Banerjee et al., 2015; Illian et al., 2008). We denote the intensity function by $\lambda(s, t) = \lambda(x)$ where $x = [s^T, t] \in \mathscr{D}$ is a vector of spatiotemporal coordinates. In species distribution modeling applications of LGCP, the intensity, $\lambda(x)$, describes

the (relative) density of a species over regions (Illian et al., 2013; Yuan et al., 2017; Mäkinen and Vanhatalo, 2018) and the points correspond to individual specimen of a species. A GP prior for the latent function, $f(\mathrm{x}) = \log\lambda(\mathrm{x})$, is defined by its mean function, $\mu(\mathrm{x}) = \mathbb{E}\,(f(\mathrm{x}))$, and a covariance function, $k\left(\mathrm{x}, \mathrm{x}'; \theta\right) = \mathrm{Cov}\left(f(\mathrm{x}), f(\mathrm{x}')\right)$, where θ is the vector of the hyperparameters corresponding to the covariance function parameters. Note, we first consider only spatiotemporal coordinates but when analyzing the real data in the case study (Section 7), we extend the model to include also covariates. The mean and covariance functions to be used in this work are also defined in detail in the experiments.

Here, the variable of central interest is the intensity function but the posterior inference is typically conducted first for the latent function. If the study domain is fully surveyed and the exact locations, $\xi_i \in \mathscr{D}$, $i = 1, \ldots, n'$, of points are recorded, the likelihood function for the latent function is (Banerjee et al., 2015)

$$L(\xi_1, \ldots, \xi_{n'}|f(\cdot)) = e^{\int_{\mathscr{D}} e^{f(\mathrm{x})} d\,\mathrm{x}} \prod_{i=1}^{n'} e^{f(\xi_i)}. \tag{1}$$

However, often the whole study domain cannot be surveyed but the survey includes only finite number of its subsets $\mathscr{D}_i \subset \mathscr{D}$, $i = 1, \ldots, n$ to be called *survey sites* hereafter. For example, in our case study and in the species distribution studies of Yuan et al. (2017) and Vanhatalo et al. (2017) the survey sites correspond to survey transects whereas in the work by Chakraborty et al. (2011) survey sites were square plots. If the survey sites are mutually disjoint, this partially observed LGCP leads to the likelihood function

$$L(\xi_1, \ldots, \xi_{n'}|f(\cdot)) = e^{\int_{\mathscr{D}_1} e^{f(\mathrm{x})} d\,\mathrm{x} + \int_{\mathscr{D}_n} e^{f(\mathrm{x})} d\,\mathrm{x}} \prod_{i=1}^{n'} e^{f(\xi_i)} \tag{2}$$

$$= e^{\int_{\mathscr{D}} e^{f(\mathrm{x})}\pi(\mathrm{x}) d\,\mathrm{x}} \prod_{i=1}^{n'} e^{f(\xi_i)}, \tag{3}$$

where $\pi(\mathrm{x})$ is a thinning function such that $\pi(\mathrm{x}) = 1$ if $\mathrm{x} \in \cup_{i=1}^{n'} \mathscr{D}_i$ and $\pi(\mathrm{x}) = 0$ otherwise. Hence, the model for the point observations is a thinned LGCP with intensity function $\lambda(\mathrm{x})\pi(\mathrm{x})$. In principle, the value of the thinning function inside a survey site could be less than one in which case it would describe the (relative) observation probability or search effort at different locations within the survey site (Illian et al., 2012; Simpson et al., 2016). However, in this work we assume that $\pi(\mathrm{x})$ is either zero or one describing only the survey design.

In species distribution studies data typically include only the number of specimen in a survey site but not their exact locations within the sites. If the survey sites are mutually disjoint, the likelihood function (2) then reduces to a product of finite number of Poisson likelihood terms, one for each survey site. Moreover, if the survey sites are small enough so that we can reasonably approximate $\int_{\mathscr{D}_i} \lambda(\mathrm{x}) d\,\mathrm{x} \approx V_i e^{f(\mathrm{x}_i)}$, where V_i is the volume/area and x_i the centroid of the i'th survey site, the likelihood function is

$$\begin{aligned} L(y|f(\cdot)) &\approx L(y_1, \ldots, y_n|\mathrm{f}) \\ &= \prod_{i=1}^{n} \mathrm{Poisson}\left(y_i|V_i e^{f(\mathrm{x}_i)}\right) \end{aligned} \tag{4}$$

where n is the number of survey sites, $y = [y_1, \ldots, y_n]^T$ is the vector of count observations in the survey sites and $\mathrm{f} = [f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n)]^T$ is a vector of latent variables. A likelihood function similar to (4) is typically used with exact point observations as well. In that case, the domain is discretized into mutually disjoint cells so that the integral over the domain in the exact likelihood function, (1) or (2), can be approximated by a finite sum leading to a product of Poisson likelihood terms as in (4) (Møller et al., 1998; Banerjee et al., 2015). Each term would then correspond to one cell, n would be the total number of cells and V_i the volume/area of the cell. The goodness of the approximation (4) depends on the resolution of the discretization (size of V_i) compared to the

variability of the intensity function. If the intensity function is expected to vary significantly within the survey sites the discretization resolution should be increased to grid cells smaller than the size of survey sites. In practice, the optimal choice of the discretization resolution is a compromise between accuracy and computational burden (see Møller et al., 1998; Møller and Waagepetersen, 2004, for details on the discretization approximation).

Hereafter, we define a spatial *survey design*, $d_n = \{\mathrm{x}_1, \ldots, \mathrm{x}_n : \mathrm{x}_i \in \mathscr{D}\}$, for a partially observed LGCP as a collection of n survey sites with centroids x_i. We consider the sizes of the survey sites, V_i to fixed *a priori* and, hence, not part of the design problem. Moreover, we assume that the survey sites are small enough so that the integral of intensity function over a site can be approximated by $V_i e^{f_i}$ so that the likelihood for the latent function is as in (4). As discussed in Section 2, there are typically two types of posterior distributions of central interest: the posterior density of the intensity at individual locations, $p(\lambda(\mathrm{x})|y, d_n)$, and the posterior for the whole intensity function over the region of interest described by the posterior probability measure $\mathbb{P}(\lambda(\cdot)|y, d_n)$. The former informs about point wise densities and can be used to build, e.g. species distribution maps. The latter, however, is needed when calculating, for example, the posterior distribution for the total biomass over the region which requires first solving $p(\int_{\mathscr{D}} \lambda(\mathrm{x}) d\,\mathrm{x}|y, d_n)$ (Kallasvuo et al., 2017). For computational reasons, the intensity field is described on predictive lattice grid $X_* = \{\mathrm{x}_{*,1}, \ldots, \mathrm{x}_{*,N}\}$ over the study domain with N cells and centroids x_*. Hence, we need to calculate the joint posterior predictive density for $\mathrm{f}_* = [f(\mathrm{x}_{*,1}), \ldots, f(\mathrm{x}_{*,N})]^T$ and $\lambda_* = [\lambda(\mathrm{x}_{*,1}), \ldots, \lambda(\mathrm{x}_{*,N})]^T$ after which an integral over the study domain can be approximated as a finite sum.

The traditional method to infer the LGCP is Markov chain Monte Carlo (MCMC, Møller and Waagepetersen, 2004) but in recent years deterministic approximations such as the Integrated Nested Laplace approximation (Illian et al., 2012; Simpson et al., 2016) and Gaussian approximations built with expectation propagation and the Laplace method (Vanhatalo et al., 2010; Kallasvuo et al., 2017) have become popular due to their computational benefits. Here, we use the Laplace method built over the Gaussian approximation due to its simple analytical form and because it has been shown to give accurate approximation for these models (e.g., Vanhatalo et al., 2010). For a given design d_n and realization of observations y, the Laplace approximation for the posterior of the latent function at prediction locations X_*, conditional on the hyperparameters θ, is $p(\mathrm{f}_*|d_n, y, \theta) \approx N\left(\mathrm{f}_*|\mu_{*|y,\theta}, K_{*|y,\theta}\right)$, where the (approximate) posterior mean and covariance are (Vanhatalo et al., 2010)

$$\mu_{*|y,\theta} = \mu(X_*) + K(X_*, d_n)K(d_n)^{-1}(\hat{\mathrm{f}} - \mu(d_n)) \tag{5}$$

$$K_{*|y,\theta} = K(X_*) - K(X_*, d)(K(d_n) + W_y^{-1})^{-1}K(d_n, X_*) \tag{6}$$

and $K(d_n)$, $K(X_*)$, $\mu(d_n)$ and $\mu(X_*)$ are the prior covariance matrices and mean vectors at the survey sites and prediction locations; $K(X_*, d_n)$ is the prior covariance matrix between the prediction locations and the survey sites; $\hat{\mathrm{f}} = \arg\max_{\mathrm{f}} p(\mathrm{f}|y, d_n, \theta)$ is the *maximum a posteriori* (MAP) estimate of latent variables at survey sites; and $W_y = -\nabla\nabla \log L(y|\mathrm{f})|_{\mathrm{f}=\hat{\mathrm{f}}} = \mathrm{diag}(V_1 e^{\hat{f}_1}, \ldots, V_n e^{\hat{f}_n})$ is the negative Hessian matrix of the log likelihood. The Laplace method provides also an approximation for the marginal likelihood of the hyperparameters (Vanhatalo et al., 2010)

$$L(y|\theta, d_n) = \int p(y|\mathrm{f}) \mathrm{d}p(\mathrm{f}|\theta, d_n) \approx p(y|\hat{\mathrm{f}})|K(d_n)(K(d_n)^{-1} + W_y)|^{-1/2} e^{-\frac{1}{2}\hat{f}^T K(d_n)^{-1}\hat{f}}. \tag{7}$$

The Laplace approximation for the marginal likelihood can then be used to optimize them to their approximate MAP estimate $\hat{\theta} = \arg\max_\theta L(y|\theta, d_n)p(\theta)$ as described by Vanhatalo et al. (2010) after which the posterior of the latent function can be approximated with $p(\mathrm{f}_*|y, d_n) \approx N\left(\mathrm{f}_*|\mu_{*|y,\hat{\theta}}, K_{*|y,\hat{\theta}}\right)$. In this work, we refer to this approximation as the Laplace method.

4. Survey design for partially observed LGCP

4.1. Expected utility and loss of a design

From survey design point of view, the key question is where should we select the survey sites $\mathrm{x}_i \in d_n$. Typically, as in our case study, there exists earlier data or prior knowledge about the

intensity that can help plan the future surveys. We follow the Bayesian decision theoretic framework and define a utility function to measure the goodness of a design (Eidsvik et al., 2015).

We denote by $D_n = \{d_n\}$ the set of all possible designs d_n of size n in domain $\mathscr{D}$ and by $U(d_n, Y, f(\cdot), \theta)$ a utility function where $Y = [Y_1, \ldots, Y_n]^T$ is a random vector denoting the new data to be collected at survey sites. Alternatively, we may define a loss function $L(\cdot) = -U(\cdot)$. In a more general treatment where surveys are used to inform decision making, utility and loss should depend also on the decisions (Lindley, 2003; Eidsvik et al., 2015). However, we do not consider decision making here and omit decisions from our notation. A design should be evaluated according to its expected utility which in the case of partially observed LGCP is

$$\bar{U}(d_n) = \sum_{y \in \mathbb{N}^n} p(y|d_n) \int_f \int_\theta U(d_n, y, f(\cdot), \theta) d\mathbb{P}(f(\cdot)|d_n, y, \theta) d\mathbb{P}(\theta|y, d_n), \tag{8}$$

where $\mathbb{P}(f(\cdot)|d_n, y, \theta)$ and $\mathbb{P}(\theta|y, d_n)$ are the posterior probability measures of the latent function and the hyper-parameters given a realization $y = [y_1, \ldots, y_n]^T$ from the design d_n and $p(y|d_n) = \int p(y|f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n)) dp(f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n))$ is the (prior) predictive density of y. Hence, the outer summation corresponds to expectation over the prior predictive distribution of Y. Next we introduce the utility functions to be used in this work and analyze some of their properties with partially observed LGCP.

