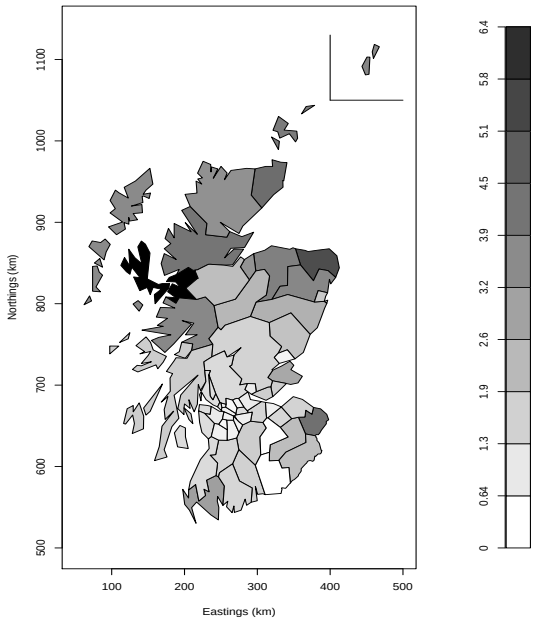
Statistical Modelling Approaches to Disease Mapping

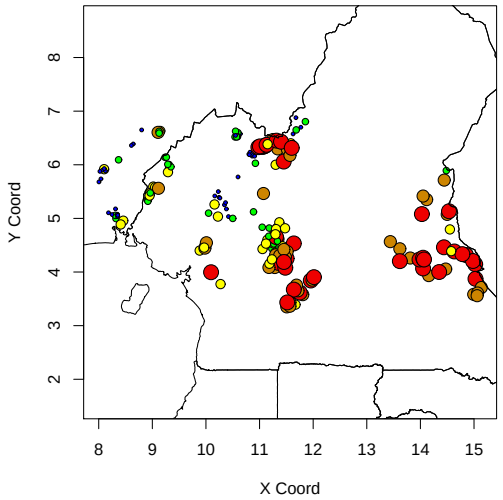
Peter J Diggle

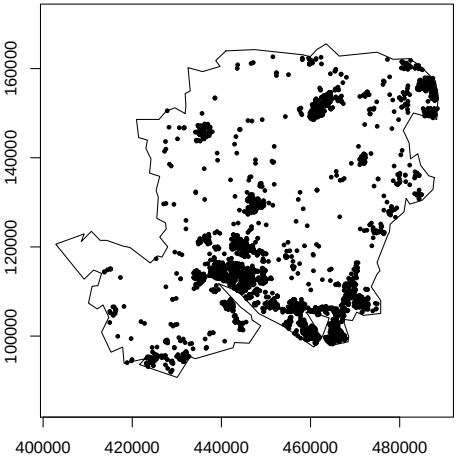
Lancaster University and University of Liverpool

UNIVERSITY OF LIVERPOOL | INSTITUTE OF INFECTION AND GLOBAL HEALTH

Spatial statistics according to Cressie (1991)



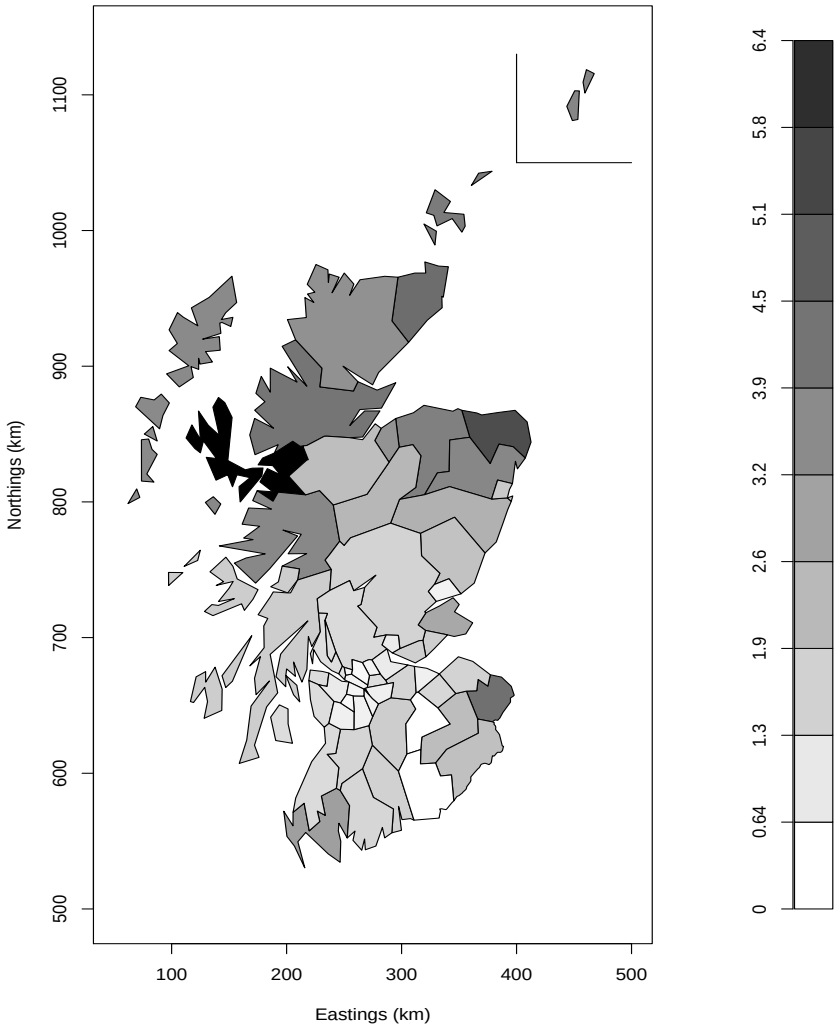




Lattice data **Geostatistics** **Point patterns**

Cressie, N.A.C. (1991). Statistics for Spatial Data. Wiley.

