Log-Gaussian Cox Process for London crime data

Jan Povala

May 3, 2018

Outline

Motivation

Methodology

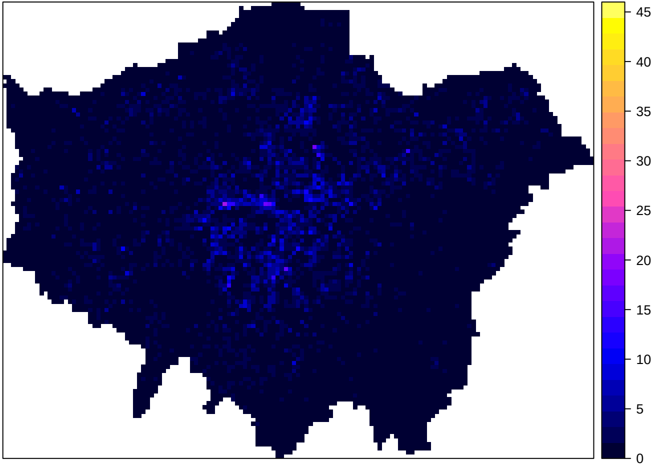
Results

Current work, Next steps

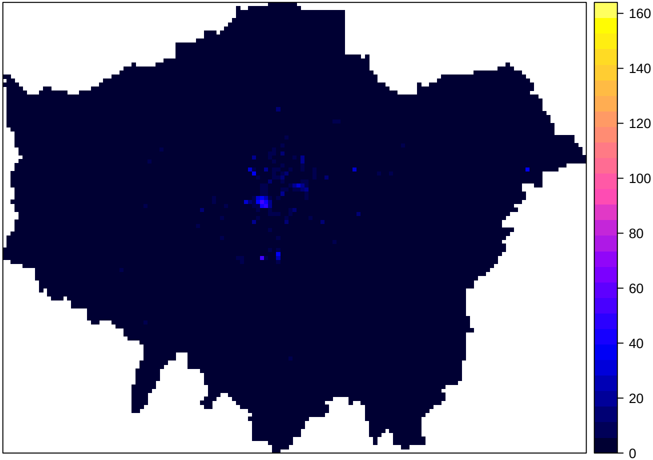
Aims and Objectives

- Modelling of crime and short-term forecasting.
- Two stages involved:
    1. *inference* - what is the underlying process that generated the observations?
    2. *prediction* - use the inferred process's properties to forecast future values.

Burglary



Theft from the person



Outline

Motivation

Methodology

Results

Current work, Next steps

Cox Process

Cox process is a natural choice for an environmentally driven point process (Diggle et al., 2013).

Definition

Cox process Y$(\boldsymbol{x})$ is defined by two postulates:

1. $\Lambda(\boldsymbol{x})$ is a nonnegative-valued stochastic process;
2. conditional on the realisation $\lambda(\boldsymbol{x})$ of the process $\Lambda(\boldsymbol{x})$, the point process Y$(\boldsymbol{x})$ is an inhomogeneous Poisson process with intensity $\lambda(\boldsymbol{x})$.

Log-Gaussian Cox Process

- Cox process with intensity driven by a Gaussian Process $f(\boldsymbol{x})$:

$$\Lambda(\boldsymbol{x}) = \exp\left(f(\boldsymbol{x})\right).$$

- The latent surface f is modelled by placing a GP prior:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(0, k_\theta(\cdot, \cdot)).$$

- Discretised version of the model over a regular grid on the observation window is:

$$\mathsf{y}_i | f(\boldsymbol{x}_i) \sim \text{Poisson}(\exp\left[f(\boldsymbol{x}_i)\right]).$$

Field inference

Given the observations $\mathbf{y}$ on the grid X, our goal is to find the distribution of the latent field $\mathbf{f}$:

$$p(\mathbf{f}|\mathbf{y}, X, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{f}, X, \boldsymbol{\theta})p(\mathbf{f}|X, \boldsymbol{\theta})}{p(\mathbf{y}|X, \boldsymbol{\theta})},$$

where

$$p(\mathbf{y}|X, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f}, X, \boldsymbol{\theta})p(\mathbf{f}|X, \boldsymbol{\theta})d\mathbf{f}$$

which is intractable.

