Stat 534 Project:
Extrapolation from Poisson Process Intensity Surface Models

Kenny Flagg

March 15, 2017

1 Introduction

One little-studied application of inhomogeneous point process intensity estimation is the use of an intensity surface estimated from events in a subregion to predict the intensity over the entire region of interest. This procedure would be relevant whenver it is known or suspected that some type of plant, animal, or other item occurs in the region, and the goal of the analysis is to map the trend in where these tend to be located rather than estimate parameters of some process at work across hypothetical replicates of similar regions. It may be prohibitively expensive or difficult to observe all events over the entire region, so a sample of subregions is taken. For example, conservationists want to study the spatial distribution of an endangered plant across a large region of thick jungle so that they can establish a preserve where the plant is protected. They cannot search the entire jungle, so they take a simple random sample of quadrats and record the locations of the plants in those quadrats. They will then want to fit a model describing the trends in intensity of these plants extrapolated over the entire jungle. In this setting, the objective is to map the realized spatial intensity of the plant over this jungle so that these plants can be protected, not to estimate parameters of the process that arranges plants of this species at jungles like this one.

Another situation where it is useful to extrapolate point pattern intensity outside of the observed region is in mapping subsurface geomagnetic anomalies, such as the munitions debris found at former military test ranges. This is frequently done in the early stages of an unexploded ordnance (UXO) remediation, where the intensity surface of inert munitions fragments is used to identify the locations of targets so that the search for UXO can be focused on the sections of the site most likely to contain it. The anomalies are only observed when detection equipment passes directly over them, but the project leaders are concerened with finding UXO and do not want to waste resources finding every inert fragment that could be found by metal detectors. The most common data collection method is to take a systematic sample of straight-line transects and observe anomalies in rectangular regions centered along the transects. Frequently, moving averages of the intensity are computed in circular windows and an intensity map is produced using ordinary kriging to predict the intensity in a grid of these windows. There are several problems with this approach: it assumes stationarity when the moving averages are believed to be non-stationary, it ignores the point process nature of the data, and it is very sensitive to the window size (Flagg 2016; Matzke et al. 2014). In this project, I explore the use of polynomial trend surface models for the log-intensity at a simulated UXO site.

2 Surface Fitting

It is theoretically possible to use maximum likelihood or maximum pseudo-likelihood methods to fit trend surface models or regression models (with covariates) for the intensity of an inhomogeneous point process, but these are not widely used because, in the most general cases, the problems are "notoriously intractable" (Diggle 2013). However, for spatial Poisson processes, trend surface models are tenable with the help of some numerical methods.

The log-likelihood of an unmarked Poisson process on a region D with intensity $\lambda(\mathbf{s})$ is found by conditioning on the number of events in the following manner. The number of events in D follows a Poisson distribution with mean $\mu = \int_D \lambda(\mathbf{s})d\mathbf{s}$. Given that n events occured, their locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are independent and identically distributed with density $\lambda(\mathbf{s})/\mu$. Then the log-likelihood of the intensity is

$$\begin{aligned}\ell(\lambda) &= (-\mu + n\log(\mu) - \log(n!)) + \left(\sum_{i=1}^{n}\left[\log\left(\lambda(\mathbf{s}_i)\right) - \log(\mu)\right]\right)\\ &= \sum_{i=1}^{n}\log\left(\lambda(\mathbf{s}_i)\right) - \int_D \lambda(\mathbf{s})d\mathbf{s} - \log(n!).\end{aligned}$$

A natural way to model the intensity is to define a log-linear model

$$\log(\lambda(\mathbf{s})) = \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta}$$

where $\mathbf{x}(\mathbf{s})$ can, in principle, include functions of the spatial coordinates and also covariates. Unfortunately, the likelihood becomes

$$\sum_{i=1}^{n}\mathbf{x}(\mathbf{s})^T\boldsymbol{\beta} - \int_D \exp\left(\mathbf{x}(\mathbf{s})^T\boldsymbol{\beta}\right)d\mathbf{s} - \log(n!)$$

so $\mathbf{x}(\mathbf{s})$ must be known (or predicted) either across the entire region D or at enough locations to approximate the integral numerically. Thus, covariates require some extra data collection or modeling effort to incorporate, but trend surfaces are feasible.