Lattice data: Scottish lip cancer incidence



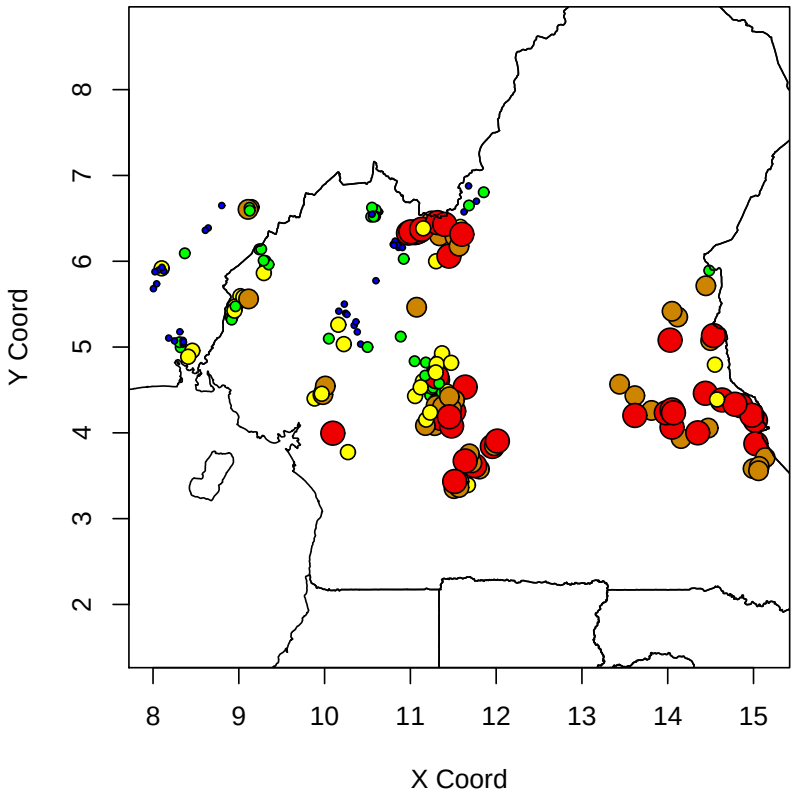
Data: county-level incidences
$Y_i : i = 1, ..., n$

Model: Markov random field:
$[Y_i | \{Y_j : j \neq i\}] : i = 1, ..., n$

- risks in near-neighbouring counties are positively correlated
- incidences Y_i are noisy versions of risk $\times$ population

Scientific interest confined to specified set of counties?

Geostatistics: Loa loa prevalence in Cameroon



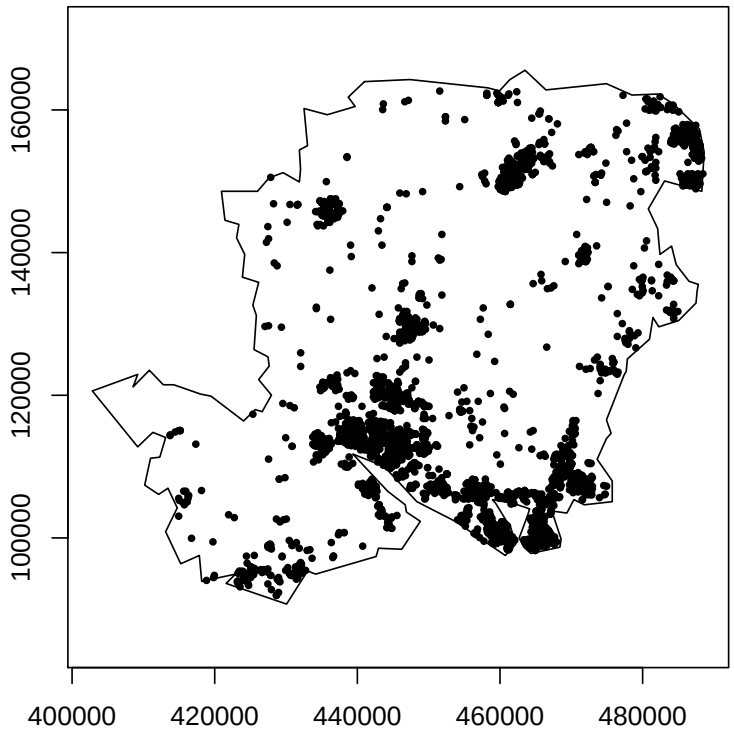
Data: empirical prevalences Y_i at sample locations $x_i : i = 1, ..., n$

Model: spatially continuous stochastic process, $S(x) : x \in \mathbb{R}^2$

- correlation between S(u) and S(v) specified as a function of distance between u and v
- $Y_i | S(x_i) \sim$ Binomial

Scientific interest extends to S(x) at non-sampled locations

Point pattern: gastro-enteric illness in Hampshire



Data: outcomes (x_i, t_i) are locations and dates of calls to NHS Direct recorded as "vomiting and/or diarrhoea"

Model: $(x_i, t_i) : i = 1, 2, \ldots$ a stochastic point process

- intensity $\lambda(x, t)$
- successive cases independent?

