

Statistical Modelling for Real-time Epidemiology

Peter J Diggle

*School of Health and Medicine, Lancaster University
and
Department of Biostatistics, Johns Hopkins University*

With help from:

Ciprian Crainiceanu, Barry Rowlingson, Ines Sousa, Alex Rodrigues

March 2009

Outline

- real-time epidemiology
- motivating examples:
 - syndromic surveillance for gastroenteric illness
 - tropical disease prevalence mapping
 - detecting incipient renal failure
- spatio-temporal stochastic processes
 - real-valued processes $S(x, t)$
 - separability of spatio-temporal correlation
- examples re-visited
- outlook

Real-time epidemiology

Epidemiology: the study of health outcomes in their natural setting

- not just epidemics
- also social outcomes

Real-time epidemiology: analysis of health outcome data as they accrue

- time-scale is context-specific
- often concerned also with spatial variation

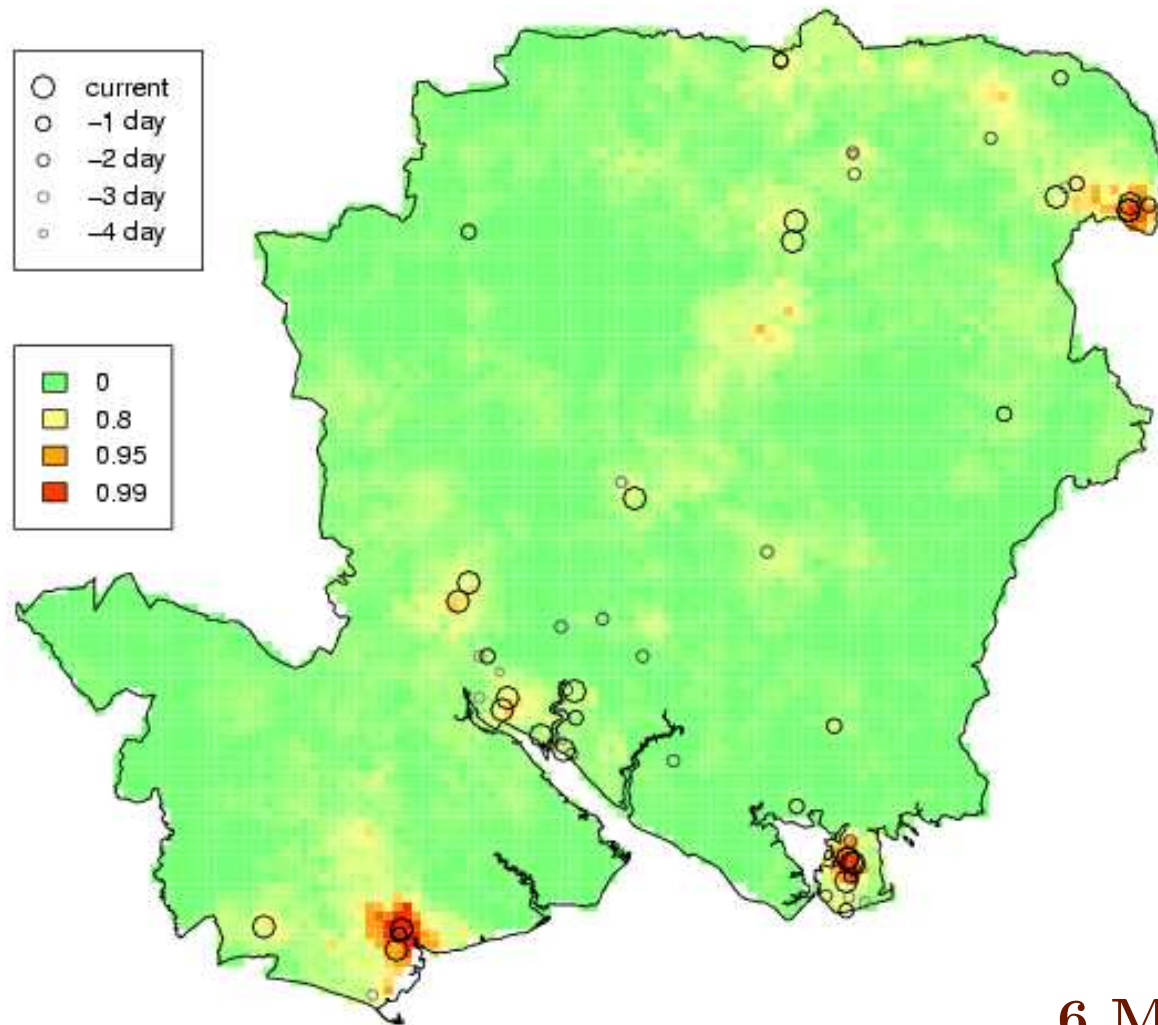
Epidemic vs endemic patterns of incidence

Two animations:

- foot-and-mouth in Cumbria, UK
- campylobacteriosis in Preston, Lancashire, UK

<http://www.lancaster.ac.uk/staff/diggle/>

Gastroenteric illness in Hampshire, UK



6 March, 2003

Loa loa in equatorial Africa

Observed prevalence of loa loa (IRD-TDR)

- 0 - 5%
- 5 - 10%
- 10 - 15%
- 15 - 20%
- >20%

Probability of [high risk]

- 0.95 - 1
- 0.9 - 0.95
- 0.8 - 0.9
- 0.7 - 0.8
- 0.6 - 0.7
- 0.5 - 0.6
- 0.4 - 0.5
- 0.3 - 0.4
- 0.2 - 0.3
- 0.1 - 0.2
- 0.05 - 0.1
- 0 - 0.05
- No Data