Laplace Approximation

Flaxman et al. (2015)

- One approach to overcome intractability is *Laplace approximation*.
- Approximate the posterior distribution of the latent surface by:

$$p(\mathbf{f}|\mathbf{y}, X, \boldsymbol{\theta}) \approx \mathcal{N}\left(\hat{\mathbf{f}}, -\left(\nabla\nabla\Psi(\mathbf{f})|_{\hat{\mathbf{f}}}\right)^{-1}\right),$$

  where $\Psi(\mathbf{f}) \coloneqq \log p(\mathbf{f}|\mathbf{y}, X, \boldsymbol{\theta}) \overset{\text{const}}{=} \log p(\mathbf{y}|\mathbf{f}, X, \boldsymbol{\theta}) + \log p(\mathbf{f}|X, \boldsymbol{\theta})$ is unnormalised log posterior, and $\hat{\mathbf{f}}$ is the mode of the distribution.
- Newton's method to find $\hat{\mathbf{f}}$.

Hyperparameters - Marginal Likelihood

Flaxman et al. (2015)

- Accurate inferences/predictions require knowing $\boldsymbol{\theta}$.
- Marginal log-likelihood:

$$\begin{aligned}\log p(\mathbf{y}|X,\boldsymbol{\theta}) &= \log \int \exp\left[\Psi(\mathbf{f})\right] d\mathbf{f} \\ &\approx \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\hat{\mathbf{f}}^{\top}\boldsymbol{K}^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log\left|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}\right|,\end{aligned}$$

where $K_{ij} = k_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ describes covariance between pairwise locations, and $\boldsymbol{W} \coloneqq -\nabla\nabla \log p(\mathbf{y}|\hat{\mathbf{f}}, X, \boldsymbol{\theta})$.

Computation I

Flaxman et al. (2015)

- The calculations above require $\mathcal{O}\big(n^3\big)$ operations and $\mathcal{O}\big(n^2\big)$ space.
- Cheaper linear algebra available if separable kernel functions are assumed, *e.g.* in $D = 2$ dimensions:

$$k((x_1, x_2), (x_1', x_2')) = k_1(x_1, x_1')k_2(x_2, x_2')$$

  implies that $\boldsymbol{K} = \boldsymbol{K}_1 \otimes \boldsymbol{K}_2$.
- Determinant approximation due to Fiedler (1971):

$$\begin{aligned} \log|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}| &= \log\big(|\boldsymbol{K} + \boldsymbol{W}^{-1}||\boldsymbol{W}|\big) \\ &\leq \log\left\{\prod_i \big(e_i + W_{ii}^{-1}\big)\prod_i W_{ii}\right\} \\ &= \sum_i \log\left(1 + e_i W_{ii}\right), \end{aligned}$$

  where $e_1, \ldots, e_n$ are sorted eigenvalues of $\boldsymbol{K}$.

Computation II

Flaxman et al. (2015)

Applying the above properties, the inference and predictions can be computed using $\mathcal{O}\left(Dn^{\frac{D+1}{D}}\right)$ operations and $\mathcal{O}\left(Dn^{\frac{2}{D}}\right)$ space thanks to:

- Conjugate gradient for solving $\boldsymbol{K}^{-1}\boldsymbol{b} = \boldsymbol{x}$, where matrix-vector multiplication is efficient due to Kronecker structure.
- Eigendecomposition utilising Kronecker structure.

Outline

Motivation

Methodology

Results

Current work, Next steps

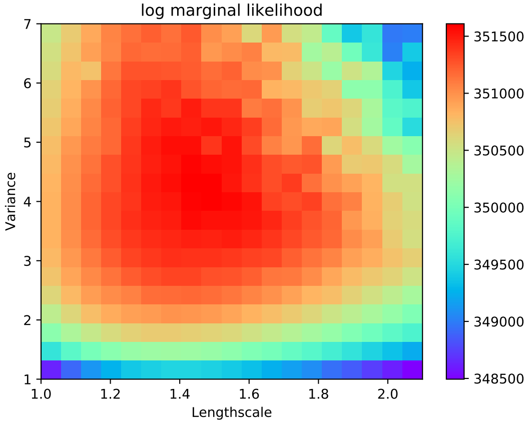
Experiment

Spatial model with isotropic Matérn covariance function:

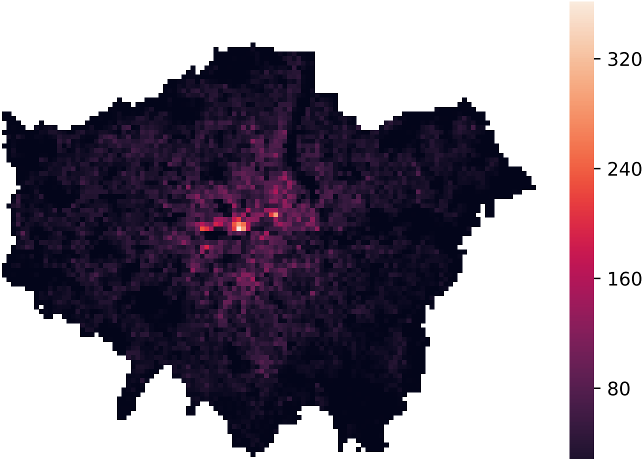
- Dataset used: 2016 data
- Crime types: Burglary, Theft from the person
- Grid: 117x91, one cell is an area of 500m by 500m.
- Missing locations were treated as imaginary with a special noise model.
- Two hyperparameters inferred: lengthscale(ℓ), marginal variance (σ^2)

Burglary - inferred hyperparameters

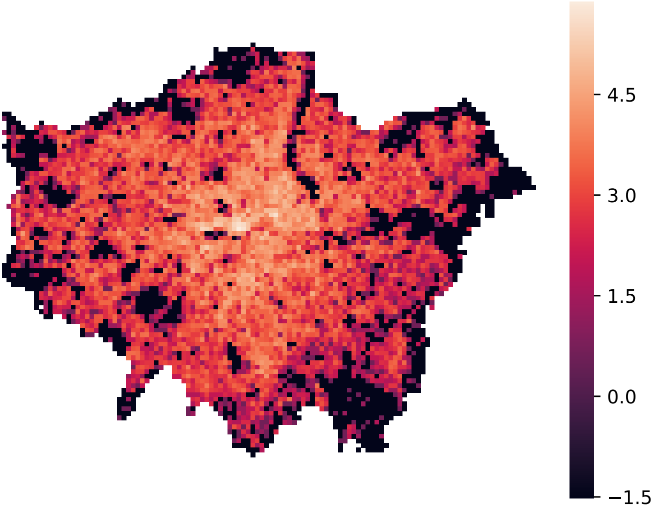
Inferred hyperparameters: $\ell = 1.41$, and $\sigma^2 = 4.16$



