

Log-Gaussian Cox processes and sampling paths: towards optimal design

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 (placeholder abstract). Evaluate a wide variety of path designs in terms design-based heuristics and model-based criteria for spatial prediction using Bayesian LGCP models. Identify promising initial designs for later optimization and sequential design (not actually optimizing yet). Illuminate any relationships among design characteristics and predictive criteria that will be helpful for constrained optimization. Discuss sequential construction of paths as a precursor to online sequential design.

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

1. Introduction

Spatial point process 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 accommodate variable sampling effort in these models using modern computing tools (Yuan et al., 2017). However,

*Corresponding author

Email address: `kenneth.flagg@montana.edu` (Kenneth Flagg)

the literature on optimal sampling for spatial point process models is in its infancy (Liu and Vanhatalo, 2020). In this article, we present a variety of sampling path designs and assess their optimality for LGCP models.

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. Aggregation has the downside of introducing arbitrary structure into the data by the choice of gridding scheme or averaging window, and requires unnecessary 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; R Core Team, 2019).

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 (USACE, 2015; Flagg et al., 2020). 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. In this paper, we adapt nearest-neighbor criteria to the path design setting and introduce a sequential path construction scheme as a starting point for future optimization. We then compare a variety of path design schemes with respect to a suite of model-based and design-based criteria for simulated point pattern data.

1.1. Spatial design

Design-based sampling. Most classical sampling work has been done for points or small quadrats approximated as points. Space-filling criteria may be good starting points (Borkowski and Piepel, 2009). Latin hypercube sampling has space-filling properties McKay et al. (1979); Husslage et al. (2011).

Space-filling curves. Used in design of dense or stretchable circuits (Ogorzałek, 2009; Ma and Zhang, 2016) and high-dimensional data visualization in bioinformatics (Anders, 2009). Peano curve is very flexible for filling irregular shapes (Fan et al., 2014).

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. The Hilbert curve is fast and simple to construct.

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).

1.2. Paths as sampling designs

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.

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

The Visual Sample Plan software includes features to create systematic transect plans and augment plans with additional transects in regions lacking spatial coverage (Matzke et al., 2014). It helps the user choose the transect spacing to maximize the probability of detecting the presence of a hotspot of specified size and intensity. However, it does not employ criteria to optimize spatial prediction.

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.

2. Materials and methods

Heuristics of a good path 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

We give attention to some design-based criteria that tie directly to practical considerations of data collection.

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

Corners. Data collection equipment (e.g. metal detectors) may have limited mobility, requiring minimizing the number or angle of turns.

Nearest neighbor distance. A common criterion for space-filling designs, we adapt it to be meaningfully calculated for any point on a path. Define the k th-order nearest neighbor distance as $\text{nnd}_k(u) = \min |u - v|$ 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 $\text{avg}[\text{nnd}_1(u)]$.

(Move details to appendix.)

95 2.2. Model-based criteria

Mean-squared prediction error. Minimize MSPE for the GP.

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

2.3. Sampling schemes

100 *These are the focus, move them earlier?*

Parallel line transects. Parallel straight-line transects are common in ordnance response studies and in ecological studies using distance sampling. Systematic designs are common because they provides good spatial coverage in the sense that any point in the study region has a known maximum distance from the path.
 105 For point designs, systematic designs are optimal for prediction, simple random samples are optimal for estimation, and inhibitory with close pair designs are becoming a popular compromise. We adapt all of these to the parallel line transect setting. We use line transects running north-south, with three ways of choosing the horizontal coordinate: simple random sample (SRS), systematic
 110 with a random starting point and even spacing, inhibitory plus close pairs. To vary the length, we use designs with 10, 25, 50, and 70 transects. Figure 1 shows an example of each scheme with 25 transects.

For the inhibitory plus close pairs, we vary the numbers of paired and unpaired transects. The total number of transects is 10, 25, 50, or 70, with 10%
 115 and 20% of the transects (rounded to the nearest integer) as redundant members of a pair. The remaining primary transects were placed according to a one-dimensional Strauss process with $\gamma = 0.05$ and a radius of 80. Then each redundant transect was randomly paired to a primary transect, an placed within an 80 unit radius of the primary transect according to a uniform distribution.

120 We should note that our we will be using an isotropic covariance function in the model. If an anisotropic model is used, parallel transects may need to be augmented with perpendicular transects so that the observed data provide the full range of pairwise distances in other directions.