4.2. Average predictive variance (APV) loss

Spatial designs are commonly compared with the average predictive variance (APV) loss over the study domain. It is a widely used measure for the marginal accuracy of point wise predictions (see, e.g., Diggle and Lophaven, 2006; Müller, 2007; Ryan et al., 2016; Chipeta et al., 2016, 2017). It is, hence, a natural measure of the goodness of a design when the aim is to reduce the overall uncertainty in the point wise predictions needed in, for example, species distribution maps. The APV loss of the latent function depends only on the predictive variance of the latent function $f(\cdot)$ so that the loss function and the corresponding expected loss are

$$L_{\mathrm{APV}}(d_n, Y) = \frac{1}{|\mathscr{D}|} \int_{\mathrm{x}_* \in \mathscr{D}} \mathrm{Var}\{f(\mathrm{x}_*)|d_n, Y\} d\,\mathrm{x}_*, \tag{9}$$

$$\bar{L}_{\mathrm{APV}}(d_n) = \frac{1}{|\mathscr{D}|} \sum_{y \in \mathbb{N}^n} p(y|d_n) \int_{\mathrm{x}_* \in \mathscr{D}} \mathrm{Var}\{f(\mathrm{x}_*)|d_n, y\} d\,\mathrm{x}_*\,, \tag{10}$$

where $|\mathscr{D}|$ is the size (area or volume) of the study domain. The expected APV loss of the intensity is $\bar{L}_{\mathrm{APV}\lambda}(d_n) = \frac{1}{|\mathscr{D}|} \sum_{y \in \mathbb{N}^n} p(y|d_n) \int_{\mathrm{x}_* \in \mathscr{D}} \mathrm{Var}\{\lambda(\mathrm{x}_*)|d_n, y\} d\,\mathrm{x}_*$ where, due to the log transformation $\lambda(\mathrm{x}_*) = \exp(f(\mathrm{x}_*))$, the variance of $\lambda(\mathrm{x}_*)$ is

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

In the experiments, we approximate the integral over $\mathscr{D}$ by a finite sum over the lattice grid cells in X_*. The expectation over Y is approximated by Monte Carlo approximation

$$\bar{L}_{\mathrm{APV}}(d_n) \approx \frac{1}{M} \sum_{j=1}^{M} \Bigg[\frac{1}{N} \sum_{\mathrm{x}_* \in X_*} \mathrm{Var}\{f_j(\mathrm{x}_*)|d_n, Y_j\}\Bigg],$$

where $Y_j \in \mathbb{N}^n$ is the j'th Monte Carlo draw from $\mathbb{P}(Y|d_n)$ and M is the number of Monte Carlo samples.

4.3. The expected Kullback–Leibler divergence

When estimating the total biomass, point wise predictions are not enough but we need to know the posterior measure of the intensity function within that region (see Section 3 and Kallasvuo et al.,

2017). The aim of data collection should then be to increase the information concerning the full intensity function. In the discretized domain this corresponds to increasing the information about the joint distribution of f_*. In this case, a natural choice for the utility function is the Kullback–Leibler (KL) divergence from prior to posterior, since it is by definition a measure of information provided by data (Lindley, 1956; O'Hagan and Kingman, 1978; Kullback, 1987; Lindley, 2003; Ryan et al., 2016). For a partially observed LGCP, it is

$$U_{\mathrm{KL}}(d_n, Y) = \mathrm{KL}\Big(d\mathbb{P}(f(\cdot)|d_n, Y) \parallel d\mathbb{P}(f(\cdot)) \Big) \tag{11}$$

$$= \int \log \frac{d\mathbb{P}(f(\cdot)|d_n, Y)}{d\mathbb{P}(\cdot)} d\mathbb{P}(f(\cdot)|d_n, Y), \tag{12}$$

and the expected KL-divergence is

$$\bar{U}_{\mathrm{KL}}(d_n) = \sum_{y \in \mathbb{N}^n} p(y|d_n) \mathrm{KL}\Big(d\mathbb{P}(f(\cdot)|d_n, y) \parallel d\mathbb{P}(f(\cdot)) \Big), \tag{13}$$

where we use again Monte Carlo approximation to numerically solve the expectation over Y in (13). The KL-divergence has a simple form when the observations $Y_1, \ldots, Y_n$ are conditionally independent given the corresponding latent variables:

Lemma 1. *Assume $f(\cdot)$ is a latent function with Gaussian process prior probability measure $\mathbb{P}(f(\cdot))$. Assume further that we have finite data (d_n, Y), where $d_n = [\mathrm{x}_1^T, \ldots, \mathrm{x}_n^T]$ are covariates and $Y = [Y_1, \ldots, Y_n]$ are observations that are conditionally independent given the corresponding latent variables, that is $p(Y|f(\cdot)) = \prod_{i=1}^n p(Y_i|f(\mathrm{x}_i))$. Denote by $\mathbb{P}(f(\cdot)|d_n, Y)$ the posterior probability measure of $f(\cdot)$. The KL-divergence from the prior to the posterior for $f(\cdot)$ is*

$$KL\Big(d\mathbb{P}(f(\cdot)|d_n, Y) \parallel d\mathbb{P}(f(\cdot)) \Big)$$

$$= \sum_{i=1}^n \int p(f(\mathrm{x}_i)|d_n, Y) \log p(Y_i|f(\mathrm{x}_i)) df(\mathrm{x}_i) - \log p(Y), \tag{14}$$

where $\log p(Y) = \log \int p(Y|\mathrm{f}) p(\mathrm{f}) d\,\mathrm{f}$ is the log marginal likelihood.

See Appendix A for a proof. Moreover, the KL divergence from prior to posterior for intensity function $\lambda(\cdot)$ is the same as that of the latent function $f(\cdot)$ (see Appendix A).

In order to calculate the KL divergence from the prior process to the posterior process over $\mathscr{D}$, we need to calculate the marginal likelihood $p(Y)$ and n one dimensional integrals. Conditional on fixed hyperparameters, we can use the Laplace method (Section 3) to approximate $p(Y|\theta)$ and $p(f(\mathrm{x}_i)|d_n, Y, \theta)$. Hence, when using the Laplace method for inference, the KL-divergence has a particularly simple form. When conducting full posterior inference for both latent variables and hyperparameters with MCMC, we can directly approximate the first term in (14) using Monte Carlo. For the log marginal likelihood we use the Laplace–Metropolis estimator (Kass and Raftery, 1995)

$$\log p(Y) = \log \int p(Y|\mathrm{f}) p(\mathrm{f}|\theta) p(\theta) d\,\mathrm{f}\, d\theta$$

$$\approx \frac{d}{2} \log 2\pi + \frac{1}{2} \log |\hat{\Sigma}_\theta| + \log p(Y|\hat{\theta}) + \log p(\hat{\theta}), \tag{15}$$

where $\hat{\theta}$ is the MAP estimate of hyperparameters, $\hat{\Sigma}_\theta$ is the Monte Carlo estimator for the posterior covariance of the hyperparameters and d is the number of hyperparameters.

Sometimes the interest is to predict the latent function and intensity only over a subdomain $\tilde{\mathscr{D}} \subset \mathscr{D}$ that contain partial observation locations (see Sections 2 and 7). Let us denote by $f_{\tilde{\mathscr{D}}}$ the marginal latent field over subdomain $\tilde{\mathscr{D}}$. The KL-divergence from prior to posterior for $f_{\tilde{\mathscr{D}}}$ is

$$\mathrm{KL}\Big(d\mathbb{P}(f_{\tilde{\mathscr{D}}}|d_n, Y) \parallel d\mathbb{P}(f_{\tilde{\mathscr{D}}}) \Big) = \int \log p(y|f_{\tilde{\mathscr{D}}}) d\mathbb{P}(f_{\tilde{\mathscr{D}}}|_n, Y) - \log p_{\tilde{\mathscr{D}}}(Y), \tag{16}$$

where $p_{\tilde{\mathscr{D}}}(y) = \int p(y|f_{\tilde{\mathscr{D}}})d\mathbb{P}(f_{\tilde{\mathscr{D}}})$. Since $p(y|f_{\tilde{\mathscr{D}}}) = \int p(y|\mathrm{f})d\mathbb{P}(\mathrm{f}|f_{\tilde{\mathscr{D}}})$, calculating the KL-divergence becomes more difficult than in (14) if any $\mathrm{x}_i \notin \tilde{\mathscr{D}}$ (Appendix A).

4.4. Properties of the APV loss and KL divergence utility in partially observed LGCP

We study first the behavior of prior predictive probability of non-zero count observations in the survey sites. Conditional on the hyperparameters, the prior predictive mean of the future observations $\mathbb{E}[Y_i|d_n,\theta] = V_i e^{\mu(d_n)_i + K(d_n)_{ii}/2}$ and, by the law of total variance, $\mathrm{Var}[Y_i|d_n,\theta] = V_i e^{\mu(d_n)_i+K(d_n)_{ii}/2} + V_i^2\left(e^{K(d_n)_{ii}}-1\right)e^{2\mu(d_n)_i+K(d_n)_{ii}}$. If the survey area (volume), V_i, is fixed, both $\mathbb{E}[Y_i|d_n,\theta] \to 0$ and $\mathrm{Var}[Y_i|d_n,\theta] \to 0$ when $\mu(d_n)_i + K(d_n)_{ii} \to -\infty$. The prior predictive mean and variance of future counts approach zero also when $V_i \to 0$. Hence, $\Pr(Y_i > 0|d_n,\theta) \approx 0$ if the survey site is very small or if $\mu(d_n)_i + K(d_n)_{ii} \ll 0$. This is intuitively reasonable: Because the number of points within the study region is finite, as the proportion of the surveyed area compared to the total area approaches zero the expected number of observed points approaches zero as well. Similarly, if we have strong *a priori* expectation that the intensity is approximately zero throughout the survey site we expect to observe zero counts regardless of the size of the survey site. However, if we observe only zeros we cannot make inference on the intensity function.

Recall then the Laplace approximation for the conditional posterior mean and variance of the latent variables at prediction locations (5)–(6) and consider that θ is fixed. Both the posterior mean and variance are functions of the MAP estimates of the latent variables at survey sites

$$\hat{\mathrm{f}}|Y, d_n, \theta = \arg\max_{\mathrm{f}} -(\mathrm{f}-\mu)^T K(d_n)^{-1}(\mathrm{f}-\mu) + \sum_{i=1}^{n}\left(Y_i - V_i e^{f_i}\right), \tag{17}$$

which depend on the future observations Y through terms $Y_i - V_i e^{f_i}$. Now since $\Pr(Y > 0|d_n,\theta) \approx 0$ when $\mu(d_n) + K(d_n)_{ii} \ll 0$ or when $V_i \approx 0$, it follows that $\mathbb{E}_{p(Y|d_n,\theta)}[\hat{\mathrm{f}}] \approx \mu(d_n)$ under the same conditions. In this situation, the posterior mean and covariance are not *a priori* expected to change from the prior mean and covariance. For the mean this can be seen by plugging in $\hat{\mathrm{f}} = \mu(d_n)$ to (5). In case of covariance we first use the Woodbury–Sherman–Morrison Lemma to write

$$(K(d_n) + W_y^{-1})^{-1} = W_y - W_y(K(d_n)^{-1} + W_y)^{-1}W_y.$$

Now, the elements of the Hessian matrix $W_y = \mathrm{diag}(V_1 e^{\hat{f}_1}, \ldots, V_n e^{\hat{f}_n})$ decrease with decreasing $\hat{\mathrm{f}}$ and V_i so that $W_{y,i} \to 0$ as $\hat{\mathrm{f}}_i \to -\infty$ or $V_i \to 0$. Hence, at these limits $(K(d_n) + W_y^{-1})^{-1} \to 0$ and when plugging this in (6) we see that the posterior covariance reduces to the prior covariance $K(X_*)$ as well.

