Log-Gaussian Cox processes and line transect sampling: optimal design critera

Kenneth Flagg[a,*], John Borkowski[a], Andrew Hoegh[a]

[a] *Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717*

Abstract

Goal of this paper. Evaluate line-transect designs in terms of many design-based and model-based criteria for spatial prediction using Bayesian LGCP models. Identify promising criteria and initial designs for later optimization and sequential design (not actually optimizing yet). Illuminate any relationships among criteria that will be helpful for constrained optimization. Other innovations are to compute design-based criteria with respect to the path (i.e. as path integrals) and to introduce a model-based APV criterion with a variable weight that more strongly penalizes errors near a decision threshold.

Keywords: log-Gaussian Cox process, optimal sampling, model-based design, spatial sampling design

1. Introduction

Spatial point processes models have long been considered generally infeasible because of their computational demands, but recent advances in Bayesian computing have made the Log-Gaussian Cox process an attainable model in practice (Rue et al., 2009; Lindgren et al., 2011; Illian et al., 2012; Simpson et al., 2016). Variable sampling effort leads to a degraded point pattern Chakraborty et al. (2011) and it is relatively simple to accomodate variable sampling effort in these

[*] Corresponding author
Email address: `kenneth.flagg@montana.edu` (Kenneth Flagg)

models using modern computing tools (Yuan et al., 2017). However, the literature on optimal sampling for spatial point process models is in its infancy (Liu and Vanhatalo, 2020).

Point pattern data are routinely collected in species distribution studies and ordnance response projects. These applications may use quadrat sampling or line-transect sampling, with transect sampling being more common. When the objective is mapping where events occur in space, various spatial mapping procedures have been used. Traditionally these have involved aggregating the data to grid cell counts or computing moving averages. These have the downside of introducing arbitrary structure into the data by the choice of gridding scheme or averaging window, and require uneccessary computation effort (Simpson et al., 2016). Software is now available to fit spatial point process models to data acquired via distance sampling and simultaneously estimate the detection function (Johnson et al., 2014).

In ecological settings, sampling plans are often designed around the goal of estimating total abundance. Ordnance response surveys are typically designed with the objective of detecting (but not necessarily mapping) intensity hotspots. However, to our knowledge, there has been very little work done in deciding *where* to collect data when the goal is to map the intensity using a spatial point process model.

1.1. Paths as sampling designs

Liu and Vanhatalo (2020) used narrow quadrats (swaths along line-transects) as their sampling units. The transects were short relative to the size of the study region and not connected into a path.

Pollard et al. (2002) adaptively zigzagged their line transects in a species abundance survey.

Add relevant VSP references. VSP can create parallel line-transect plans and add infill based on the actual course-over-ground, but I do not recall what criteria this may optimize.

While some ideas about the characteristics of a good point design apply to paths, creating an optimal path design is not as simple as connecting the points of a point design with line segments. There are many ways to connect points into a path, so optimal design criteria must apply to the whole path and not only to the waypoints.

1.2. Design-based sampling

Most work done for points. (Or quadrats approximated as points?) Space-filling criteria may be good starting points (e.g nearest-neighbor distance).

Not using space-filling point desings? (i.e. designs that have nonzero area as sample size goes to infinity)

1.3. Space-filling curves

Used in circuit design (Fan et al., 2014) (*find more citations*) and high-dimensional data visualization in bioinformatics (Anders, 2009). Peano curve is very flexible for filling irregular shapes (Fan et al., 2014). Hilbert curve is easy to construct.

Space-filling curves are one-dimensional paths constructed iteratively; as the number of iterations goes to infinity, the limiting path has nonzero area and actually fills the space (Sagan, 1994). For applications we stop after a finite number of iterations.

1.4. Model-based spatial design

Regularity is optimal for spatial prediction but randomness and a variety of interpoint distances are best for parameter estimation (Diggle and Lophaven, 2006). Inhibitory plus close pairs is a good compromise (Chipeta et al., 2017).

1.5. Multi-objective optimization

summarize Lark (2016) and related

we use a large suite of criteria to explore the relationships among them

1.6. Notation and Terminology

- need to add notation for paths!
- process defined on $\mathcal{D} \subset \mathbb{R}^2$, domain of the intensity function, define $d = \dim(\mathcal{D})$
- observation window $\mathcal{S} \subset \mathcal{D}$
- define three regions:
  - the domain $\mathcal{D}$ over which the process mathematically operates
  - the study region $\mathcal{R}$ over which inferences are desired
  - the observed/sampled observation window $\mathcal{S}$
- general relationship is $\mathcal{S} \subset \mathcal{R} \subset \mathcal{D} \subset \mathbb{R}^d$ where all of the subset symbols taken to mean "subset or equal"
- $\mathcal{D}$ can be unbounded or bounded (often $\mathbb{R}^d$), $\mathcal{S}$ practically always bounded, $\mathcal{R}$ bounded or unbounded depending on application and inferential goals
- the "fully surveyed" situation is $\mathcal{S} = \mathcal{R}$
- $\mathbf{X}$ point process on $\mathcal{R}$, $\mathbf{x} = \{x_1, \ldots, x_n\}$ realized point pattern
- point $x \in \mathbf{x}$ called an event
- intensity function $\lambda(u)$
- types of "points" in space:
  - x event in the point pattern
  - s numerical integration node
  - u arbitrary location in $\mathcal{D}$ used to index intensity function and predictors
- $z(u)$ a column vector of covariates/predictors at u
- "point" refers to a u unless clearly stated otherwise

