# Statistical Methods for real-time monitoring of health outcomes

**Peter J Diggle**

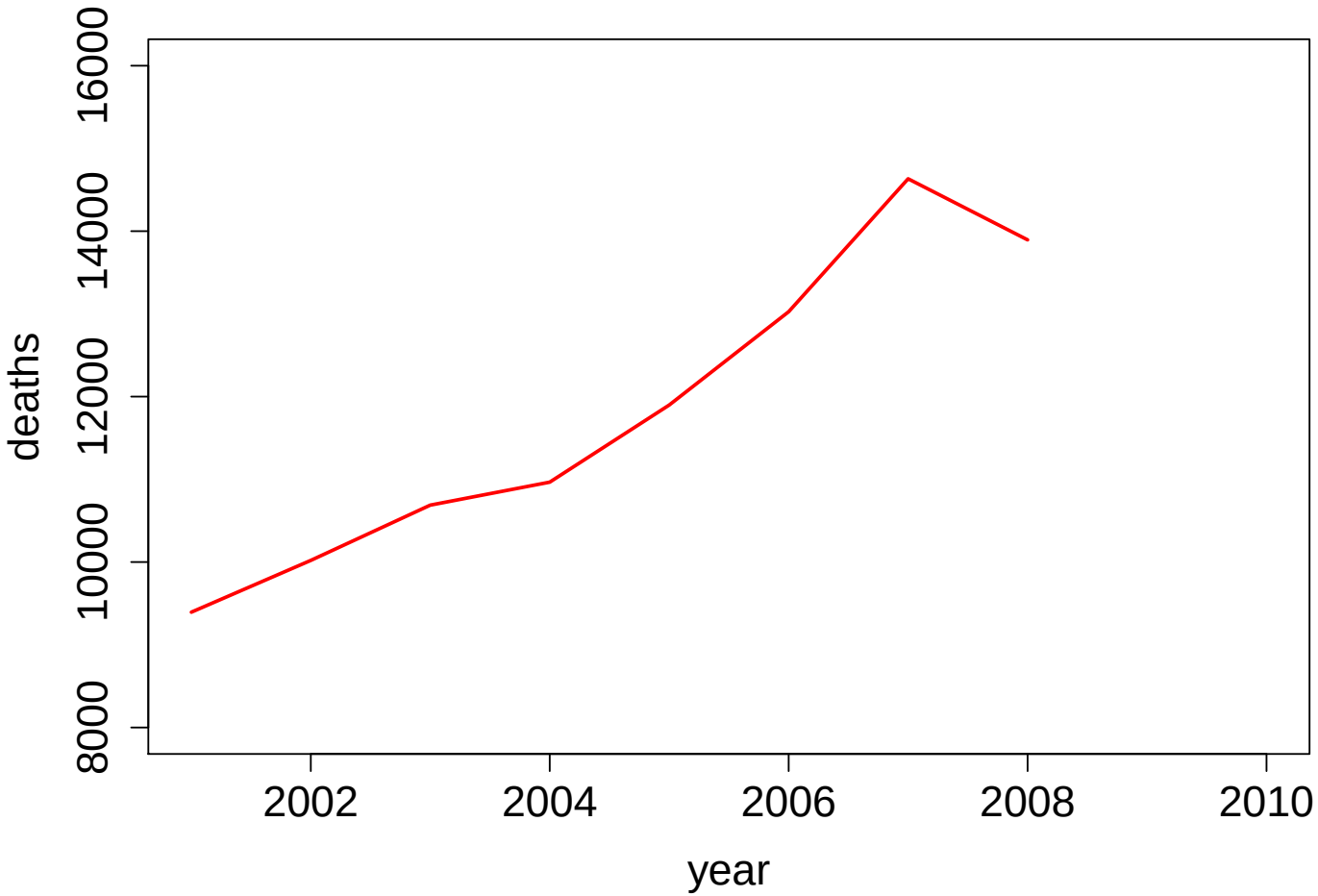
**CHICAS, Faculty of Health and Medicine, Lancaster University**

**September 2015**

# Context

- **increasing availability of electronically recorded health outcome data**
- **community and/or individual level**
- **accruing in "real-time"**
- **often spatially referenced**
- **prediction and/or explanation**
- **case-studies:**
  - **monitoring progression towards end-stage renal failure**
  - **human and veterinary surveillance of gastro-enteric illness**
  - **local-scale malaria prevalence mapping**

# Chronic renal failure: UK mortality data




http://www.endoflifecare-intelligence.org.uk

# Diagnosis, treatment and survival

## Diagnosis

- Serum creatinine $\Rightarrow$ estimated glomerular filtration rate

$$\text{eGFR} = 186 \times \left(\frac{\text{SCr}}{88.4}\right)^{-1.154} \times \text{age}^{-0.203} (\times 0.742 \text{ if female})$$

- progression can be asymptomatic for many years
- SCr easy to measure from blood-sample

## Treatment and survival

- aggressive control of blood-pressure
- renal replacement therapy: dialysis and transplantation
- early diagnosis can slow rate of progression

| | Survival rate (%) to year | | | |
|---|---|---|---|---|
| | 1 | 2 | 5 | 10 |
| Dialysis | 79.3 | 64.7 | 33.6 | 10.2 |
| Transplant (living) | 98.4 | 96.5 | 90.0 | 76.0 |

# Royal Salford Hospital, NW England

## Clinical guideline

Loss of $> 5\%$ eGFR per year $\Rightarrow$ refer to secondary care

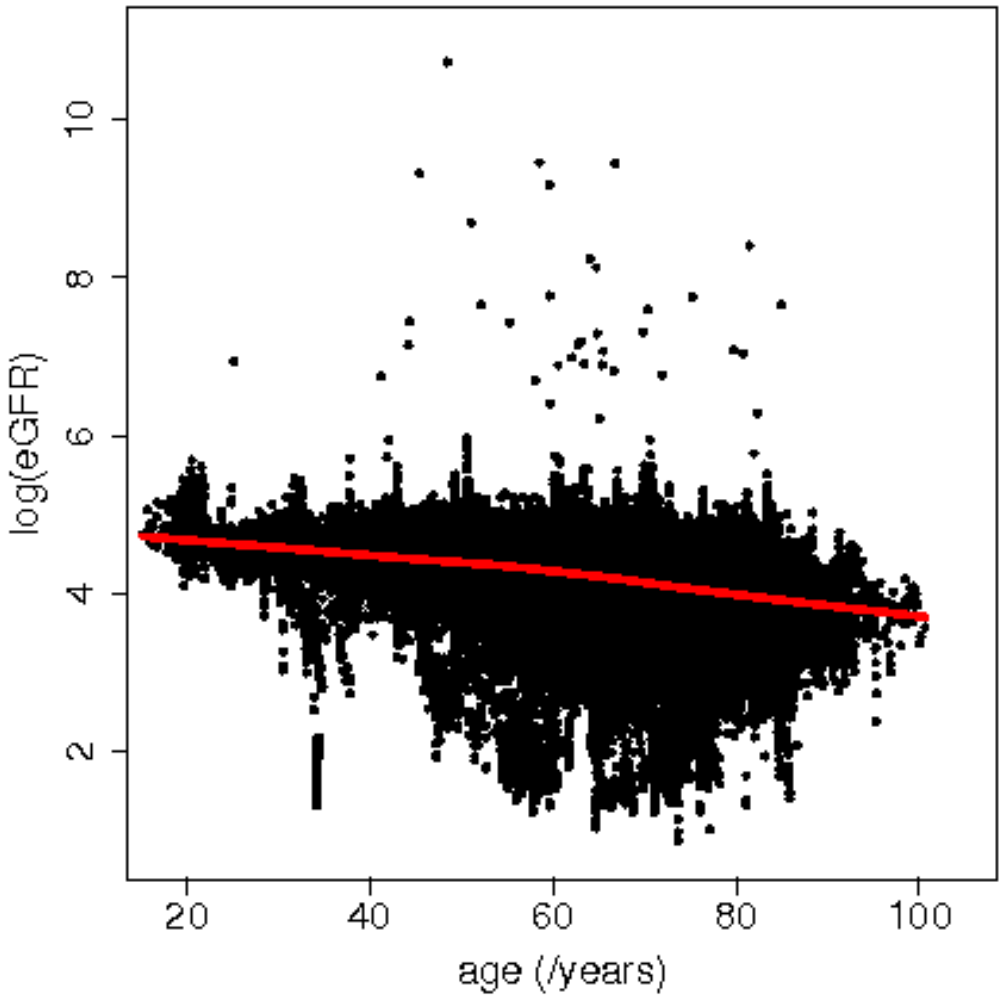
## Data

- measurements $Y_{ij} = \log \text{eGFR}$ at times $t_{ij}$,
  explanatory variables $x_i$ (age, sex)
  - $i = 1, \ldots, m = 22,910$ "at-risk" primary care patients
  - $j = 1, \ldots, n_i \leq 305$ (median $n_i = 12$)
  - $0 \leq 10.02$ years follow-up (median 4.46)
- $\mathcal{H}_i(t) = \{x_i, (t_{ij}, y_{ij}) : t_{ij} \leq t\}$