The posterior for $f(\cdot)$ is, thus, not *a priori* expected to differ from the prior if in survey sites $\mu(\mathrm{x}_i)$ or V_i is so small that the prior predictive probabilities $\Pr(Y_i > 0|\mathrm{x}_i,\theta) \approx 0$. Moreover, the difference between prior and posterior should be the larger the more design points are located in places where the prior predictive probability $\Pr(Y_i > 0|\mathrm{x}_i,\theta)$ is significantly above zero. To put it another way, both the APV loss and the KL divergence utility are functions of the prior mean and covariance; that is we could write $\bar{L}_{\mathrm{APV}}(d_n) = \bar{L}_{\mathrm{APV}}(d_n, \mu(\cdot), k(\mathrm{x},\mathrm{x}'))$. Hence, if we have prior information on intensity function, for example, from earlier data, it can be used to construct better survey designs. This property is very different from the properties of a Gaussian process with Gaussian observation model $y_i|f(\mathrm{x}_i) \sim N(f(\mathrm{x}_i),\sigma^2)$ where the posterior predictive mean and variance are

$$\mu_{*|y,\theta} = \mu_* + K(X_*, d_n)(K(d_n) + \sigma_n^2 I)^{-1}(y - \mu(d_n)), \tag{18}$$

$$K_{*|y,\theta} = K(X_*) - K(X_*, d_n)(K(d_n) + \sigma^2 I)^{-1}K(d_n, X_*). \tag{19}$$

Under a Gaussian observation model the APV loss depends only on the sampling locations and prior covariance but not on prior mean $\mu(\mathrm{x})$, so that $\bar{L}_{\mathrm{APV}}(d_n) = \bar{L}_{\mathrm{APV}}(d_n, k(\mathrm{x},\mathrm{x}'))$. For this reason uniform space filling designs typically work well with traditional Gaussian model, but we do not expect them to work equally well with partially observed LGCP models. Also in terms of the KL-divergence

utility, the locations where $\Pr(y_i > 0) \approx 0$ are *a priori* expected to be less informative about the latent and intensity functions than locations where the prior predictive probability for non-zero observations is significantly above zero.

Above we assumed that the hyperparameters, θ, are fixed. The same argument, that locations with $\Pr(Y_i > 0|\mathrm{x}_i) \approx 0$ are expected to be less informative than locations with $\Pr(Y_i > 0|\mathrm{x}_i) > 0$, applies also in full Bayesian analysis where we marginalize over the posterior of hyperparameters as well. To see this, consider the Laplace approximation for the marginal likelihood (7) which also depends on data through $\hat{\mathrm{f}}$ and W_y. Similarly as above, $\Pr(Y_i > 0|\mathrm{x}_i) \approx 0$ if $V_i \approx 0$, or if $\mu(d_n)_i + K(d_n)_{ii} \ll 0$ for all θ that have significant prior probability. On the other hand, $\hat{f} \approx \mu(d_n)$ for any θ for which $\Pr(Y_i > 0|\mathrm{x}_i) \approx 0$. Hence, we expect to learn about θ and f_* more, if we survey at sites where $\Pr(Y_i > 0|\mathrm{x}_i)$ is significantly above zero than at sites where $\Pr(Y_i > 0|\mathrm{x}_i) \approx 0$; that is the utility function depends again also on $\mu(d_n)$ and $K(d_n)$.

5. Survey designs

5.1. Spatially balanced and random designs

Our goal is to develop an algorithm that generates reasonable designs without numerically hard and time consuming optimization of the expected utility (or loss). We begin the development from studying and implementing the common random and spatially balanced designs, which are summarized here. In the next section, we introduce our extensions to them to achieve better survey designs for LGCP modeling.

We denote by *Random* the uniform random sampling of survey sites within the study domain. Since the random sampling does not typically lead to uniform coverage of survey sites within the prediction domain, several spatially balanced designs have been introduced as alternatives to it (Cambanis, 1985; Müller, 2007). Here, we test two common quasi-random number sequences, the Sobol and Halton sequences (Sobol, 1976) to be denoted by *Sobol* and *Halton*, and one of the most popular spatially balanced designs, the Fibonacci lattice. For 2-D square it is detailed by, for example, Koehler and Owen (1996) and Pei et al. (2009) define an algorithm to construct the Fibonacci lattice in 3D setting. We use their algorithm here and denote the corresponding design *Fibo_lat*.

We include into comparison also two recent distance-based design methods; the simple inhibitory and the inhibitory plus close pairs designs (Chipeta et al., 2017), denoted here by *min_dran* and *close_pair* respectively. These designs introduce a minimum dispersion threshold in the random sampling of survey site locations. For example, under the simple inhibitory design, the distance between any two locations should be greater than or equal to the threshold. Chipeta et al. (2017) showed that the designs generated by these methods have good performance in parameter estimation and spatial prediction with Gaussian observation model. Their algorithm was tailored for continuous covariate space so we extended it for discretized locations to be denoted by *min_dist*. The distance thresholds and number of close pair points in these designs could be optimized (Chipeta et al., 2016). However, we fixed them based on preliminary test runs. All distance-based designs were constructed in unit cube and then scaled to the actual size of the domain. Three different sizes of the design were studied. The distance threshold in the cube was $\delta = 0.21$ for design size $n = 50$, $\delta = 0.15$ for $n = 100$ and $\delta = 0.1$ for $n = 150$. For the close_pair design we set the number of close pair points, k, to $0.5 \times n$ and the distance threshold to $\delta_k = \delta * \sqrt{n/(n-k)}$.

The above designs are based on either random or quasi random sequences and can be easily used as proposal algorithms in rejection sampling. We add into comparison also one deterministic space-filling design. Space-filling designs fall into a class of purely geometrical designs using distance-based criteria to search for uniform spatial coverage (Royle and Nychka, 1998; Nychka and Saltzman, 1998; Müller, 2007; Johnson et al., 1990). Müller (2001) summarizes these designs by a numerical search algorithm called "Coffee-house" which is used in this work and called *space_fill*. These alternative random and spatially balanced designs are visualized in Fig. 2.

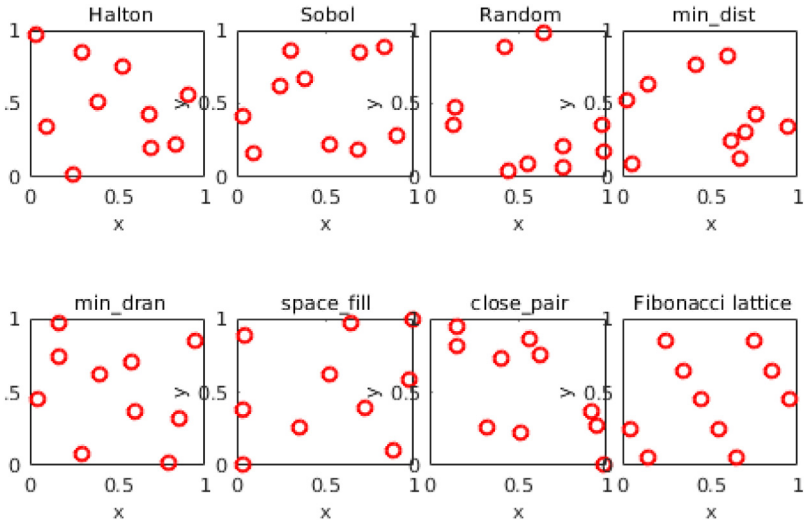


Fig. 2. Eight alternative spatial designs of size 10. Each circle shows the centroid of a survey site. The hyperparameter $\delta = 0.4$ was used for the distance-based designs.

5.2. Rejection sampling designs

Balanced designs perform well in maintaining the spatial regularity. This may, however, be suboptimal if some locations are expected to be more informative than others. In order to account for the specific properties of the expected utility under the LGCP model (Section 4.4), we extend them so that on average more survey sites are located to places which are expected to increase the utility the most. In practice, we extend the idea of balanced acceptance sampling (Robertson et al., 2013) and propose a new design method and name it *rejection sampling design*, where a random or spatially balanced design is thinned with an inclusion probability that is a function of the prior mean and variance of the latent function or intensity function in LGCP.

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

1. Generate a location x^* within the study domain (here any of the above random or quasi-random sequence can be used);
2. Calculate an inclusion probability $0 \leq p(\mathrm{x}^*) \leq 1$;
3. Accept the location with probability $p(\mathrm{x}^*)$. If accepted, set $\mathrm{x}_j = \mathrm{x}^*$ 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.

The inclusion probability can be linked to prior knowledge of the intensity function and its choice governs how much weight is assigned to sample higher intensity areas. The above algorithm cannot be directly used with the deterministic space-filling design for which reason we developed a modified coffee-house algorithm for rejection sampling as detailed in Appendix B. For the 3-D Fibonacci lattice design, we used dynamic scaling (Family and Vicsek, 1985) to obtain a design with inclusion probability restricted to unit cube.

We tested three inclusion probability functions: an inclusion probability proportional to the expectation of the latent function $p(\mathrm{x}) \propto \mu(\mathrm{x})$, an inclusion probability proportional to the expected intensity, $p(\mathrm{x}) \propto e^{\mu(x)+2\sigma^2(x)}$, and an inclusion probability proportional to truncated expected intensity $p(\mathrm{x}) \propto \min\left(p_{\max}, e^{\mu(x)+2\sigma^2(x)}\right)$ where $\mu(\mathrm{x})$ and $\sigma^2(x)$ are the prior mean and variance of the GP and $p_{\max}$ is a tuning parameter. If $\mu(\mathrm{x})$ is negative at some x, proper scaling on $\mu(\mathrm{x})$ is necessary to keep $\{\mu(\mathrm{x}) \geq 0,\ \forall \mathrm{x} \in \mathscr{D}\}$ in the first inclusion probability function. Each of these

inclusion probabilities gives more weight to locations with higher $\mathbb{E}[Y]$ which should provide more informative data as discussed in Section 4.4. Moreover, if μ and $\sigma^2(x)$ are constant, the rejection sampling design reverts to the underlying space filling design.

The inclusion probability proportional to $\mu(x)$ weights the high intensity locations least and, hence, modifies the underlying random or balanced design the least. The inclusion probability proportional to expected intensity weights the high intensity locations the most. If there are large differences in the intensity function this inclusion probability can lead to survey designs that are too concentrated in only a small portion of the study domain. For this reason, we introduced above an inclusion probability proportional to the truncated expected intensity. The tuning parameter $p_{\max}$ governs how much weight is given to the highest intensity locations. For example, in one of the case studies over 90% of the prediction domain would have inclusion probability less than 5% if the probability was formed proportional to the expected intensity. By the choice of $p_{\max}$ we can control the proportion of design points within the low intensity region which forms the majority of the study domain.

6. Simulation studies

6.1. Study setting

In this section, we study the properties of spatially balanced designs and their rejection sampling versions introduced in Section 5 with simulation study. The study was carried out in the unit cube $[0, 1]^3$ with both a separable and an additive GP prior for the log intensity:

$$\text{Separable model: } \log\lambda(\mathrm{x}) = f(\mathrm{s}, t) \sim GP\left(\mu(\mathrm{x}), k(\mathrm{s}, \mathrm{s}')k(t, t')\right) \tag{20}$$

$$\text{Additive model: } \log\lambda(\mathrm{x}) = f(\mathrm{s}, t) \sim GP\left(\mu(\mathrm{x}), k(\mathrm{s}, \mathrm{s}') + k(t, t')\right). \tag{21}$$

The rationale for considering these two models is the following. The separable model is a commonly used "general purpose" spatiotemporal model whose covariance structure allows joint effects of space and time (Schmidt and O'Hagan, 2003; Kyriakidis and Journel, 1999). The additive model corresponds to $f(\mathrm{s}, t) = \mu(\mathrm{s}, t) + g(\mathrm{s}) + h(t)$ where the additive terms are mutually independent Gaussian processes $g(\mathrm{s}) \sim GP(0, k(\mathrm{s}, \mathrm{s}'))$ and $h(t) \sim GP(0, k(t, t'))$. In the additive model, the spatial pattern of intensity is stable in time but there are temporal relative changes in intensity. This can be used to represent, for example, distributions of species that are present in their stable distribution area only at certain time of the year as in our case study (see Section 7). In this case, the component $g(\mathrm{s})$ has the interpretation of distribution area and $h(t)$ explains the temporal changes in their abundance. In both models, the mean, $\mu(\mathrm{x})$, is a deterministic function that represents prior information about expected temporal changes in species intensity across the study domain.

We used a Matérn (2013) covariance function with 3/2 degrees of freedom $k_{\nu=3/2}(\mathrm{s}, \mathrm{s}') = \sigma_s^2\left(1 + \frac{\sqrt{3}|\mathrm{s}-\mathrm{s}'|}{l_\mathrm{s}}\right)\exp\left(\frac{\sqrt{3}|\mathrm{s}-\mathrm{s}'|}{l_\mathrm{s}}\right)$ in the spatial domain and a Gaussian covariance function $k_t(t, t') = \sigma_t^2 e^{-(t-t')^2/l_t^2}$ in the temporal domain. The positive parameters l_s and l_t are characteristic length-scales (Banerjee et al., 2015), which affect the correlation structures, and the positive parameters σ_s^2 and σ_t^2 are variance parameters that govern the magnitude of process variations. In the separable model, we set $\sigma_s^2 = 1$ for identifiability. We used a constant area of survey site $V_i = 1$ and concave temporal mean function $\mu(\mathrm{s}, t) = \mu(t) = a - c(t - b)^2$ with parameters $a = 2$, $b = 0.5$ and $c = 30$ so that the prior predictive probability of $y > 0$ is almost zero at the start and at the end of the time period $t \in [0, 1]$ but $\mathbb{E}[y]$ is clearly above zero in the middle of the time period. See an example in Fig. 3.