Parallel serpentine transects. One simple way to observe a greater variety of
 125 locations and different directions is to add lateral zigzags to transects. We include alternate right and left turns at right angles to create serpentine transects. This could decrease prediction variance because many points in the study area will be closer to the path than they would be under a similar plan of line transects. (*Check the math—is this true? And does similar mean same length or*
 130 *same number of transects?*) They will also improve estimation of covariance parameters in the presence of anisotropy.

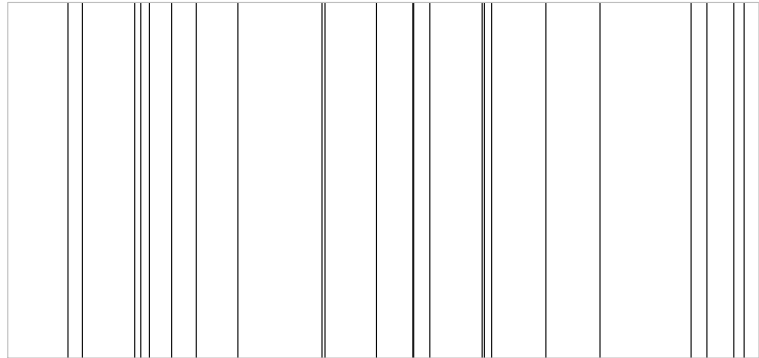
We generate designs with 7, 22, 47, and 67 serpentine transects. We vary the complexity of the serpentines by using versions with 5 and 8 zigzags. We define a zigzag as a single north-south segment or a pair of connected north-south and
 135 east-west segments. The lengths of the east-west segments are set so that the total distance equals the total distance of a line transect design with three more transects. Figure 2 shows examples.

Latin hypercube sampling. Random Latin hypercube design connected by shortest path. Waypoints generated by the `lhs` R package Carnell (2020). Connected
 140 into a the shortest path by the `TSP` package (Hahsler and Hornik, 2020).

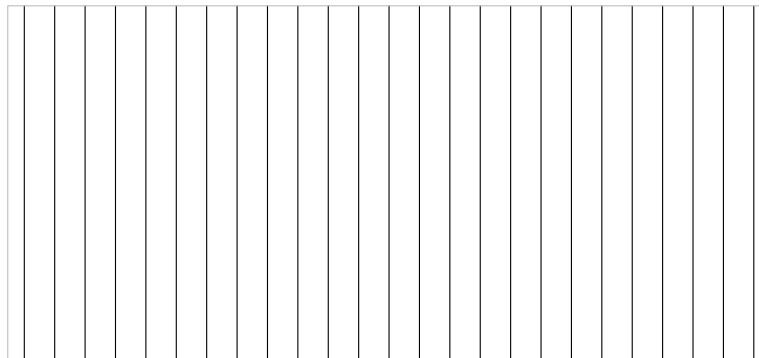
Space-filling curves. Hilbert curve generated by HilbertVis package (Anders, 2009). This is a deterministic design, so a random offset is added.

Particle movement model. Models the way data are actually collected. Waypoints generated sequentially by generating a jump distance and a direction.
 145 The jump distance is generated from a scaled beta distribution, and negatively correlated with previous jump distance. This behavior should approximately alternate between a short “transition” and a long transect. The negative correlation was achieved by applying a $1 - x$ transformation to a beta autoregressive