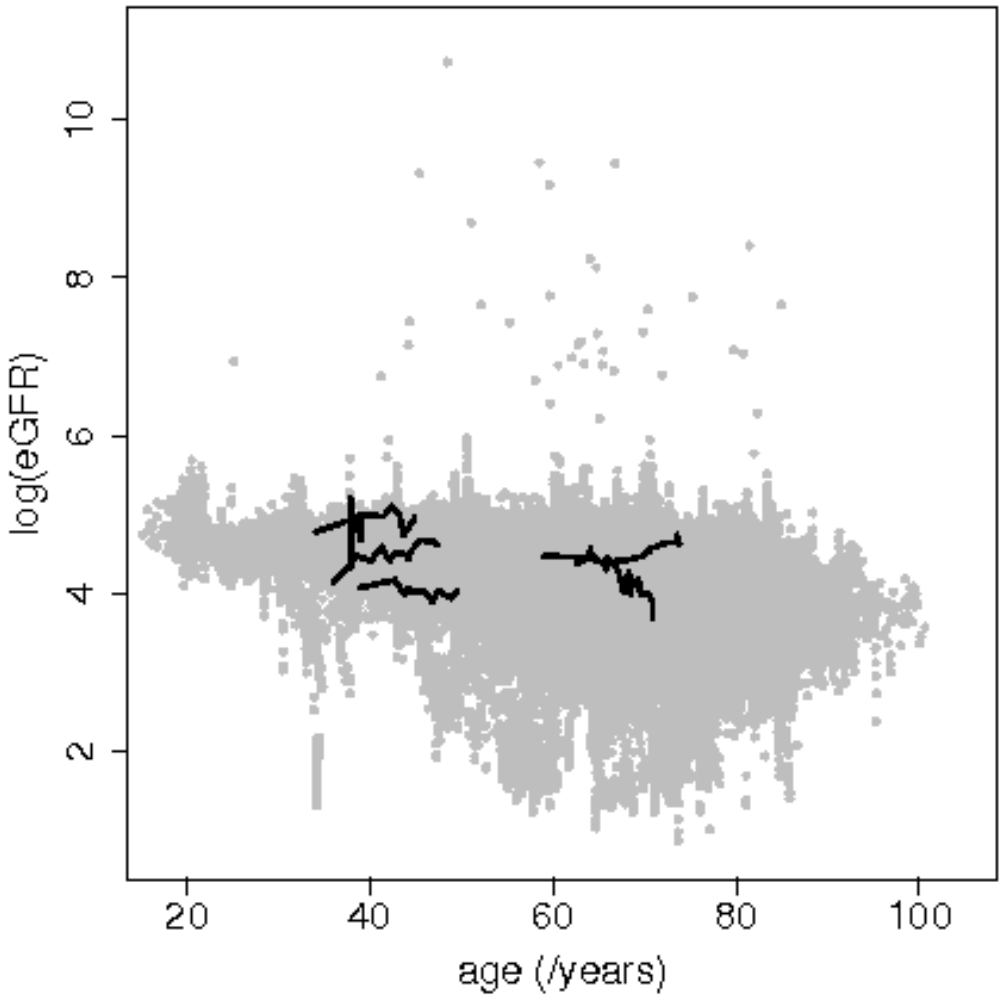
## Statistical objective

$$\mathrm{P}\left(\frac{d}{dt}\log \text{GFR} < -0.05 | \mathcal{H}_i(t)\right) = \quad ?$$

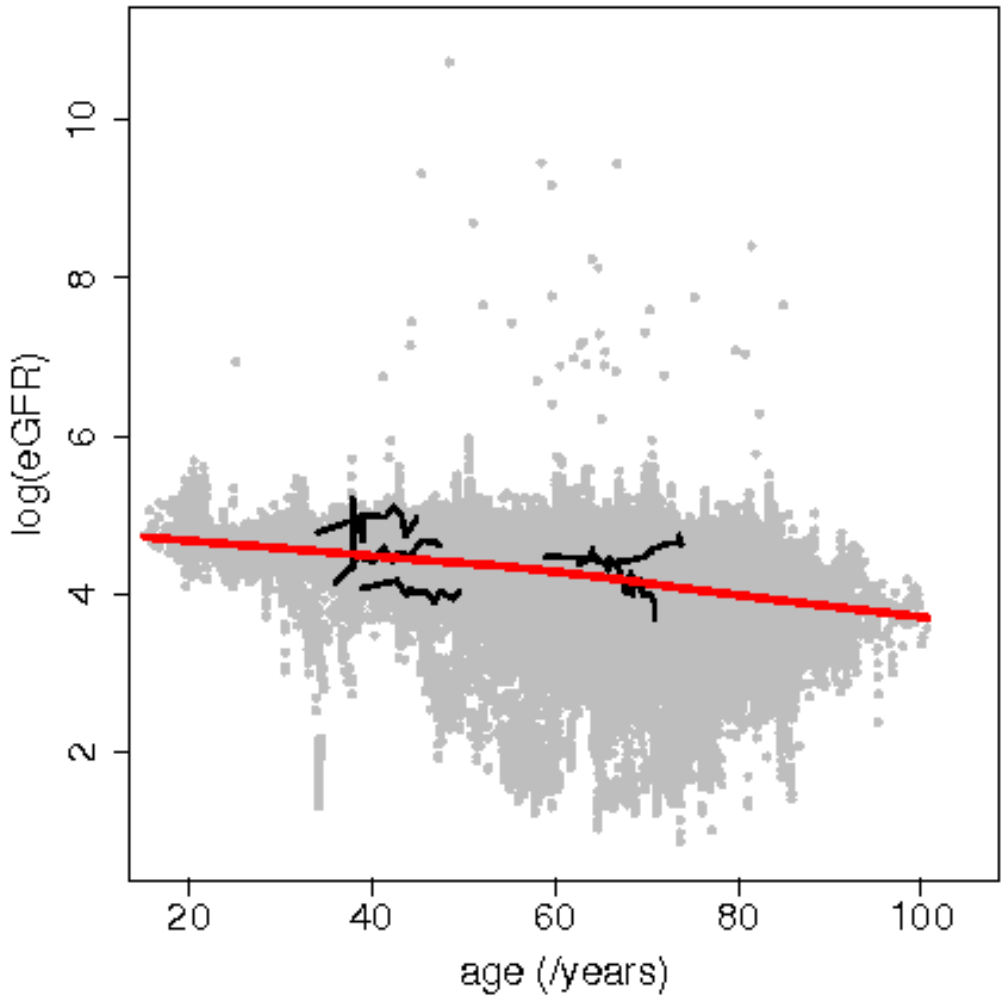
# Data: all cross-sectional



# Data: all cross-sectional and selected longitudinal



# Data: all cross-sectional and selected longitudinal



# Dynamic regression model

- **subjects** $i = 1, \ldots, n$ observed at times $t_{ij}, j = 1, \ldots n_i$

$$Y_{ij} = \log(\text{eGFR})$$

- **expected value** of $Y_{ij}$ linear in initial age and time since recruitment
- **rate of progression** varies randomly:
  - **between subjects:** random effect $U_i$
  - **within subjects:** random effect $C_i(t_{ij})$

# Dynamic Regression Model

$$
\begin{aligned}
Y_{ij} &= \alpha_0 + \alpha_1 \times I(\text{female}) \\
&+ \alpha_2 \times age_{i1} + \alpha_3 \times (age_{ij} - age_{i1}) + \alpha_4 \times \max(0, age_{ij} - 56.5) \\
&+ U_i + C_i(t_{ij}) + Z_{ij}
\end{aligned}
$$

- $Z_{ij}$: measurement error, $N(0, \tau^2)$
- $U_i$: between-subject random intercept, $N(0, \omega^2)$
- $C_i(t)$: within-subject stochastic process

Model $C_i(t)$ as integrated Brownian motion

$$
\begin{aligned}
C_i(t) &= \int_0^t B_i(u)du \\
B_i(u)|B_i(s) &\sim N\left(B_i(u), (u-s)\sigma^2\right)
\end{aligned}
$$

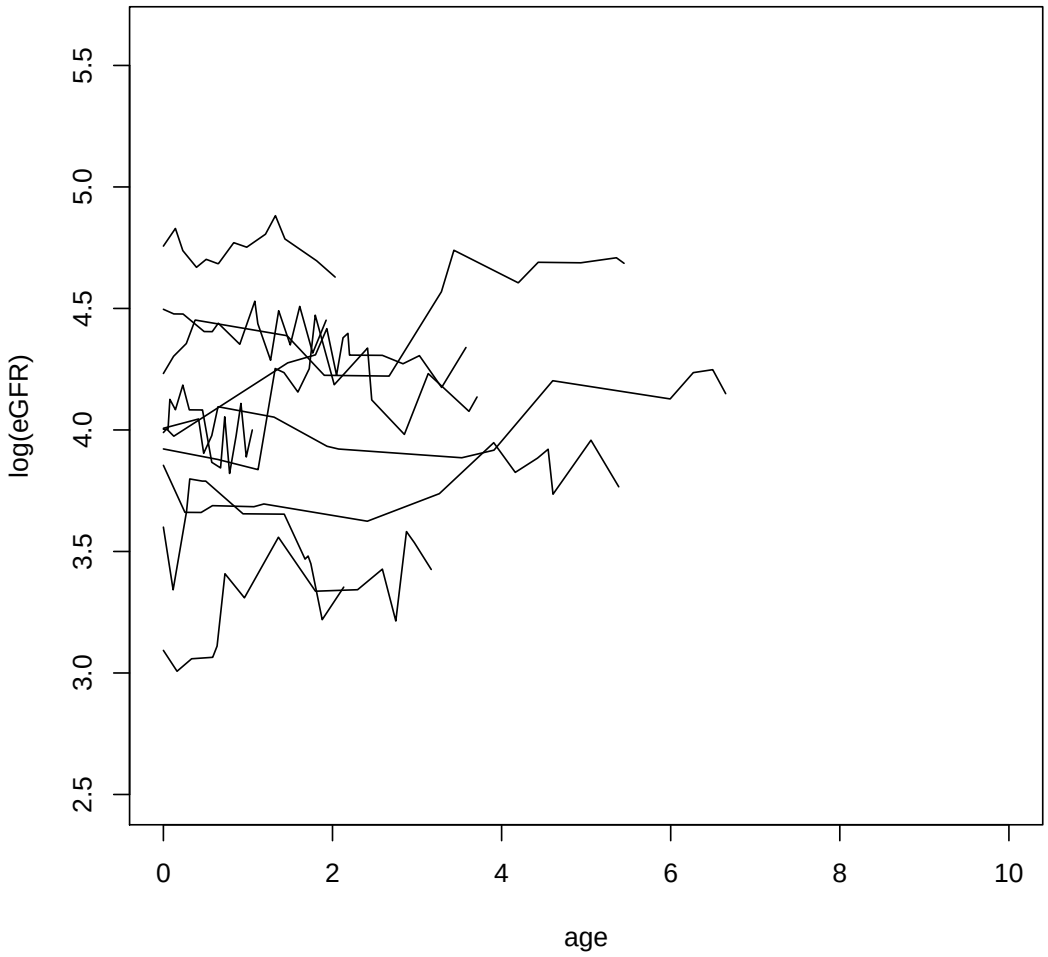
$B_i(u)$ is rate of progression for subject i at time t