In order to gain understanding on performance of alternative designs with different kinds of spatiotemporal random effects we tested first the designs with a set of alternative fixed covariance function parameter values. The tested temporal range parameters were $l_t = \{0.2, 0.85, 1.5\}$, temporal variances were $\sigma_t^2 = \{0.5, 1, 2\}$ and spatial range parameters were $l_\mathrm{s} \in \{0.2, 0.4, \ldots, 1.6\}$. The spatial variance was fixed at $\sigma_\mathrm{s}^2 = 2$ in all experiments. The spatiotemporal random effects corresponding to these hyperparameter values range from very fast varying autocorrelated noise to a nearly monotonic trend within the study region and the magnitude of the random effect ranges

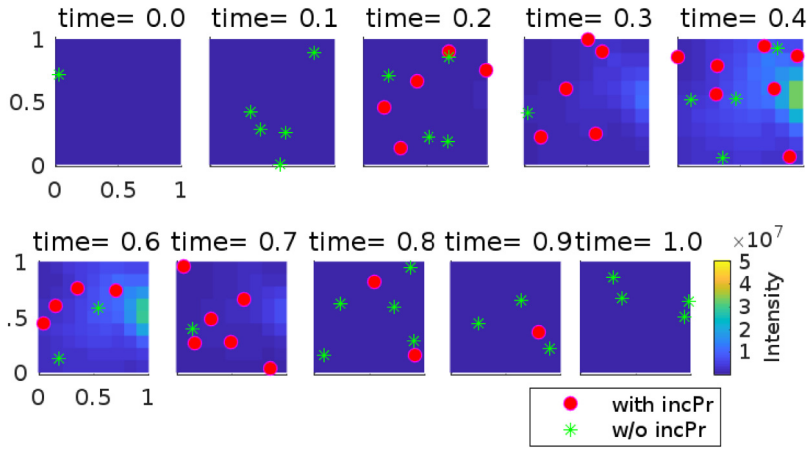


Fig. 3. A random draw from an additive GP with unimodal mean function along time (color surface) and samples from Sobol design ($n = 30$) with (red dots) and without (green asterisks) rejection sampling. The inclusion probability, $p(t) \propto \mu(t)$.

from near negligible to the same order of magnitude compared to the variation in the mean function. Since the hyperparameters are fixed, the evaluation criterion in this first experiment is the expected utility conditional on θ

$$\bar{U}(d_n) = \bar{U}(d_n, \theta = \tilde{\theta}) = \sum_{y \in \mathbb{N}^n} p(y|d_n) \int U(d_n, y, f(\cdot)) d\mathbb{P}(f(\cdot)|d_n, y, \tilde{\theta}). \tag{22}$$

Each design was evaluated with design sizes $n = 50, 100$ and 150 using the expected APV loss (9) and the expected KL-divergence (11) utility calculated using Lemma 1 and the Laplace approximation for the conditional posterior for latent variables.

As a second simulation study, we consider full posterior inference for both the hyperparameters and the latent variables. We tested two sets of priors. In *informative prior* case, we gave a Gaussian prior, $\mathcal{N}(0.85, 0.05)$, for s_l and t_l, and a Gamma prior with shape and inverse scale parameters equal to 20 for σ_t^2. These priors were set so that they were centered approximately at the mean of the alternative hyperparameter values in the first simulation study and that approximately 90% of the prior probability covered the range of the respective hyperparameter values. The *weakly informative priors* were inflated version of the informative priors with Gaussian prior, $\mathcal{N}(0.85, 0.09)$, for s_l and t_l, and a Gamma prior with shape and inverse scale parameters equal to 7 for σ_t^2. The utility/loss was computed with (8) where the KL-divergence was approximated using Lemma 1 and (15).

The inclusion probability used in the simulation studies is proportional to the expectation of the latent function; that is $p(\mathrm{x}) \propto \mu(t)$. An example, of a random draw from an additive GP ($l_t = 1, l_s = 1$ and $\sigma_s^2 = 2, \sigma_t^2 = 1$), together with Sobol design with and without rejection sampling is presented in Fig. 3. It depicts the allocation of more samples to times with high inclusion probabilities using rejection sampling. Both designs cover the whole unit square. In order to compare the effect of Poisson likelihood to the optimal design we evaluated the designs also with equal GP models with a Gaussian observation model and fixed hyperparameters. We only show the results of $n = 100$. Results with other design sizes were similar but the expected losses were smaller and expected utilities larger with increasing design size. We applied the GPstuff toolbox (Vanhatalo et al., 2013) in the calculations here and in the case study.

6.2. Results, average predictive variance

The differences between designs were qualitatively similar whether considering the APV of latent function or intensity so we show only the former here. Fig. 4 and Fig. 1 in the supplement

show the expected APV of intensity for designs with and without rejection sampling averaged over $l_s \in \{0.2, 0.4, \ldots, 1.6\}$ for different values of l_t and σ_t^2. Results for each l_s separately are given in Supplement. In general, the expected APV loss decreases when using rejection sampling compared to not using rejection sampling.

The Halton and minimum distance designs (min_dist and min_dran) have smaller loss than the alternatives, while the Fibonacci lattice design (fibo_lat) has the highest loss. The Sobol design has in general the second highest loss. The decrease in expected APV loss in designs with rejection sampling compared to corresponding designs without rejection sampling ranges from nearly zero (space_fill design) to approximately 20%–30% (rest of the designs).

The relative difference in expected APV between designs with and without rejection sampling increases with increasing l_t and decreasing l_s (see also figures in Supplement). This is reasonable since when temporal length-scale, l_t, increases the temporal variation of $f(\mathrm{x})$ around $\mu(t)$ gets smaller and observations at times around the prior predicted peak intensity time inform more about the spatial variation around $\mu(t)$ at other times as well. This is especially evident in the additive model where the spatial structure of the spatiotemporal random effect is the same, $g(\mathrm{s})$, throughout the time interval and only its level, $h(t)$, changes. With increasing l_t the spatiotemporal random effect approaches temporally constant spatial random effect in which case sampling at times when we expect to see most spatial variation in observations inform about the structure of spatiotemporal random effect at other times as well. With increasing spatial length-scale l_s the spatial variation in $\lambda(\mathrm{x})$ decreases and the prior uncertainty about spatial structure decreases as well.

In the second simulation study, with full MCMC inference, the expected APV losses of the latent function and intensity function increase considerably compared to losses with fixed hyperparameters. As a result, the difference in expected APV loss of designs with and without rejection sampling is practically negligible compared to the total expected APV (see Fig. 3 in Supplement) and smaller than the Monte Carlo error in approximate integration over future data.

To summarize, the more random spatial variation around $\mu(t)$ (that is the smaller l_s) and the less temporal variation (the longer l_t) we expect $f(\mathrm{x})$ to have, the more beneficial the rejection sampling algorithm is expected to be compared to its non-rejection sampling alternative. Similarly, if the prior mean has only spatial structure so that $\mu(\mathrm{x}) = \mu(\mathrm{s})$ the rejection sampling with inclusion probability proportional to $\mu(\mathrm{s})$ is expected to be the more beneficial the more random temporal variation (the smaller l_t) and the less spatial variation (the longer l_s) we expect $f(\mathrm{x})$ to have. When the prior mean varies in both space and time, rejection sampling with inclusion probability proportional to $\mu(\mathrm{x})$ is expected to decrease expected APV compared to designs without rejection sampling. This is illustrated in the case study experiments in Section 7.

6.3. Results, KL-divergence

Fig. 5 and Fig. 2 in the supplement show the expected KL-divergence $\bar{U}_{\mathrm{KL}}$ from prior to posterior under the different designs and alternative fixed hyperparameter values. The designs with rejection sampling work again better than the designs without rejection sampling. The expected KL-divergence is approximately 20% smaller in designs without rejection sampling compared to designs with rejection sampling. With full Bayesian inference, the expected KL-divergence utilities of designs with rejection sampling are approximately 15% larger compared to designs without rejection sampling in case of informative prior for hyperparameters. The rejection sampling designs outperform the corresponding balanced and random designs also with weakly informative priors but their relative differences are smaller. See Fig. 6 and Fig. 4 in Supplement. The relative differences between alternative designs are also larger in the expected KL-divergence than in the expected APV loss with both fixed hyperparameters and full MCMC approach. With both fixed hyperparameters and full Bayesian inference, Halton is again the best and Fibonacci design is the worst rejection sampling design; other well performing rejection sampling designs are Random and minimum distance (min_dist and min_dran) designs. The Sobol design is among the second best rejection sampling designs with fixed hyperparameters but among the worst with full Bayesian inference, similarly as it was among the worst in APV loss (see Fig. 4); however, it is among the worst non-rejection sampling design in all experiments. Hence, Sobol sequence seems to work reasonably well as a

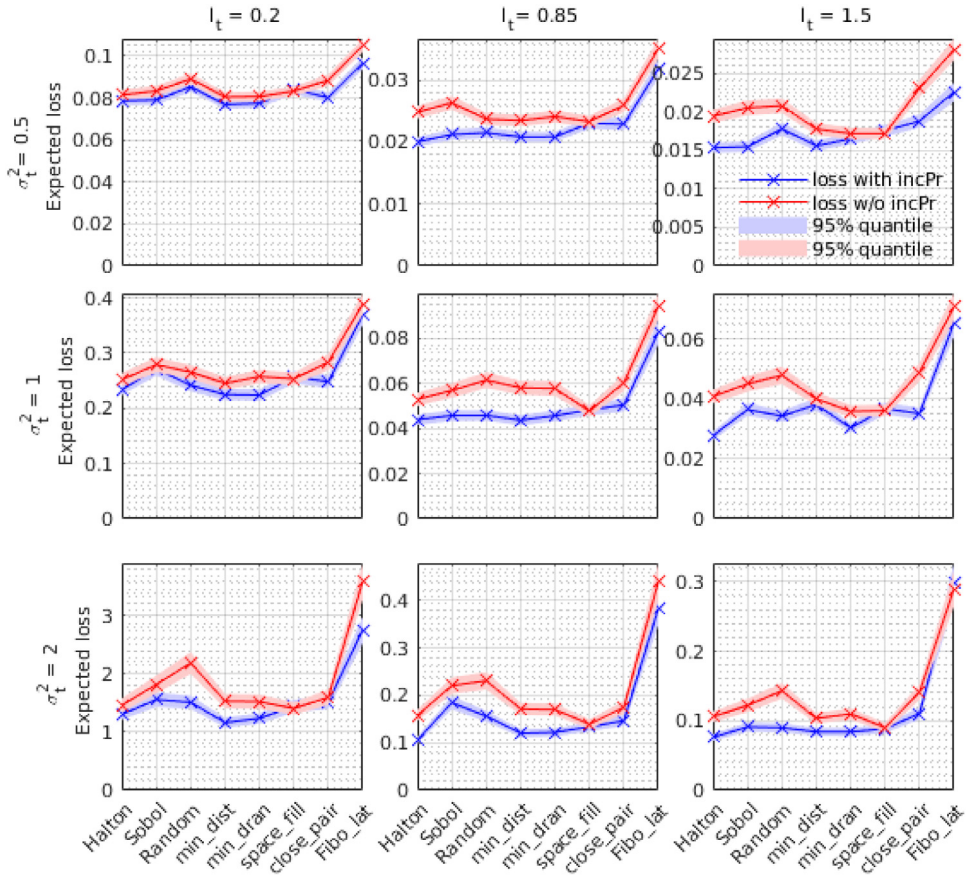


Fig. 4. The expected APV of Poisson intensity ($\bar{L}_{\text{APV}\lambda}$) of a model with separable covariance function at different values of l_t and σ_t^2 averaged over $l_s \in \{0.2, 0.4, \ldots, 1.6\}$ when using designs with and without rejection sampling (denoted as incPr in the legend) and $n = 100$. The crosses connected with solid lines show the Monte Carlo estimate of the loss and the highlighted regions show its 95% credible interval estimated as $\pm$ twice the standard error.

proposal distribution for rejection sampling when the aim is to update the joint posterior probability of latent variables with known hyperparameters. However, the Sobol sequence neither works well itself nor as a proposal distribution for rejection sampling to reduce the overall uncertainty in latent variables and hyperparameters. The best performing rejection sampling algorithms use balanced designs that produce spatially the most uniform allocation of survey sites among the alternatives. They also provide the most information about the posterior covariance structure of $f(\mathrm{x})$.