Berman and Turner (1992) proposed a practical method for fitting these models with generalized linear models software using quadratic approximations to the likelihood, even generalizing to other link functions besides the log link. Their method is based on partitioning D such that each subset in the partition contains at most one event, and then using a binomial or Poission distribution for the count of events in each subset. This was followed up by Baddeley et al. (2014), who developed a logistic regression-based approach that avoids the quadratic approximation, thereby reducing bias in the coefficient estimates. Both methods are available in the `spatstat` package (Baddeley, Rubak, and Turner 2015) for R (R Core Team 2017). The `spatstat` function `ppm` implements these methods using the base R `glm` function to maximize the likelihood; I will use the Berman and Turner method because it is the default in `ppm` and therefore better reflects the experience of a naïve practitioner who is not an expert on spatial point processes.

3 Simple Examples

log-linear example from HW4

simulate from

$$\log(\lambda(x, y)) = 5x + 2y; \quad (x, y) \in (0, 1)^2$$

but do not observe any events in $(0.5, 0.8)^2$. Then fit

$$\log(\lambda(x, y)) = \beta_0 + \beta_1 x + \beta_2 y$$

and predict on $(-0.5, 1.5)^2$.

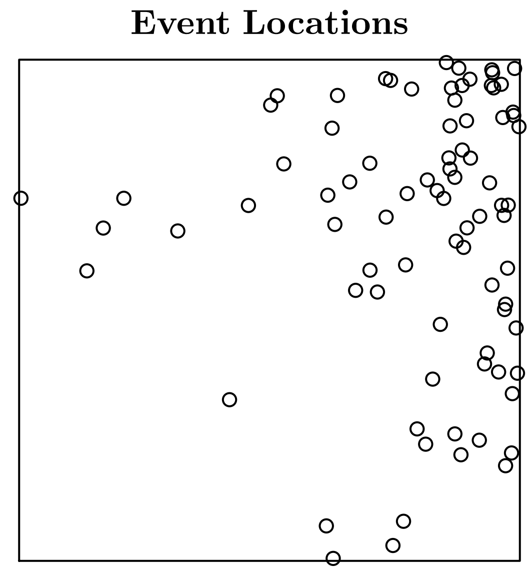

Figure 1: One realization of a Poisson process with log-linear intensity on the unit square.

	Estimate	S.E.
$\widehat{\beta}_0$	0.16	0.54
$\widehat{\beta}_1$	4.61	0.58
$\widehat{\beta}_2$	2.02	0.42

Table 1: Estimated coefficients for the log-linear trend model fit using the full region.

	Estimate	S.E.
$\widehat{\beta}_0$	0.05	0.58
$\widehat{\beta}_1$	4.72	0.61
$\widehat{\beta}_2$	1.99	0.43

Table 2: Estimated coefficients for the log-linear trend model fit using a subset of the region.

easy site, realization 2,000, 390 ft spacing

use `polynom()` not `poly()`!