# Maximum likelihood estimates of model parameters

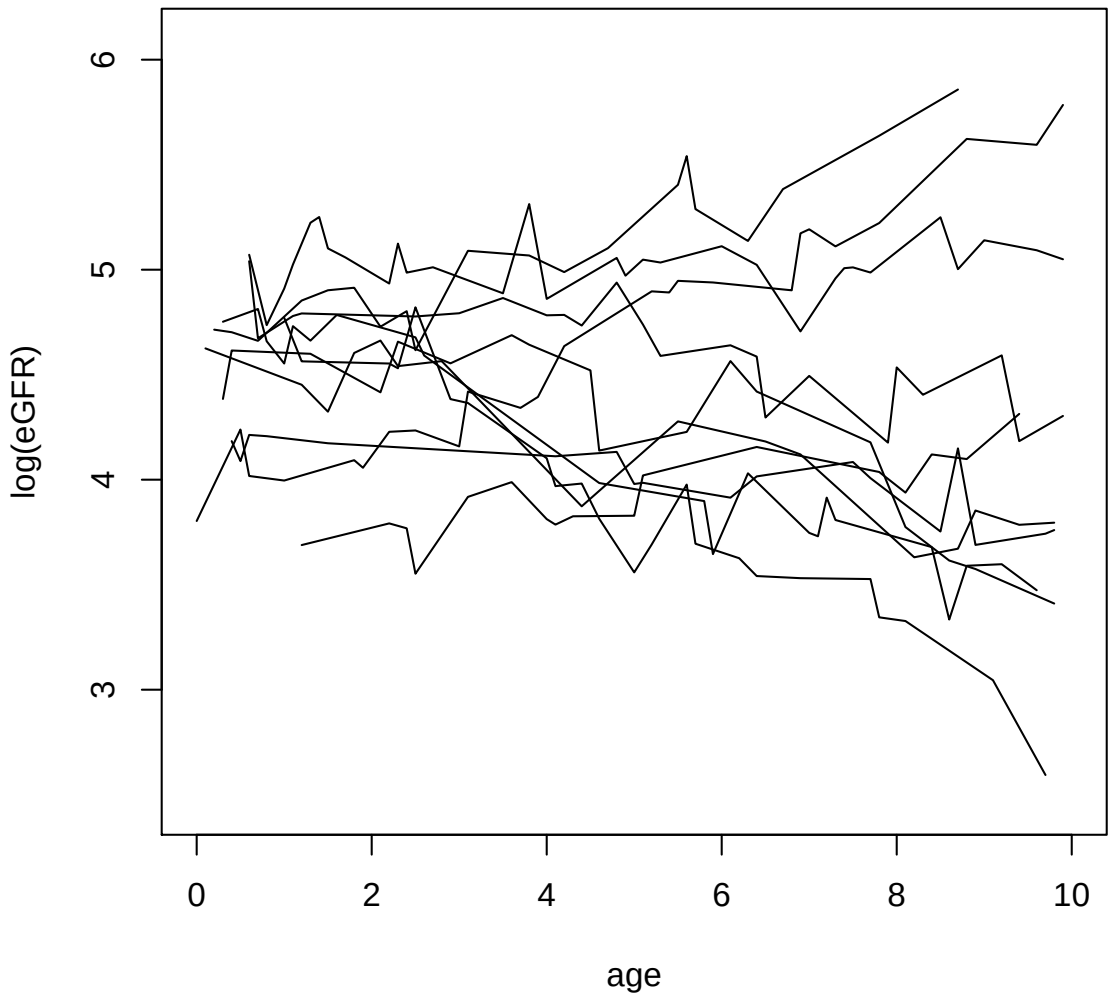
**RE(%)= $100(\exp(\hat{\alpha}) - 1)$ corresponds to estimated annual percentage change in renal function.**

| Parameter | | Estimate | SE | RE(%) |
|---|---|---|---|---|
| $\alpha_0$ | intercept | 4.6006 | 0.0203 | |
| $\alpha_1$ | female | -0.0877 | 0.0048 | -8.4 |
| $\alpha_2$ | age on entry | -0.0048 | 0.0004 | -0.5 |
| $\alpha_3$ | follow-up | -0.0232 | 0.0011 | -2.3 |
| $\alpha_4$ | age>56.5 | -0.0075 | 0.0006 | -0.6 |
| | | | | |
| $\omega^2$ | intercept | 0.1111 | 0.0012 | |
| $\sigma^2$ | signal | 0.0141 | 0.0002 | |
| $\tau^2$ | noise | 0.0469 | 0.0001 | |

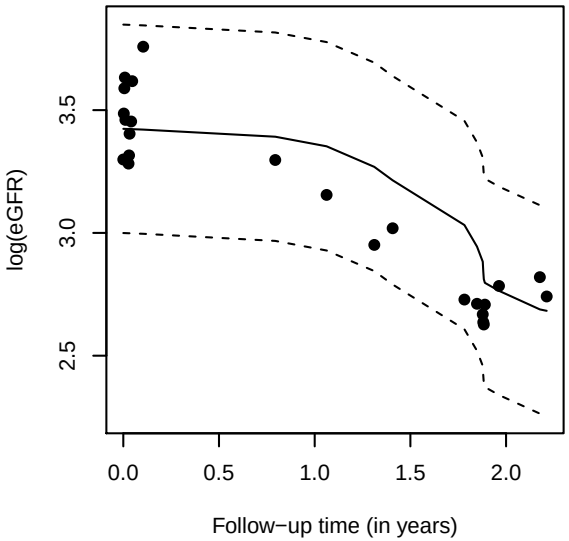
# Sample data-sequences

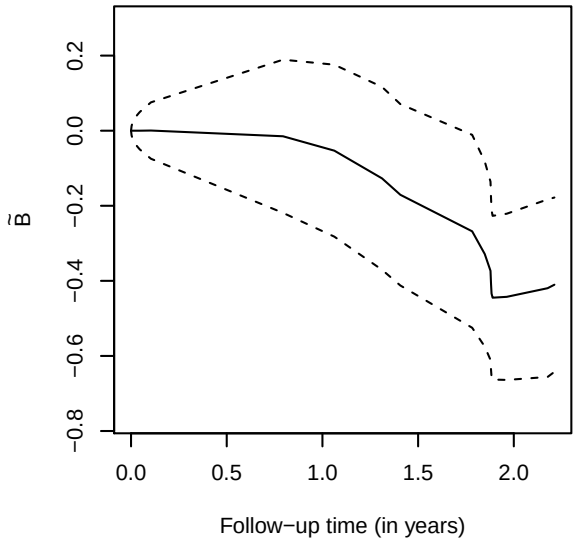


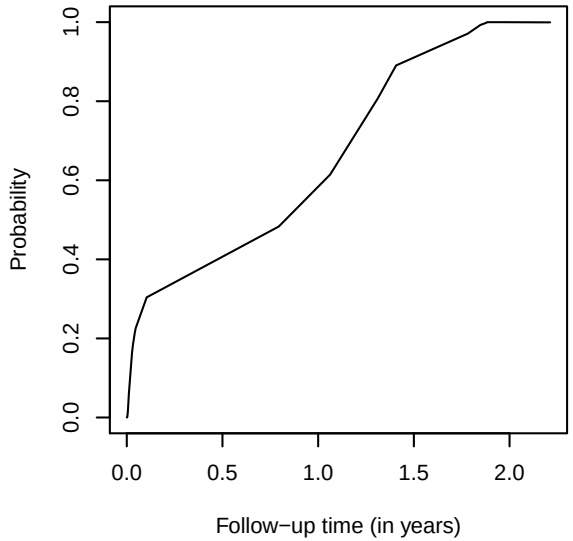
# Simulations



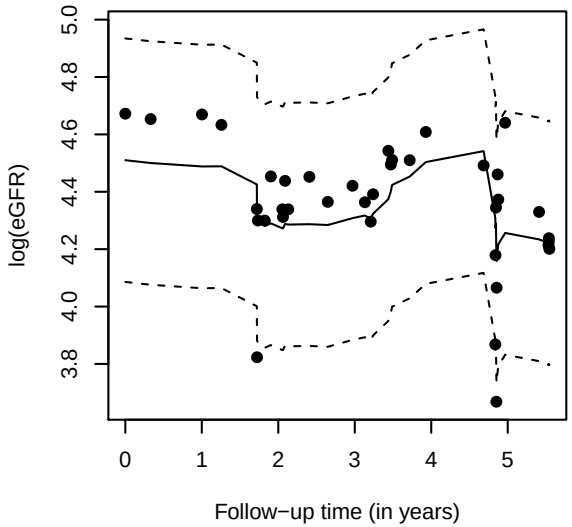
# Prediction: classic progression pattern