6.4. Comparison to models with the Gaussian likelihood

Contrary to the results concerning the Poisson likelihood, with the Gaussian likelihood the expected APV loss increases and the expected KL-divergence decreases when using rejection sampling compared to not using rejection sampling (see Supplement). This is well in line with earlier results on optimal spatial designs (Diggle and Lophaven, 2006) since with the Gaussian likelihood, data are equally informative everywhere in the whole study domain regardless of $\mu(\mathrm{x})$ and we learn the most about the latent function by sampling the domain "uniformly" (see Section 4.4). Hence, the optimal survey designs can be very different under the Gaussian and LGCP models.

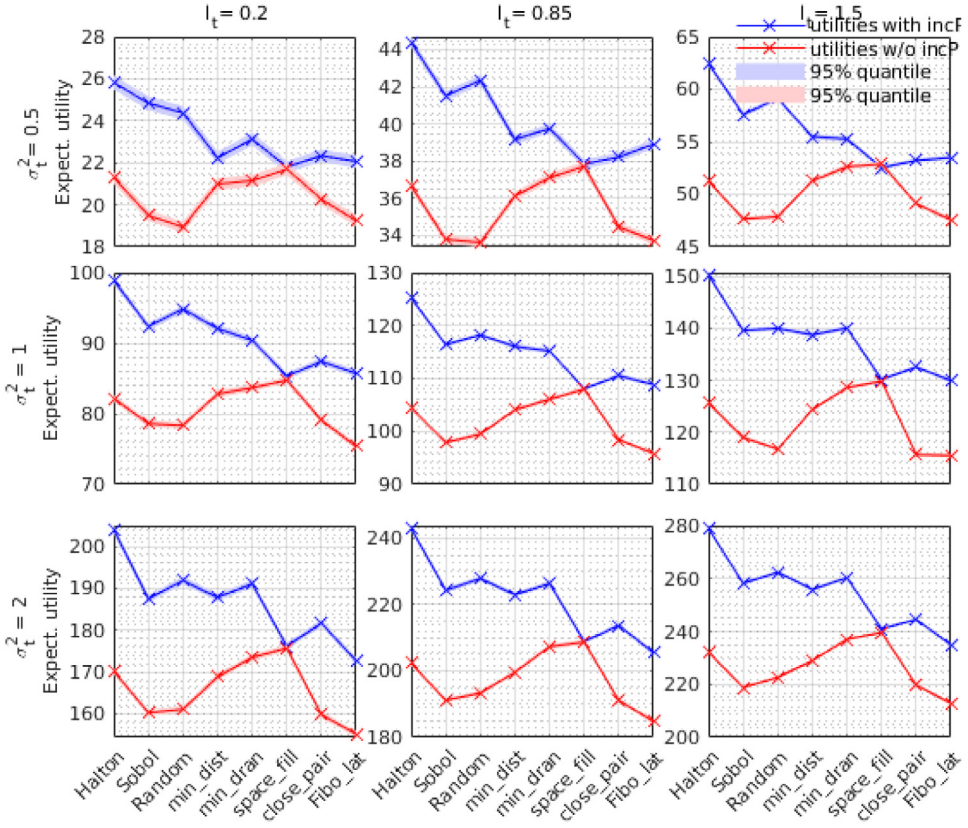


Fig. 5. The expected KL-divergence utility of latent function and intensity ($\bar{U}_{\mathrm{KL}\lambda}$) of a model with separable covariance function at different values of l_t and σ_t^2 averaged over $l_s \in \{0.2, 0.4, \ldots, 1.6\}$ when using designs with and without rejection sampling (denoted as incPr in the legend) and $n = 100$. The crosses connected with solid lines show the Monte Carlo estimate of the loss and the highlighted regions show its 95% credible interval estimated as $\pm$ twice the standard error. The estimated MC error becomes ignorable compared with the scale of utility when l_t and σ_t^2 increase.

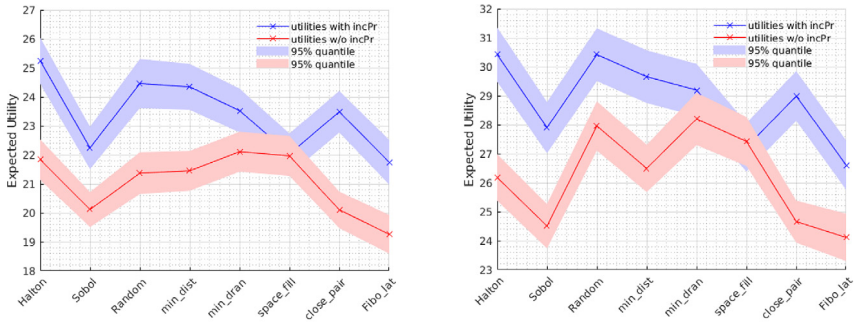


Fig. 6. The expected KL-divergence utility of latent function and intensity ($\bar{U}_{\mathrm{KL}\lambda}$) after full Bayesian inference when using designs with and without rejection sampling (denoted as incPr in the legend) and $n = 100$. The crosses connected with solid lines show the Monte Carlo estimate of the loss and the highlighted regions show its 95% credible interval estimated as $\pm$ twice the standard error.

7. Case study on fish reproduction areas

7.1. Species distribution model

In our case study, the aim is to construct a new spatiotemporal survey design to improve the distribution estimates in a region $\tilde{\mathscr{A}} \subset \mathscr{A}$ that is not included in the earlier data collection (Fig. 1, Section 2). The survey times should lie between calendar days 100–240 (from early April to the end of August) leading to the prediction domain $\tilde{\mathscr{D}} = \tilde{\mathscr{A}} \times [100, 204]$. The case study is based on the earlier data and model developed and validated by Kallasvuo et al. (2017). We modeled the observed larval counts as overdispersed Poisson process. Due to the small size of survey sites (transects) compared to total study region the intensity within survey sites could be treated as a constant leading to likelihood function (4). Then, at i'th survey site the observed number of larvae is $Y_i \sim \text{Poisson}(V_i\lambda(\mathrm{x}_i)\epsilon_i)$. Here, V_i is the sampled volume of water, $\lambda(\mathrm{x}_i)$ corresponds to the intensity of the Poisson point process in survey site at location x_i and ϵ_i is an independent random effect. The random effects describe, for example, non-structured stochasticity due to environmental conditions during the data collection. Since volumes V_i are approximately equal we gave a joint prior for the random effects with $\epsilon_i \sim \text{Gamma}(r, 1/r)$ where the Gamma distribution is parameterized with scale and shape so that $\mathbb{E}[\epsilon_i] = r\frac{1}{r} = 1$ and $Var[\epsilon_i] = 1/r$. We can now write $Y_i \sim \text{Poisson}(\tilde{\lambda}_i(\mathrm{x}))$ where $\tilde{\lambda}_i(\mathrm{x}) \sim \text{Gamma}(r, V_i\lambda_i/r)$ and marginalize over $\tilde{\lambda}_i(\mathrm{x})$ to get $Y_i \sim \text{Negative-Binomial}(V_i\lambda_i, r)$ where the Negative-Binomial distribution is parameterized so that $\mathbb{E}[Y_i] = V_i\lambda_i$ and $Var[Y_i] = \mathbb{E}[Y_i] + \mathbb{E}[Y_i]^2/r$. Hence, r is an overdispersion parameter corresponding to, for example, multiplicative independent random errors in observations (Lindén and Mäntyniemi, 2011). When the dispersion parameter $r \to +\infty$, the Negative Binomial approaches Poisson distribution. The likelihood function is now

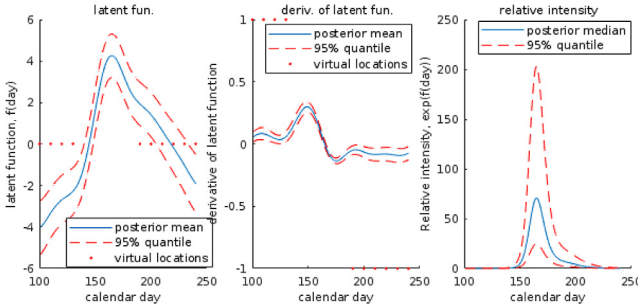
$$L(\mathrm{y}|f(\cdot), V, r) = \prod_{i=1}^{n} \text{Negative-Binomial}\,(V_i e^{f_i}, r). \tag{23}$$

The log intensity was given a zero mean additive GP prior

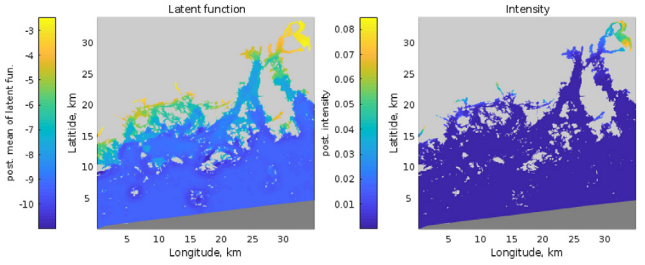
$$f(\boldsymbol{z}, \mathrm{s}, t, \tau) \sim GP\left(0, \sigma_\alpha^2 + \sum_{j=1}^{7} k_j(z_j, z_j') + k_8(\boldsymbol{z}, \boldsymbol{z}') + k_9(\mathrm{s}, \mathrm{s}') + k_{10}(\tau, \tau')\right),$$

where $\boldsymbol{z} \in \Re^7$ is the vector of environmental covariates and τ corresponds to the day of a year. The additive components are: σ_α^2 is the prior variance of intercept, $k_1, \ldots, k_7$ are Gaussian covariance functions related to additive covariate effects, k_8 is a Gaussian covariance function of joint covariate effect, k_9 is a Mátern, $\nu = 3/2$, covariance function of spatial random effect, and k_{10} is a Gaussian covariance function for the effect of a day within a year. Kallasvuo et al. (2017) modeled larval distribution only during their (approximate) peak abundance and did not include the last additive term. It was included here in order to model the development of larval abundance within a year which then provides information when the future surveys should be done. We gave weakly informative priors for the covariance function parameters so that inverse of length-scales and variance parameters were given half Student$_{\nu=4}$-$t(\mu = 0, s^2 = 1)$ prior distributions.

The survey days in the existing data are distributed rather sparsely from early May to the end of July. A zero mean GP prior for the calendar day effect thus gives ecologically unreasonable results due to the property of radial basis covariance functions to revert the GP prediction to the prior mean far from data. For this reason, we imposed a functional constraint for the calendar day effect that forces it to have positive (negative) derivative at the beginning (the end) of the potential survey period. The joint distribution of the latent function and its derivative $df(\boldsymbol{z}, \mathrm{s}, t, \tau)/d\tau$ is a Gaussian process (Rasmussen and Williams, 2006), so we can impose the monotonicity constraint by using virtual derivative observations (Riihimäki and Vehtari, 2010; Shively et al., 2009). We set in total 10 virtual observations for pike perch every ten days between calender days $[100, 130]$ and $[190, 240]$, whereas for Baltic herring we use 7 virtual observations between days $[100, 120]$ and $[210, 240]$. At the virtual observation locations within the former limits the derivative of the



(a) The calendar day effect with monotonicity information.



(b) The latent and intensity function at peak larval week.

Fig. 7. Subplot (a) shows the calendar day effect on larval abundance of pike perch, its derivative, and the corresponding relative intensity changes in larval abundance. The plots show also the virtual observation locations used to code the monotonicity information. Subplot (b) shows the expected posterior mean of the latent function and its corresponding intensity in the prediction region $\tilde{\mathscr{A}}$ on calendar day 165.

latent function was given a probit likelihood $\Phi(\rho^{-1}df/d\tau)$ and within the latter limits the derivative was given a likelihood $\Phi(-\rho^{-1}df/d\tau)$. The scaling parameter ρ governs how closely the standard Gaussian cumulative distribution function ($\Phi(\cdot)$) approximates the step function (Riihimäki and Vehtari, 2010) and it was set to $\rho = 10^{-6}$.