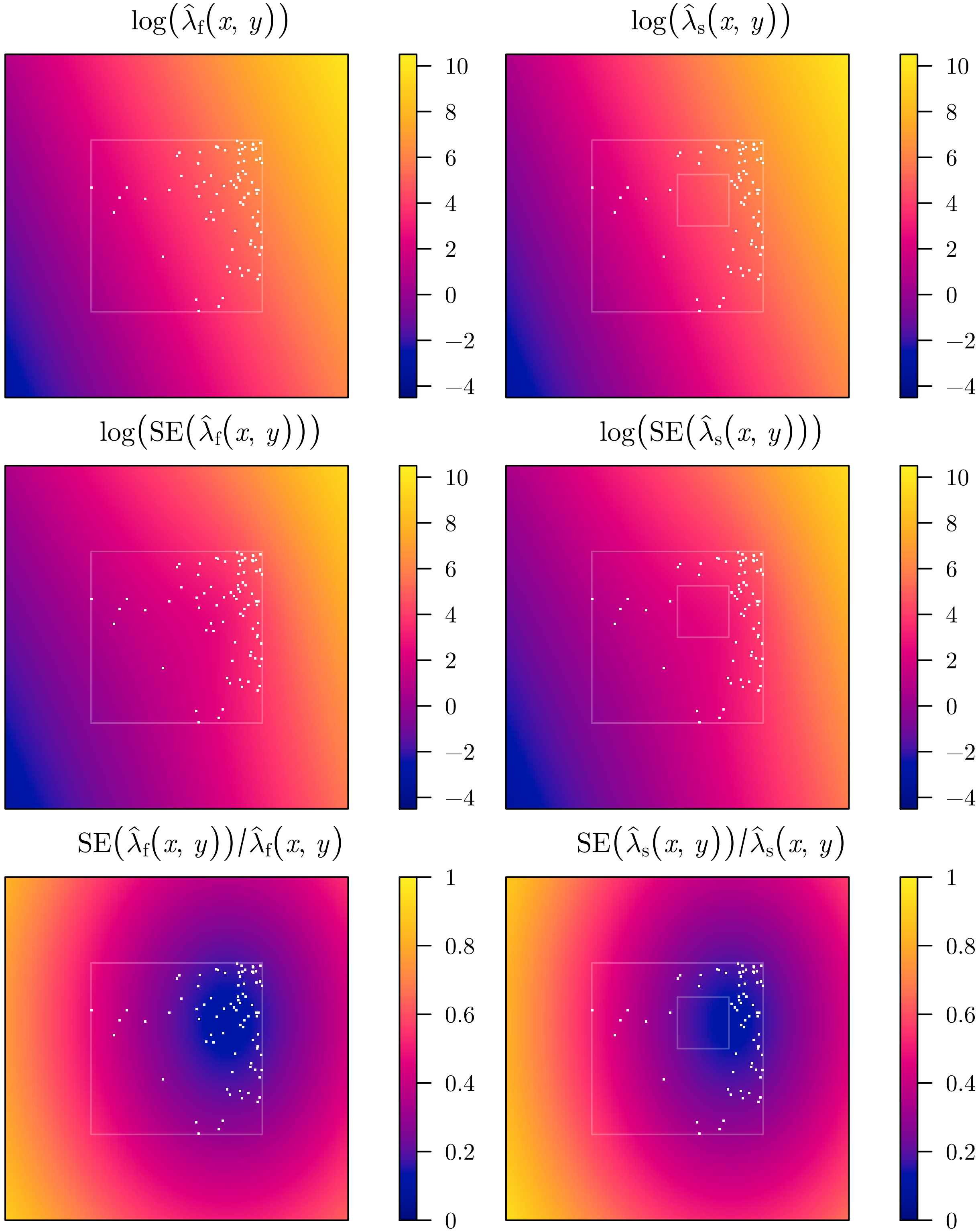

Figure 2: Estimated log-intensity surface, log-scale prediction standard errors, and prediction coeffiecient of variation from models fit using all events in the full simulation region (left) and events in a subset of the simulation region (right). The regions and event locations are overlaid in white.

4 Simulation Study

What to compare? MSPE for this method vs the moving average kriging? I'd like to cross-validate (could leave one transect out) but I don't know how to contruct a criterion. Maybe compare modeled intensity within the rectangle to a kernel intensity estimate?

5 Discussion and Conclusion

smallish SEs outside observed region because of faith in the model form – need model checking

how much of the region do you need to observe to trust your model?

A R Code Appendix

References

Baddeley, Adrian, Ege Rubak, and Rolf Turner (2015). *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman and Hall/CRC Press.

Baddeley, Adrian et al. (2014). "Logistic regression for spatial Gibbs point processes". In: *Biometrika* 101.2, pp. 377–392.

Berman, Mark and Rolf Turner (1992). "Approximating point process likelihoods with GLIM". In: *Applied Statistics*, pp. 31–38.

Diggle, Peter J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. 3rd ed. CRC Press.

Flagg, KA (2016). "Visual Sample Plan and Unexploded Ordnance: What do we need to know to find UXO?" M.S. writing project. Montana State University, Bozeman. URL: `https://github.com/kflagg/vspuxo`.

Matzke, Brett et al. (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`.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.