Burglary - counts

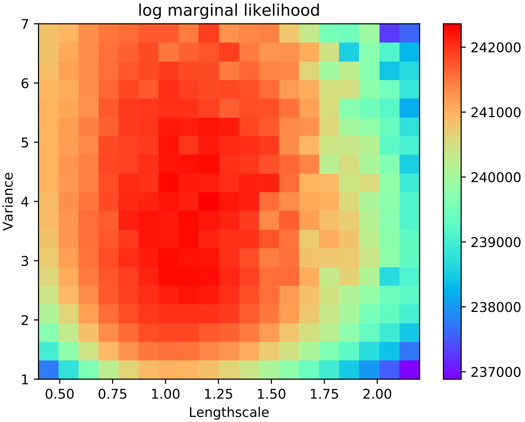


Burglary - latent field

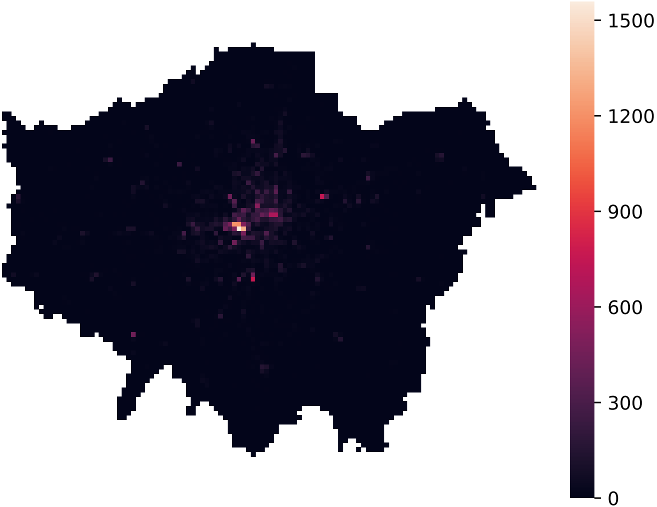


Theft from the person - inferred hyperparameters

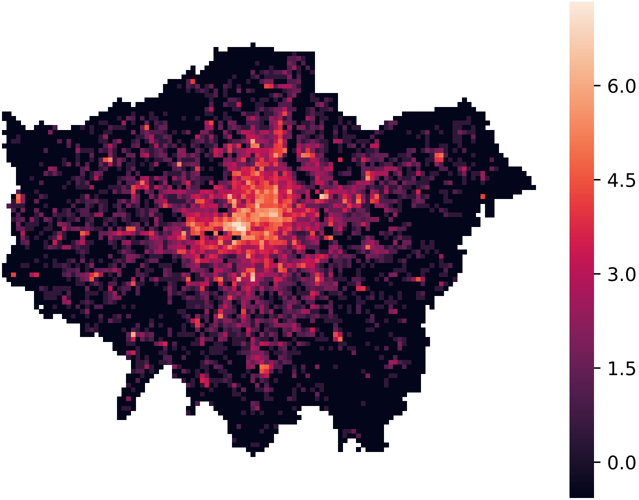
Inferred hyperparameters: $\ell = 1.16$, and $\sigma^2 = 3.84$



Theft from the person - counts



Theft from the person - latent field



Comments

- The inference confirmed that number of occurrences in a cell influences neighbouring locations.
- The process driving Burglary is ‘smoother’ than the process driving Theft from the person.

Outline

Motivation

Methodology

Results

Current work, Next steps

Forecasting

The domain will now be $X_1 \times X_2 \times T$ and the kernel will be of the form

$$k((x_1, x_2, t), (x_1', y_2', t')) = k_1(x_1, x_1') k_2(x_2, x_2') k_t(|t - t'|)$$

with $k_1(\cdot, \cdot)$, $k_2(\cdot, \cdot)$ as before and $k_t(\cdot)$ as one of the below:

- A kernel with period of 12 months for seasonal variation (Flaxman, 2014):
$$k_t(\tau) = \exp\left(-\frac{2\sin^2\left(\frac{\tau\pi}{12}\right)}{\ell^2}\right)$$
- Spectral mixture kernel with Q components (Flaxman et al., 2015):
$$k_t(\tau) = \sum_{q=1}^{Q} w_q \exp\left(-2\pi^2\tau^2 v_q\right) \cos\left(2\pi\tau\mu_q\right)$$

Stochastic PDEs

Another computationally tractable, and more mechanistic, approach is describing the crime activity using stochastic PDEs:

- Finite Element Method to solve SPDEs as described in Lindgren, Rue, and Lindström (2011).
- Sigrist, Künsch, and Stahel (2015) solve transport-diffusion SPDE using spectral methods on a grid.

Bibliography I

Diggle, Peter J. et al. (2013). "Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm". en. In: *Statistical Science* 28.4, pp. 542–563. ISSN: 0883-4237. DOI: 10.1214/13-STS441. URL: http://projecteuclid.org/euclid.ss/1386078878.

Fiedler, Miroslav (1971). "Bounds for the Determinant of the Sum of Hermitian Matrices". In: *Proceedings of the American Mathematical Society* 30.1, p. 27. ISSN: 00029939. DOI: 10.2307/2038212. URL: http://www.jstor.org/stable/2038212?origin=crossref.

Flaxman, Seth et al. (2015). "Fast Kronecker Inference in Gaussian Processes with non-Gaussian Likelihoods". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning*. Vol. 37. ICML'15. Lille, France: JMLR.org, pp. 607–616.

Bibliography II

Flaxman, Seth R. (2014). *A General Approach to Prediction and Forecasting Crime Rates with Gaussian Processes*. Tech. rep. Heinz College Technical Report, 2014. URL https://www. ml. cmu. edu/research/dap-papers/dap_flaxman. pdf.

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4, pp. 423–498. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2011.00777.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2011.00777.x/abstract.

Saatçi, Yunus (2012). "Scalable inference for structured Gaussian process models". PhD Thesis. Citeseer.

Bibliography III

- Sigrist, Fabio, Hans R. Künsch, and Werner A. Stahel (2015). "Stochastic partial differential equation based modelling of large space-time data sets". en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.1, pp. 3–33. ISSN: 13697412. DOI: 10.1111/rssb.12061. URL: http://doi.wiley.com/10.1111/rssb.12061.
- Wilson, Andrew Gordon et al. (2014). "Fast Kernel Learning for Multidimensional Pattern Extrapolation". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Cambridge, MA, USA: MIT Press, pp. 3626–3634. URL: http://dl.acm.org/citation.cfm?id=2969033.2969231.