7.2. Survey designs

Since we have existing data to inform about the intensity function we base the choice of the rejection sampling design on posterior instead of prior predictive utility/loss. The prior predictive distribution $p(Y|d_n)$ in Eqs. (10) and (13) was replaced with the posterior predictive distribution $p(Y|\mathbf{y}, d_n)$, where $\mathbf{y}$ denotes the existing data. Similarly, $Var(f|d_n, Y)$ and $p(f|Y, d_n)$ were replaced by $Var(f|d_n, Y, \mathbf{y})$ and $p(f|Y, d_n, \mathbf{y})$. In the rejection sampling designs, we tested all three inclusion probabilities introduced in Section 5.2. The rejection sampling worked better than the corresponding balanced or random design without rejection sampling in each case. We report the results only for the best inclusion probability that was proportional to truncated expected intensity

$$p(\boldsymbol{z}, \mathrm{s}, \tau) \propto \min\left(p_{\max}, e^{\mathbb{E}[f(\boldsymbol{z},\mathrm{s},\tau)|\mathbf{y}]+2\mathrm{Var}\mathbb{E}[f(\boldsymbol{z},\mathrm{s},\tau)|\mathbf{y}]}\right). \tag{24}$$

The truncation threshold $p_{\max}$ was set to 0.15 for pike perch and to 0.5 for herring. The larval intensity changes significantly with calendar day and spatial location (Fig. 7). Hence, without threshold large proportion of the prediction domain would have practically zero acceptance probability during rejection sampling. With the chosen threshold values 90% of the prediction domain had 0.05 or

larger inclusion probability which results in good spatial coverage within the prediction domain. We used the Laplace method to form approximation for the posterior of the latent function so that we optimized the hyperparameters to their (approximate) MAP estimate and conditional on this estimate approximated the posterior of latent function (Vanhatalo et al., 2010).

We compared the alternative spatially balanced and random designs with and without rejection sampling using the expected APV loss and the expected KL-divergence with $n = 100$ design points. When constructing designs, we scaled the design space to the unit cube and used the same distance thresholds as in the simulation study. The prediction domain is not continuous but includes land areas that need to be ruled out (Fig. 7). First we used the traditional design sampling methods to generate candidate points from a cube that covers the subdomain $\tilde{\mathscr{D}}$ and the time interval of interest. We then applied the Branch-and-Prune method (Kubica, 2014) to rule out the land areas (and the sea area out-of scope of the study, see Fig. 10). In rejection sampling, the reject rule was then applied on each candidate point left. These steps were continued so long that we had as many design points as wanted.

The expected APV loss was calculated similarly as in the simulation studies. With temporal resolution of one week and spatial resolution of 50 meters the total number of 3D grid cells was 4 588 580. Since the prediction domain $\tilde{\mathscr{D}}$ does not include all data points (the old data falls outsize the study region), we would need to use (16) to calculate the KL-divergence instead of (14) that was used in the simulation studies. For this reason we would need to approximate also the KL-divergence on the 3D grid. Due to the size of the grid the required covariance matrix inversion was infeasible for which reason we report results only for APV loss.

7.3. Results

Fig. 7 summarizes the posterior distribution of the calendar day effect and the intensity function across the prediction region on the peak larval season of pike perch. Fig. 8 shows the weekly inclusion probability surfaces for the rejection sampling design in case of pike perch (the corresponding figures for Baltic herring are in Supplement). There are clear spatial and temporal differences in the intensity function that are transferred to inclusion probability. Due to strong variation in larval density, the inclusion probability is significantly above zero only during and near the peak larval period and decreases to practically zero ($<0.05\%$) over most of the region in the beginning and in the end of the study period.

The expected APV losses of the latent function are shown in Fig. 9. The results for APV loss of the intensity function were qualitatively similar so we omitted them here. In general, the designs with rejection sampling have the lowest APV loss. However, there are clear differences in the performance of alternative designs. Contrary to simulation studies and herring sampling, Halton and Sobol designs are not expected to be as good as other designs for pike perch. The reason for this is likely the qualitative difference between the inclusion probability surface of pike perch case study compared to that of herring case study and simulation study. In the pike perch study only in very small proportion of the prediction domain the inclusion probability is large (Fig. 8) whereas for herring and simulation studies the inclusion probability surface varies more moderately (Fig. 3 and Supplement). Hence, the Halton and Sobol designs do not seem to be good proposal distributions in case of heavily concentrated inclusion probability surface.

Fig. 10 shows the spatiotemporal configuration of the best rejection designs and the corresponding balanced and random designs without rejection sampling for pike perch (Random) and Baltic herring (Halton). Table 1 summarizes the weekly distribution of number of survey sites for these same designs. In the rejection sampling designs, the survey sites are clearly concentrated on the weeks around the peak larval period (Fig. 7 and Supplement). The survey design covers the whole spatial study area only during predicted peak larval period whereas on other weeks the survey sites concentrate on locations with expected high larval intensity (Fig. 7). This is reasonable since the high larval intensity spatial locations are the most informative on calendar day effect, and the peak larval period is expected to be the most informative on the spatial distribution of larvae. The survey sites are more evenly distributed throughout the prediction domain for herring than for pike perch. The even distribution is due to less variability of the intensity function for Baltic herring with a less peaked calendar day effect than the pike perch function.

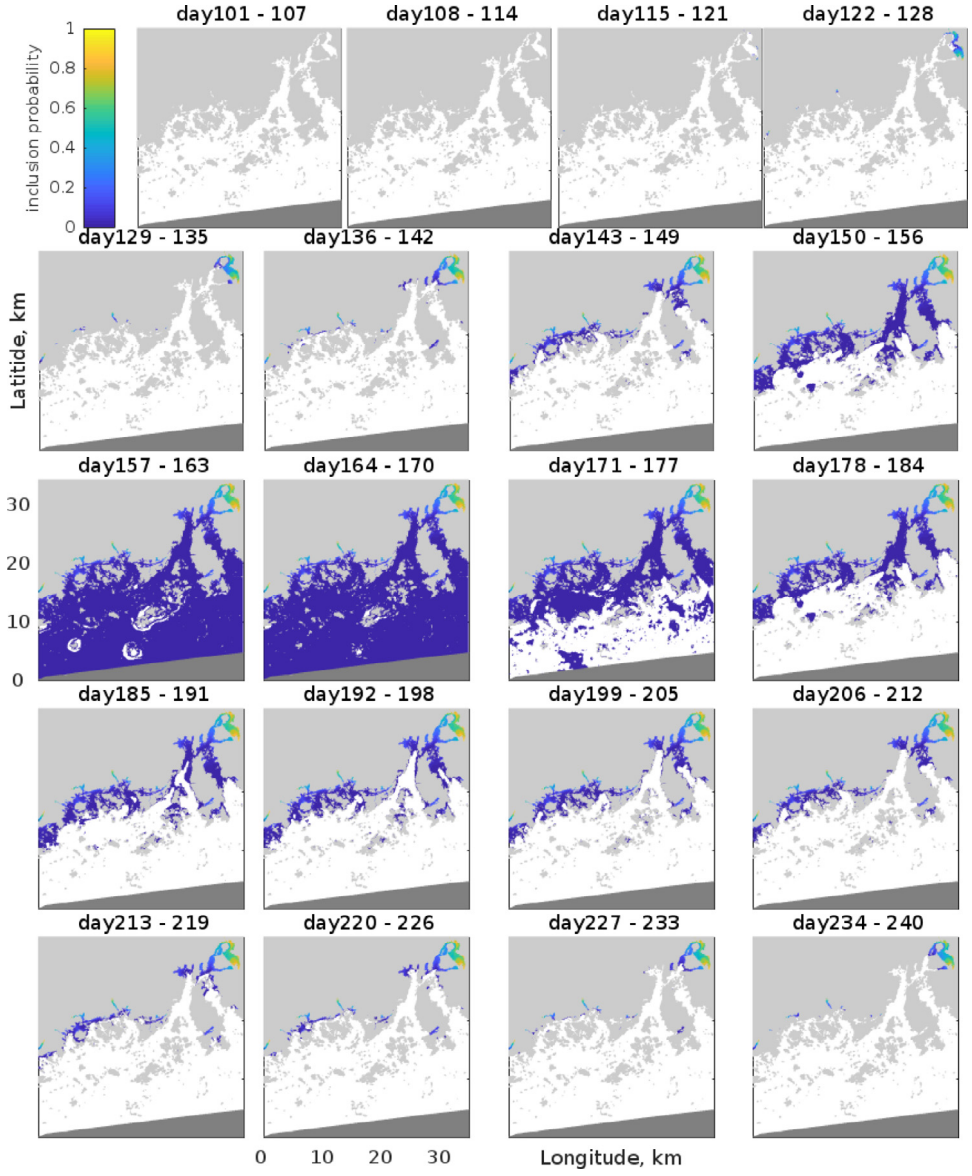
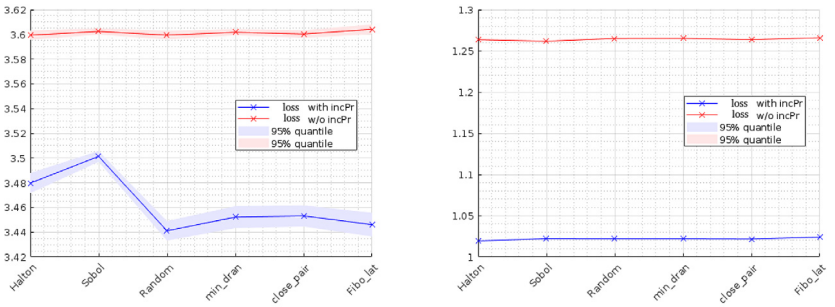


Fig. 8. The weekly inclusion probabilities (24) for rejection sampling design. The light gray areas are land, the dark gray color indicates sea area out of scope of this study and the white color shows the sea areas where inclusion probability is less than 0.05%.

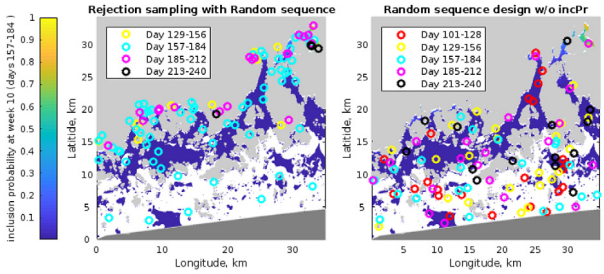
8. Discussion and conclusion

The LGCP is a widely used point process model in many practical applications. In ecology it has gained increasing interest since it is an efficient and theoretically valid approach to build species distribution models (Simpson et al., 2016; Yuan et al., 2017; Vanhatalo et al., 2017; Mäkinen and Vanhatalo, 2018). In these applications, collecting data is typically time consuming and expensive.

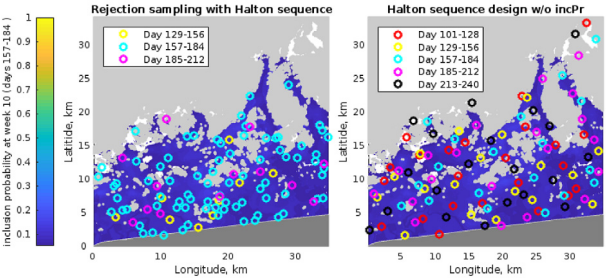


a) pike perch comparison b) herring comparison

Fig. 9. The APV loss of latent function for alternative designs in case of pike perch (a) and Baltic herring (b). The crosses connected with solid lines show the Monte Carlo estimates and the highlighted regions show the 95% credible interval of these estimates.



a) pike perch survey design



b) herring survey design

Fig. 10. The spatial configuration of design points with and without rejection sampling for each week during the survey period. Light gray areas are land and dark gray is sea area out of the prediction region. The water areas are colored according to the inclusion probability at day 165/171 with white corresponding to less than 0.1% / 5% inclusion probability for pike perch and herring, respectively.