Simple random sample line transect design



Systematic line transect design



Inhibitory plus close pairs line transect design

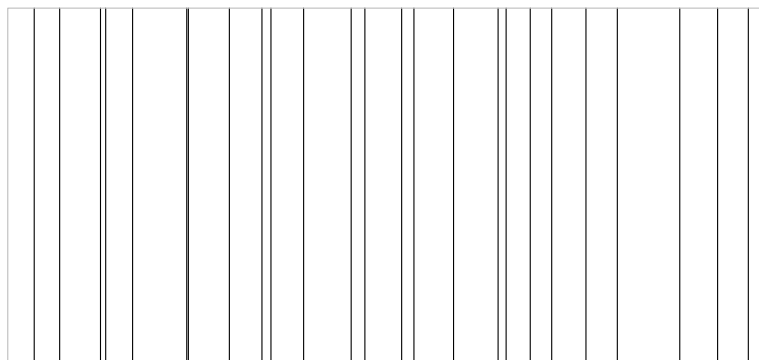
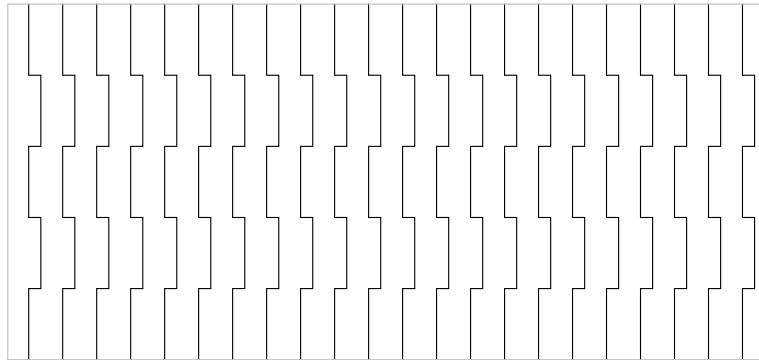


Figure 1: Examples of three different parallel line transect designs with the same number of transects.

Serpentine transect design with 5 zigzags



Serpentine transect design with 8 zigzags

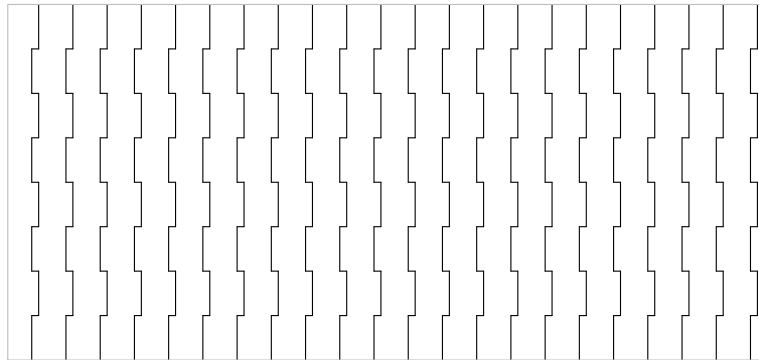


Figure 2: Examples of systematic serpentine transect designs of the same distance. The number of zigzags is the number of north-south segments per transect.

LHS-TSP design

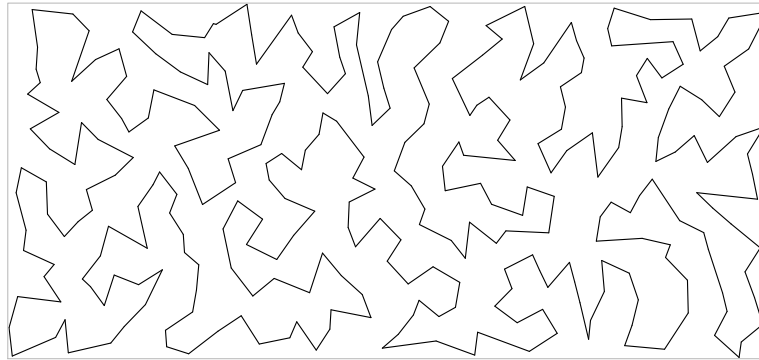


Figure 3: Example of a shortest path through a Latin hypercube sampling design.

Hilbert design of order 4

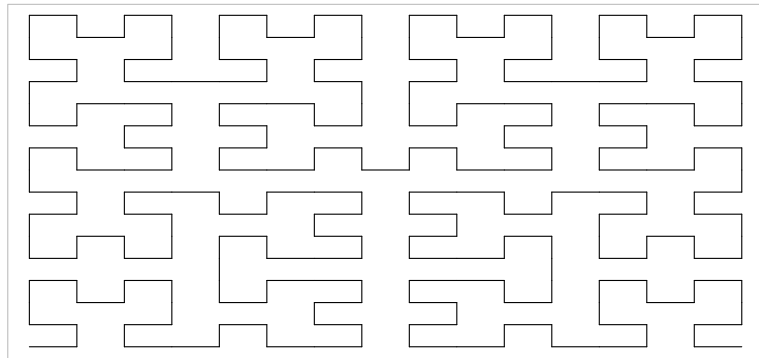


Figure 4: Example of a Hilbert curve design.