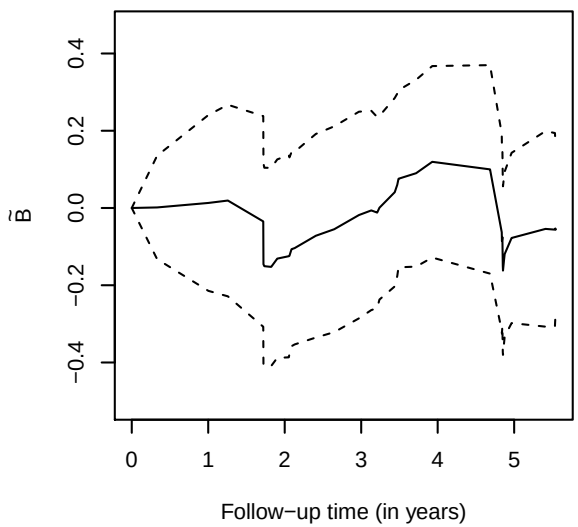


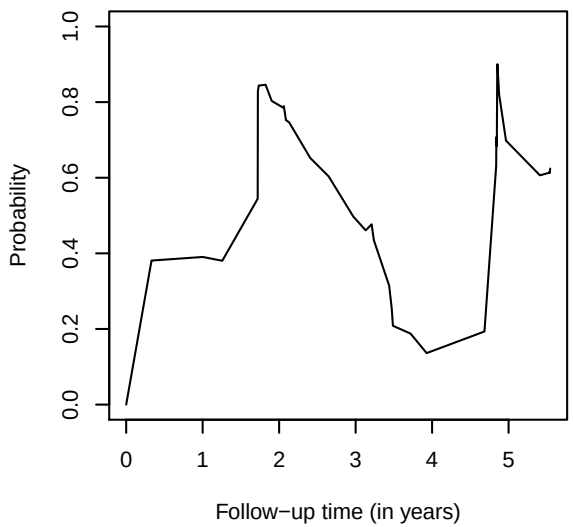



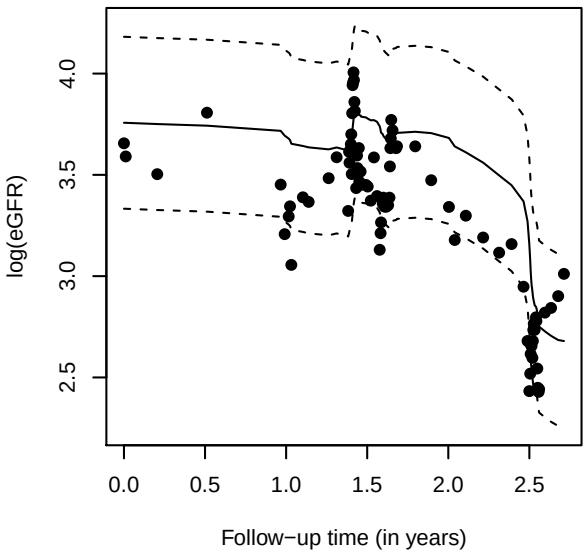
# Prediction: AKI (Acute Kidney Injury) recovery

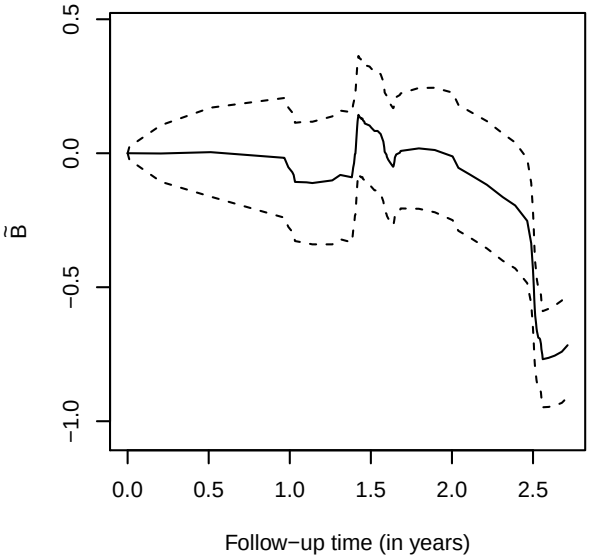


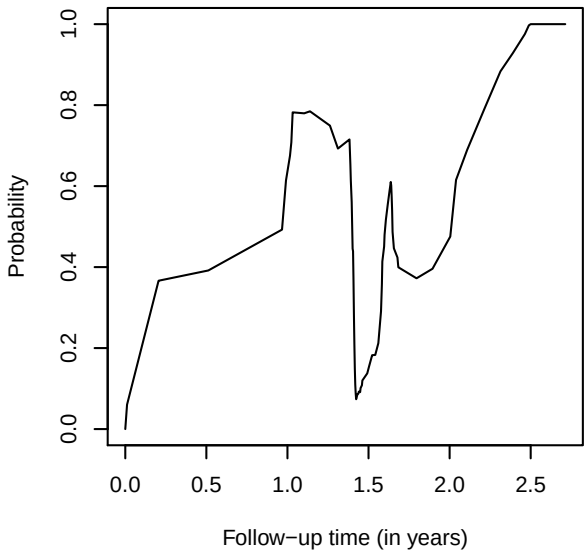



# Prediction: non-recovery from AKI



# Work-in-prospect

**Field-testing:** comparative evaluation against current methods

- eye-balling
- OLS fit to three most recent values

**Informative follow-up:** eGFR more likely to be measured when subject is in poor health

$\Rightarrow$ joint modelling of eGFR measurements and follow-up times

**Implementation:** in clinical practice...needs informatics expertise

# Gastro-intestinal disease

**Reported UK annual incidence**

| | |
|---|---|
| **Campylobacter** | **50,000** |
| **Salmonella** | **10,000** |
| **Cryptosporidium** | **5,000** |
| **Giardia** | **3,000** |
| **E Coli,...** | **?** |

# Real-time spatial surveillance, ca 2003

**AEGISS: 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: Cox process model

$$\begin{aligned} \textbf{actual} &= \textbf{expected} \times \textbf{unexpected} \\ \lambda(x,t) &= \lambda_0(x,t) \times R(x,t) \end{aligned}$$

**Scientific objective**

- use incident data up to time t to construct predictive distribution for current "risk" surface, $R(x,t)$,
- hence identify anomalies, for further investigation.