For example, the small scale sampling for 100 new data points in our case study would cost approximately 50 000 euros and the costs of larger scale applications, such as the distance sampling for whale counting (Yuan et al., 2017) are easily in millions of euros. Hence, there is a real need for efficient survey designs in species distribution studies and their development has been active in recent years (Foster et al., 2017; Williams et al., 2018; Reich et al., 2018). There are no previous

Table 1
Weekly distribution of the number of survey sites for pike perch (Random design) and herring (Halton design).

Calender Day	With rejection sampling (pike perch/herring)	Without rejection sampling (pike perch/herring)
101–128	0/0	23/23
129–156	11/9	22/18
157–184	65/78	21/19
185–212	19/13	20 / 20
213–240	4/0	14 / 20

works on model-based survey designs for LGCPs though, and most of the existing data used in LGCP analysis are based on the classical balanced or stratified survey designs which are optimized for (linear) Gaussian models.

Our results show that in the presence of prior information on intensity function, survey designs that are expected to be most informative for partially observed LGCPs are different from traditional designs used for Gaussian models. This highlights that classical spatial designs can be inefficient for partially observed LGCPs. The difference is caused by larger spread of the Poisson distribution with increasing intensity. The closer to zero the intensity is the less uncertainty there is on the outcome of the future data and the less information new data is expected to provide. For this reason, we proposed a new method to construct survey designs, a spatially balanced or random design with rejection sampling, which gives more weight to prior predictive high intensity areas than low intensity areas. Our extensive simulation and case study experiments showed that when analyzed with APV loss and KL-divergence utility, the rejection sampling designs consistently outperformed the corresponding balanced and random designs. The relative performance of the rejection sampling designs versus balanced and random designs without rejection sampling was not sensitive to the variance and length-scale of the spatiotemporal Gaussian process. With all tested combinations of fixed length-scale and variance as well as with full MCMC the rejection sampling design performed better than the corresponding balanced or random design. The benefits from rejection sampling increasing for larger temporal and spatial length-scales. The inclusion probability in our new design algorithm is based on the prior (or a current posterior) mean of the intensity. The inclusion probability function can be also formulated differently but we leave more thorough studies on this for future.

One potential concern related to the rejection sampling design raised by a reviewer of this work is preferential sampling. Diggle et al. (2010) define preferential sampling to arise when the process that determines the data locations and process being modeled are dependent. They demonstrate that conventional geostatistical methods, which assume that survey designs are non-preferential, may then produce biased estimates. However, as mentioned by several discussants of their work (e.g. J. I. Illian, A. P. Dawid and R. D. Wilkinson), preferential sampling, as defined by Diggle et al. (2010), practically never occurs since building survey design cannot be related to the underlying process itself but it can depend only on survey designer's prior information on the process. This, however, can be taken into account in the prior of a Bayesian model. Hence, rejection sampling design introduced here is a specific example of general sequential data collection schemes in which prior information, I, provided by data collected so far is used to inform the future data collection (Lindley, 1956; Eidsvik et al., 2015). In sequential data collection schemes, the locations of new data, d_n, depend on the current prior information through $p(d_n|I)$ but the model for the new data is also conditional to that prior information $p(Y|d_n, I)$. Conditioning to the current information in the model automatically corrects for the biases that Diggle et al. (2010) were concerned about. In our simulation studies this prior information, I, corresponds to the mean of the GP model and in the case study I corresponds to the posterior of the intensity function conditional on the current data.

Our inferential interest was in the latent function and the intensity function. In light of this objective, the rejection sampling designs worked better than the corresponding balanced and random designs with both fixed hyperparameters and the full Bayesian analysis where hyperparameters

were inferred as well. One obvious future research direction is, however, to study which designs are best for learning most about the hyperparameters of partially observed LGCPs instead of or alongside the intensity function. Our analysis in Section 4.4 shows also that the expected utility of survey sites decreases as the size of the survey sites, V_i decreases. However, we leave more thorough analysis and experiments on the effects of V_i for future studies.

Our case study has a direct relevance to pike perch and herring fisheries management. The rejection sampling method introduced here is straightforward to implement and, hence, can easily be applied in other regions as well. The new data can then be used to revise species distribution maps that are used in regional marine spatial planning and coastal land use management. The results illustrated that the rejection sampling method can produce very different survey designs for different species. The survey sites for herring were more uniformly distributed than the survey sites for pike perch. If these two species were to be surveyed at the same time, a good joint design should be a compromise between them. In our application, reaching the survey sites was not an issue but in larger survey domains the design should take into account also logistic and financial constraints. In theory, these could be included naturally into the Bayesian model-based design by redefining the utility and loss functions to account for the survey costs or equivalently by defining a constraint functions for designs. In this case, we could define an inclusion probability function that is weighted by these constraints.

In this work, our goal was to develop a design algorithm, which offers improvements over existing balanced and random designs and is computationally easy to implement. Our specific interest was in partially observed LGCP for which the proposed rejection sampling method is straightforward to implement and improves the existing balanced and random designs. However, the core idea behind the rejection sampling method to build survey designs is very general and it could be applied to other models as well.

Acknowledgments

This research was funded by Academy of Finland (Grants 304531 and 317255) and the research funds of University of Helsinki, Finland (decision No. 465/51/2014). The authors acknowledge CSC–IT Center for Science, Finland, for computational resources. The authors are grateful to the two anonymous reviewers for their insightful and constructive reviews, which are noticeably in help of the improvements of the manuscript.

Appendix A. Results on KL-divergence

Proof of Lemma 1. Assume $f(\cdot) : \mathscr{D} \rightarrow \Re$ is a latent function with a Gaussian process prior and denote the prior probability measure of $f(\cdot)$ in the domain $\mathscr{D}$ by $\mathbb{P}(f(\cdot))$. Denote by $\mathrm{f} = [f(\mathrm{x}_1), f(\mathrm{x}_2), \ldots, f(\mathrm{x}_n)]$ a vector of latent variables, at finite number of locations $d_n = [\mathrm{x}_1^T, \ldots, \mathrm{x}_n^T]$ where $\mathrm{x}_i \in \mathscr{D}$, and by $Y = [Y_1, \ldots, Y_n]$ a random vector of observations at these locations d_n. The observations are assumed to be conditionally independent given the latent variables; that is $p(Y = y|f(\cdot)) = p(Y = y|\mathrm{f}) = \prod_{i=1}^{n} p(y_i|f(\mathrm{x}_i))$, where $y = [y_1, \ldots, y_n]$ is a realization of Y. Denote by $\mathbb{P}(f(\cdot)|d_n, Y)$ the posterior probability measure. By Bayes theorem $\frac{d\mathbb{P}(f(\cdot)|d_n,Y)}{d\mathbb{P}(f(\cdot))} = \frac{p(y|f(\cdot))}{p(y)}$. Since the posterior probability measure is absolutely continuous with respect to the prior probability measure (Schervish, 1995), we can calculate the KL-divergence from the prior to the posterior by

$$
\begin{aligned}
&\mathrm{KL}\Big(d\mathbb{P}(f(\cdot)|d_n, y) \parallel d\mathbb{P}(f(\cdot))\Big) = \int \log \frac{d\mathbb{P}(f(\cdot)|d_n, y)}{d\mathbb{P}(\cdot)} d\mathbb{P}(f(\cdot)|d_n, y) \\
&= \int \log \frac{p(y|f(\cdot))d\mathbb{P}(f(\cdot))}{d\mathbb{P}(f(\cdot)) \int p(y|f(\cdot))d\mathbb{P}(f(\cdot))} d\mathbb{P}(f(\cdot)|d_n, y) \\
&= \int \log p(y|f(\cdot))d\mathbb{P}(f(\cdot)|d_n, y) - \log p(y) \\
&= \int \log p(y|\mathrm{f})d\mathbb{P}(\mathrm{f}|y) - \log p(y), \qquad \text{(A.1)}
\end{aligned}
$$

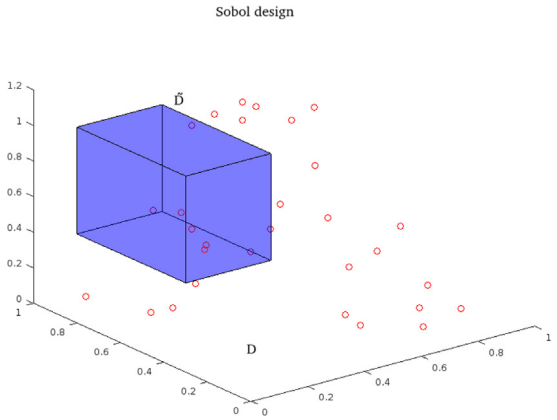


Fig. A.1. d_n includes 30 locations in a unit cube ($\mathscr{D}$). We are interested in a region $\tilde{\mathscr{D}}$ marked by a blue cube which is a subset of $\mathscr{D}$.

where $p(y) = \int p(y|f(\cdot))d\mathbb{P}(f(\cdot)) = \int p(y|\mathrm{f})p(\mathrm{f})d\,\mathrm{f}$. The last equality holds because $d_n \subset \mathscr{D}$. In case of Gaussian observation model $p(y_i|f_i) \sim N(f_i, \sigma^2)$, this simplifies to KL divergence between two multivariate Gaussian distributions

$$\mathrm{KL}\left(d\mathbb{P}(f(\cdot)|d_n, y) \parallel d\mathbb{P}(f(\cdot))\right) = \frac{1}{2}\left[\log|KK_1^{-1}| + \mathrm{Tr}(K_1K^{-1}) + \boldsymbol{\mu}_1^T K^{-1}\boldsymbol{\mu}_1 - n\right],$$

where $\boldsymbol{\mu}_1 = K(K+\sigma_n^2 I)^{-1}y$ and $K_1 = K - K(K+\sigma_n^2 I)^{-1}K$ are the posterior mean and covariance of f.

Let us next consider the KL divergence from the prior to posterior of $f(\cdot)$ over a region (or subset) of locations of $\mathscr{D}$ that does not contain all the observations. This is illustrated in Fig. A.1. We denote by $\tilde{\mathscr{D}} \subset \mathscr{D}$ this subregion and by $f_{\tilde{\mathscr{D}}}(\cdot) : \tilde{\mathscr{D}} \to \Re$ the latent function restricted to this subregion. The KL divergence for $f_{\tilde{\mathscr{D}}}$ is

$$\mathrm{KL}\left(d\mathbb{P}(f_{\tilde{\mathscr{D}}}(\cdot)|d_n, y) \parallel d\mathbb{P}(f_{\tilde{\mathscr{D}}}(\cdot))\right) = \int \log p(y|f_{\tilde{\mathscr{D}}})d\mathbb{P}(f_{\tilde{\mathscr{D}}}|d_n, y) - \log p_{\tilde{\mathscr{D}}}(y),$$

where $p_{\tilde{\mathscr{D}}}(y) = \int p(y|f_{\tilde{\mathscr{D}}})d\mathbb{P}(f_{\tilde{\mathscr{D}}}(\cdot))$. Note that $p(y|f_{\tilde{\mathscr{D}}}) = \int p(y|\mathrm{f})d\mathbb{P}(\mathrm{f}|f_{\tilde{\mathscr{D}}})$. The challenging part of this equation is the conditional probability measure $d\mathbb{P}(\mathrm{f}|f_{\tilde{\mathscr{D}}})$. In practice we need to discretize the subdomain $\tilde{D}$ with fine grid cells indexed by x_{*j}, and approximate the conditional measure by a conditional distribution $p(\mathrm{f}|\mathrm{f}_{\tilde{\mathscr{D}}})$ where $\mathrm{f}_{\tilde{\mathscr{D}}} = \{\mathrm{f}(x_{*,1}), \mathrm{f}(x_{*,2}), \ldots, \mathrm{f}(x_{*,N})\}$. Hence all the above representations can be approximated numerically but with large $\tilde{\mathscr{D}}$ and fine discretization the calculations might become infeasible.