Scientific interest is in locations themselves

Disease mapping

Context

- region of interest **A**
- disease risk $\rho(\mathbf{x})$: $\mathbf{x} \in \mathbf{A}$
- data relating to variation in disease prevalence over **A**

Objective

- estimate $\rho(\mathbf{x})$?
- calculate $\mathrm{P}\{\rho(\mathbf{x}) > \mathbf{c}|\mathbf{data}\}$?

The answer to any prediction problem is a probability distribution

Peter McCullagh, FRS

Markov Random Field (MRF) models

- **Random variables** $S = (S_1, ..., S_n)$
- **Joint distribution** $[S]$ **fully specified by full conditionals,**

$$[S_i|\{S_j : j \neq i\}] : i = 1, ..., n$$

- **Neighbourhood of** i **is** $\mathcal{N}(i) \subset \{1, 2, ..., n\}$

$$[S_i|\{S_j : j \neq i\}] = [S_i|S_j : j \in \mathcal{N}(i)] : i = 1, ..., n$$

Hierarchical Poisson/Gaussian MRF

- latent Gaussian MRF $S = (S_1, ..., S_n)$,

$$S_i | \{S_j : j \neq i\} \sim N(\bar{S}_i, \tau^2/m_i)$$

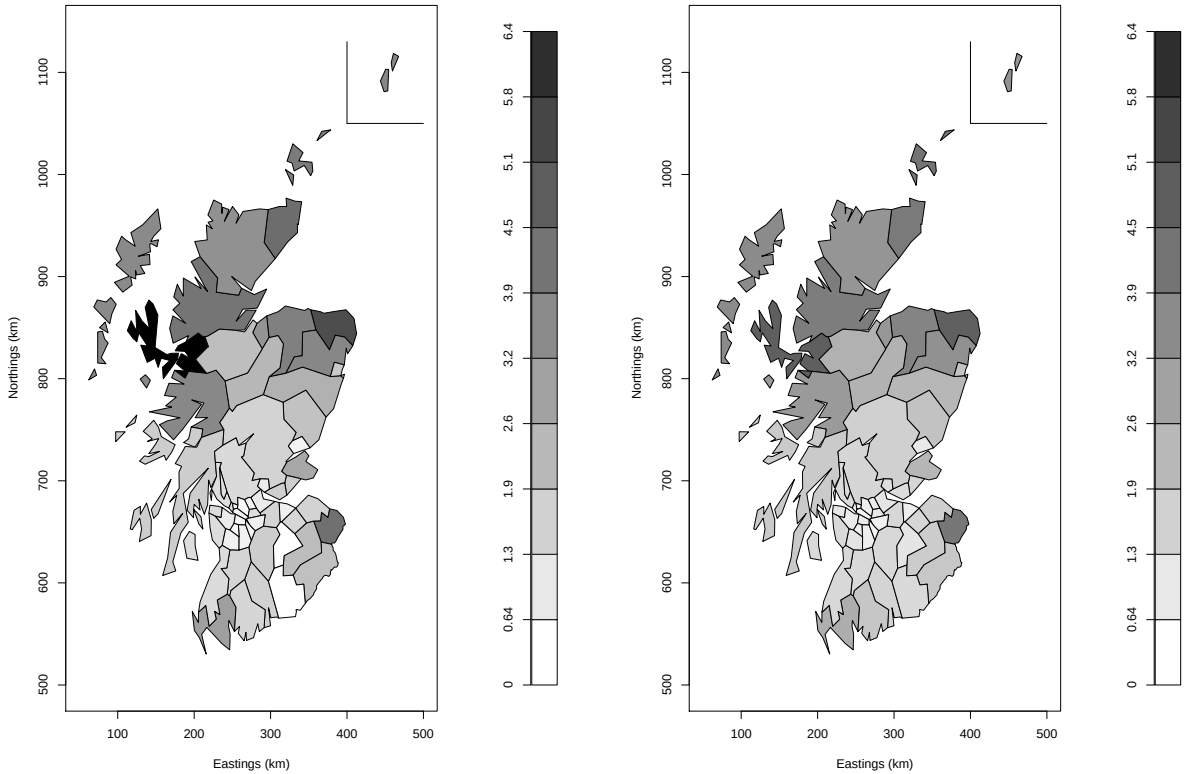
- conditionally independent $Y_i | S \sim \text{Poiss}(z_i'\beta + \gamma S_i)$

- risk map: $E[S_i|Y]$

Besag, York and Mollié, 1991

Cancer atlases

Raw and spatially smoothed relative risk estimates for lip cancer in 56 Scottish counties



Wakefield (2007)

Limitations of MRF models for spatial data

MRF's are just multivariate probability distributions

- **parameterised in a way that has a spatial interpretation**
- **but specific to a fixed set of locations $x_1, \ldots, x_n$**