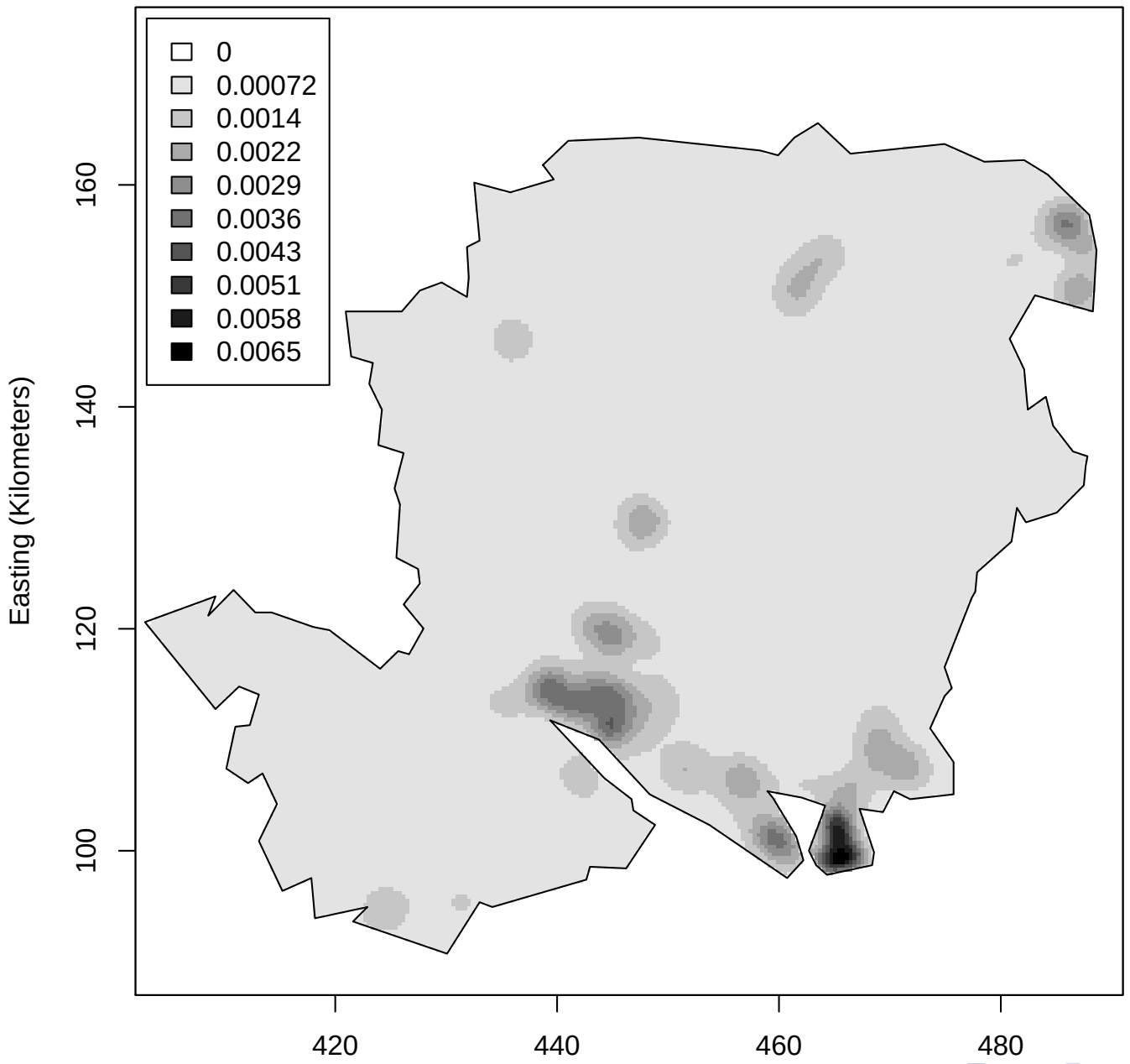
# AEGISS model formulation

$$\lambda(x,t) = \lambda_0(x,t)R(x,t)$$

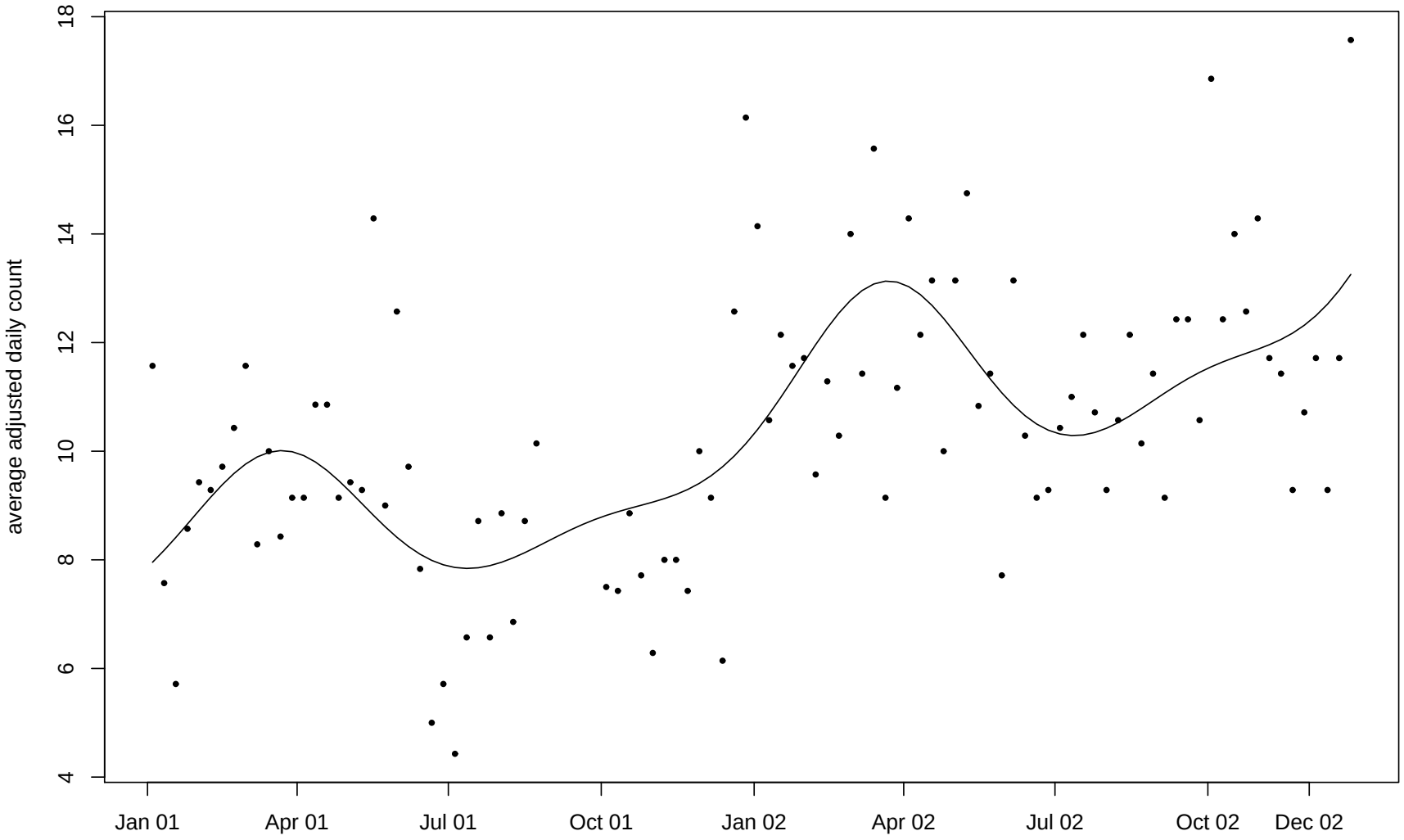
- $\lambda_0(x,t) = \lambda_0(x)\mu_0(t)$
- $R(x,t) = \exp\{S(x,t)\}$
- $S(x,t)$ = spatio-temporal Gaussian process

**Conditional on $R(x,t)$, incident cases form an inhomogeneous Poisson process with intensity $\lambda(x,t)$**

# $\hat{\lambda}_0(\mathbf{x})$: adaptive kernel smoothing

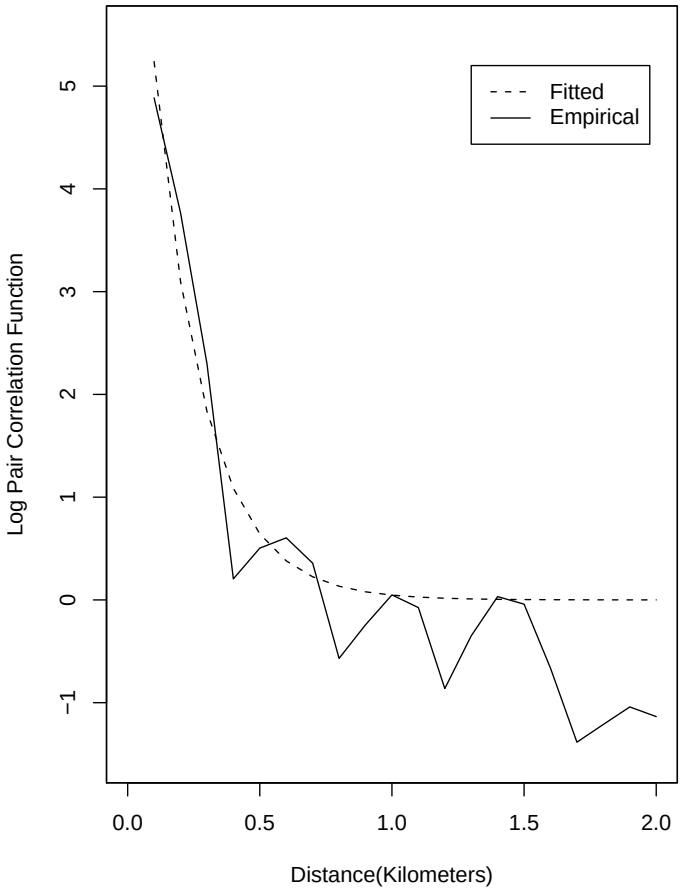


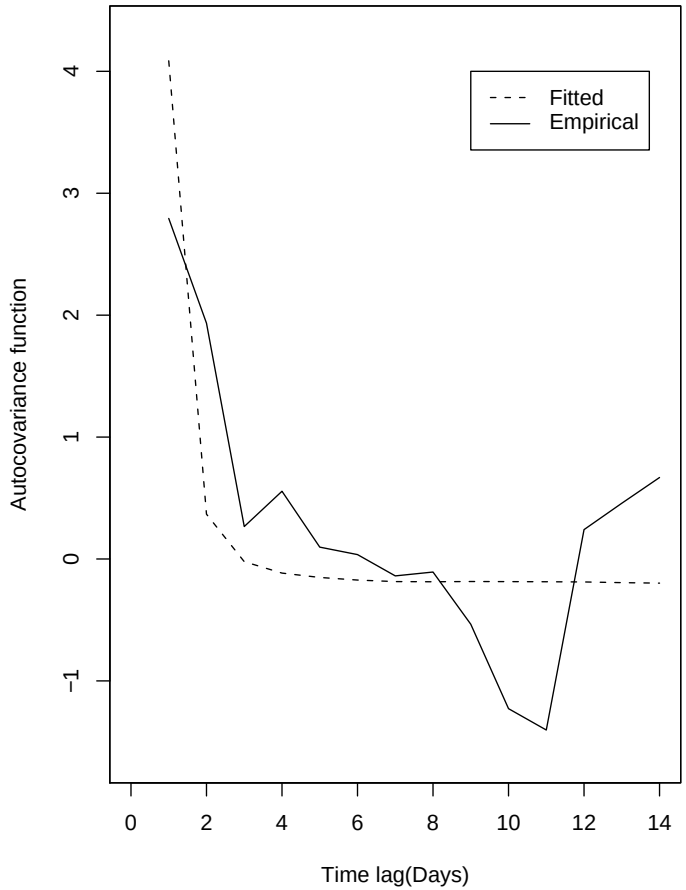
# $\hat{\mu}_0(\mathbf{t})$: Poisson log-linear model



# Spatio-temporal covariance

$$\rho(u, v) = \rho_x(u)\rho_t(v)$$



# Spatial prediction

- **plug-in for estimated model parameters**
- **MCMC to generate samples from conditional distribution of $S(x,t)$ given data up to time $t$**
- **choose critical threshold value $c > 1$**
- **map empirical exceedance probabilities,**

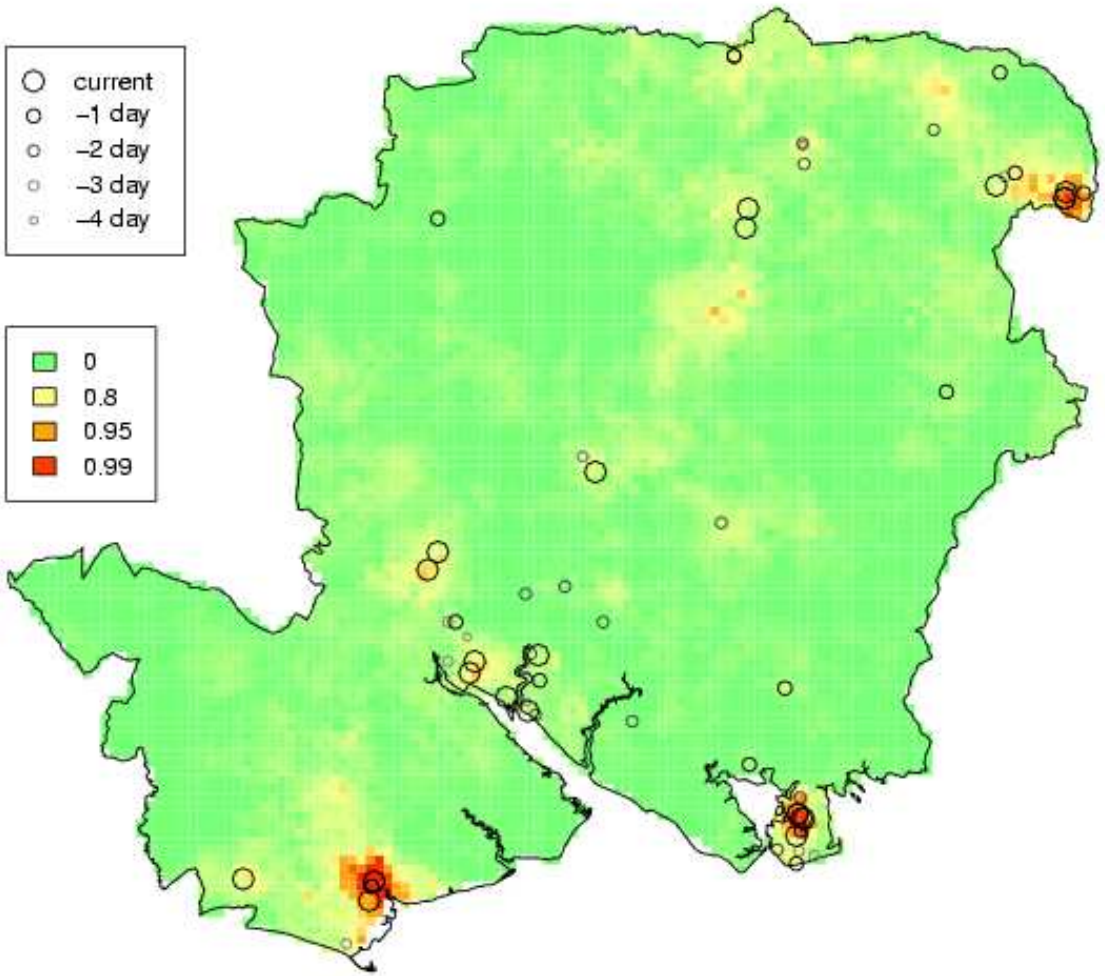
$$p_t(x) = P\left(\exp\{S(x,t)\} > c | \text{data}\right)$$