- bold for sets and spatial processes, normal italics for spatial vectors
- y and variations will be used for objects derived from the point pattern, e.g. marks, pseudodata

2. Material and methods

Heuristics of a good transect design:

- Should start with a sparse design with regular spacing, then refine with infill
  - Provides good spatial coverage even if aborted early
  - Imagine downloading a high-resolution intensity jpeg over 56k
- Path should avoid sharp turns but is allowed to cross itself
- One option is to generate two segments at a time, first a short-to-medium length segment to get to the start of the next transect, then a medium-to-long segment for the transect
- Could have new segment length be negatively correlated with the previous segment length

2.1. Design-Based Criteria

Path length. The total distance that traveled is often a constraint. Minimize it.

Nearest neighbor distance. A common criterion for space-filling designs, we adapt it to be meaningfully calculated for any point on a path. Define $\text{nnd}_k(u) = \min |u - v|$ where v is any point in the set of path segments at least k steps away from the segment containing u. If $k = 0$, this includes v in the same segment as u so trivially $\text{nnd}_0(u) = 0$ for all u in the path. $\text{nnd}_1(u)$ includes all segments except the one containing u. $\text{nnd}_2(u)$ excludes the segment containing u and segments with which it shares vertices. Segments not accessible by a connected path starting at u are always included. Maximize $\min[\text{nnd}_2(u)]$, $\text{avg}[\text{nnd}_2(u)]$, and $\text{avg}[\text{nnd}_1(u)]$.

2.2. Model-Based Criteria

Posterior prediction variance. Minimize maximum prediction variance and average prediction variance for GP.

Posterior parameter variance. Minimize posterior variance for each model parameter (intercept, variance, spatial scale).

Decision-based criteria. Maximize error rates and AUC of thresholding the intensity at an action level. Minimize the threshold-penalized average predictive variance (TPAPV),

$$\text{TPAPV} = \int_{\mathcal{R}} \text{Var}\left[\lambda(u)\right] p^{|\lambda(u)-A|} \mathrm{d}u,$$

where A is the action level/decision threshold and $0 < p < 1$ penalizes uncertainty about the boundary in used for thresholding.

2.3. Sampling Situations

- SRS of parallel transects
- systematic sample of parallel transects
- inhibitory plus close pairs of parallel transects
- random Latin hypercube design connected by shortest path (look at `lhs` package)
- fractal curves with random starting points
- movement model
  - generate sequentially, two waypoints at a time
  - generate a jump distance and a direction
  - distance negatively correlated with previous distance (should approximately alternate between a short "transition" and a long "transect")
  - direction bimodal with modes near $\pm\pi/2$

- if using location/scale beta for direction, allow any direction when the support

set up to think about adaptive sampling (adding a transect at a time or stopping early but don't actually do it here)

3. Results

look at examples of designs that minimize each criterion

look at examples of designs along the Pareto front

4. Discussion

discuss starting points for optimization and sequential design

5. Conclusions

References

Anders, S., 2009. Visualisation of genomic data with the hilbert curve. Bioinformatics doi:10.1093/bioinformatics/btp152.

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. Journal of the Royal Statistical Society: Series C (Applied Statistics) 60, 757–776.

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

Diggle, P., Lophaven, S., 2006. Bayesian geostatistical design. Scandinavian Journal of Statistics 33, 53–64.

Fan, J.A., Yeo, W.H., Su, Y., Hattori, Y., Lee, W., Jung, S.Y., Zhang, Y., Liu, Z., Cheng, H., Falgout, L., Bajema, M., Coleman, T., Gregoire, D., Larsen, R.J., Huang, Y., Rogers, J.A., 2014. Fractal design concepts for stretchable electronics. Nature communications 5, 3266.

Illian, J.B., Sørbye, S.H., Rue, H., 2012. A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (inla). The Annals of Applied Statistics , 1499–1530.

Johnson, D., Laake, J., VerHoef, J., 2014. DSpat: Spatial Modelling for Distance Sampling Data. URL: https://CRAN.R-project.org/package=DSpat. r package version 0.1.6.

Lark, R., 2016. Multi-objective optimization of spatial sampling. Spatial Statistics 18, 412–430.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 423–498.

Liu, J., Vanhatalo, J., 2020. Bayesian model based spatiotemporal survey designs and partially observed log Gaussian Cox process. Spatial statistics 35, 100392.

Pollard, J., Palka, D., Buckland, S., 2002. Adaptive line transect sampling. Biometrics 58, 862–870.

Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71, 319–392.

Sagan, H., 1994. Space-filling curves. Springer.

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.

Yuan, Y., Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B., Buckland, S.T., Rue, H., Gerrodette, T., et al., 2017. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. The Annals of Applied Statistics 11, 2270–2297.