Neighbourhood specification can be problematic

- **natural hierarchy of models on regular lattices**
- **not so for irregular lattices**
- **and arguably un-natural for spatially aggregated data,**

$$Y_i = \int_{A_i} Y(x)dx$$

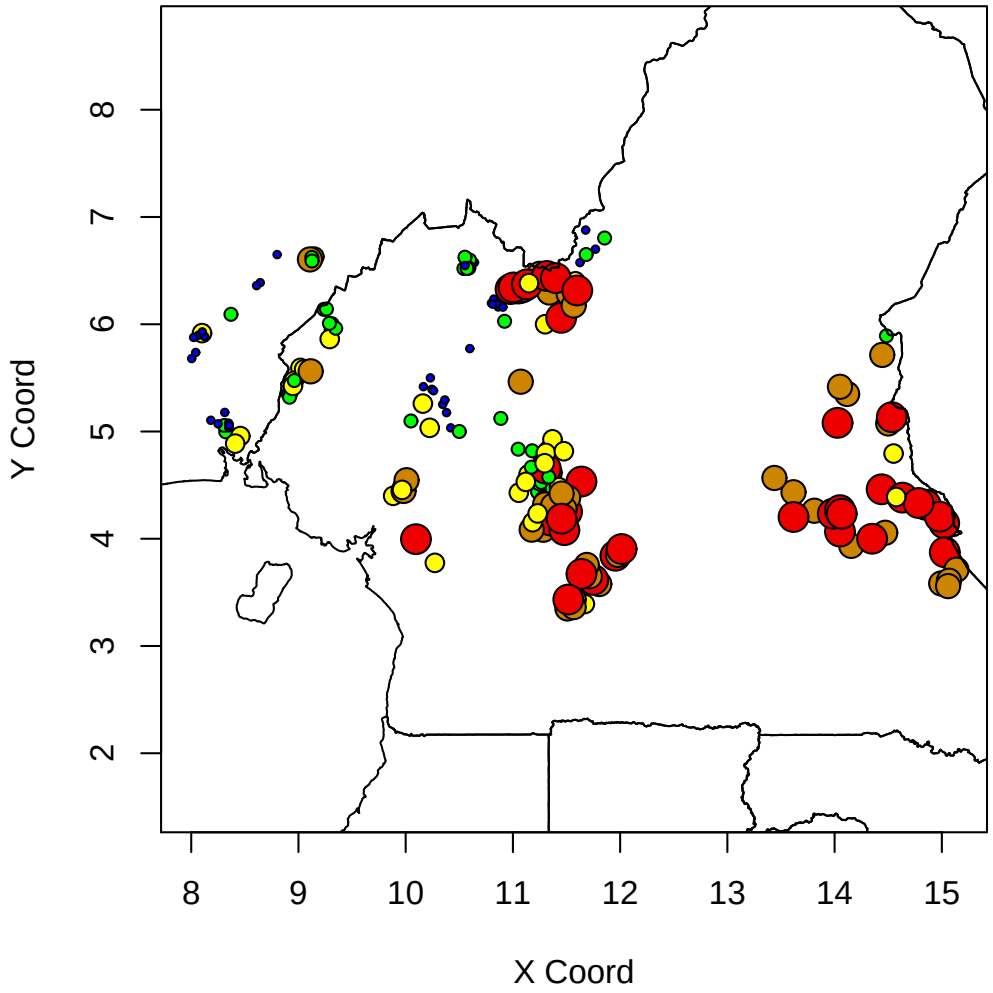
Geostatistical models

- **Stochastic process** $S(x) : x \in A \subset \mathbb{R}^2$
- **Data** $\{(Y_i, x_i) : i = 1, ..., n\}$
- **Stationary Gaussian model**

$$E[S(x) = 0] \quad Cov\{S(x), S(x-u)\} = \sigma^2 \rho(u)$$

$$[Y|S] = [Y_1|S(x_1)]...[Y_n|S(x_n)]$$

A geostatistical data-set: Loa loa prevalence surveys



Loa loa: generalised linear model

- **Latent spatially correlated process**

  $S(x) \sim \mathrm{SGP}\{0, \sigma^2, \rho(u))\}$
  $\rho(u) = \exp(-|u|/\phi)$

- **Linear predictor (regression model)**

  $d(x) =$ environmental variables at location x
  $\eta(x) = d(x)'\beta + S(x)$
  $p(x) = \log[\eta(x)/\{1 - \eta(x)\}]$

- **Conditional distribution for positive proportion Y_i/n_i**

  $Y_i|S(\cdot) \sim \mathrm{Bin}\{n_i, p(x_i)\}$ (binomial sampling)

Probabilistic exceedance map for Cameroon (Diggle et al, 2007)