- **web-based reporting with daily updates**

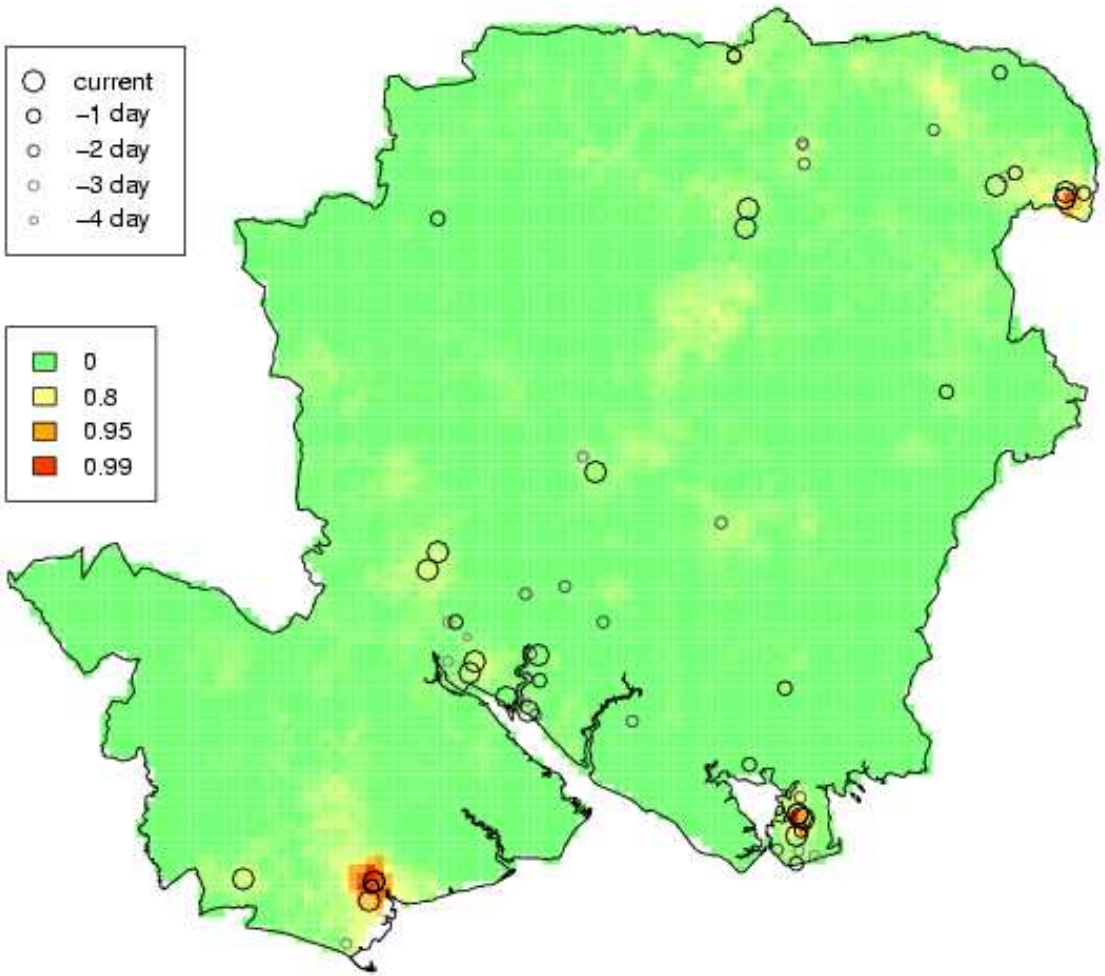
  (`www.lancs.ac.uk/staff/diggle/aegiss/`)

# Spatial prediction: 6 March 2003

$c = 2$



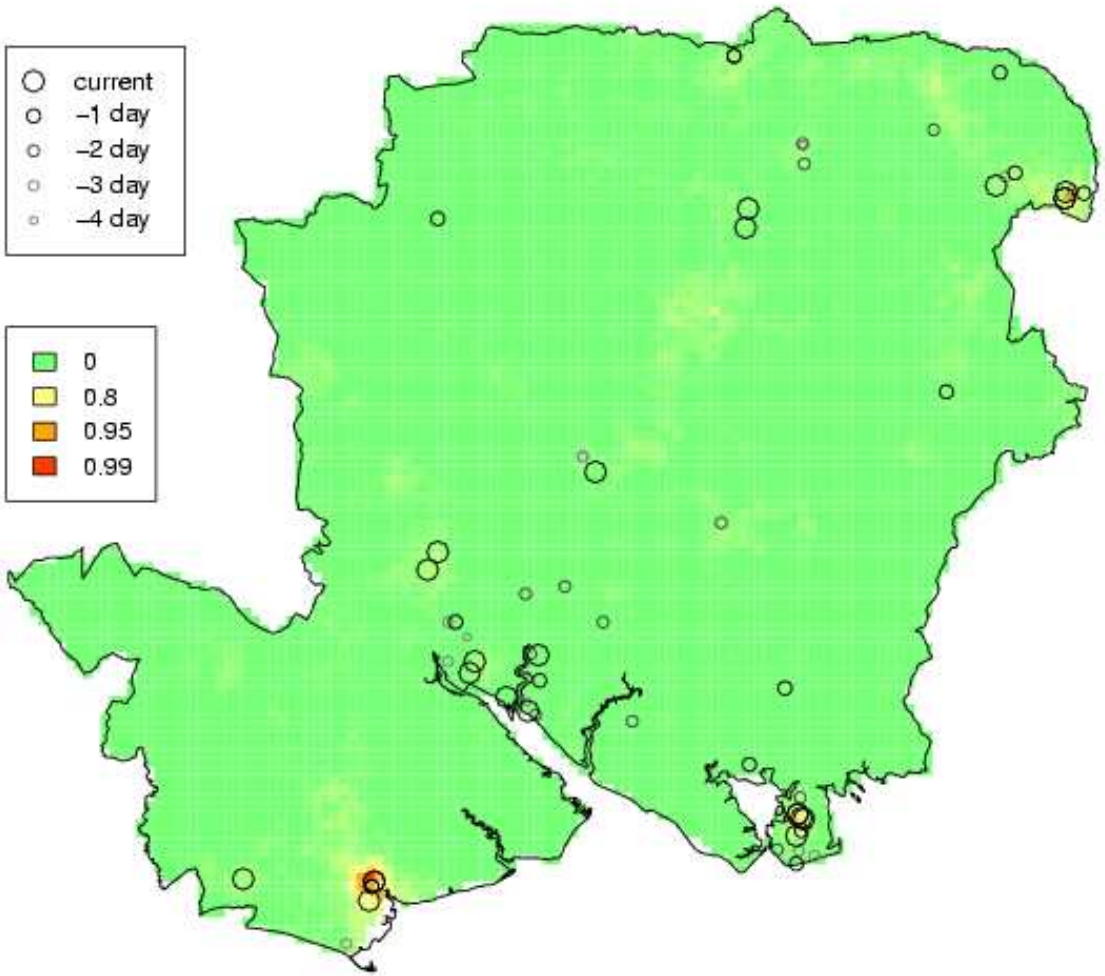
# Spatial prediction: 6 March 2003




$c = 4$

# Spatial prediction: 6 March 2003

$c = 8$



# Fast-forward to 2015

- **expand to national coverage**
- **integrate human and small-animal veterinary surveillance**

**BUT...**

- **replacement of single NHS Direct by multiple NHS111 services**
- **full post-code data no longer available!**

**SAVSNET: real-time data-feed from network of small-animal vet practices:**

- **practice location**
- **species (cat or dog)**
- **diagnosis**

`http://www.savsnet.co.uk/realtimedata/`

# Work-in-prospect

- **re-calibration of AEGISS model**
- **coarser spatial resolution...fitting spatially continuous models to spatially discrete data**
- **joint modelling of human and animal incidence**
- **implementation as part of routine surveillance systems**

# Malaria prevalence mapping

# Prevalence mapping 1

**Single prevalence survey**

**Sample n individuals, observe Y positives**

$$Y \sim \mathrm{Bin}(n, p)$$

**Multiple prevalence surveys**

**Sample $n_i$ individuals, observe $Y_i$ positives, $i = 1, \ldots, m$**

$$Y_i \sim \mathrm{Bin}(n_i, p_i) \ ?$$

# Prevalence mapping 2

**Extra-binomial variation**

**Sample $n_i$ individuals, observe $Y_i$ positives, $i = 1, \ldots, m$**

$$Y_i | d_i, U_i \sim \mathrm{Bin}(n_i, p_i) \quad \log\{p_i/(1-p_i)\} = d_i'\beta + U_i$$