Random particle movement design

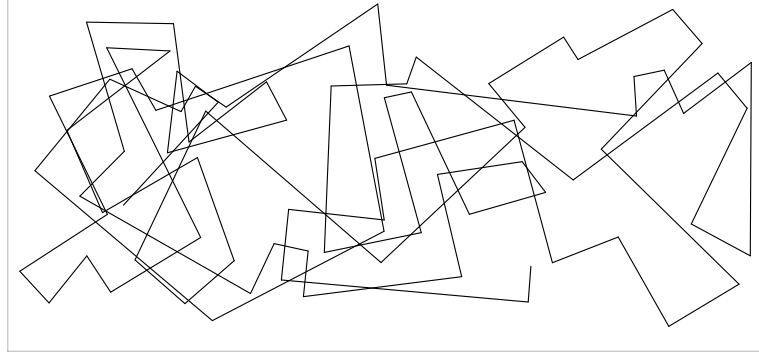


Figure 5: Example of a random particle movement design.

process (McKenzie, 1985). The direction angle is drawn from a bimodal dis-
tribution that is symmetric around 0 (a normal distribution reflected about 0).

explain the Strauss part

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

2.4. Model fitting

INLA (Rue et al., 2009), SPDE (Lindgren et al., 2011), off-grid (Simpson
et al., 2016)

2.5. Simulation procedure

We consider a fictitious site \mathcal{R} with the simple shape of a 1500 unit by 700
unit rectangle. In this site, we will simulate two data generating models meant
to produce random intensity functions with hotspots. First, a LGCP with
latent GP mean $\mu = \log(250/|\mathcal{R}|)$ and a Matérn covariance with $\nu = 1$, $\sigma = 2$,
and range = 200. This model produces relatively unstructured hotspots due to
large variability in the GP.

Second, the superposition of a two-stage cluster process superposed and a
LGCP. The cluster process (a Neyman-Scott or, more specifically, a Thomas
process) is constructed as follows. The number of clusters is Poisson with mean

3. The number of events per cluster is Poisson with mean 200. The cluster centers distributed uniformly over \mathcal{R} . Events come from a bivariate normal distribution with mean equal to the cluster center and variance $\Sigma = \tau^2 \mathbf{I}$, $\tau = 50$.
 170 The LGCP has $\mu = \log(250/|\mathcal{R}|)$ and Matérn covariance with $\nu = 1$, $\sigma = 1$, and range = 200. This model is based upon the typical conceptual model of a firing range, with a background process (represented by the LGCP) and a small number of higher-intensity foreground clusters containing the events of interest.

Path design schemes:

- 175 • Simple random sample of north-south line transects
 - Number of transects = 10, 25, 50, 70
 - Expect high variance, large prediction error in big gaps.
- Systematic sample of north-south line transects
 - Number of transects = 10, 25, 50, 70
 - 180 – Uniformly distributed starting point
 - Constant spacing
 - Expect low bias and ok variance, can miss structures at certain sizes, may not have best space-filling properties.
- Systematic sample of north-south serpentine transects
 - 185 – Number of transects = 7, 22, 47, 67
 - Uniformly distributed starting point
 - Constant spacing
 - Number of zigzags = 5, 8
 - Horizontal zigzag distance set so that the total horizontal distance traveled equals 2100 units (the length of 3 non-zigzag line transects)
 - 190 – Expect better space-filling properties than line-transect designs, lower bias/variance farther from path, would be better at estimating anisotropic covariance than line-transects.

- Inhibitory plus close pairs sample of north-south line transects
 - 195 – Total number of transects (including pairs) = 10, 25, 50, 70
 - Number of pairs = 0.1, 0.2 times the total number of transects (rounded up or down to nearest whole number)
 - Pairs uniformly distributed within radius of primaries, max pair radius = 1500/total number of transects
 - 200 – Position of primaries generated from a 1-dimensional Strauss process with $\gamma = 0.05$
 - A compromise between SRS and systematic in every way.
- Latin Hypercube Sampling waypoints
 - Number of bins = 50, 300, 1200, 2400
 - 205 – Expect low bias/variance per unit distance traveled, many sharp corners, some big open areas.
- Hilbert curve
 - Order = 3, 4, 5, 6
 - Created in square and then scaled to fit in \mathcal{R}
 - 210 – A uniform random offset added equal to spacing between segments
 - Expect good space filling, good bias and variance, lots of short segments.
- Random particle movement
 - Distance cutoff = 6700, 17200, 34700, 49700
 - 215 – Segment lengths uniform 50 to 500 units
 - Adjacent segments uncorrelated or $\rho = -0.8$
 - Turn angle $N(\mu = \pi/3, \sigma = \pi/6)$ or $N(\mu = \pi/2, \sigma = \pi/12)$
 - Angle multiplied by discrete uniform over $\{-1, 1\}$