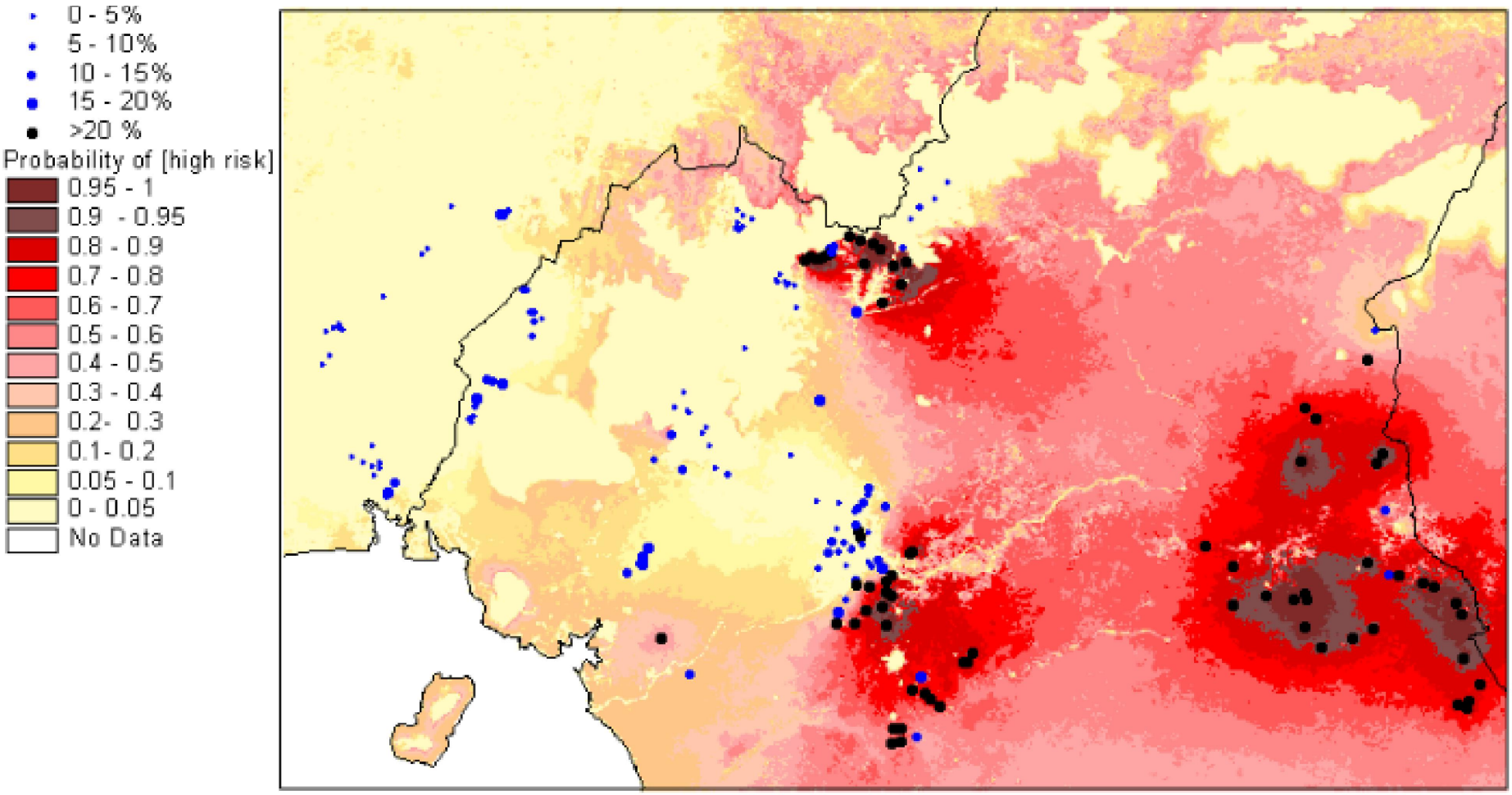


Figure 6: PCM for [high risk] in Cameroon based on ERMr with ground truth data.

Point process models (log-Gaussian Cox processes)

- **Stochastic process** $S(x) : x \in A \subset \mathbb{R}^2$
- **Data** $\mathcal{X} = \{x_i : i = 1, ..., n\}$
- **Stationary Gaussian model**

$$E[S(x) = 0] \quad \mathrm{Cov}\{S(x), S(x-u)\} = \sigma^2 \rho(u)$$

$$[\mathcal{X}|S] = \text{Poisson process, intensity } \Lambda(x) = \exp\{S(x)\}$$

Real-time spatial surveillance: spatio-temporal point process

Ascertainment and
Enhancement of
Gastroenteric
Infection
Surveillance
Statistics

- **largely sporadic incidence pattern**
- **concentration in population centres**
- **occasional "clusters" of cases**

Can spatial statistical modelling enable earlier detection of "clusters"?