The KL divergence of intensity. Let $\mathbb{P}(\lambda(\cdot))$ and $\mathbb{P}(\lambda(\cdot)|d_n, y)$ denote the prior and posterior probability measure of the intensity function. As shown in Appendix A, the KL-divergence from the prior to the posterior of the intensity is

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}(\lambda(\cdot)|d_n, y) \parallel \mathbb{P}(\lambda(\cdot))\right) &= \int \log p(y|\lambda)\frac{p(y|\lambda)p(\lambda)}{p(y)}d\lambda - \log p(y) \\
&= \int \log p(y|e^{f_1}, \ldots, e^{f_n})\frac{p(y|e^{f_1}, \ldots, e^{f_n})p_f(\log\lambda)}{p(y)\prod_{i=1}^n \lambda_i}d\lambda \\
&\quad - \log p(y) \\
&= \int \log p(y|\mathrm{f})p(\mathrm{f}|d_n, y)d\,\mathrm{f} - \log p(y), \qquad \text{(A.2)}
\end{aligned}$$

where $\lambda = [e^{f_1}, \ldots, e^{f_n}]$ and by change of variables $p(\lambda) = p_f(\log(\lambda))/\prod_{i=1}^n \lambda_i$ and $df_i = d\lambda_i/\lambda_i$. Since $p(y|e^x)$ and $p(y|x)$ have same σ algebra, we have that $p(y|e^{f_1}, \ldots, e^{f_n}) = p(y|\mathrm{f})$.

Appendix B. Space filling rejection sampling design

The space filling rejection sampling design on discrete space is generated as follows:

0. Generate a set of candidate design locations $C = \{\bar{\mathrm{x}}_j : \bar{\mathrm{x}}_j \in \mathscr{D}\}$; for example a dense grid or a Halton/Sobol sequence.
1. Pick up a location $\mathrm{x}_1 \in C$ from a corner of the domain and include that into the design $d_1 = \{\mathrm{x}_1\}$. Set $k = 1$ and $i = 1$.
2. Search the location $\mathrm{x}_{i+1} = \arg\max^{k}_{\mathrm{x}^* \in C} \min_{\mathrm{x}_j \in d_i} \| \mathrm{x}_j - \mathrm{x}_* \|$ where $\arg\max^{k}_{\mathrm{x}^* \in C}$ denotes the k'th largest value.
3. Apply rejection sampling for x_{i+1}. If x_{i+1} is rejected, set $k = k + 1$ and return to Step 2. Otherwise include x_{i+1} in to the design $d_{i+1} = d_i \cup \{\mathrm{x}_{i+1}\}$ and set $i = i + 1$ and $k = 1$;
4. Repeat Steps 2 and 3 until the n'th location has been found.

Appendix C. Availability and implementation

A MATLAB toolbox named *Experimental-design* related to this article is available online, and can be downloaded from GitHub https://github.com/jialiuGit/Experimental-design.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2019.100392.

References

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2015. Hierarchical Modelling and Analysis for Spatial Data, second ed. Chapman Hall/CRC.

Cambanis, S., 1985. 13 sampling designs for time series. Handbook of Statist. 5, 337–362.

Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M., Silander, J.A., 2011. Point pattern modelling for degraded presence-only data over large regions. J. R. Stat. Soc. Ser. C. Appl. Stat. 60, 757–776.

Chipeta, M.G., Terlouw, D.J., Phiri, K., Diggle, P.J., 2016. Adaptive geostatistical design and analysis for sequential prevalence surveys. Spatial Stat. 15, 70–84.

Chipeta, M., Terlouw, D., Phiri, K., Diggle, P., 2017. Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. Environmetrics 28, 1–11.

Clark, J.S., Nemergut, D., Seyednasrollah, B., Turner, P.J., Zhang, S., 2016. Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. Ecol. Monograph 87, 34–56.

Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley.

Diggle, P., Lophaven, S., 2006. Bayesian geostatistical design. Scand. J. Stat. 33, 53–64.

Diggle, P.J., Menezes, R., li Su, T., 2010. Geostatistical inference under preferential sampling. J. R. Stat. Soc. Ser. C. Appl. Stat. 59, 191–232. http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x.

Eidsvik, J., Mukerji, T., Bhattacharjya, D., 2015. Value of Information in the Earth Sciences. Cambridge University Press.

Elith, J., Leathwich, J.R., 2009. Species distributions models: Ecological explanation and predictions across space and time. Annu. Rev. Ecol. Evol. Syst. 40.

Family, F., Vicsek, T., 1985. Scaling of the active zone in the Eden process on percolation networks and the ballistic deposition model. J. Phys. A: Math. Gen. 18, L75.

Foster, S.D., Hosack, G.R., Lawrence, E., Przeslawski, R., Hedge, P., Caley, M.J., Barrett, N.S., Williams, A., Li, J., Lynch, T., Dambacher, J.M., Sweatman, H.P., Hayes, K.R., 2017. Spatially balanced designs that incorporate legacy sites. Methods Ecol. Evol..

Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P., 2010. Handbook of Spatial Statistics. CRC press.

Gelfand, A.E., Jr, J.A.S., Wu, S., Latimer, A., Lewis, P.O., Rebelo, A.G., Holder, M., 2006. Explaining species distribution patterns through hierarchical modelling. Bayesian Anal. 1, 41–92.

Grafström, A., Lundström, N.L., Schelin, L., 2012. Spatially balanced sampling through the pivotal method. Biometrics 68, 514–520.

Illian, J.B., Martino, S., Sørbye, S.H., Gallego-Fernández, J.B., Zunzunegui, M., Esquivias, M.P., Travis, J.M.J., 2013. Fitting complex ecological point process models with integrated nested laplace approximation. Methods Ecol. Evol. 4, 305–315.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. Statistical Analysis and Modelling of Spatial Point Patterns. Wiley.

Illian, J.B., Sørbye, S.H., Rue, H., 2012. A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (INLA). Ann. Appl. Stat. 6, 1499–1530.

Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. J. Statist. Plann. Inference 26, 131–148.

Kallasvuo, M., Vanhatalo, J., Veneranta, L., 2017. Modeling the spatial distribution of larval fish abundance provides essential information for management. Can. J. Fish. Aquat. Sci. 74.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795. http://dx.doi.org/10.2307/2291091.
Koehler, J., Owen, A., 1996. 9 computer experiments. Handbook of Statist. 13, 261–308.
Kotta, J., Vanhatalo, J., Jänes, H., Orav-Kotta, H., Rugiu, L., Jormalainen, V., Bobsien, I., Viitasalo, M., Virtanen, E., Sandman, A.N., Isaeus, M., Leidenberger, S., Jonsson, P., Johannesson, K., 2019. Integrating experimental and distribution data to predict future species patterns. Sci. Rep. 9, 1821. http://dx.doi.org/10.1038/s41598-018-38416-3.
Kubica, B.J., 2014. Excluding regions using sobol sequences in an interval branch-and-prune method for nonlinear systems. Reliab. Comput. 19.
Kullback, S., 1987. Letter to the editor: The Kullback-Leibler distance. Am. Stat. 41, 340–341.
Kyriakidis, P.C., Journel, A.G., 1999. Geostatistical space–time models: a review. Math. Geol. 31, 651–684.
Lindén, A., Mäntyniemi, S., 2011. Using the negative binomial distribution to model overdispersion in ecological count data. Ecology 92, 1414–1421.
Lindley, D.V., 1956. On a measure of the information provided by an experiment. Ann. Math. Stat. 27, 986–1005.
Lindley, D.V., 2003. Making Decisions, second ed. John Wiley & Sons.
Lombardo, L., Opitz, T., Huser, R., 2018. Point process-based modeling of multiple debris flow landslides using INLA: an application to the 2009 messina disaster. Stoch. Environ. Res. Risk Assess. 32, 2179–2198.
Mäkinen, J., Vanhatalo, J., 2018. Hierarchical bayesian model reveals the distributional shifts of Arctic marine mammals. Divers. Distrib. 24, 1381–1394.
Matérn, B., 2013. Spatial Variation, volume 36. Springer Science & Business Media.
Møller, J., Syversveen, A.R., Waagepetersen, R.P., 1998. Log Gaussian Cox processes. Scand. J. Stat. 25, 451–482.
Møller, J., Waagepetersen, R.P., 2004. Statistical Inference and Simulation for Spatial Point Processes, one hundred ed. In: Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, Boca Raton, FL.
Møller, J., Waagepetersen, R.P., 2007. Modern statistics for spatial point processes. Scand. J. Stat. 34, 643–684.
Müller, P., 1999. Simulation based optimal design. Bayesian Stat. 25, 459–474.
Müller, W.G., 2001. Coffee-house designs. In: Optimum Design 2000. Springer, pp. 241–248.
Müller, W.G., 2007. Collecting spatial data. In: Optimum Design of Experiments for Random Fields.
Nychka, D., Saltzman, N., 1998. Design of air-quality monitoring networks. In: Case Studies in Environmental Statistics. Springer, pp. 51–76.
O'Hagan, A., Kingman, J.F.C., 1978. Curve fitting and optimal design for prediction. J. R. Stat. Soc. Ser. B Stat. Methodol. 1–42.
Pei, S.-C., Liu, H.-H., Chen, J.-H., 2009. Color quantization by 3D spherical fibonacci lattices. In: Image Processing, ICIP, 16th IEEE International Conference on. IEEE, pp. 489–492.
Rasmussen, C.E., Williams, C.K., 2006. Gaussian Processes for Machine Learning. Citeseer.
Reich, B.J., Pacifici, K., Stallings, J.W., 2018. Integrating auxiliary data in optimal spatial design for species distribution modelling. Methods Ecol. Evol. 9, 1626–1637.
Renner, I.W., Warton, D.I., 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 69, 274–281.
Riihimäki, J., Vehtari, A., 2010. Gaussian processes with monotonicity information. In: Artificial Intelligence and Statistic, 13th International Conference on, vol. 9, pp. 645–652.
Robert, C.P., 2004. Monte Carlo Methods. Wiley Online Library.
Robertson, B., Brown, J., McDonald, T., Jaksons, P., 2013. BAS: Balanced acceptance sampling of natural resources. Biometrics 69, 776–784.
Royle, J., Nychka, D., 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Comput. Geosci. 24, 479–488.
Russo, D., 1984. Design of an optimal sampling network for estimating the variogram. Soil Sci. Am. J. 48, 708–716.
Ryan, E.G., Drovandi, C.C., McGree, J.M., Pettitt, A.N., 2016. A review of modern computational algorithms for Bayesian optimal design. Internat. Statist. Rev. 84, 128–154.
Schervish, M.J., 1995. Theory of Statistics. Springer Series in Statistics.
Schmidt, A.M., O'Hagan, A., 2003. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. J. R. Stat. Soc. Ser. B Stat. Methodol. 65, 743–758.
Shively, T.S., Sager, T.W., Walker, S.G., 2009. A Bayesian approach to non-parametric monotone function estimation. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 159–175.
Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., Rue, H., 2016. Going off grid: Computationally efficient inference for log-Gaussian cox processes. Biometrika 103, 49–70.
Sobol, I.M., 1976. Uniformly distributed sequences with an additional uniform property. USSR Comput. Math. Math. Phys. 16, 236–242.
Stein, M.L., Handcock, M.S., 1989. Some asymptotic properties of kriging when the covariance function is misspecified. Math. Geol. 21, 171–190.
Stevens, Jr., D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. J. Amer. Statist. Assoc. 99, 262–278.
Van Groenigen, J., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. J. Environ. Qual. 27, 1078–1086.
Vanhatalo, J., Hartmann, M., Veneranta, L., 2019. Additive multivariate gaussian processes for joint species distribution modeling with heterogeneous data. Bayesian Anal. http://dx.doi.org/10.1214/19-BA1158.
Vanhatalo, J., Hosack, G.R., Sweatman, H., 2017. Spatio-temporal modelling of crown-of-thorns starfish outbreaks on the great barrier reef to inform control strategies. Ecol. Appl. 54, 188–197.

Vanhatalo, J., Pietiläinen, V., Vehtari, A., 2010. Approximate inference for disease mapping with sparse Gaussian processes. Stat. Med. 29, 1580–1607.
Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., Vehtari, A., 2013. GPstuff: Bayesian modeling with Gaussian processes. J. Mach. Learn. Res. 14, 1175–1179.
Vlachos, P., Gelfand, A., 1996. Bayesian Decision Theoretic Design for Group Sequential Medical Trials Having Multivariate Patient Response. Technical Report. University of Connecticut.
Warton, D.I., Shepherd, L.C., 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. Ann. Appl. Stat. 4, 1383–1402.
Williams, P.J., Hooten, M.B., Womble, J.N., Esslinger, G.G., Bower, M.R., 2018. Monitoring dynamic spatio-temporal ecological processes optimally. Ecology 99, 524–535, arXiv:1707.03047.
Yuan, Y., E. Bachl, F., Lindgren, F., Borchers, D., B. Illian, J., T. Buckland, S., Rue, H., Gerrodette, T., 2017. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. Ann. Appl. Stat. 11, 2270–2297.