- Strauss-esque thinning, antirepulsion = 0.8, pair radius = 80
- 220 – All combinations of the above, plus pair distance of 300 for 6700 distance cutoff
- Expect variation in all characteristics due to extreme randomness, but some near-optimality that could be harnessed by search algorithms, should see exploration followed by infill, negative ρ with turns
- 225 centered tightly on $\pi/2$ should mimic zigzagging among parallel transects.

(Probably should move explanations of schemes to appendix.)

100 designs from each scheme. All events within a 2 unit radius of the path are observed. Whole experiment repeated for 5 realizations from each data
 230 generating model.

Model:

- \mathbf{X} is a Poisson process on \mathcal{R} with intensity $\lambda(u)$
- $\log[\lambda(u)] = \mu + \mathbf{e}(u)$
- $\mu \sim \text{Unif}(-\infty, \infty)$
- 235 • \mathbf{e} is a Gaussian process with mean $\mathbf{0}$ and a Matérn covariance function with fixed $\nu = 1$
- PC prior on σ and ρ with $\Pr(\sigma > 3) = 0.1$ and $\Pr(\rho < 100) = 0.1$ (Fuglstad et al., 2019; Simpson et al., 2017)
- SPDE approach of Lindgren et al. (2011) using mesh in Figure 6
- 240 • Likelihood factorization of Simpson et al. (2016)

3. Results

look at examples of designs that minimize each criterion
 look at examples of designs along the Pareto front

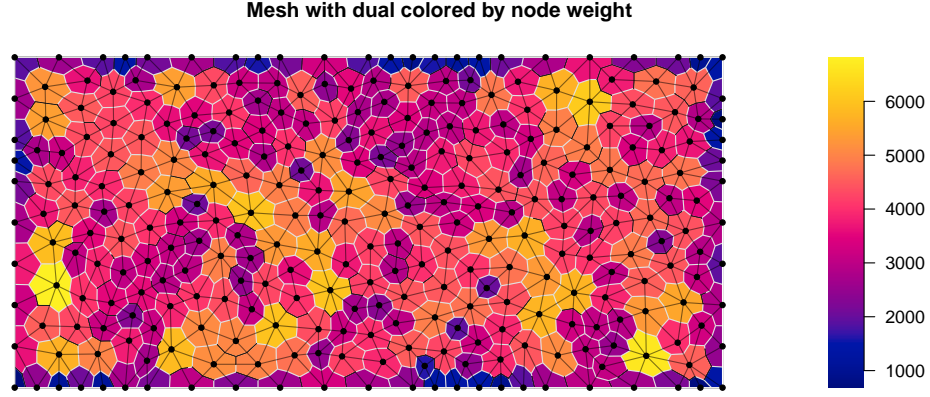


Figure 6: Illustration of the mesh and associated numerical integration scheme used to approximate the latent GP.

4. Discussion

245 discuss starting points for optimization and sequential design
practical issue: path will be smoothed, no instantaneous direction changes
at corners, equipment may have limitations which is why we looked at number
and distribution or turn angles
could incorporate turns into loss function or use multi-objective optimization
250 tion (Lark, 2016)

5. Conclusions

Appendix A. Notation and Terminology

- process defined on $\mathcal{D} \subset \mathbb{R}^d$, domain of the intensity function, in this manuscript $d = 2$
- 255 • 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}
- 260 • 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 bounded or unbounded (often equal to \mathbb{R}^d), \mathcal{S} practically always bounded, \mathcal{R} bounded or unbounded depending on application and inferential goals
- 265 • the “fully surveyed” (censused) situation is $\mathcal{S} = \mathcal{R}$
- survey path \mathcal{P} is a one-dimensional subset of \mathcal{R}
 - set of one or more sequences of waypoints connected by line segments
 - \mathcal{S} is the set of all points within a fixed (and assumed known) radius of \mathcal{P}
- 270 • \mathbf{X} point process on \mathcal{R} , $\mathbf{x} = \{x_1, \dots, x_n\}$ realized point pattern
 - $\mathbf{X}_{\mathcal{S}} = \mathbf{X} \cap \mathcal{S}$ the restriction of \mathbf{X} to \mathcal{S} , $\mathbf{x} = \mathbf{X} \cap \mathcal{S}$ the realized observeable point pattern
- point $x \in \mathbf{x}$ called an event
- intensity function $\lambda(u)$
- 275 • 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
- 280 • $\mathbf{z}(u)$ a column vector of covariates/predictors at u (not used in this manuscript)
- “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
- 285 • distance sampling fits into the framework with expansion of notation to include a (nontrivial) detection function and differentiate between the observed and observable point patterns