AEGISS: log-Gaussian Cox process model

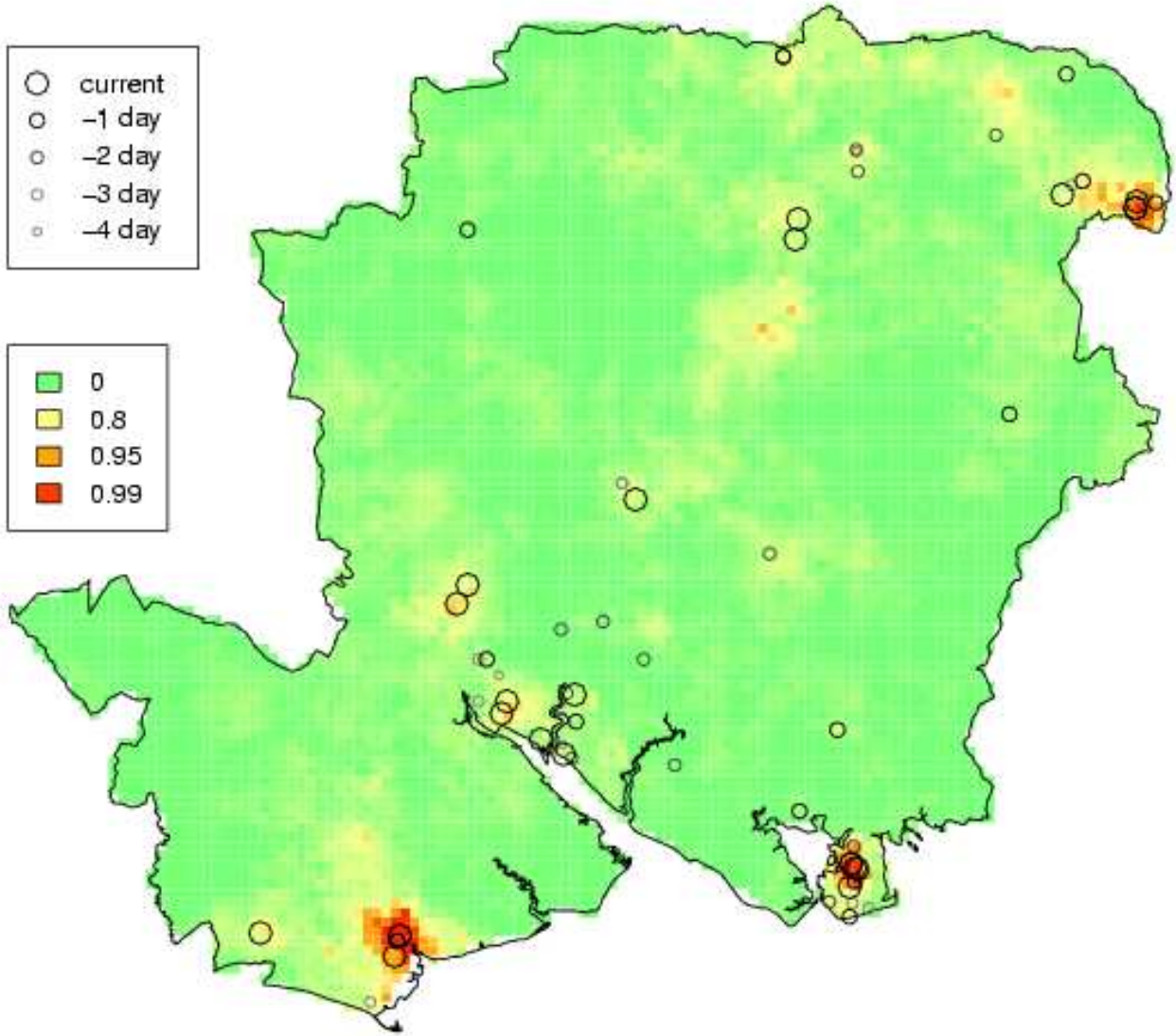
$$\begin{aligned} \text{intensity} &= \text{expected} \times \text{unexpected} \\ \Lambda(x,t) &= \lambda_0(x) \times \mu_0(t) \times R(x,t) \end{aligned}$$

Objective: use incident data up to time t to construct predictive distribution for current "anomaly" surface, $R(x,t)$

Model

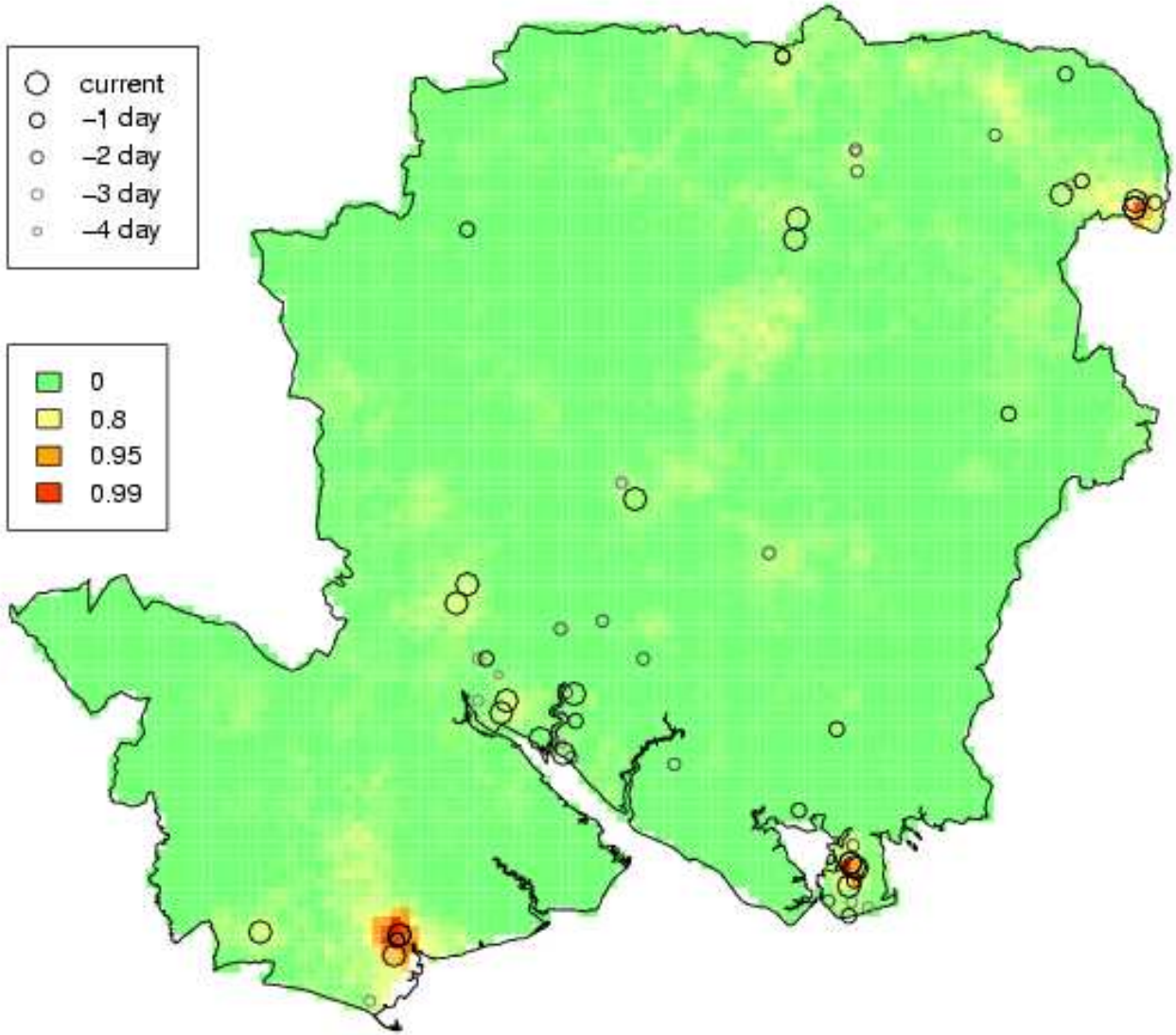
- spatio-temporal point process $\mathcal{P}$
- $\log R(x,t) \sim$ latent Gaussian process
- $\mathcal{P}|R \sim$ Poisson process