**Question: What to do if the $d_i$ and/or the $U_i$ are spatially structured**

# Geostatistical model for prevalence data

- **Latent spatially correlated process**

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

- **Latent spatially independent random effects**

  $U_i \sim \mathrm{iidN}(0, \nu^2)$

- **Linear predictor (regression model)**

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

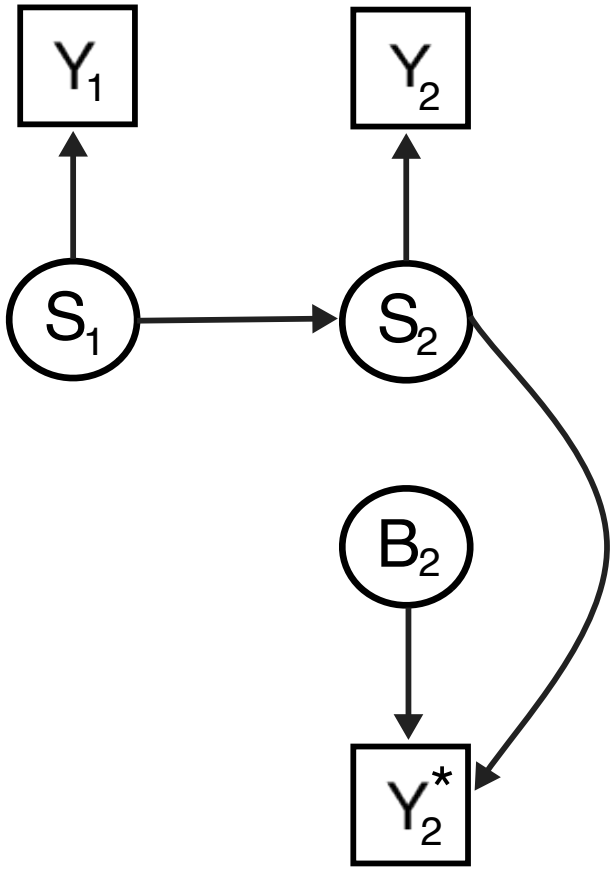
- **Conditional distribution for positive proportion $Y_i/n_i$**

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

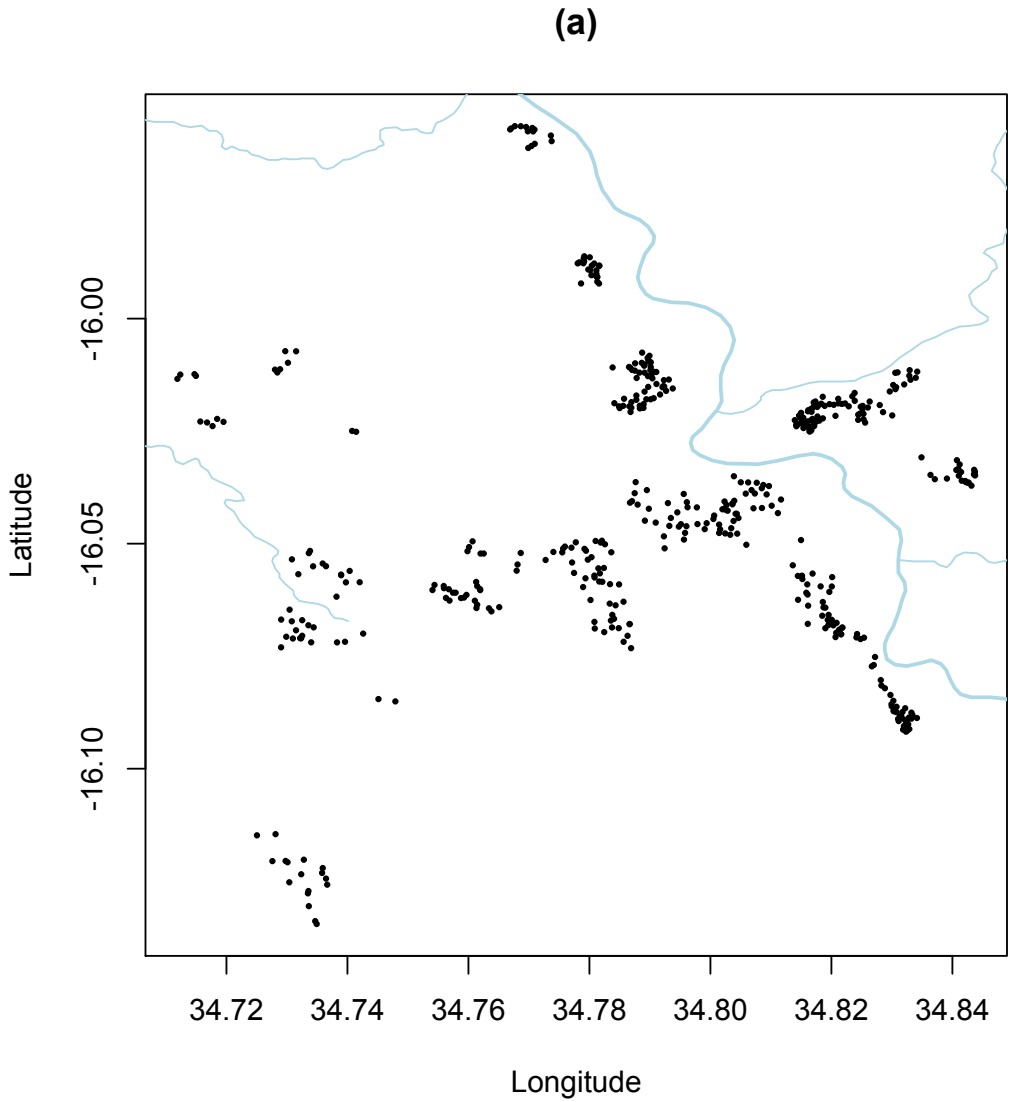
# Multiple surveys (Giorgi et al, 2015)

**Surveys:** $i = 1, \ldots, r$ **locations** $x_{ij} : j = 1, \ldots, n_i$

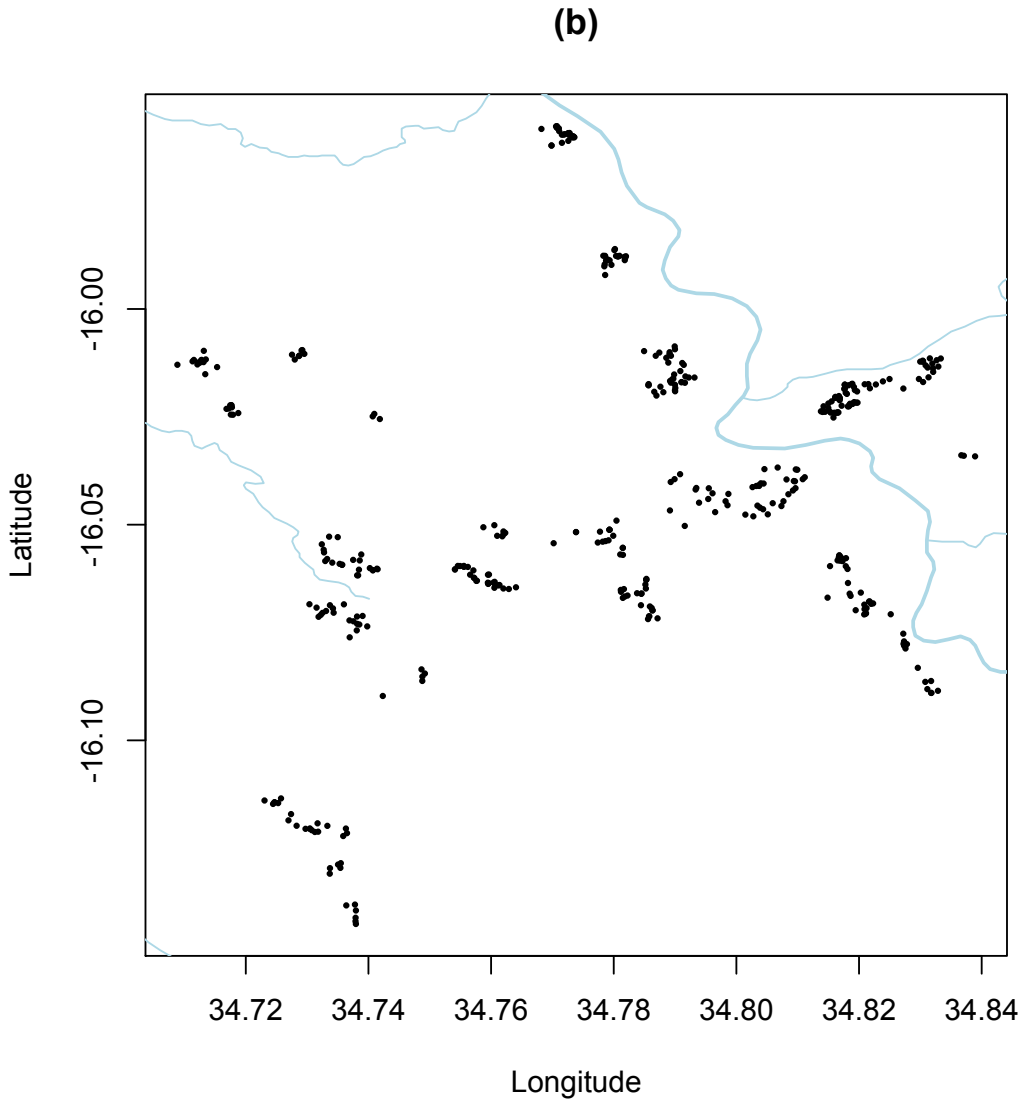
$$\eta_{ij} = d(x_{ij})^\top \beta_1 + S_i(x_{ij}) + I(i \in \mathcal{B})[B_i(x_{ij}) + d(x_{ij})'\beta_i] + U_{ij}$$