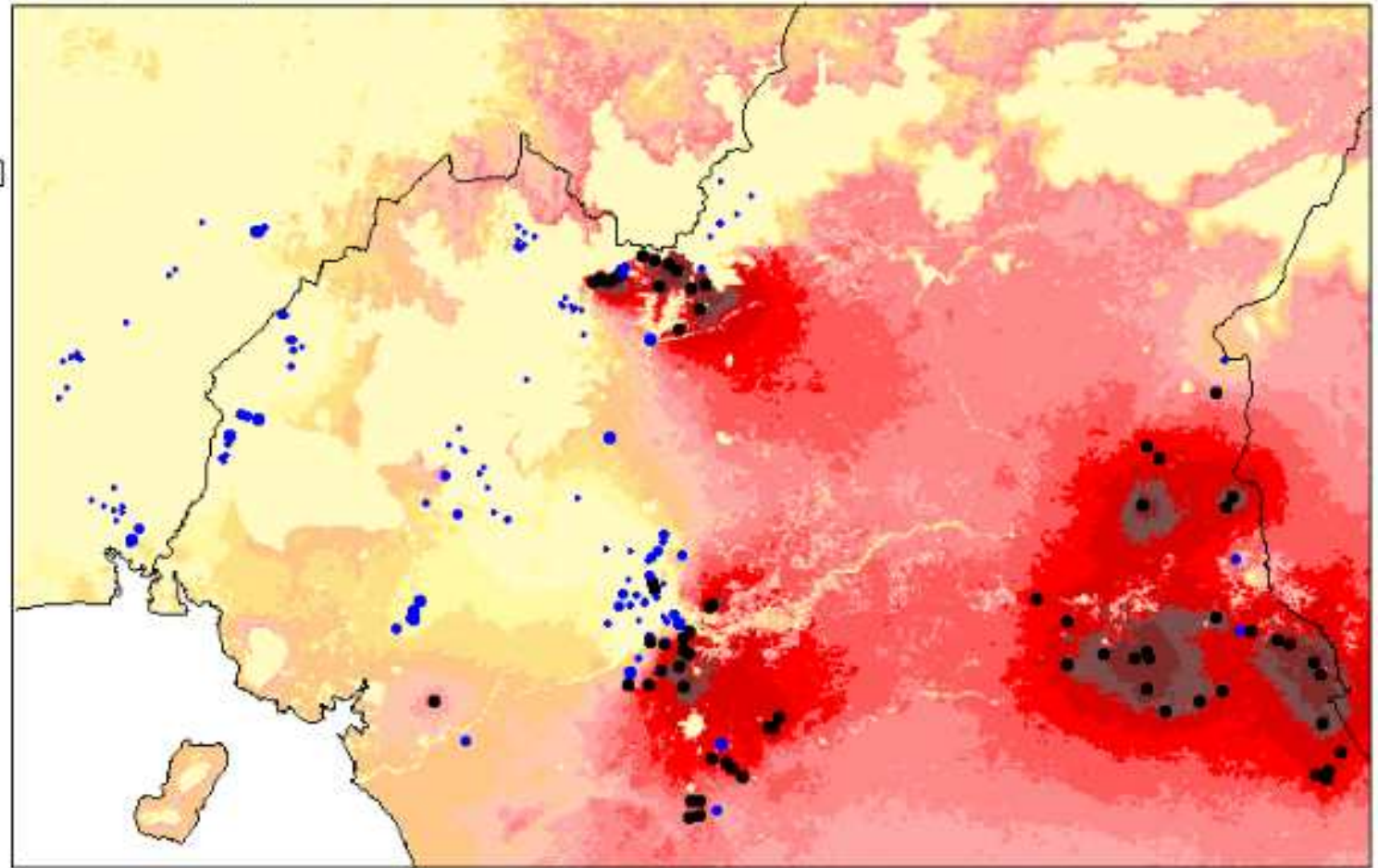
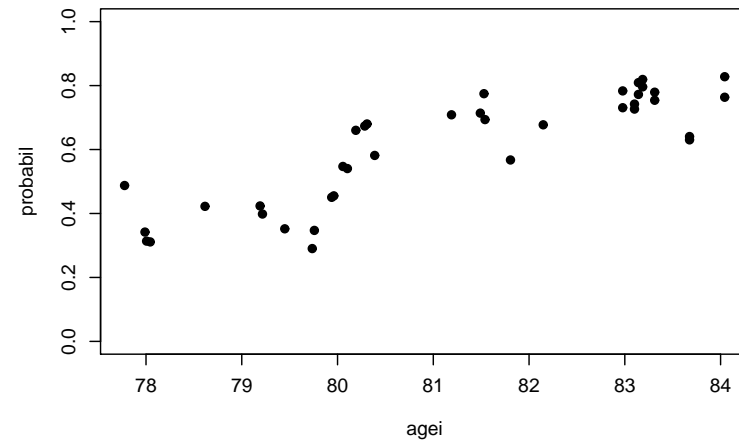
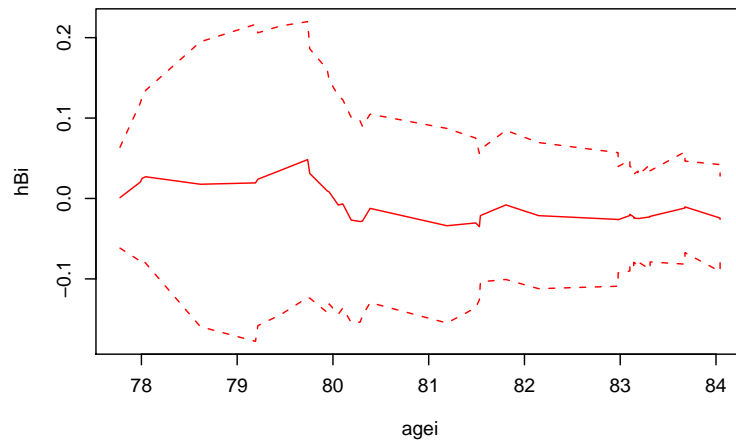
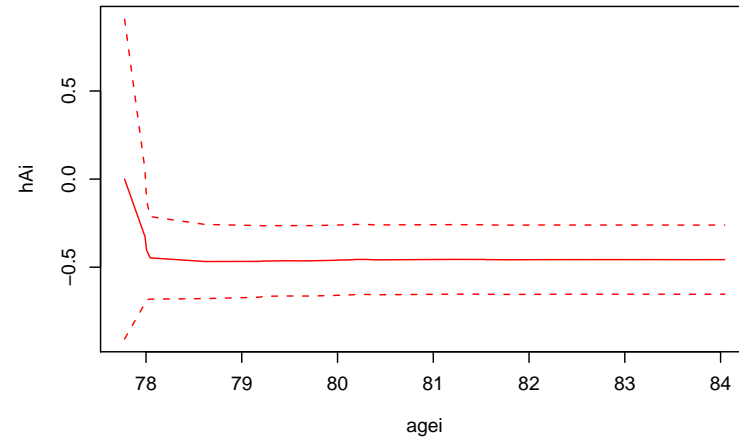
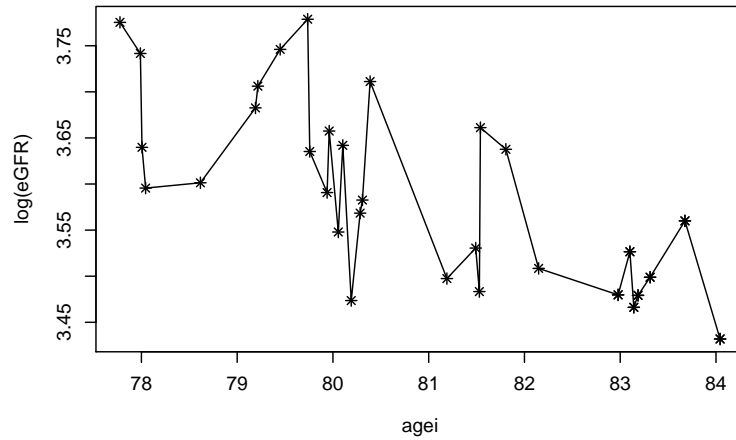


Figure 6: PCM for [high risk] in Cameroon based on ERM with ground truth data.

Incipient renal failure in Salford, UK



Spatio-temporal stochastic processes

Real-valued stochastic process $\{S(x, t) : x \in \mathbb{R}^2; t \in \mathbb{R}^+\}$

- $\mu(x, t) = \mathbb{E}[S(x, t)]$ (mean function, or trend)
- $\gamma(x, x', t, t') = \text{Cov}\{S(x, t), S(x', t')\}$ (covariance function)

Stationarity:

- mean stationary: $\mu(x, t) = \mu$
- covariance stationary: $\gamma(x, x', t, t') = \gamma(u, v)$
($u = \|x - x'\|, v = |t - t'|$)

Gaussian process:

- $\{S(x_i, t_i) : i = 1, \dots, m\} \sim$ multivariate Gaussian

Modelling covariance structure

- positive-definiteness requirement imposes non-obvious constraints
- spectral representation easier, but maybe less intuitive
- some general considerations:
 - separability: $\gamma(u, v) = \sigma^2 \rho(u, v) = \sigma^2 \rho_x(u) \rho_t(v)$
 - standard families for $\rho_x(\cdot), \rho_t(\cdot)$
 - constructive definitions using convolutions
 - low-rank approximations to ease computation

Separability

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

Conditioning on the past:

- $S_t = \{S(x, t) : x \in \mathbb{R}^2\}$
- model S_t conditional on $\{S_v : v < t\}$
- add Markov assumption,

$$[S_t | \{S_v : v < t\}] = [S_t | S_{t-1}]$$

Separability implies

$$[S(x, t) | \{S(u, t-1) : u \in \mathbb{R}^2\}] = [S(x, t) | S(x, t-1)]$$

Standard families

Matérn family widely used in purely spatial case

$$\rho(u) = 2^{\kappa-1} (u/\phi)^\kappa K_\kappa(u/\phi)$$

- shape parameter $\kappa > 0$, scale parameter $\phi > 0$
- $K_\kappa(\cdot)$: modified Bessel function of order κ
 - integer part of κ determines mean-square differentiability of $S(\cdot)$
 - but estimation of κ is difficult
- no obvious non-separable spatio-temporal extension

Spatial convolution-based models

$$S(x) = \int k(x - u)B(u)du$$

- $\{B(u) : u \in \mathbb{R}^2\}$ spatially continuous white noise
- kernel function $k(v)$, such that

$$\int k(v)dv = 1 \quad \int k(v)^2 v dv < \infty$$

Covariance structure of $S(\cdot)$ given by

$$\text{Cov}\{S(x), S(x - u)\} \propto \int k(u - v)k(v)dv$$

Low-rank approximations

$$S(x) = \sum_{i=1}^m k(x - s_i) B_i, \quad x \in A,$$

- $B = (B_1, \dots, B_m) : i = 1, \dots, m$ (coefficients)
- $s_i \in A : i = 1, \dots, m$ pre-specified locations (knots)
- $k(\cdot) : i = 1, \dots, m$ real-valued function (kernel)

Practical simplification (stationary case):

- white noise coefficients
- knots to form a regular lattice

Extension to spatio-temporal setting

$$\gamma(u, v) = \text{Cov}\{S(x, t), S(x', t')\} \propto \int \int k(r-u, s-v)k(u, v)drds$$

- for separable correlation, specify $k(u, v) = k_1(u)k_2(v)$
- positive and negative non-separability:
 - positively non-separable if $\rho(u, v) > \rho(u, 0)\rho(0, v)$
 - negatively non-separable if $\rho(u, v) < \rho(u, 0)\rho(0, v)$

A non-separable kernel

$$k(u, v) = \rho(v)^c \exp\{-\rho(v)^\beta (u/\tau)^2\}$$

Requires $\beta \leq 1$, $c > 0$, $\rho(\cdot)$ any valid correlation function

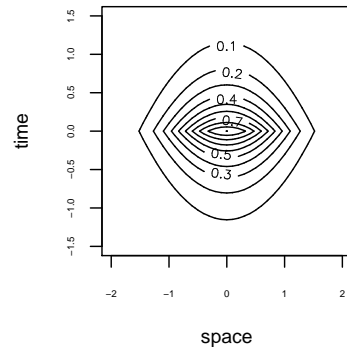
- negatively non-separable when $\beta < 0$
- separable when $\beta = 0$
- positively non-separable when $\beta > 0$

Use as

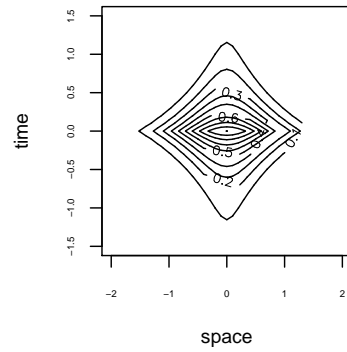
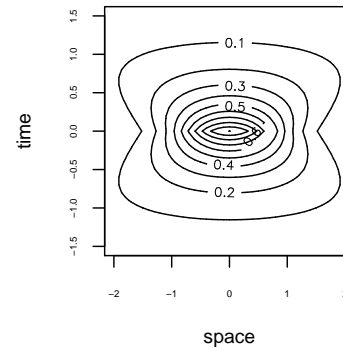
- extant model for measured (geostatistical) data
- latent model for point process data

The space-time kernel: $c = 1$

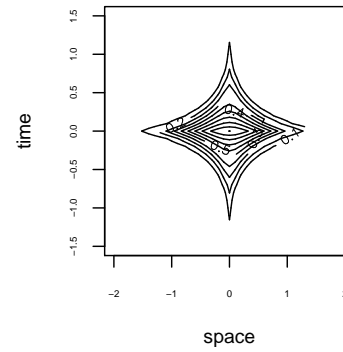
$$\beta = 0$$



$$\beta = 1$$



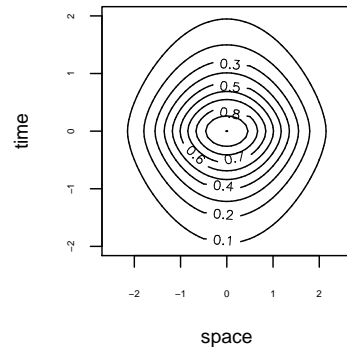
$$\beta = -1$$



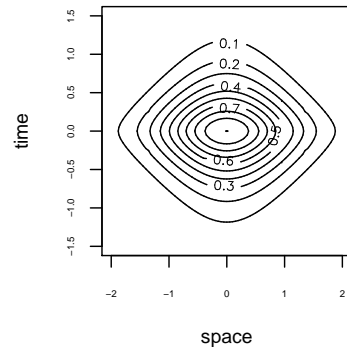
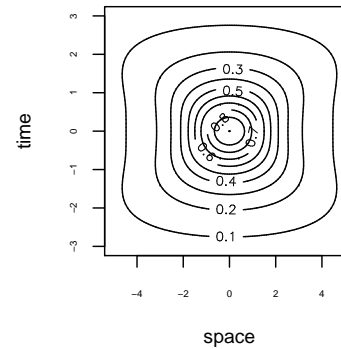
$$\beta = -3$$

The induced covariance structure: $c = 1$

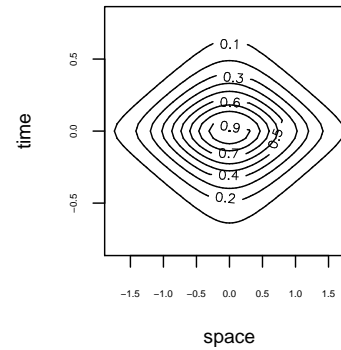
$$\beta = 0$$



$$\beta = 1$$



$$\beta = -1$$



$$\beta = -3$$

AEGISS

Problems with current surveillance system in UK include

- under-reporting by general practitioners (GP's)
- inconsistencies in reporting rates between GP's
- delays between onset and confirmation of cases

The AEGISS project: can spatio-temporal statistical modelling improve the early detection of anomalies in the observed pattern of incident cases?

Possible data-sources

GP-reported data

- GP locations form a discrete spatial network
- known number of patients registered with each GP
- possibility of individual follow-up
- problems with inconsistency of GP reporting

NHS Direct data

- a relatively new, 24hr phone-in advisory service
- date and location (post-code) recorded for each call
- but spatial and temporal pattern of usage is unknown
- data on approx 10,000 incident cases, 2001 to 2003

Statistical formulation

Notation

$\lambda_0(x, t) =$ expected intensity of incident cases

$\lambda(x, t) =$ actual intensity of incident cases

$R(x, t) =$ spatio-temporal variation from normal pattern

$$\begin{aligned}\lambda(x, t) &= \lambda_0(x, t)R(x, t) \\ &= \lambda_0(x)\mu_0(t)\exp\{S(x, t)\}\end{aligned}$$

$S(x, t)$ modelled as stationary Gaussian process

Scientific objective

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

Diggle et al (2003), Diggle, Rowlingson and Su (2004)

Spatial prediction

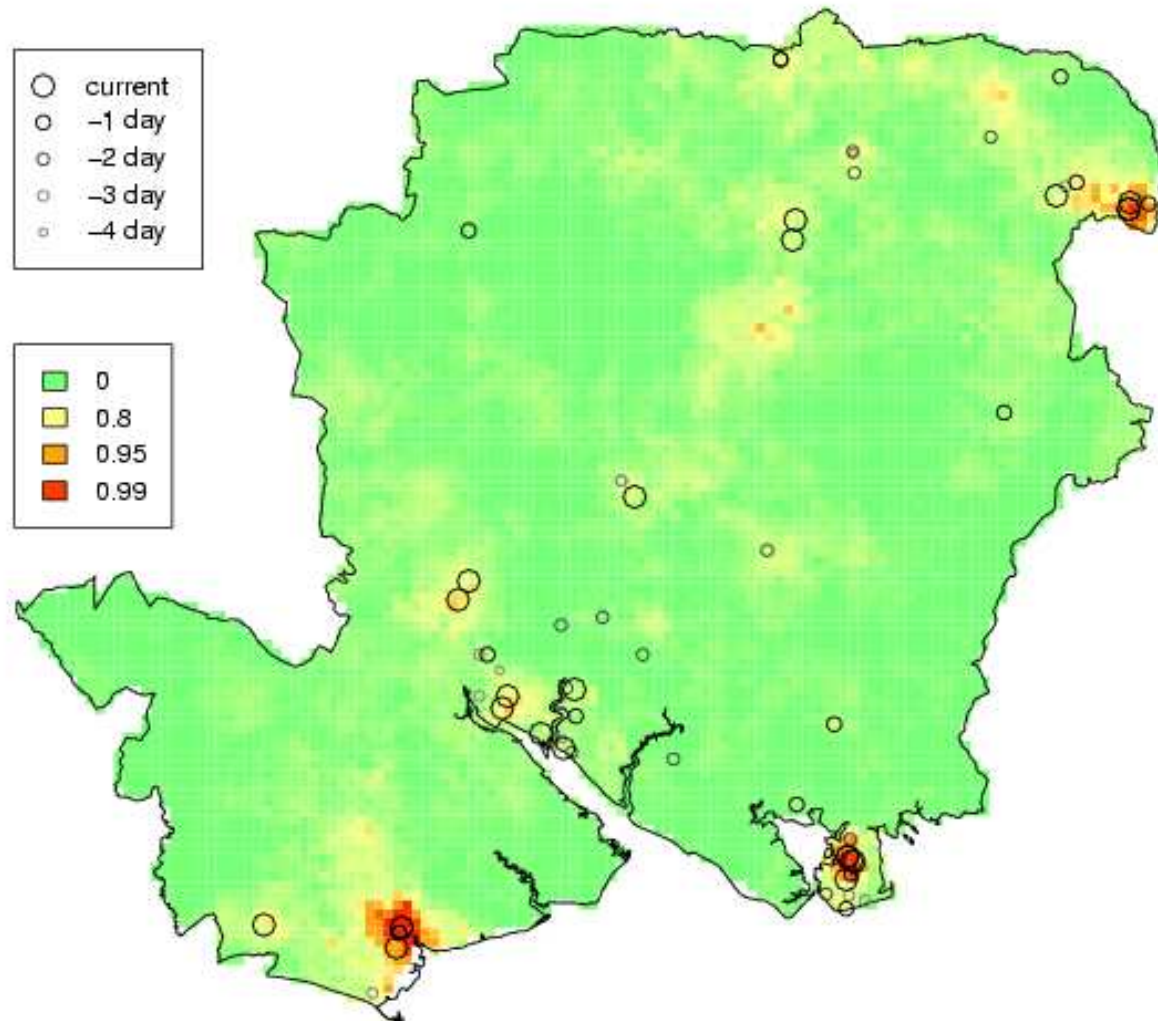
- plug-in for estimated model parameters
- MCMC to generate samples from conditional distribution of $S(x, t)$ given data up to time t
- choose critical threshold value $c > 1$
- map empirical exceedance probabilities,

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

- web-reporting with daily updates, demo at:

<http://www.lancaster.ac.uk/staff/diggle/>

Spatial prediction : results for 6 March 2003

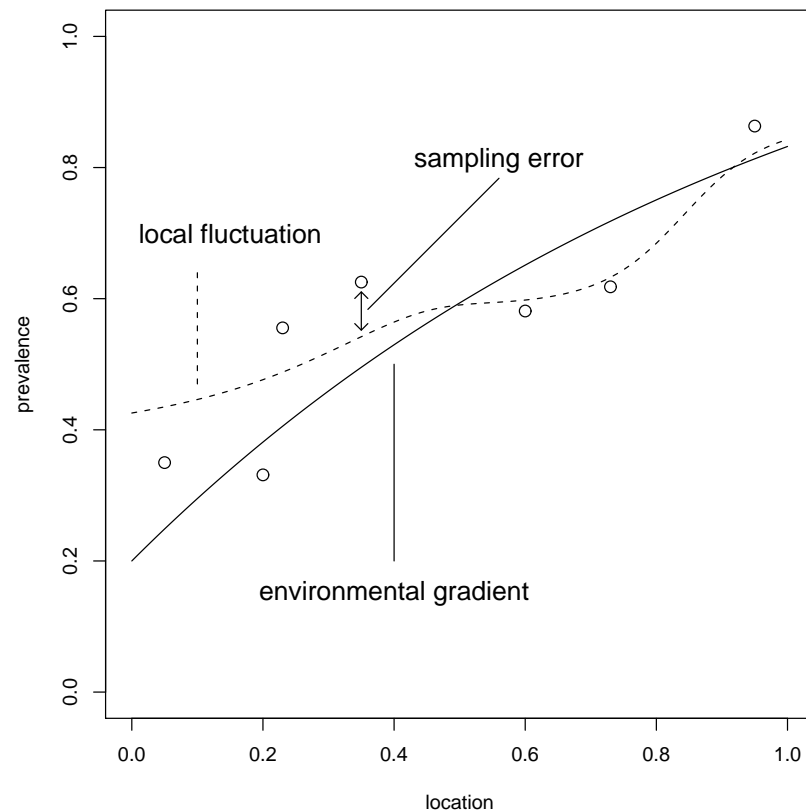


$c = 2$

Loa loa

WHO policy requirement: identify areas where prevalence exceeds 20%

Model:



Loa loa result

Predictive map: $P(\text{prevalence} > 0.2 | \text{data})$

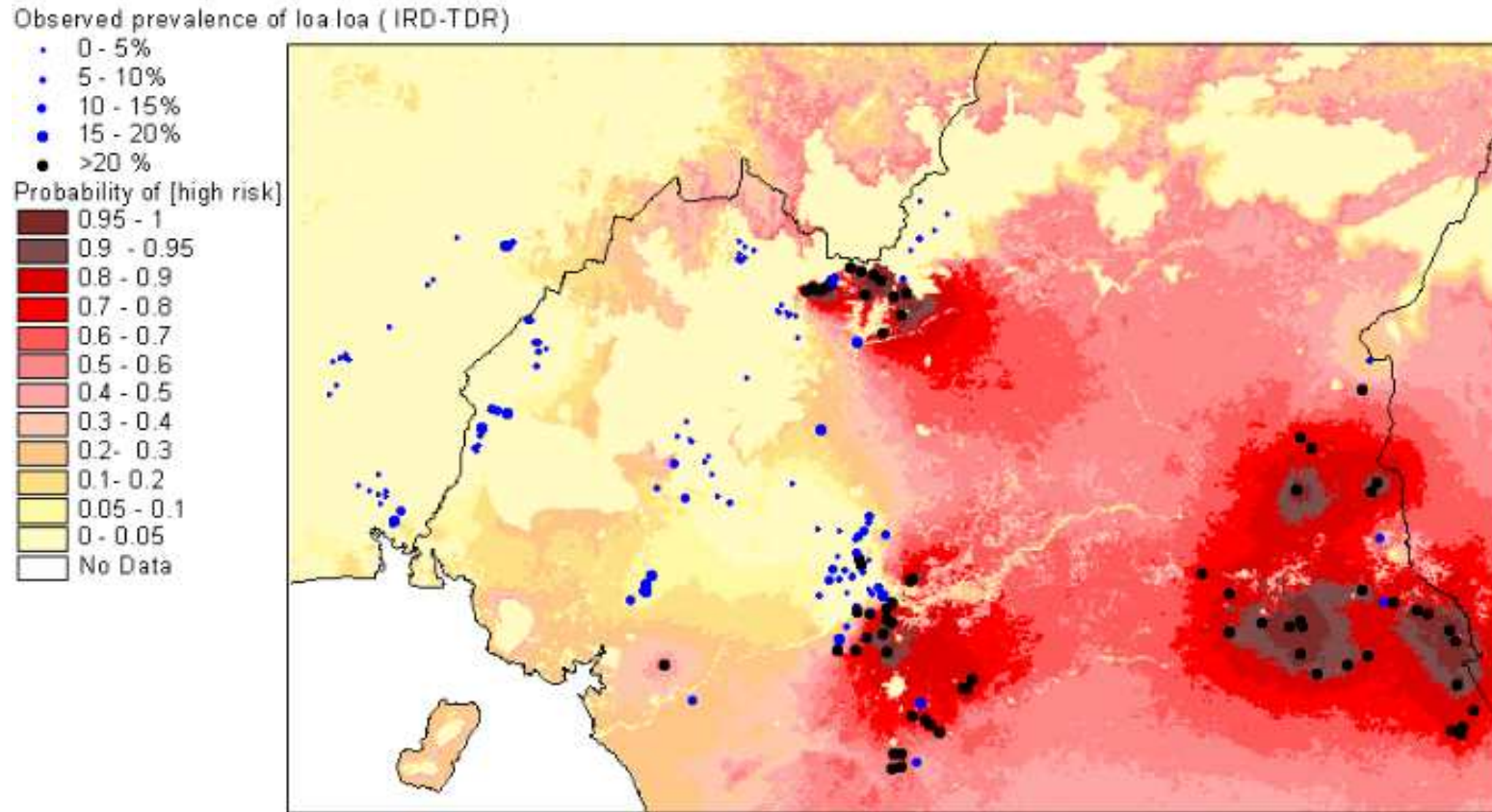
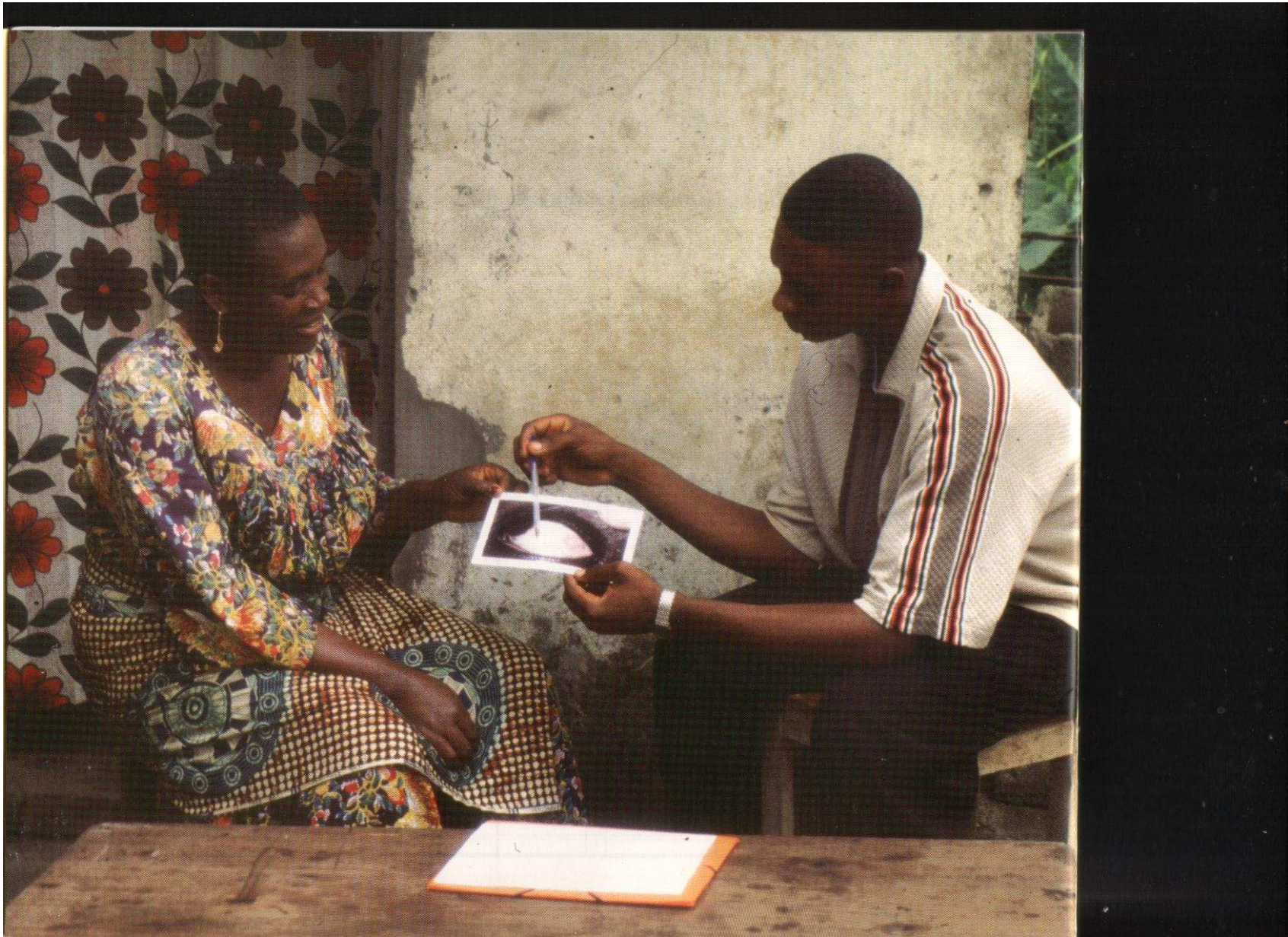


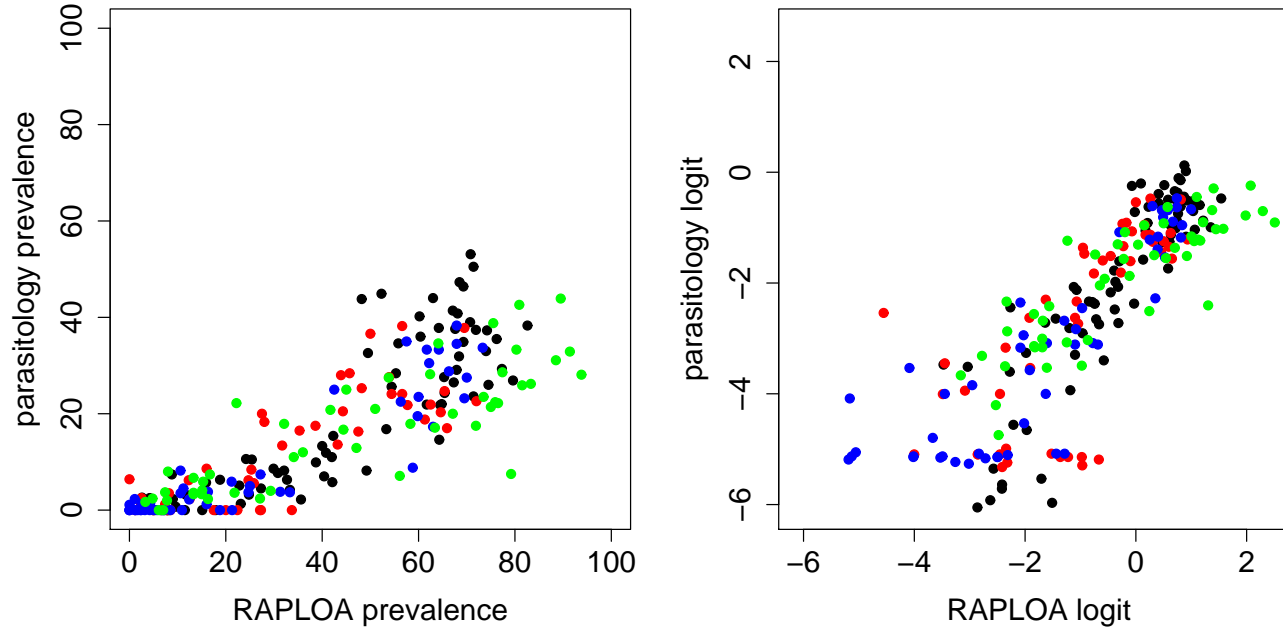
Figure 6: PCM for [high risk] in Cameroon based on ERM with ground truth data.

RAPLOA

- A cheaper alternative to parasitological sampling:
 - have you ever experienced eye-worm?
 - did it look like this photograph?
 - did it go away within a week?
- RAPLOA data collected:
 - in sample of villages previously surveyed
(to calibrate parasitology vs RAPLOA estimates)
 - in villages not previously surveyed
(to reduce local uncertainty)
- Calibration model needed to reconcile parasitological and RAPLOA prevalence estimates



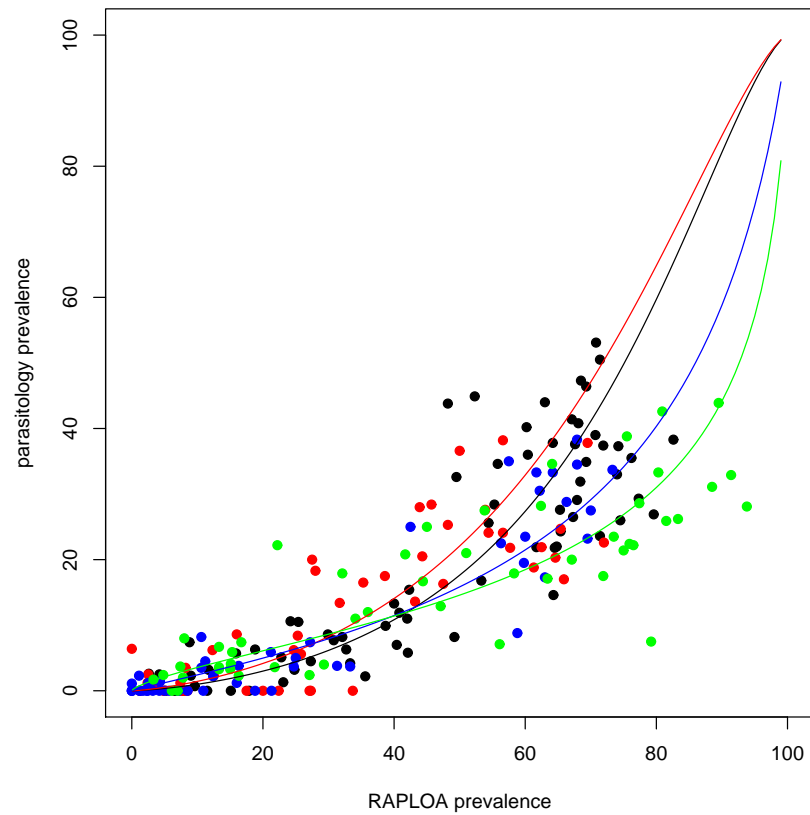
RAPLOA calibration



Empirical logit transformation linearises relationship
Colour-coding corresponds to four surveys in different regions

RAPLOA calibration (ctd)

Fit linear functional relationship on logit scale and back-transform



ARLAT

A RapLoa Analysis Tool

- plug-in for QGIS (open-source GIS) by Barry Rowlingson
- updates predictive exceedance maps as and when new RAPLOA data acquired
- portable to lap-top PC's for use in the field
- periodic off-line updating of base-map

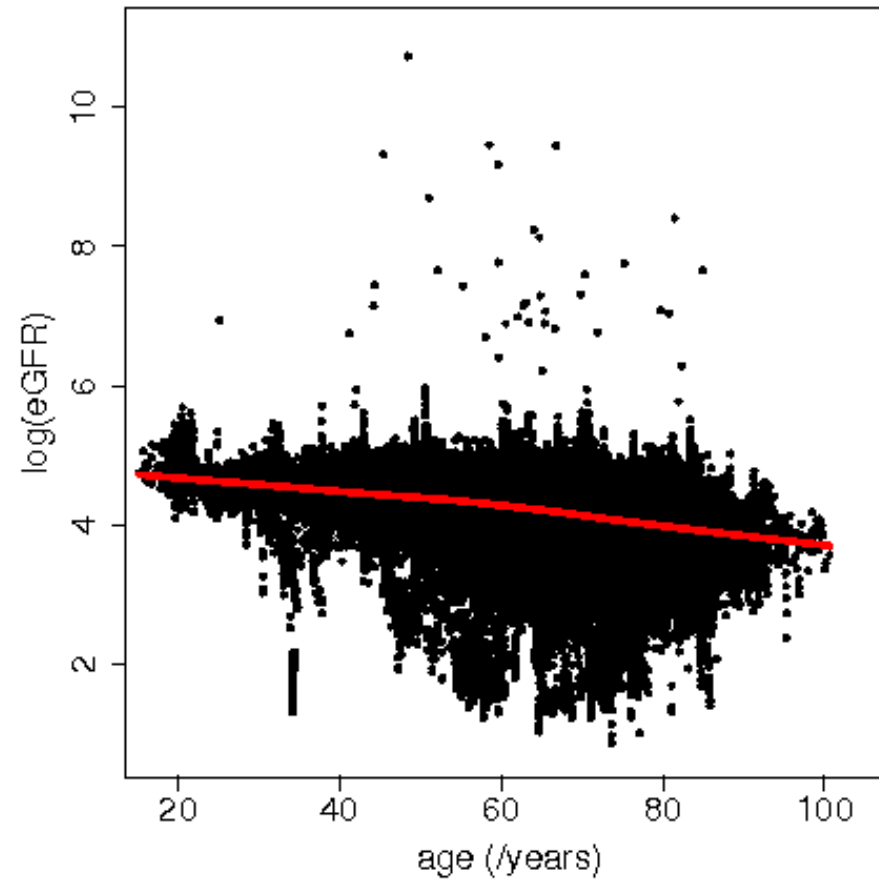
Demo at <http://www.lancaster.ac.uk/staff/diggle/>

Predicting incipient renal failure

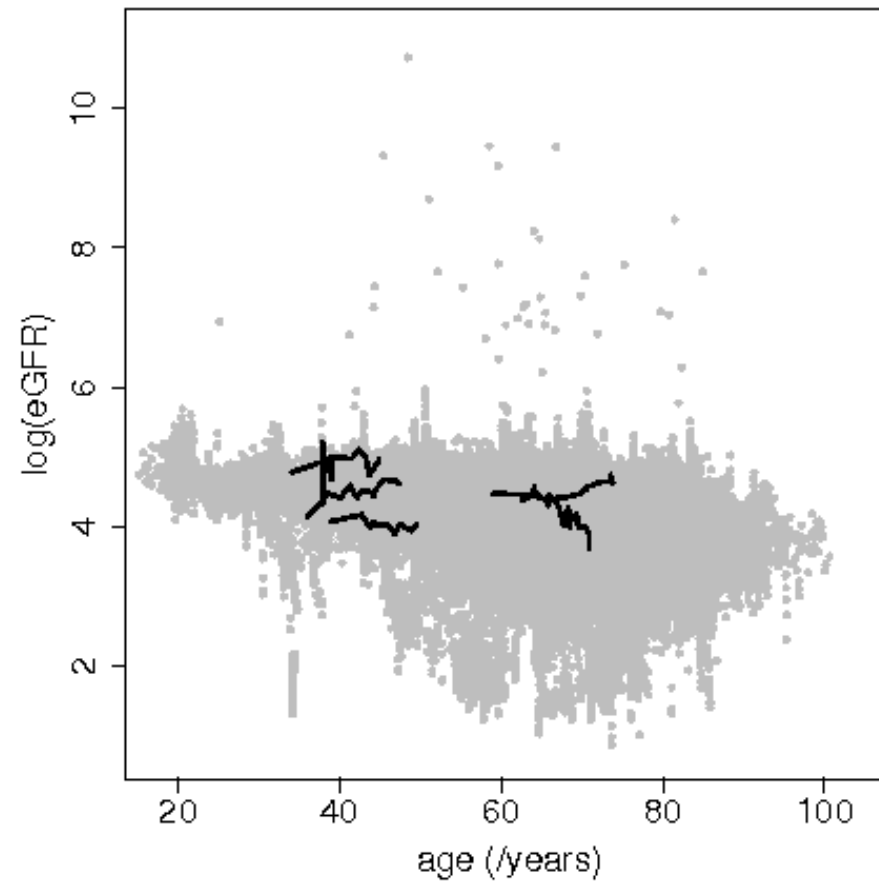
- kidney failure can be asymptomatic for many years
- kidney function measured by estimated glomerular filtration rate (eGFR)
- eGFR declines naturally with increasing age
- but unusually rapid rate of decline suggests need for specialist treatment

Goal: early detection of primary care patients whose annual rate of decline in eGFR is greater than 5%

Data: cross-sectional



Data: cross-sectional vs longitudinal



Dynamic Regression Model

- $Y_{ij} = \log(\text{eGFR})$ for patient i at time t_{ij}
- $E[Y_{ij}] = \text{linear function of covariates, initial age and time}$

$$Y_{ij} = \alpha_0(x_i) + \alpha_1 t_{ij} + U_i + C_i(t_{ij}) + Z_{ij}$$

- x_i base-line covariates (age, co-morbidities,...)
- U_i subject-specific random effect (intercept)
- $C_i(t_{ij})$ subject-specific random effect (time-varying)
- Z_{ij} measurement error and/or short-term fluctuations

Model for $C_i(t)$ is integrated Brownian Motion

$$C_i(t_j) = \int_0^{t_j} B_i(u) du, \quad B_i(u) | B_i(s) \sim \text{N}(B_i(s) + (u - s)\sigma^2, (u - s)\sigma^2)$$

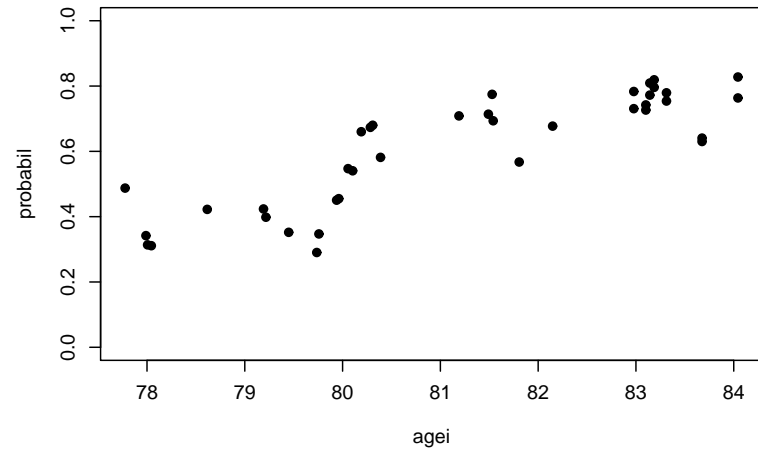
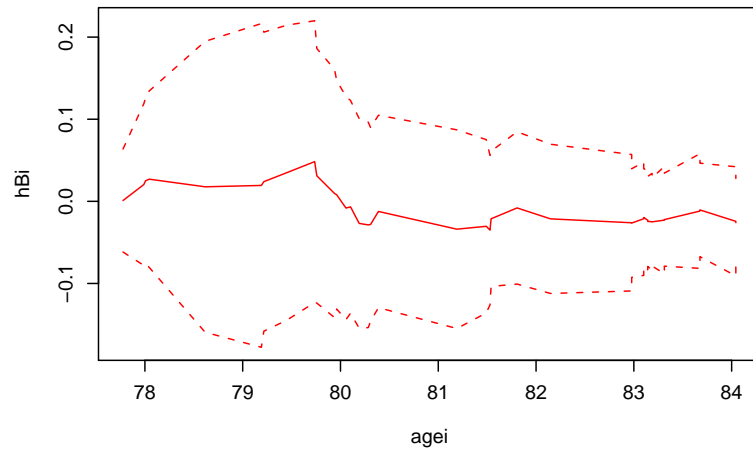
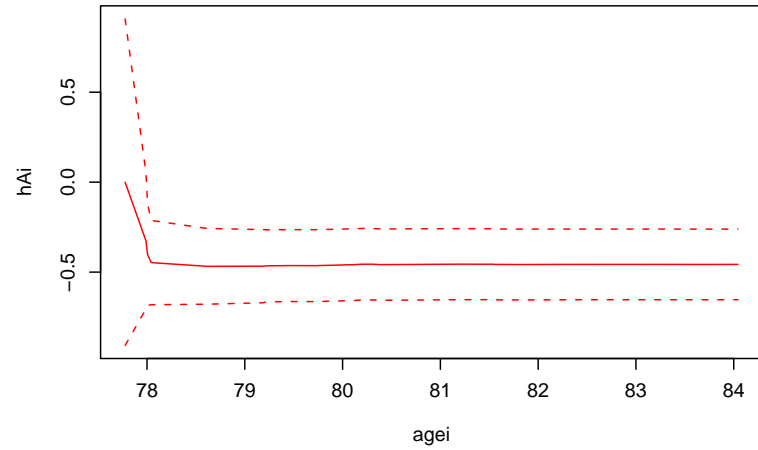
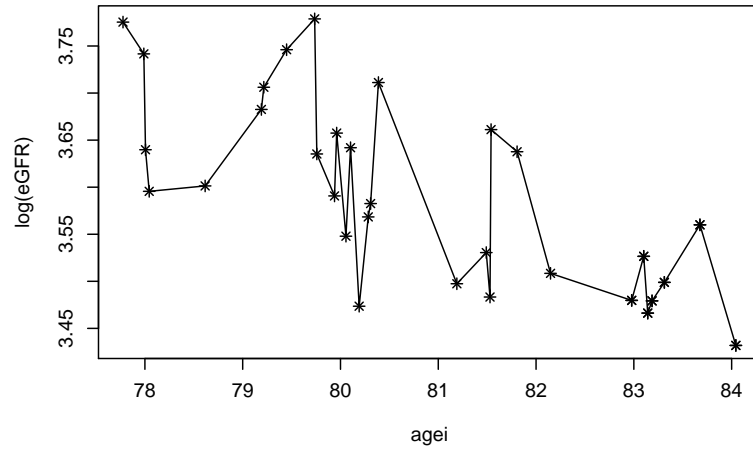
Prediction of subject-specific rate of change

Goal: predict $[B_i(t_{ij})|Y_{i1}, \dots, Y_{i,j-1}, Y_{ij}]$

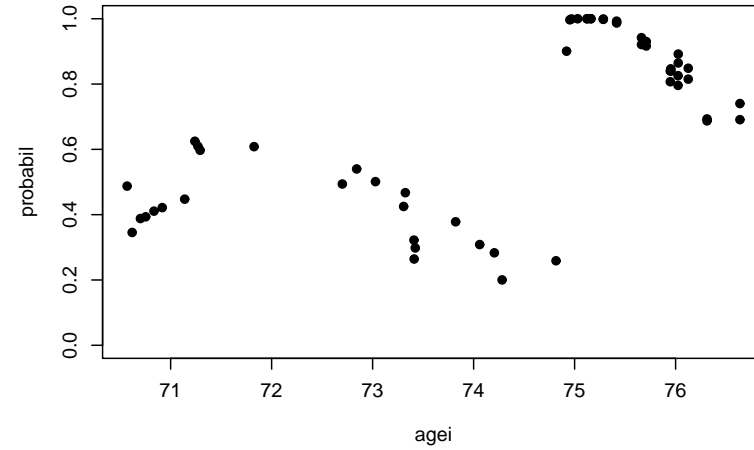
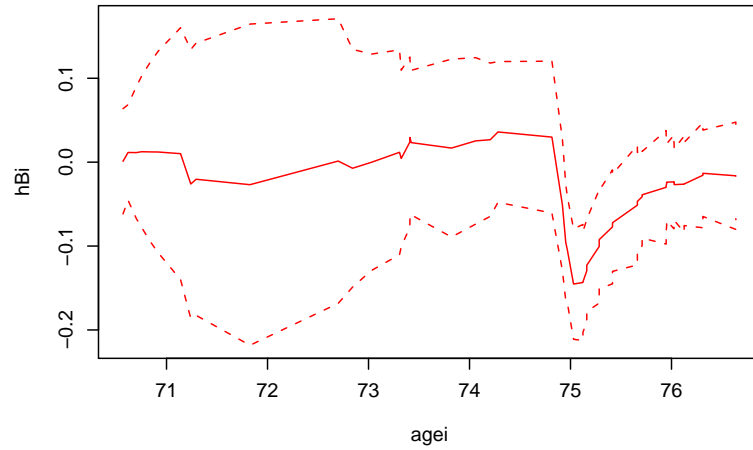
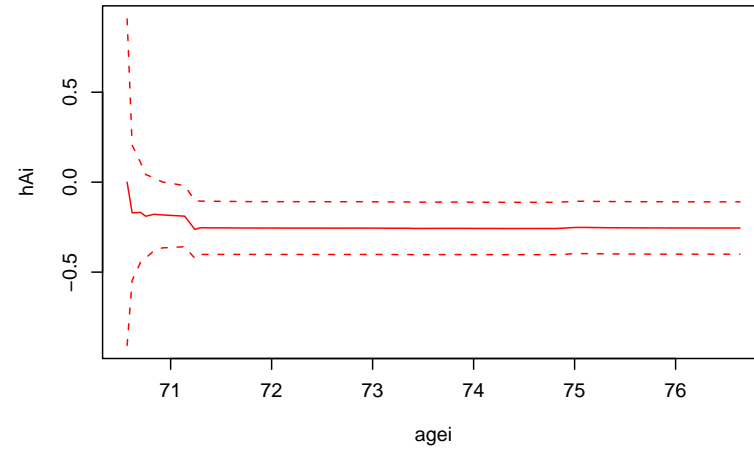
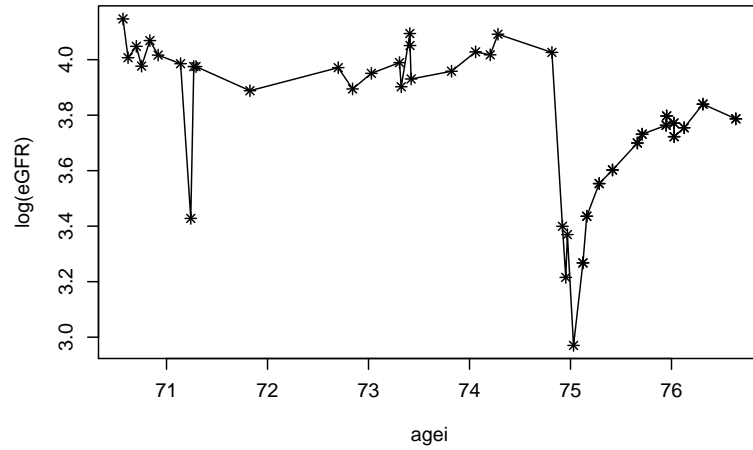
- $B_i(t)$ measures rate of change of eGFR relative to expectation
- Kalman filter algorithm allows efficient updating of predictive distribution whenever new eGFR measurement becomes available

Predictive inference: $P(B_i(t_{ij}) < -0.05|Y_{i1}, \dots, Y_{i,j-1}, Y_{ij})$

Results for diabetic patient A in primary care



Results for diabetic patient B in primary care



Outlook

- real-time health outcome data are becoming more widely available
- but are often under-utilised
- real-time data presents interesting challenges for statisticians
- statistical modelling is a two-edged sword

We buy information with assumptions

Coombs, 1964

- \Rightarrow need for close collaboration between statisticians and subject-matter experts

References

- Crainiceanu, C., Diggle, P.J. and Rowlingson, B.S. (2008) Bivariate modelling and prediction of spatial variation in Loa loa prevalence in tropical Africa (with Discussion). *Journal of the American Statistical Association* **103**, 21–47.
- Diggle, P.J. (2007). Spatio-temporal point processes: methods and applications. In *Semstat2004: Statistics of Spatio-Temporal Systems*, ed B Finkenstadt, L Held, V. Isham, 1–45. London: CRC Press.
- Diggle, P.J., Knorr-Held, L., Rowlingson, B., Su, T., Hawtin, P. and Bryant, T. (2003). Towards On-line Spatial Surveillance. In *Monitoring the Health of Populations: Statistical Methods for Public Health Surveillance*, ed R. Brookmeyer and D. Stroup. Oxford : Oxford University Press.
- Diggle, P.J. and Ribeiro, P.J. (2007). *Model-based Geostatistics*. New York: Springer.
- Diggle, P., Rowlingson, B. and Su, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, **16**, 423–34.
- Diggle, P.J., Thomson, M.C., Christensen, O.F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, H., Boussinesq, M. and Molyneux, D.H. (2007). Spatial modelling and prediction of Loa loa risk: decision making under uncertainty. *Annals of Tropical Medicine and Parasitology*, **101**, 499–509.
- Rodrigues, A. and Diggle, P.J. (2009). A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scandinavian Journal of Statistics* (submitted)
- Sousa, I. and Diggle, P.J. (2009). Real-time detection of incipient renal failure in primary care patients using a dynamic time series model. *Biostatistics* (submitted)