Spatial prediction: 6 March 2003



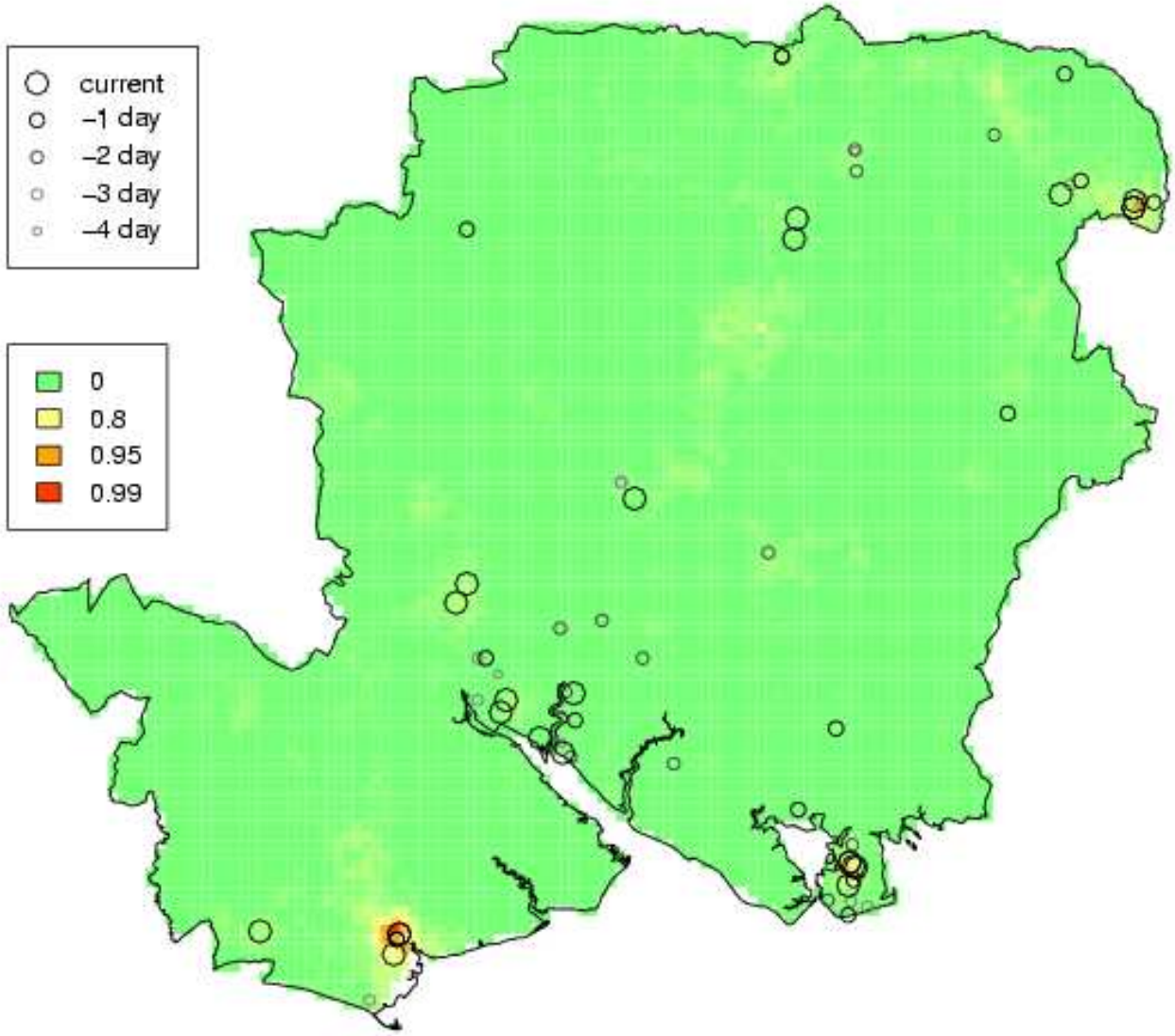
$c = 2$

Spatial prediction: 6 March 2003



$c = 4$

Spatial prediction: 6 March 2003



$c = 8$

Synthesis

$\mathbf{S}$ = state of nature
$\mathbf{Y}$ = all relevant data
$\mathbf{T}$ = $\mathcal{F}(\mathbf{S})$ = target for prediction