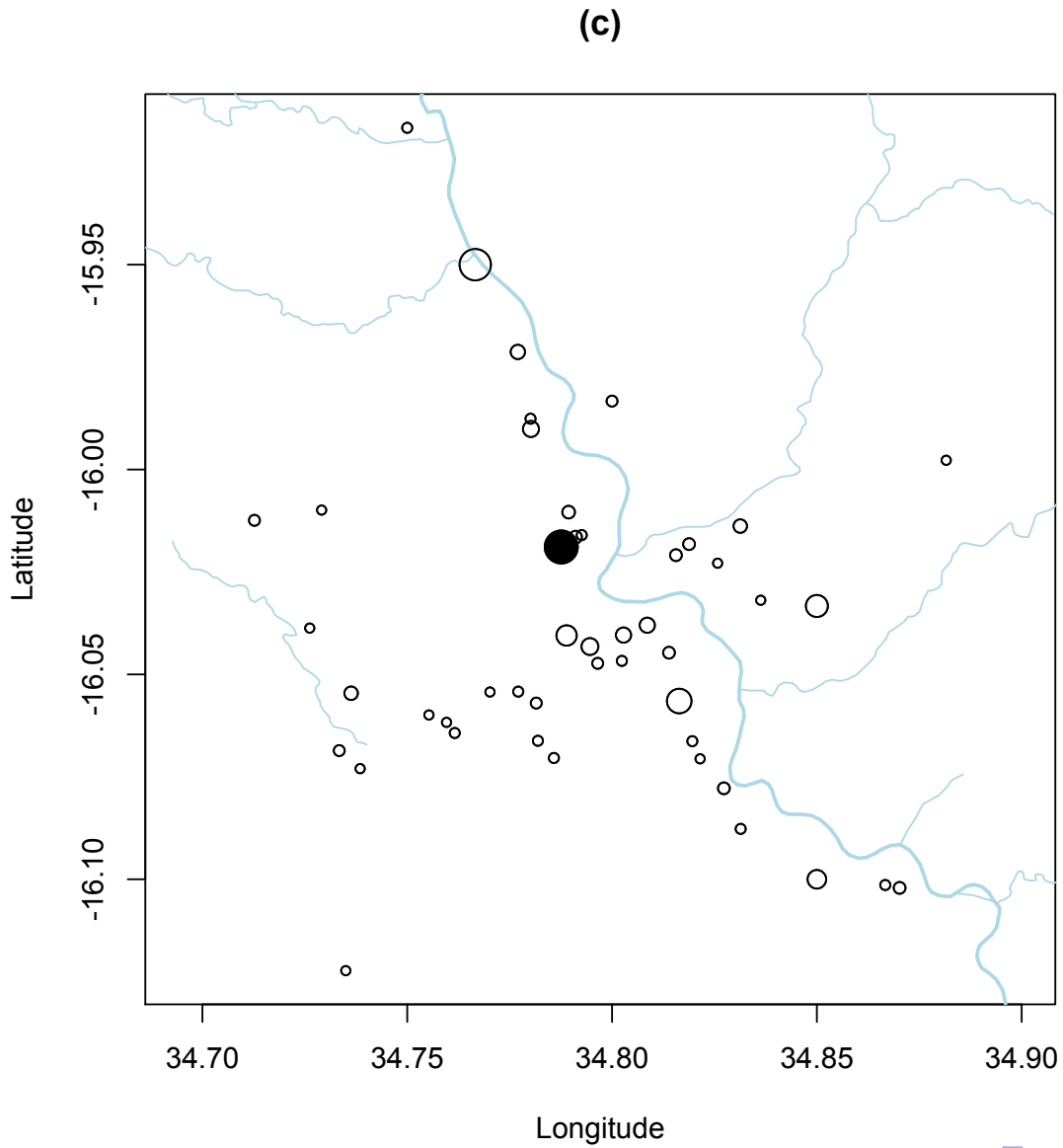
# Malaria mapping, Chikhwawa district, Malawi (Giorgi et al, 2015): rMIS individual locations



# Malaria mapping, Chikhwawa district, Malawi (Giorgi et al, 2015): eMIS individual locations

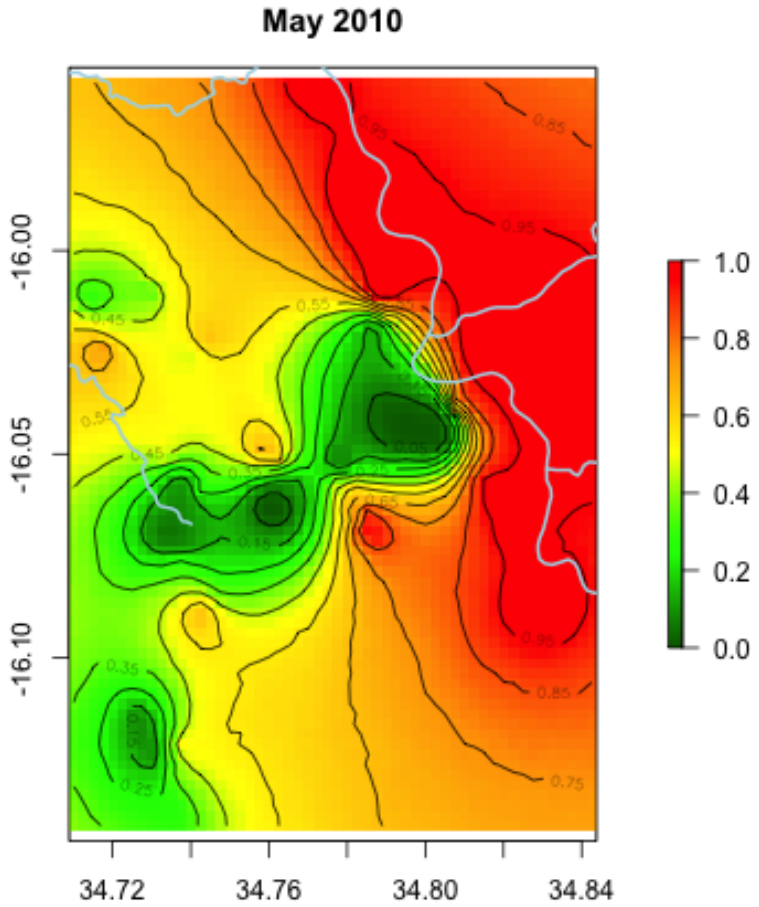


# Malaria mapping, Chikhwawa district, Malawi (Giorgi et al, 2015): EAG village locations and prevalences



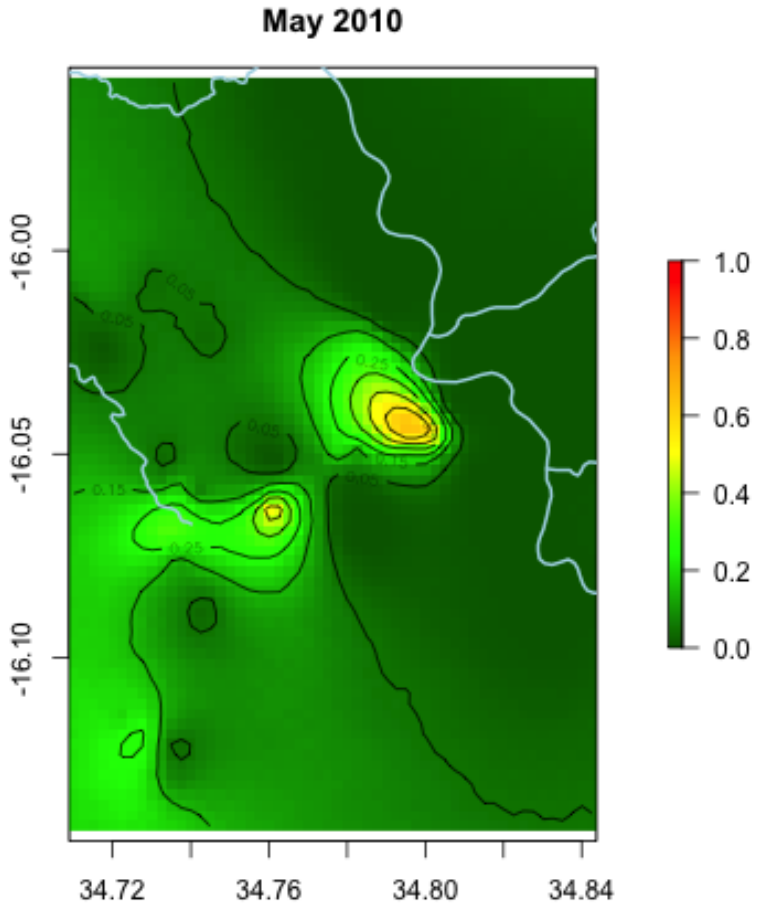
# Continuous time: rolling malaria indicator surveys

**Hotspots: P(prevalence $>$ 20%)**



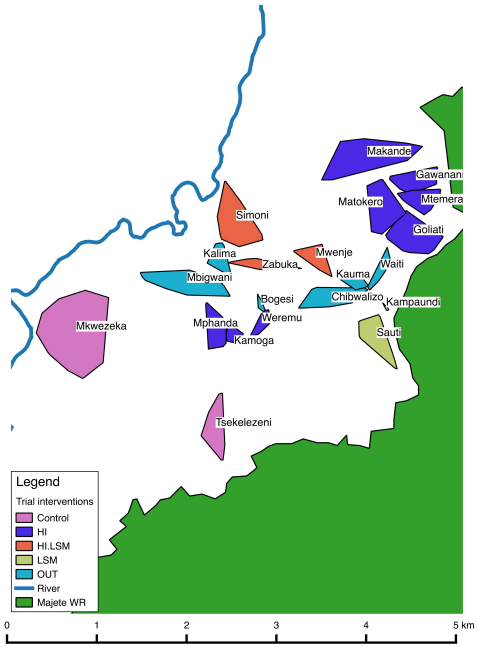
# Continuous time: rolling malaria indicator surveys

**Coldspots: P(prevalence $< 5\%$)**



# Work-in-prospect: Majete national park project, Malawi

- **adaptive design strategies**
- **embedded RCT of community-level interventions**



# Closing remarks

- **Operational issues**
  - **predictive probability of exceedance over intervention threshold**
  - **to inform, but not to over-ride, clinical judgement**
- **Methodological issues:**
  - **observational studies vs trials**
  - **long series with irregular follow-up times**
  - **informative follow-up...marked point process models**

# References

## Monitoring renal function

Diggle, P.J., Sousa, I. and Asar, Ö. (2014). Real-time monitoring of progression towards renal failure in primary care patients. (submitted)

## Gastro-enteric surveillance

Diggle, P.J., Rowlingson, B. and Su, T-L. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, **16**, 423–34.

## Malaria prevalence mapping

Giorgi, E., Sesay, S.S., Terlouw, D.J. and Diggle, P.J. (2015). Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *Journal of the Royal Statistical Society* A **178**, 445–464.