Appendix B. Extension of Nearest Neighbor Distance to Paths

References

- 290 Anders, S., 2009. Visualisation of genomic data with the Hilbert curve. *Bioinformatics* doi:10.1093/bioinformatics/btp152.
- Borkowski, J.J., Piepel, G.F., 2009. Uniform designs for highly constrained mixture experiments. *Journal of Quality Technology* 41, 35–47.
- Carnell, R., 2020. lhs: Latin Hypercube Samples. URL: <https://CRAN.R-project.org/package=lhs>. R package version 1.0.2.
- 295 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.
- 300 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.
- 305 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.

- Flagg, K.A., Hoegh, A., Borkowski, J.J., 2020. Modeling partially surveyed
 310 point process data: Inferring spatial point intensity of geomagnetic anomalies.
 Journal of Agricultural, Biological and Environmental Statistics 25, 186–205.
- Fuglstad, G.A., Simpson, D., Lindgren, F., Rue, H., 2019. Constructing pri-
 ors that penalize the complexity of Gaussian random fields. Journal of the
 American Statistical Association 114, 445–452.
- 315 Hahsler, M., Hornik, K., 2020. TSP: Traveling Salesperson Problem (TSP).
 URL: <https://CRAN.R-project.org/package=TSP>. R package version 1.1-
 10.
- Husslage, B.G., Rennen, G., Van Dam, E.R., Den Hertog, D., 2011. Space-
 filling latin hypercube designs for computer experiments. Optimization and
 320 Engineering 12, 611–630.
- 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 Dis-
 325 tance 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 Statis-
 tics 18, 412–430.
- Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian
 330 fields and Gaussian Markov random fields: the stochastic partial differen-
 tial 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 de-
 signs and partially observed log Gaussian Cox process. Spatial statistics 35,
 335 100392.

- Ma, Q., Zhang, Y., 2016. Mechanics of fractal-inspired horseshoe microstructures for applications in stretchable electronics. *Journal of Applied Mechanics* 83.
- Matzke, B., Wilson, J., Newburn, L., Dowson, S., Hathaway, J., Sego, L.,
 340 Bramer, L., Pulsipher, B., 2014. Visual Sample Plan Version 7.0 User's Guide. Pacific Northwest National Laboratory. Richland, Washington. URL: <http://vsp.pnnl.gov/docs/PNNL-23211.pdf>.
- McKay, M.D., Beckman, R.J., Conover, W.J., 1979. Comparison of three methods for selecting values of input variables in the analysis of output from a
 345 computer code. *Technometrics* 21, 239–245.
- McKenzie, E., 1985. An autoregressive process for beta random variables. *Management Science* 31, 988–997.
- Ogorzałek, M.J., 2009. Fundamentals of fractal sets, space-filling curves and their applications in electronics and communications, in: *Intelligent Computing Based on Chaos*. Springer, pp. 53–72.
 350
- Pollard, J., Palka, D., Buckland, S., 2002. Adaptive line transect sampling. *Biometrics* 58, 862–870.
- R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: [https://](https://www.R-project.org/)
 355 www.R-project.org/.
- 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.
- 360 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.

- Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., et al., 2017.
365 Penalising model component complexity: A principled, practical approach to
constructing priors. *Statistical science* 32, 1–28.
- USACE, 2015. Technical Guidance for Military Munitions Response Actions.
Technical Report EM 200-1-15. URL: [http://www.publications.usace.
army.mil/Portals/76/Publications/EngineerManuals/EM_200-1-15.
370 pdf](http://www.publications.usace.army.mil/Portals/76/Publications/EngineerManuals/EM_200-1-15.pdf).
- 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.