---

Model: $[\mathbf{S}, \mathbf{Y}] = [\mathbf{S}][\mathbf{Y}|\mathbf{S}]$
Prediction: $[\mathbf{S}, \mathbf{Y}] \Rightarrow [\mathbf{S}|\mathbf{Y}] \Rightarrow [\mathbf{T}|\mathbf{Y}]$

Diggle, P.J., Moraga, P., Rowlingson, B. and Taylor, B. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science* **(to appear)**

Pau da Lima, Salvador, Brazil

Pau da Lima, Salvador, Brazil

Leptospirosis cohort study: Pau da Lima

- **subjects** i at locations x_i, **blood-samples** taken at times $t_{ij} \approx 0, 6, 12, 18, 24$ months
- **sero-conversion** defined as change from zero to positive, or at least four-fold increase in concentration
- **data** consist of:
  - $Y_{ij} = 0/1 : j = 1, 2, 3, 4$ (seroconversion no/yes)
  - $r_i(t)$ known and hypothesised risk-factors

Leptospirosis cohort study: analysing the data

Longitudinal data, binary outcome $\Rightarrow$ standard problem?

id	Follow-up				Age
	1	2	3	4	
1	0	0	1	0	57
2	0	0	0	0	34
3	0	0	1	X	38
4	1	1	1	0	28
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
950	0	1	0	1	40

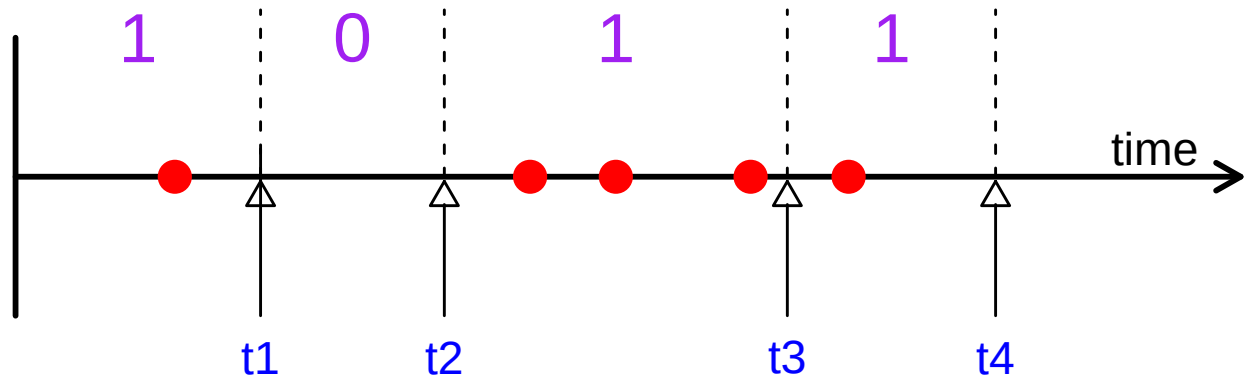
Logistic regression for binary response,

$$\log\{p_{it}/(1-p_{it})\} = \alpha + \beta \times \text{age}$$

Need to account for correlation amongst repeated outcomes on same individual

- generalized estimating equations
- generalized linear mixed models
- ...

Leptospirosis cohort study: analysing the problem



- **infection events** on each individual form a point process with time-varying intensity, $\Lambda_i(t)$
- **follow-up times** partially censor the point process record
- **reduction to binary data** represents additional censoring

Leptospirosis cohort study: model formulation

Data: $Y_{it} = 0/1 \qquad t = 1, 2, 3, 4 \qquad i = 1, 2, \ldots, n$

- $Y_{it} = 1 \Leftrightarrow$ at least one infection event
- model infection events as person-specific, inhomogeneous Cox processes,

$$\Lambda_i(t) = \exp\{r_i(t)'\beta + U_i + S(x_i)\}$$

$$P(Y_{it} = 1 | \Lambda_i(\cdot)\} = 1 - \exp\left\{ - \int_{t_{i,j-1}}^{t_{ij}} \Lambda_i(u) du \right\}$$

Inference: likelihood rules OK?

- **The likelihood principle**

  Two data-sets x and y that generate identical likelihood functions are equivalent as evidence

- **The law of likelihood**

  If $H_A \Rightarrow p_A(x)$ and $H_B \Rightarrow p_B(x)$, then data x constitutes evidence in favour of A over B iff $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$ measures the strength of the evidence

Inference: what's the question?

- **Bayesian**

  What should I believe?

- **Decision-theoretic**

  What should I do?

- **Classical:**

  What do the data tell me?

Royall, R. (1997). Statistical Evidence: a likelihood paradigm. London: Chapman and Hall.

Acknowledgements

CHICAS, Lancaster University : Paula Moraga, Barry Rowlingson, Ben Taylor

APOC Madeleine Thomson, Hans Remme, Honorat Zoure, ...

Yale University/Fiocruz, Brazil: Federico Costa, Jose Hagan, Albert Ko

MRC: Methodology Research Grant G0902153