Matérn Covariance Function

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{\ell} \right)$$

We fix $\nu = 2.5$ as it is difficult to jointly estimate ℓ and ν due to identifiability issues.

Kronecker Algebra

Saatçi (2012)

- Matrix-vector multiplication $(\otimes_d \mathbf{A}_d)\,\boldsymbol{b}$ in $\mathcal{O}(n)$ time and space.
- Matrix inverse: $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$
- Let $\boldsymbol{K}_d = \boldsymbol{Q}_d \boldsymbol{\Lambda}_d \boldsymbol{Q}_d^\top$ be the eigendecomposition of $\boldsymbol{K}_d$. Then, the eigendecomposition of $\boldsymbol{K} = \otimes_d \boldsymbol{K}_d$ is given by $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{Q} = \otimes_d \boldsymbol{Q}_d$, and $\boldsymbol{\Lambda} = \otimes_d \boldsymbol{\Lambda}_d$. The number of steps required is $\mathcal{O}\left(Dn^{\frac{3}{D}}\right)$.

Field inference - Newton Optimisation

Flaxman et al. (2015)

- The Newton optimisation step:

$$\mathbf{f}^{\mathsf{new}} \leftarrow \mathbf{f}^{\mathsf{old}} - \left(\nabla\nabla\Psi\right)^{-1} \nabla\Psi.$$

- $\nabla\nabla\Psi$ and $\nabla\Psi$ require inverting the covariance matrix of the GP:

$$\begin{aligned} \nabla\Psi(\mathbf{f}) =& \nabla \log p(\mathbf{y}|\mathbf{f}, X, \boldsymbol{\theta}) - K^{-1}\mathbf{f} \\ \nabla\nabla\Psi(\mathbf{f}) =& -\boldsymbol{W} - \boldsymbol{K}^{-1}, \end{aligned}$$

where $\boldsymbol{W} := -\nabla\nabla \log p(\mathbf{y}|\mathbf{f}, X, \boldsymbol{\theta})$.

Incomplete grids

Wilson et al. (2014)

We have that $\mathsf{y}_i \sim \mathrm{Poisson}(\exp(\mathsf{f}_i))$. For the points of the grid that are not in the domain, we let $\mathsf{y}_i \sim \mathcal{N}(\mathsf{f}_i, \epsilon^{-1})$ and $\epsilon \to 0$. Hence,

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i \in \mathcal{D}} \frac{\left(e^{\mathsf{f}_i}\right)^{\mathsf{y}_i} e^{-e^{\mathsf{f}_i}}}{\mathsf{y}_i!} \prod_{i \notin \mathcal{D}} \frac{1}{\sqrt{2\pi\epsilon^{-1}}} e^{\frac{-\epsilon(\mathbf{y}_i - \mathbf{f}_i)^2}{2}}$$

The log-likelihood is thus:

$$\sum_{i \in \mathcal{D}} [\mathsf{y}_i \mathsf{f}_i - \exp(f_i) + \mathsf{const}] - \frac{1}{2} \sum_{i \notin \mathcal{D}} \epsilon(\mathsf{y}_i - \mathsf{f}_i)^2$$

We now take the gradient of the log-likelihood as

$$\nabla \log p(\mathbf{y}|\mathbf{f})_i = \begin{cases} \mathsf{y}_i - \exp(\mathsf{f}_i), & \text{if } i \in \mathcal{D} \\ \epsilon(\mathsf{y}_i - \mathsf{f}_i), & \text{if } i \notin \mathcal{D} \end{cases}$$

and the hessian of the log-likelihood as

$$\nabla\nabla \log p(\mathbf{y}|\mathbf{f})_{ii} = \begin{cases} -\exp(\mathsf{f}_i), & \text{if } i \in \mathcal{D} \\ -\epsilon & \text{if } i \notin \mathcal{D} \end{cases}.$$

Fiedler's bound

For Hermitian positive semidefinite matrices $\boldsymbol{U}$ and $\boldsymbol{V}$:

$$\prod_i (u_i + v_i) \leq |\boldsymbol{U} + \boldsymbol{V}| \leq \prod_i (u_i + v_{n-i+1}) \, ,$$

where u_i and v_j are sorted eigenvalues of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively.