Spatio-temporal log-Gaussian Cox processes for public health data

Theresa Smith

with Peter Diggle, Ben Taylor, and Sarah O'Brien

WHAT IS GEO-LOCATED HEALTH DATA?

There are two basic kinds of spatial data:

1. Areal data: $\mathbf{x} = \{x(A_i) \mid A_i \subset W\}$ where the A_i's are a partition of the study region W.

2. Point-level data: $\mathbf{x} = \{x(s_i) \mid s_i \in W\}$ where s_i's are points in W.
   - $x(s_i)$: realization of a process at particular locations.
   - s_i: locations are the 'response'.

Point Pattern Basics

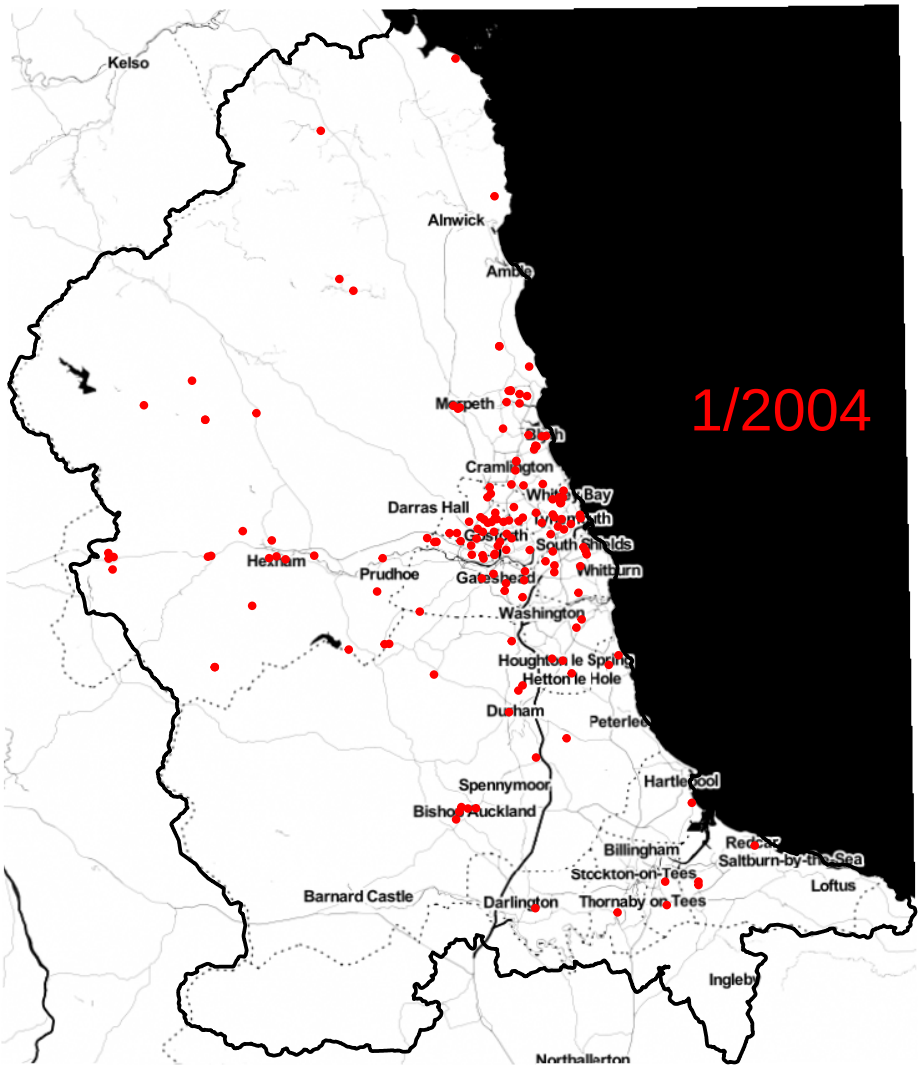
Spatial: Events can occur at any point on a window $W \subset \mathbb{R}^2$:

$$X = \{s_1, \ldots, s_n; s_i \in W\}.$$

Spatio-temporal: Events can occur at any point in $W \times (0, T)$:

$$X = \{(s_1, t_1), \ldots, (s_n, t_n); s_i \in W, t_i \in (0, T)\}.$$

Examples: tree locations, home locations of individuals with a disease, defects in a material.


1/2004
Kelso
Alnwick
Amble
Morpeth
Cramlington
Darras Hall
Hexham
Prudhoe
Gateshead
South Shields
Whitburn
Washington
Houghton le Spring
Hetton le Hole
Durham
Peterlee
Spennymoor
Hartlepool
Billingham
Stockton-on-Tees
Loftus
Barnard Castle
Darlington
Thornaby on Tees
Ingleby

Campylobacter

- Most common cause of bacterial gastroenteritis in high income countries.
- Typically self-limiting with very rare autoimmune complications.
- 1M annual cases in the US and 500K in England and Wales.
- Costs of \$1.2–4B in the US and €2.4B in the EU.
- Ubiquitous in broiler flocks and common in ruminants.

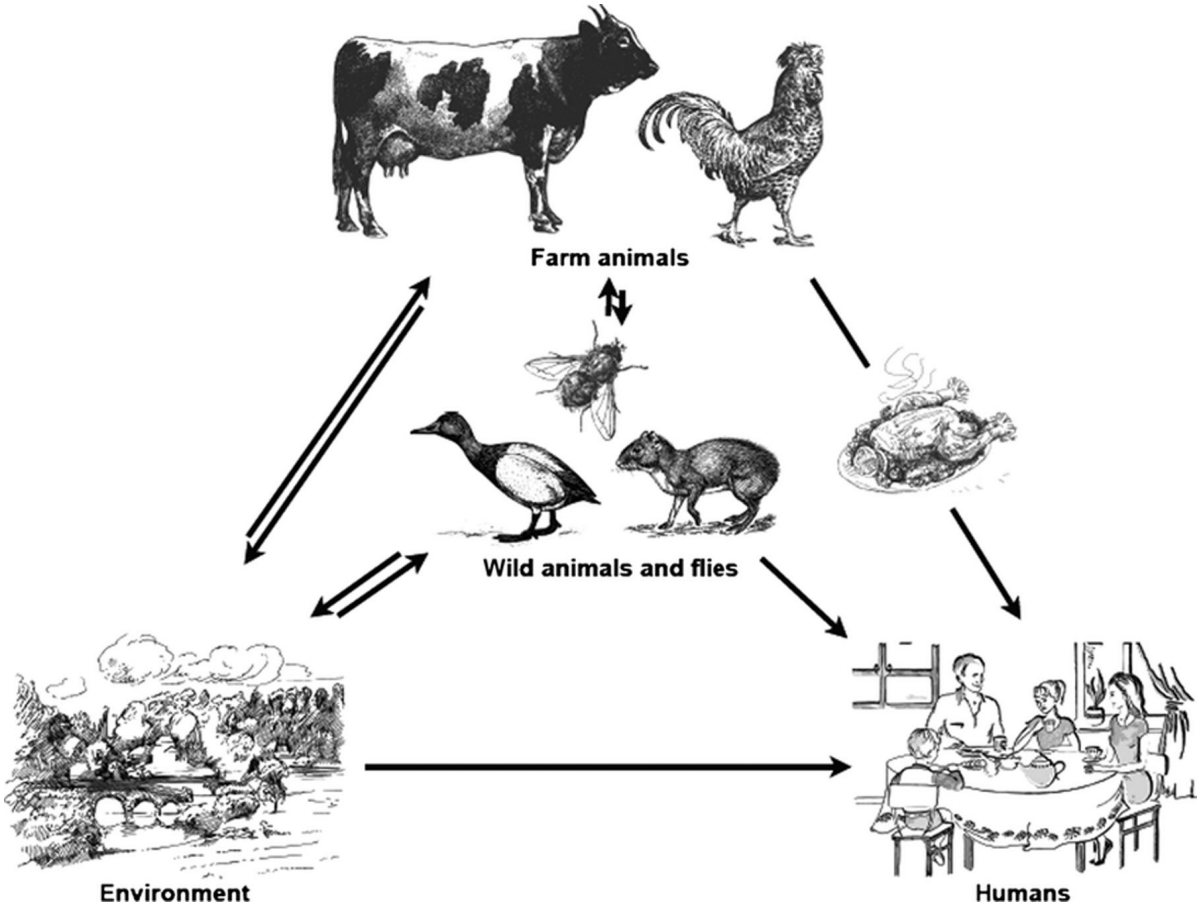
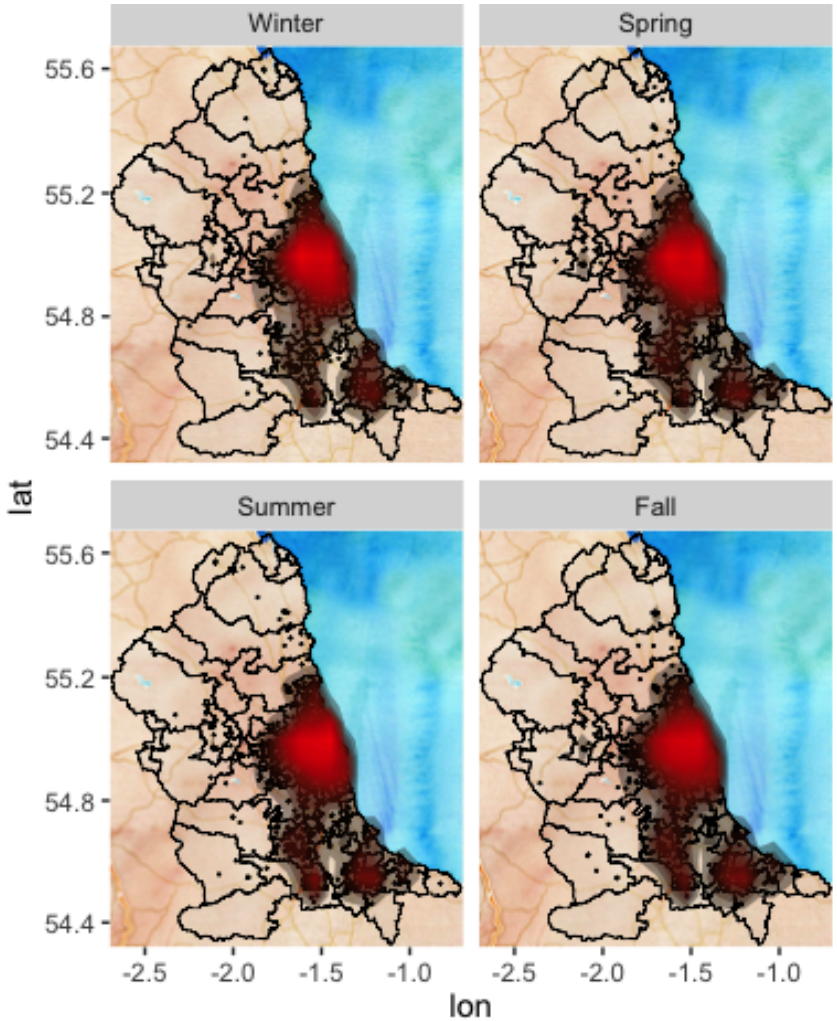


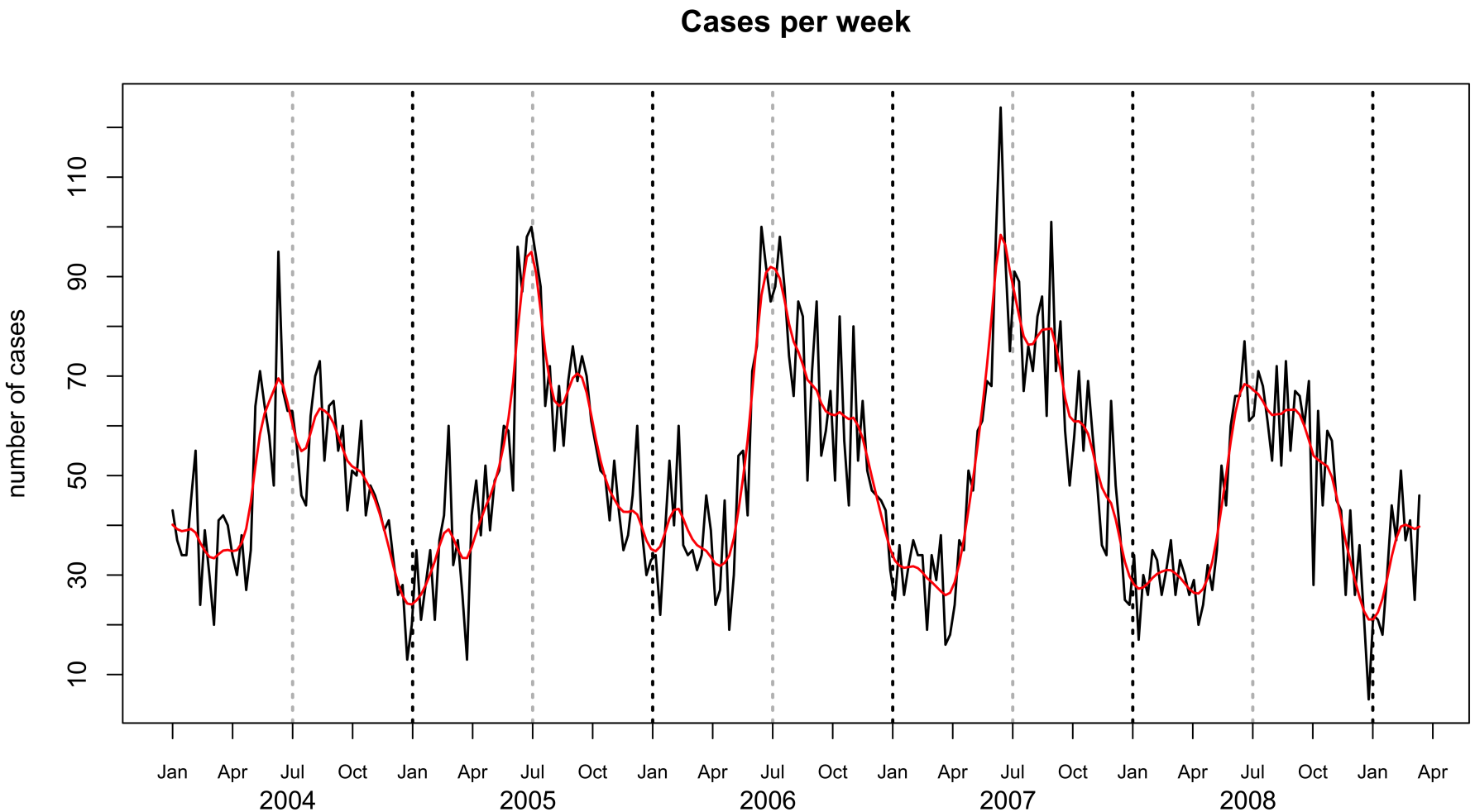
Figure: From Bronowski et al., "Role of environmental survival in transmission of Campylobacter jejuni," (2014).

Data from North East England

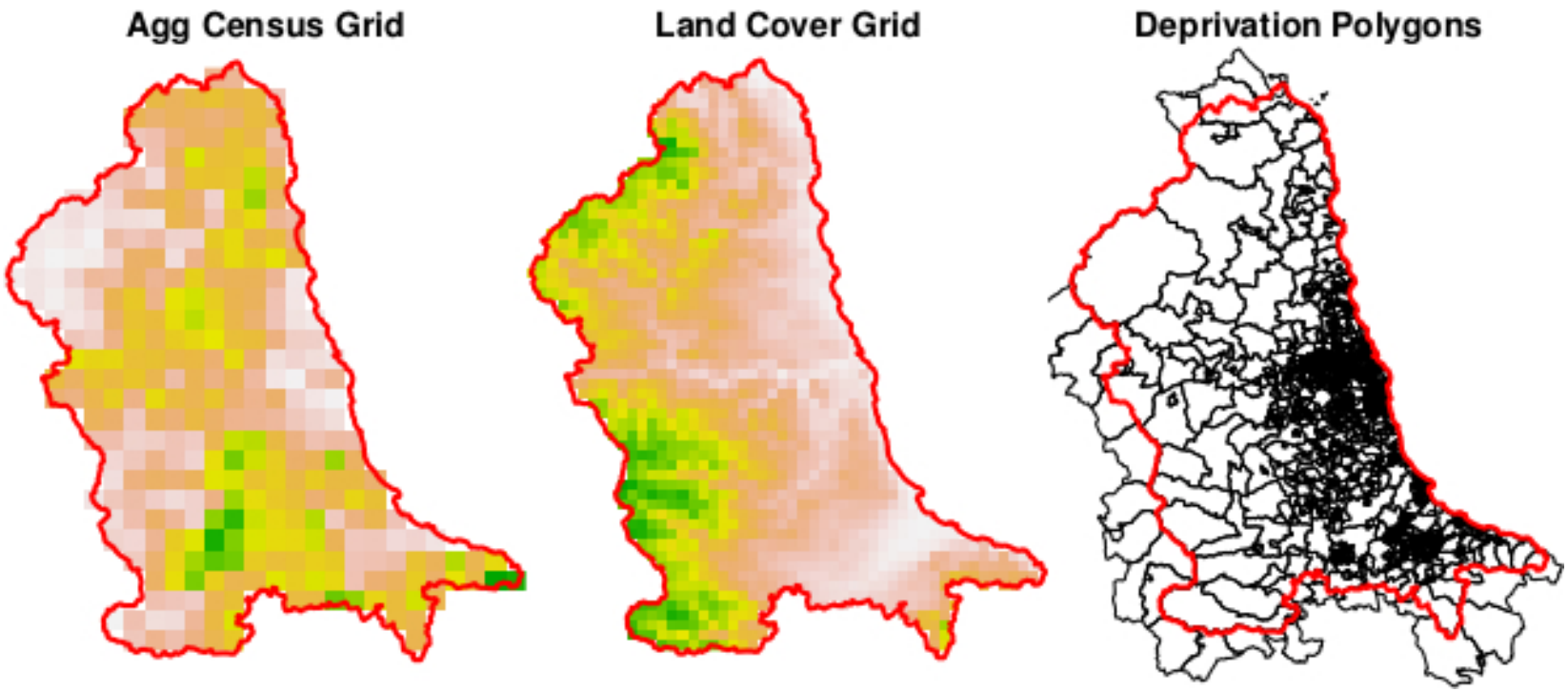
- Case data (lab confirmed cases).
  - 13,600 cases in NE England from 2004-2009.
  - Full postcode and date of sample.
  - Age and sex.
- Contextual information.
  - Demographics: population, socio-economic deprivation.
  - Land use: livestock survey, satellite imagery.
  - Weather: rainfall and temperature.



Seasonality of Campylobacter



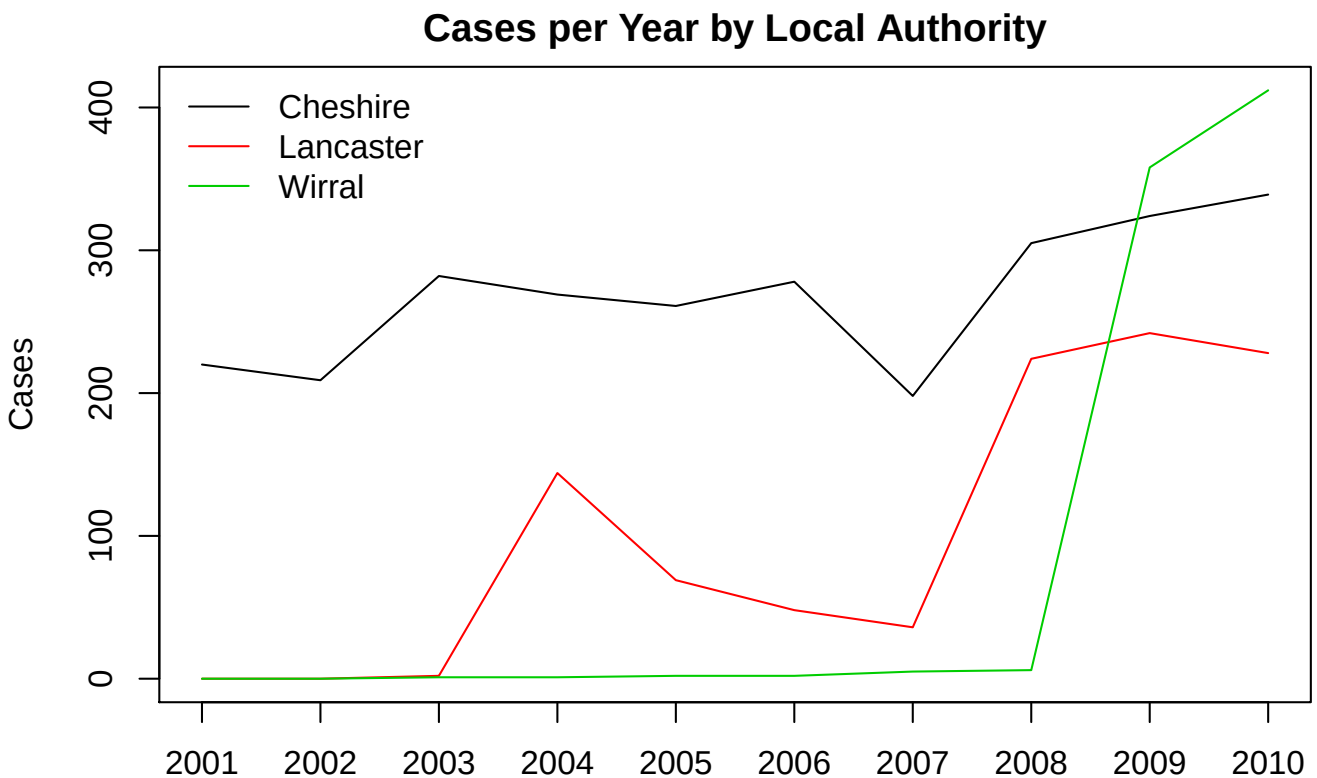
Spatial Predictors



- Census of cattle, sheep, crop land (5km by 5km raster).
- Percentage of built up area (1km by 1km raster).
- Deprivation in education and health (LSOA polygons).

UNDERREPORTING

$$\text{risk(reported campy case)} = \text{risk(campy)} \times \text{prob(reported} \mid \text{campy)}$$



Modelling Challenges

- Case data and predictor data are not on common spatial units.
- Cases may be non-uniformly underreported.
- Fitting a full spatio-temporal model is computationally challenging.
- Risk factors are correlated both with each other and in space.

Poisson Point Processes

Let $N(A)$ be the number of events in $A \subset W$ or $\subset W \times (0, T)$

$\mathcal{X} \sim$ homogeneous PPP if for constant intensity λ

- $N(A) \sim \text{Poi}(\lambda \text{area}(A))$.
- If $N(A) = n$ then the n points are uniformly distributed on A.

$\mathcal{X} \sim$ inhomogeneous PPP when $\lambda = \lambda(s)$ varies and

- $N(A) \sim \text{Poi}(\int_A \lambda(s)ds)$.
- If $N(A) = n$ then the n points form an independent random sample with pdf proportional to $\lambda(s)$.

Log Gaussian Cox Process

Inhomogeneous PPP where the log intensity is a Gaussian process

$$\Lambda = \{\lambda(s,t) : s \in W, t \in (0,T)\}$$
$$\lambda(s,t) = \mu(s,t)R(s,t)$$
$$R(s,t) = \exp\{z(s,t)\beta + Y(s,t)\}.$$

- $\mu(s,t)$ is a known offset (e.g., population density)
- $R(s,t)$ is the infection risk.
  - $z(s,t)$ are the possible explanatory variables (e.g., land use, socioeconomic deprivation, rain fall)
  - $Y(s,t)$ is a spatio-temporal Gaussian process with parameters η. We can think of these as proxies for unmeasured risk factors that are correlated in space and time.

Modelling Challenges Revisited

- Cases at high spatial resolution and risk factors have different spatial support
  $\rightarrow$ Point processes.
- Cases may be non uniformly underreported
  $\rightarrow$ LGCP.
- Fitting a full spatio-temporal latent GP is computationally challenging
- Risk factors are correlated both with each other and in space/time.

Likelihood of Log Gaussian Cox Processes

The LGCP is doubly stochastic and the likelihood is intractable:

$$\mathcal{L}(\beta, \eta; X) = \mathsf{E}_{\Lambda|\beta,\eta}\mathcal{L}(\beta, \eta; X, \Lambda)$$

$$\mathcal{L}(\beta, \eta; X, \Lambda) \propto \exp\left\{-\int_0^T \int_W \Lambda(s,t)dsdt\right\} \prod_{i=1}^{n} \Lambda(s_i, t_i)$$

Grid approximation to likelihood assuming Λ is piecewise constant on cells $g_{m,n,t}$

$$\mathcal{L}(\beta, \eta; X, \Lambda_g) \propto \exp\left\{-\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{n=1}^{N} \Lambda(g_{m,n,t})\mathsf{vol}(g_{m,n,t})\right\}$$
$$\times \prod_{t=1}^{T}\prod_{m=1}^{M}\prod_{n=1}^{N} \Lambda(g_{m,n,t})^{|X \in g_{m,n,t}|}$$

Computation for Log Gaussian Cox Processes

Approaches to inference on β, η, Y:

- Maximum Likelihood
  - GLMs to estimate β assuming independence of grid squares.
  - GAMs to estimate β with non-parametric space-time smoothers to approximate Y.
  - For η: MCMLE or match pair correlation functions.

- Bayesian Inference
  - INLA on the grid or off grid using SPDE-based approach.
  - MCMC:
    - Transform Y and use MALA +FFT (in `lgcp` package, Taylor et al., 2015).
    - Something more efficient?

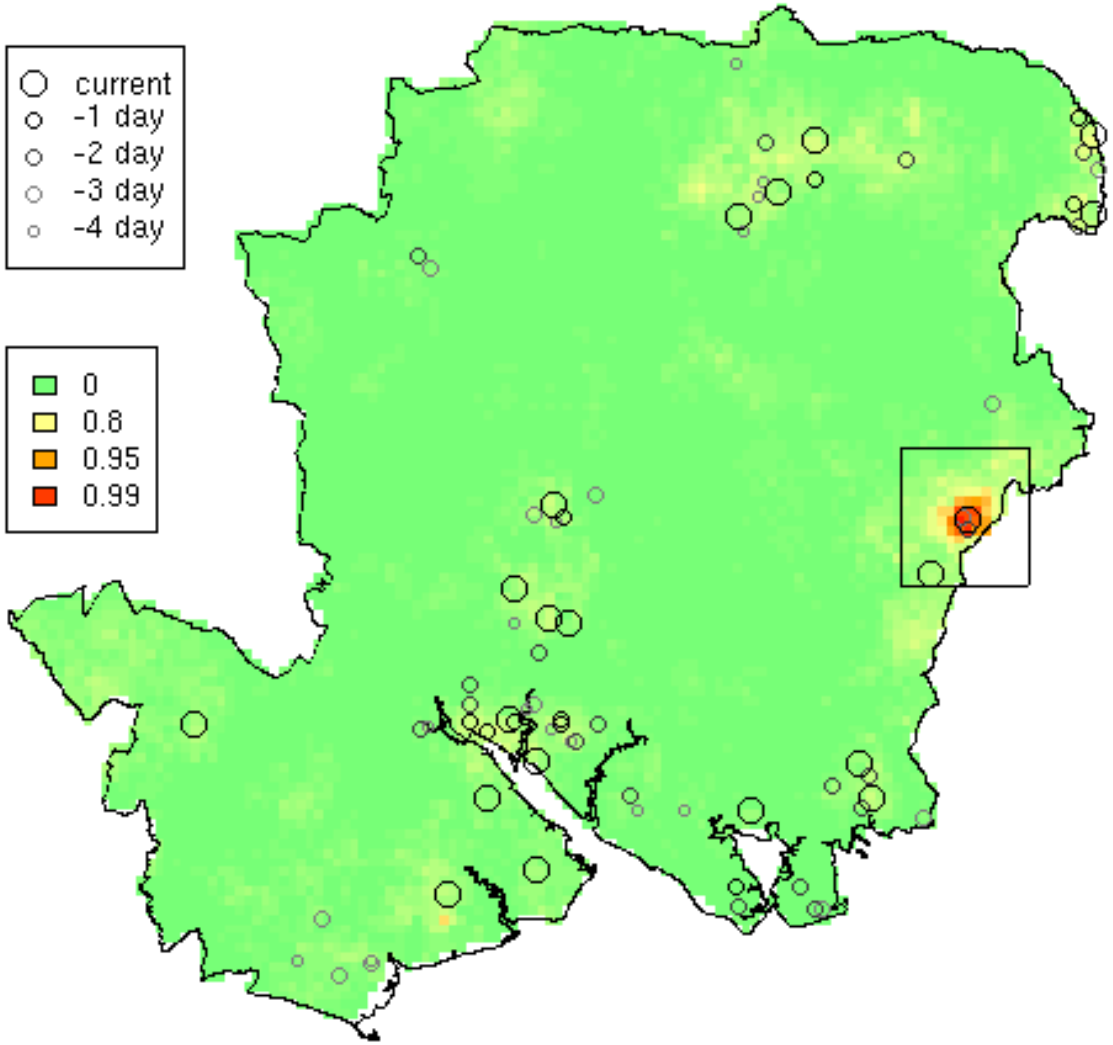
Many studies take an 'all of the above' approach.

Example: Diggle, Rowlingson and Su (2005)

Real-time surveillance for gastrointestinal illness.

$$\lambda(s,t) = \widehat{\lambda}_0(s)\widehat{\mu}_0(t)\exp Y(s,t)$$

- $\widehat{\lambda}_0(s)$ estimated using KDEs
- $\widehat{\mu}_0(t)$ estimated using GLMs (using $\widehat{\lambda}_0(s)$ as an offset)
- Hyper parameters for GP ($\widehat{\eta}$) estimated by matching theoretical and empirical pair correlation functions.
- MCMC used for Y (with everything else fixed)


current
-1 day
-2 day
-3 day
-4 day
0
0.8
0.95
0.99

MCMC for LGCPs

From Møller et al. (1998) and Brix and Diggle (2001), define Γ:

$$Y = \Sigma_\eta^{1/2}\Gamma + \mu_\eta.$$

Metropolis Hastings with MALA proposal for Γ and β and a random-walk proposal for $\log(\eta)$:

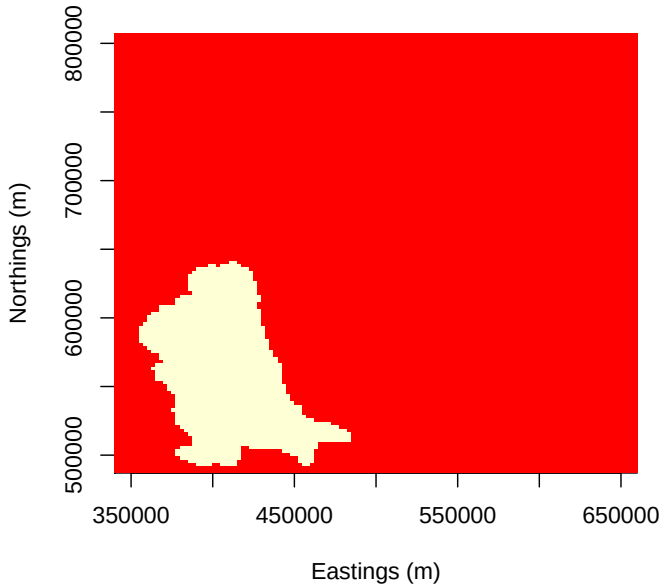
MALA proposal:

$$u' \mid u_m \sim \mathsf{N}\left(u_m + \frac{h}{2}S\nabla \log \pi(u_m), hS\right),$$

where proposal variance h can be adaptively tuned.

Bottle Necks

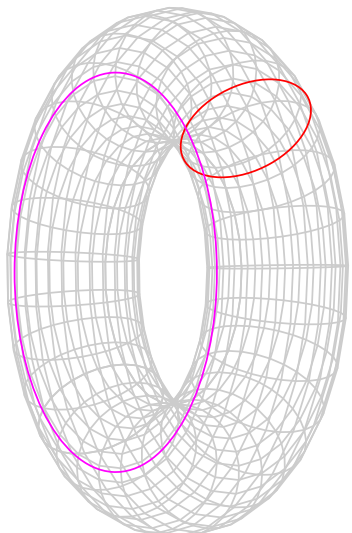
Inverting and square-rooting matrices in $O(n^3)$.



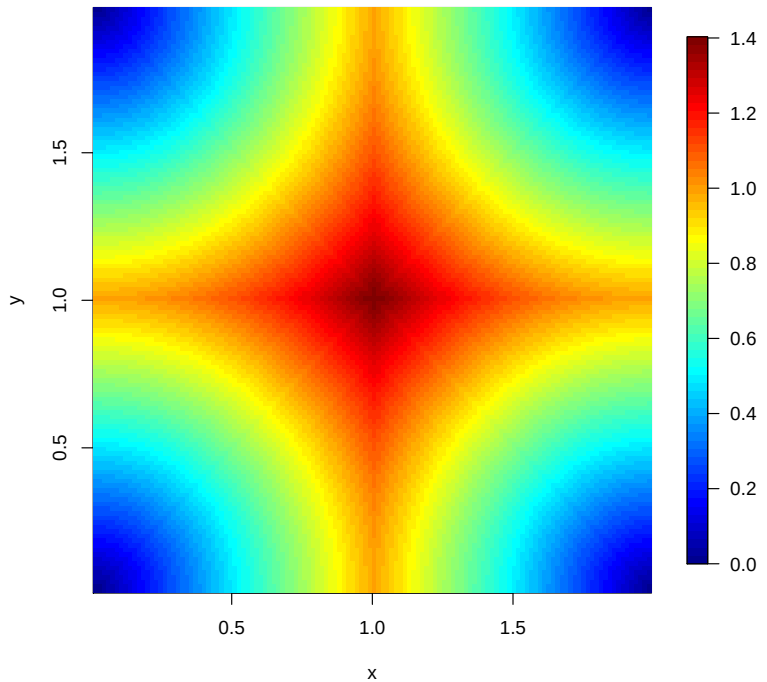
- Solution: extend and wrap grid so that Σ_η^{Ext} is block circulant and use 2-D DFT.
- Joint distribution for the points on the original grid is preserved.

$\rightarrow$FFT (and other bits of the sampler) can be implemented w/GPUs using, for example, TensorFlow.

Extended Grid



(a) 3-D torus. Unwraps to a flat grid by first cutting the red circle and then cutting along the magenta circle.



(b) Elements of the distance matrix of the grid distances are calculated across the surface of the 3-D torus.

WHAT ABOUT TIME?

Full spatio-temporal GP with separable covariance:

$$\mathsf{E}Y(t,s) = \mu(s,t) = -\sigma^2/2$$
$$\mathsf{cov}(Y(s_1,t_1), Y(s_2,t_2)) = \sigma^2 f_1(|s_1 - s_2|, \rho) f_2(|t_1 - t_2|, \phi)$$

Additive spatio-temporal: $Y = Y_1(s) + Y_2(t)$

$$\mu_1(s) = -\sigma^2/2 \qquad C_1(Y_1(s_1), Y_1(s_2)) = \sigma^2 f_1(|s_1 - s_2|, \rho)$$
$$\mu_2(t) = -\tau^2/2 \qquad C_2(Y_2(t_1), Y_2(t_2)) = \tau^2 f_2(|t_1 - t_2|, \phi)$$

Modelling Challenges Revisited

- Cases at high spatial resolution and risk factors have different spatial support
  $\rightarrow$ Point processes.
- Cases are non uniformly underreported
  $\rightarrow$ LGCP.
- Fitting a full spatio-temporal latent GP is computationally demanding
  $\rightarrow$ Grids, FFTs, Y(s,t) = Y(s) + Y(t).
- Risk factors are correlated both with each other and in space/time.

Spatial LGCPs for Campylobacter in NE England

- Spatial grid cells are 2.5km by 2.5km (need a 64×64 grid).
- Time discretized to weeks (274 weeks).
- Priors on β and η:

$$\beta \sim N(\mathbf{0}, sI) \qquad \log \eta \sim \mathsf{N}(\hat{\eta}, D)$$

- $\mu(g)$ based on population density on 1km by 1km raster.
- Assume covariates are constant on their original units and take spatially-weighted averages to get $Z(g)$.
- Initial model selection using GLMs + INLA

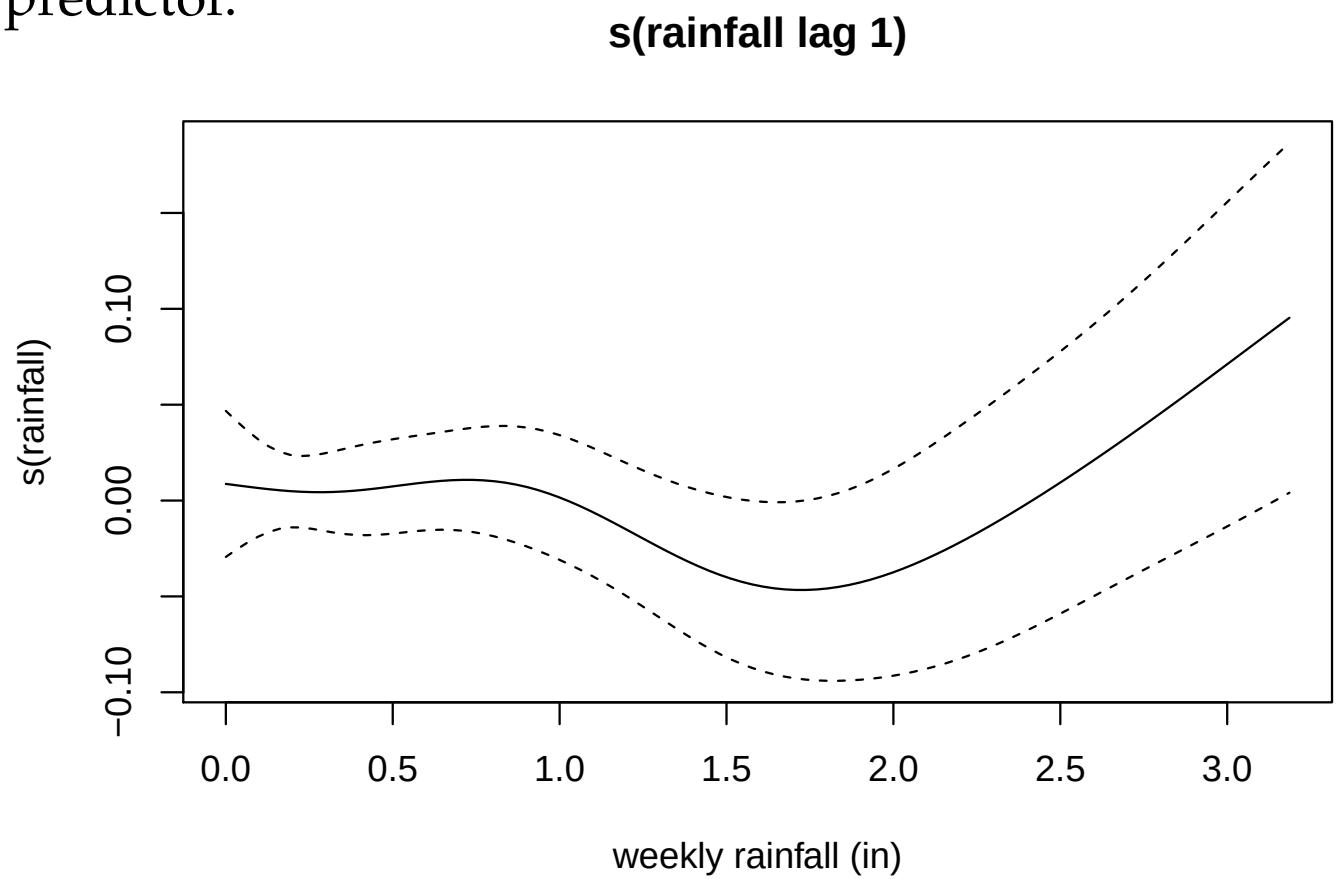
Spatial Results

	median	95% CrI
σ	0.96	(0.86, 1.11)
ρ (km)	3.43	(2.47, 5.13)
urban-rural	1.07	(1.01, 1.14)
health	1.29	(1.08 ,1.53)
education	0.86	(0.76 , 0.97)
crop area	1.00	(0.97, 1.04)
cattle	0.97	(0.94 ,1.01)
sheep	1.01	(0.96, 1.07)

Socio-economic deprivation and urban-rural are associated with Campylobacter risk.

Spatio-temporal predictors

- Harmonic functions with annual and half year periods.
- Interpolated daily minimum temperature and rainfall station data, averaged over the month.
- Non-linear relationship in rainfall $\rightarrow$ use as a categorical predictor.



Spatio-temporal Results

	median	95% CrI
σ	1.07	(0.97,1.60)
ρ (km)	3.22	(2.03,5.53)
τ	0.22	(0.185,0.26)
ϕ (w)	1.59	(1.00,2.65)
t	1.00	(0.999,1.0)
$\cos_1 t$	4.45	(1.16,16.1)
$\sin_1 t$	0.86	(0.76,0.97)
$\cos_2 t$	1.07	(1.01,1.14)
$\sin_2 t$	1.01	(0.94,1.08)
$\cos_3 t$	0.97	(0.91,1.03)
$\sin_3 t$	0.93	(0.87,0.99)
day lgth	1.46	(1.11,1.90)

	median	95% CrI
temp anomaly		
lag 1	1.01	(0.99,1.03)
Q1	0.97	(0.90,1.03)
Q2	0.99	(0.93,1.05)
rain lag 1 Q3	–	–
Q4	0.97	(0.91,1.04)
Q5	1.00	(0.92, 1.10)
urban-rural	1.01	(0.95, 1.09)
IMD	0.99	(0.95,1.01)

De-trended minimum temperature and rainfall are not important risk factors when already accounting for overall seasonality.

Modelling Challenges Revisited

- Cases at high spatial resolution and risk factors have different spatial support
  $\rightarrow$ Point processes.
- Cases are non uniformly underreported
  $\rightarrow$ LGCP.
- Fitting a full spatio-temporal latent GP is computationally demanding
  $\rightarrow$ Grids, FFTs, Y(s,t) = Y(s) + Y(t).
- Risk factors are correlated both with each other and in space/time.

Interpretation issues

- Ecological bias–we only have area-level risk factors, so we really can't say anything about how our predictors effect individual-level risk.
- Collinearity–many predictors are correlated (e.g, deprivation subindices).

- Spatial confounding–Can't guarantee that the spatial residuals aren't distorting our estimates of the effects of spatially correlated predictors.
  - Reich, Hodges, and Zadnick (2006), Paciorek (2010), Hodges and Reich (2010)

SPATIAL CONFOUNDING

We simulate a spatial LGCP on a 32 by 32 grid over the unit square with

$$\log \lambda(s) = 5 + 0.1 \cdot Z(s) + Y(s)$$
$$Y \sim GP(0, \text{Matérn}(\nu = 1, \rho))$$

Where the spatial predictor $Z(s)$ is one of

$$Z_1 = \text{distance from point source}$$
$$Z_2 \sim \mathsf{N}(0, 1)$$
$$Z_3 \sim GP(0, \text{Matérn}(\nu = 1, 0.25))$$

Simulation Results

We compare based on bias in the point estimates coverage of 95% (confidence or credible) intervals

		point source		$\mathsf{N}(0,1)$		GP	
ρ	method	Bias	Cov	Bias	Cov	Bias	Cov
$\frac{3}{32}$	GLM	0.036	0.46	0.002	0.91	-0.002	0.41
	GAM	1.25	**0.91**	**0.000**	0.92	-0.014	0.82
	INLA-sp	**-0.007**	0.79	**0.000**	**0.95**	0.002	**0.86**
	INLA-iid	0.034	0.48	0.002	0.93	**-0.001**	0.45
$\frac{1}{32}$	GLM	0.019	0.84	**-0.008**	0.85	**-0.003**	0.76
	GAM	0.447	0.97	-0.011	0.85	0.009	0.88
	INLA-sp	**0.009**	**0.95**	-0.010	**0.88**	0.004	**0.92**
	INlA-iid	0.018	0.85	-0.009	**0.88**	**-0.003**	0.80

Discussion

- LGCPs are useful for high resolution health data.
- MALA is still a useful tool for spatial LGCPs but too slow for large spatio-temporal problems.

- Socio-economic deprivation and land-use are associated with Campylobacter risk.
- De-trended minimum temperature and rainfall are not important risk factors when already accounting for overall seasonality.

- ↑ These findings are harder to interpret than you'd think!

Acknowledgements



MRC Medical Research Council
NATURAL ENVIRONMENT RESEARCH COUNCIL
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
BBSRC bioscience for the future
Food Standards Agency