APTS 2007/08: Spatial and Longitudinal Data Analysis

Peter Diggle

(Department of Medicine, Lancaster University)

Glasgow, 1-5 September 2008

Timetable

1	**Monday**	**11.15–12.45**
2		**16.00–17.00**
3	**Tuesday**	**11.15–12.45**
4		**16.00–17.00**
5	**Wednesday**	**09.15–10.45**
6	**Thursday**	**09.15–10.45**
7		**14.15–15.15**
8	**Friday**	**09.15–10.45**

Lecture topics

- **Introduction:** motivating examples
- **Review of preliminary material**
- **Longitudinal data:** linear Gaussian models; conditional and marginal models; why longitudinal and time series data are not the same thing.
- **Continuous spatial variation:** stationary Gaussian processes; variogram estimation; likelihood-based estimation; spatial prediction.
- **Discrete spatial variation:** joint versus conditional specification; Markov random field models.

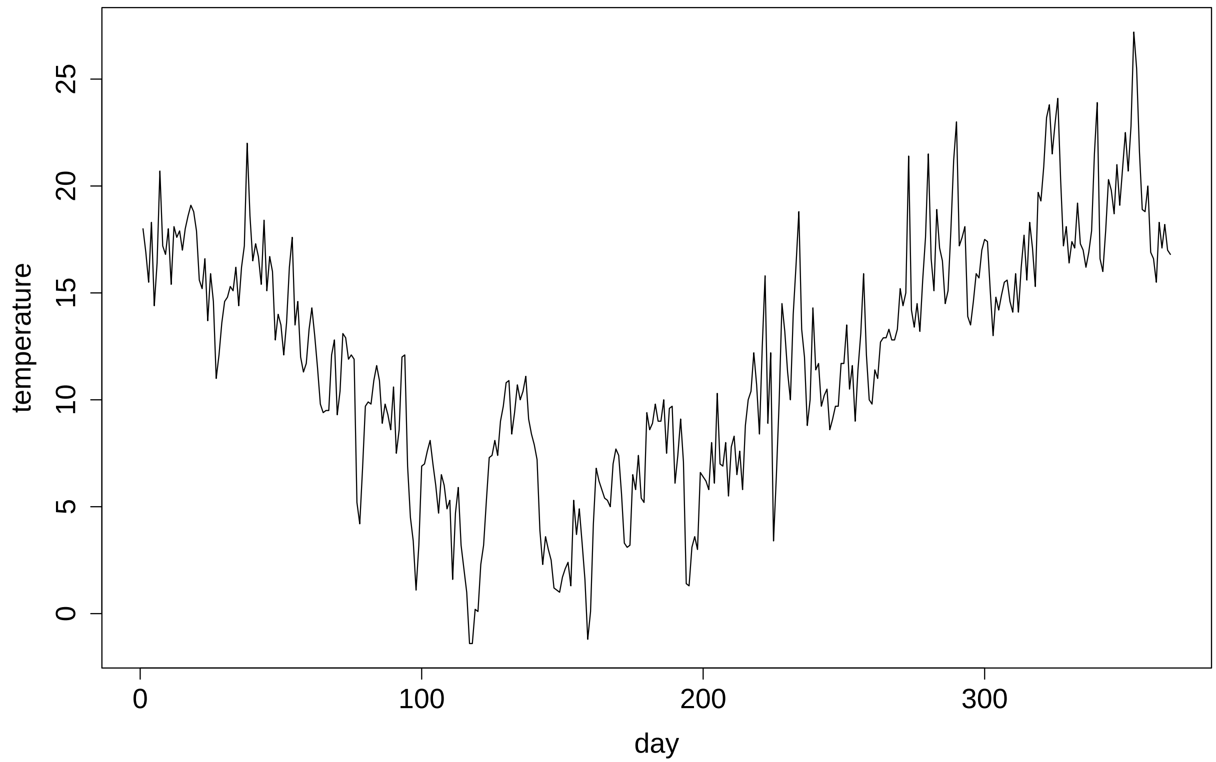
- **Spatial point patterns:** exploratory analysis; Cox processes and the link to continuous spatial variation; pairwise interaction processes and the link to discrete spatial variation.

- **Spatio-temporal modelling:** spatial time series; spatio-temporal point processes.

1. Motivating examples

Example 1.1 Bailrigg temperature records

Daily maximum temperatures, 1.09.1995 to 31.08.1996

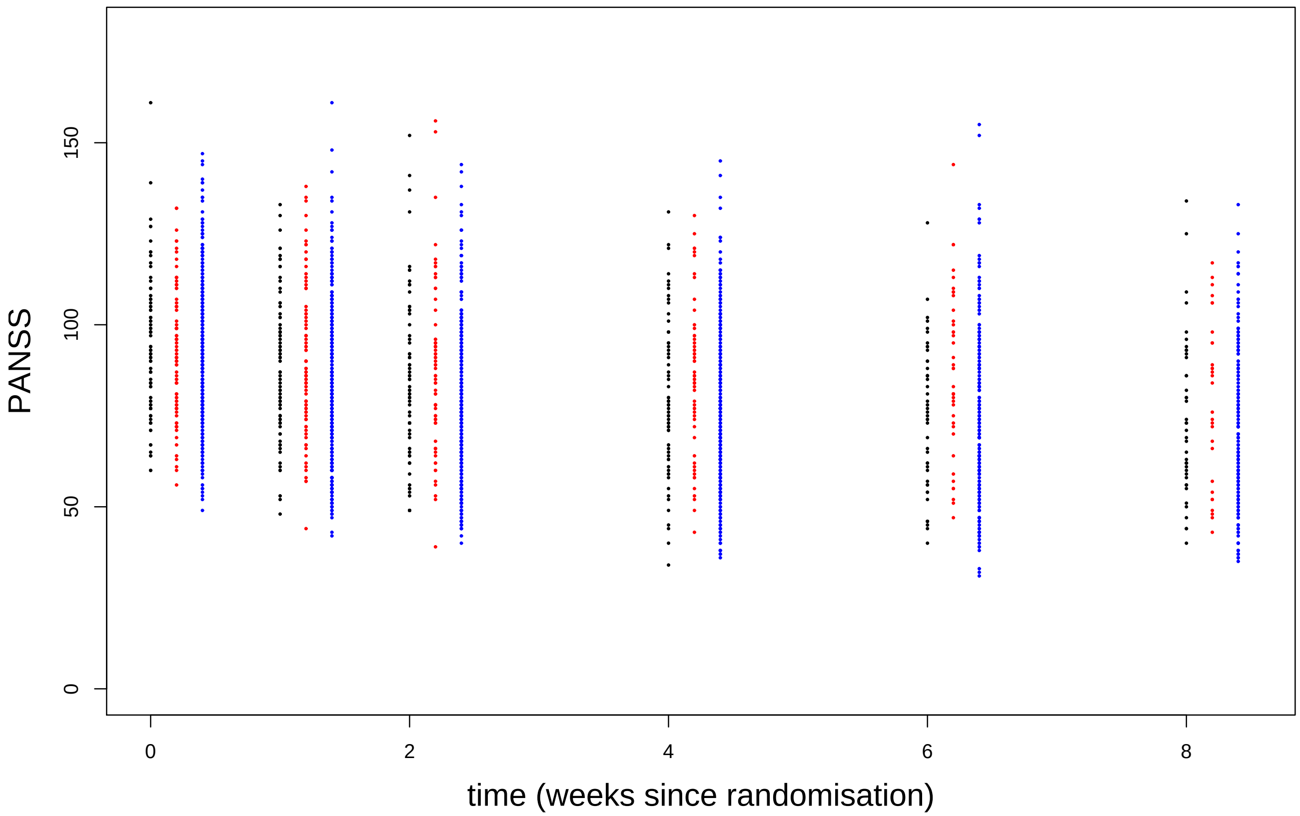


1.2 Schizophrenia clinical trial (PANSS)

- **randomised clinical trial of drug therapies**
- **three treatments:**
  - **haloperidol (standard)**
  - **placebo**
  - **risperidone (novel)**
- **dropout due to "inadequate response to treatment"**

Treatment	Number of non-dropouts at week					
	0	1	2	4	6	8
haloperidol	85	83	74	64	46	41
placebo	88	86	70	56	40	29
risperidone	345	340	307	276	229	199
total	518	509	451	396	315	269

Example 1.2: Schizophrenia trial data

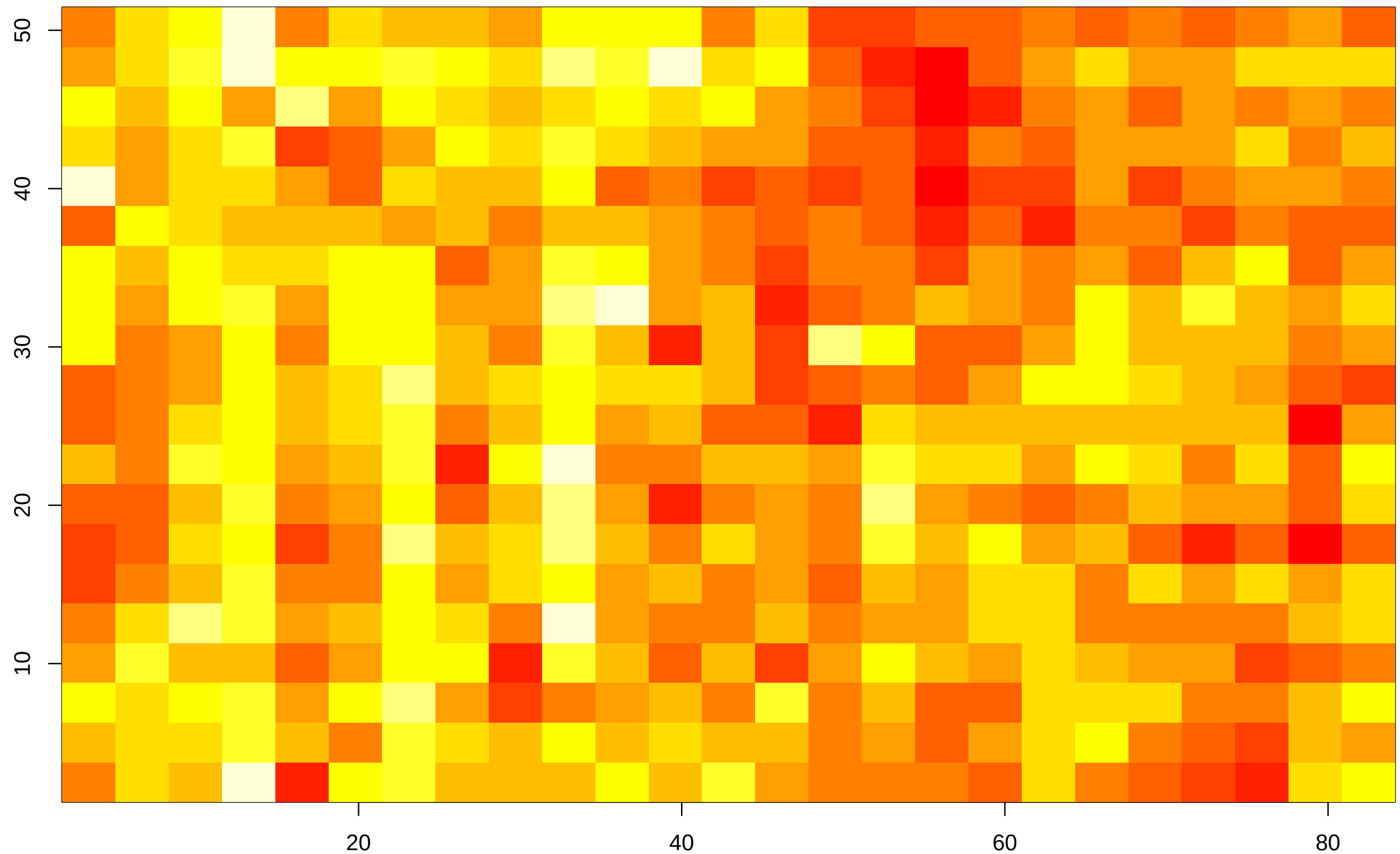


Diggle, Farewell and Henderson (2007)

Example 1.3 Wheat uniformity trial

- trial conducted at Rothamsted in summer of 1910
- wheat yield recorded in each of 500 rectangular plots (3.3m by 2.59m)
- same variety of wheat planted in all plots

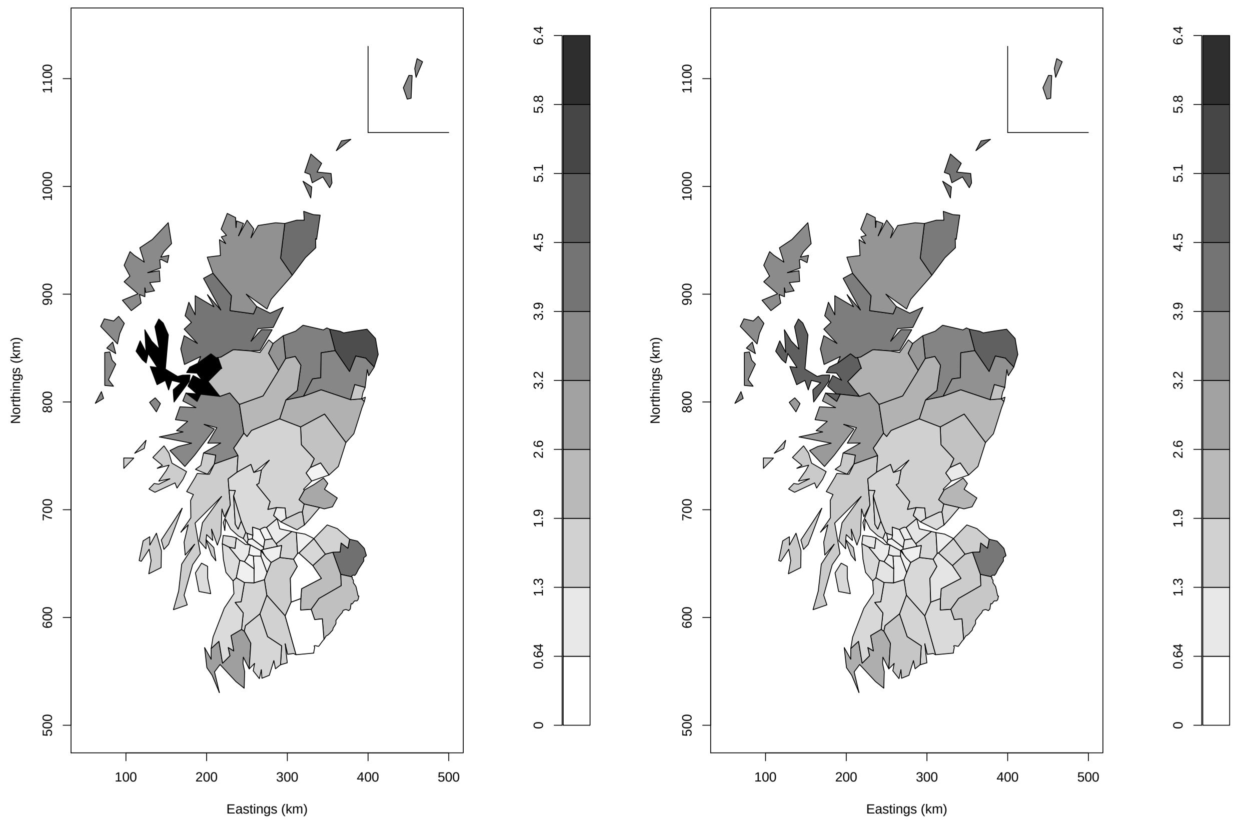
Mercer and Hall wheat yields



Mercer and Hall (1911)

1.4 Cancer atlases

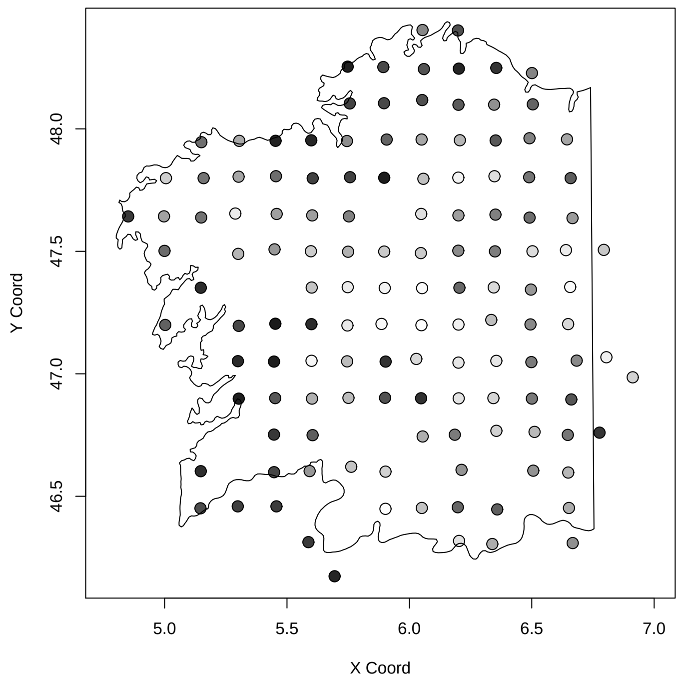
Raw and spatially smoothed relative risk estimates for lip cancer in 56 Scottish counties



Wakefield (2007)

1.5 Galicia biomonitoring study

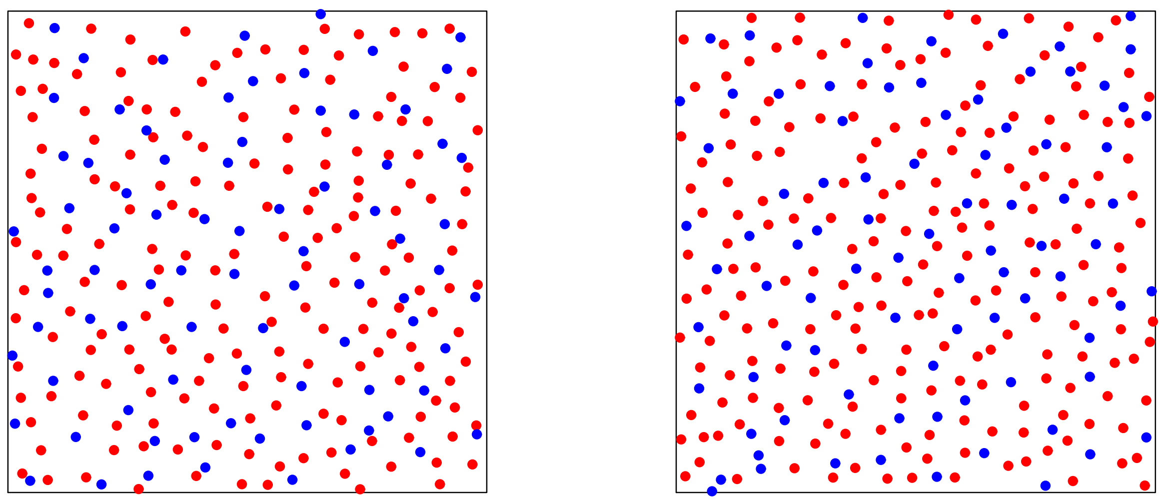
Lead concentrations measured in samples of moss, map shows locations and log-concentrations



Diggle, Menezes and Su (2008)

1.6 Retinal mosaics

Locations of two types of light-responsive cells in macaque retina (2 animals)



Eglen and Wong (2008)

2. Review of preliminary material

Time series

- trend and residual;
- autocorrelation;
- prediction;
- analysis of Bailrigg temperature data

Analysis of Bailrigg temperature data

```
data<-read.table("maxtemp.data",header=F)
temperature<-data[,4]
n<-length(temperature)
day<-1:n
plot(day,temperature,type="l",cex.lab=1.5,cex.axis=1.5)
#
# plot shows strong seasonal variation,
# try simple harmonic regression
#
```

```
c1<-cos(2*pi*day/n)
s1<-sin(2*pi*day/n)
fit1<-lm(temperature~c1+s1)
lines(day,fit1$fitted.values,col="red")
#
# looks OK, but add first harmonic of annual frequency
# to check for non-sinusoidal pattern
#
c2<-cos(4*pi*day/n)
s2<-sin(4*pi*day/n)
fit2<-lm(temperature~c1+s1+c2+s2)
lines(day,fit2$fitted.values,col="blue")
#
# two fits are almost identical - summarise by residual
# sums of squares
#
sum((temperature-mean(temperature))^2)
sum(fit1$resid^2)
sum(fit2$resid^2)
```

```
#
# examine autocorrelation properties of residuals
#
residuals<-fit1$resid
par(mfrow=c(2,2),pty="s")
for (k in 1:4) {
   plot(residuals[1:(n-k)],residuals[(k+1):n],
         pch=19,cex=0.5,xlab=" ",ylab=" ",main=k)
   }
par(mfrow=c(1,1))
acf(residuals)
#
# exponentially decaying correlation looks reasonable
#
cor(residuals[1:(n-1)],residuals[2:n])
Xmat<-cbind(rep(1,n),c1,s1)
rho<-0.01*(60:80)
profile<-AR1.profile(temperature,Xmat,rho)
```

```
#
# examine results
#
plot(rho,profile$logl,type="l",ylab="L(rho)")
Lmax<-max(profile$logl)
crit.val<-0.5*qchisq(0.95,1)
lines(c(rho[1],rho[length(rho)]),rep(Lmax-crit.val,2),lty=2)
profile
```

```
#
# profile log-likelihood function follows
#
AR1.profile<-function(y,X,rho) {
   m<-length(rho)
   logl<-rep(0,m)
   n<- length(y)
   hold<-outer(1:n,1:n,"-")
   for (i in 1:m) {
      Rmat<-rho[i]^abs(hold)
      ev<-eigen(Rmat)
      logdet<-sum(log(ev$values))
      Rinv<-ev$vectors%*%diag(1/ev$values)%*%t(ev$vectors)
      betahat<-solve(t(X)%*%Rinv%*%X)%*%t(X)%*%Rinv%*%y
      residual<- y-X%*%betahat
      logl[i]<- - logdet - n*log(c(residual)%*%Rinv%*%c(residual))
      }
   max.index<-order(logl)[m]
   Rmat<-rho[max.index]^abs(hold)
      ev<-eigen(Rmat)
      logdet<-sum(log(ev$values))
      Rinv<-ev$vectors%*%diag(1/ev$values)%*%t(ev$vectors)
      betahat<-solve(t(X)%*%Rinv%*%X)%*%t(X)%*%Rinv%*%y
      residual<- y-X%*%betahat
      sigmahat<-sqrt(c(residual)%*%Rinv%*%c(residual)/n)
   list(logl=logl,rhohat=rho[max.index],sigmahat=sigmahat,betahat=betahat)
   }
```

Longitudinal data

- **replicated time series;**
- **focus of interest often on mean values;**
- **modelling and inference can and should exploit replication**

Discrete spatial variation

- space is not like time;
- models for discrete spatial variation are tied to number of spatial units

Real-valued continuous spatial variation

- **direct specification of covariance structure;**
- **variogram as an exploratory and/or diagnostic tool**

Spatial point processes

- **the Poisson process;**
- **crude classification of processes/patterns as regular, completely random or aggregated**

3. Longitudinal data

- linear Gaussian models;
- conditional and marginal models;
- missing values

Correlation and why it matters

- different measurements on the same subject are typically correlated
- and this must be recognised in the inferential process.

Estimating the mean of a time series

$$Y_1, Y_2, ..., Y_t, ..., Y_n \qquad Y_t \sim \mathrm{N}(\mu, \sigma^2)$$

Classical result: $\bar{Y} \pm 2\sqrt{\sigma^2/n}$

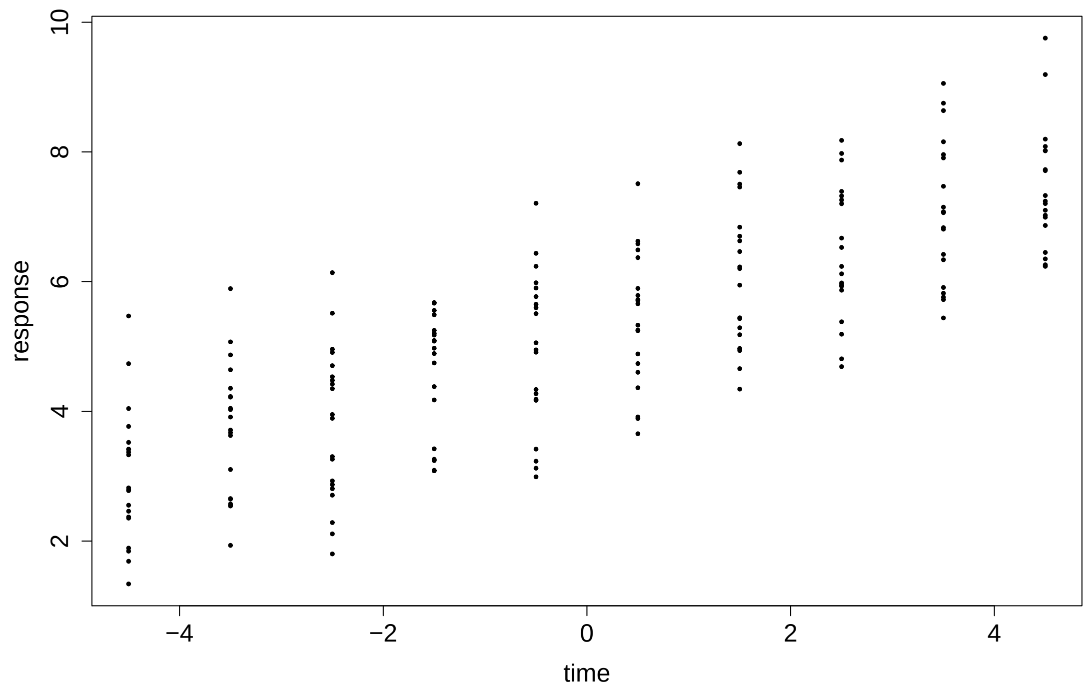
But if Y_t is a time series:

- $\mathrm{E}[\bar{Y}] = \mu$
- $\mathrm{Var}\{\bar{Y}\} = (\sigma^2/n) \times \{1 + n^{-1} \sum_{u \neq t} \mathrm{Corr}(Y_t, Y_u)\}$

Exercise: is the sample variance unbiased for $\sigma^2 = \mathrm{Var}(Y_t)$?

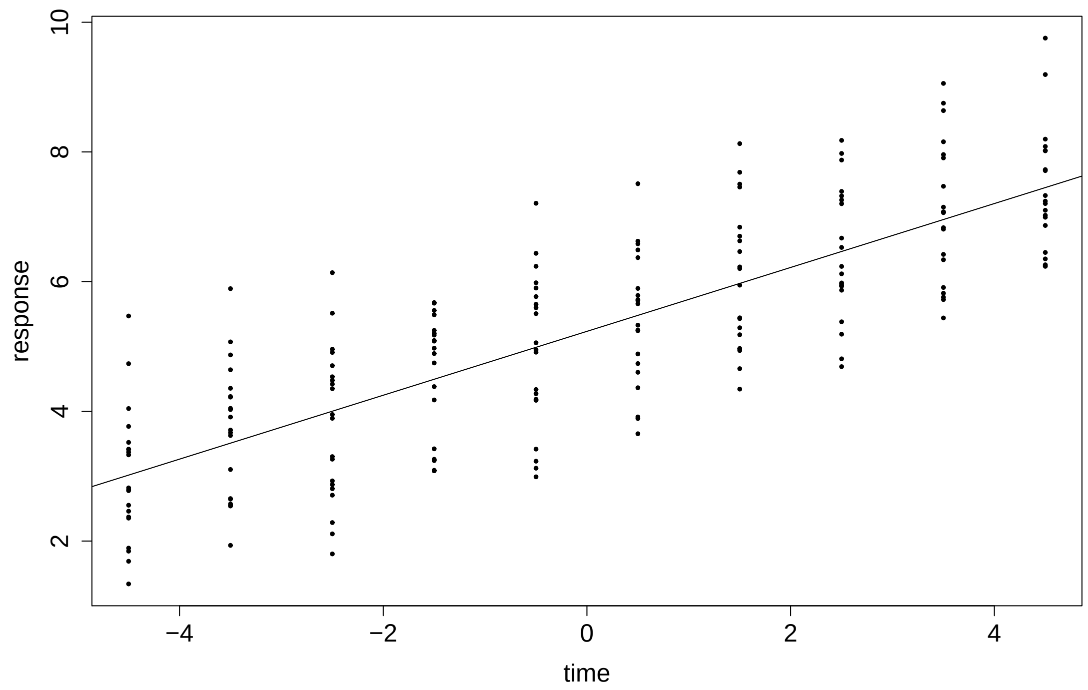
Correlation may or may not hurt you

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$



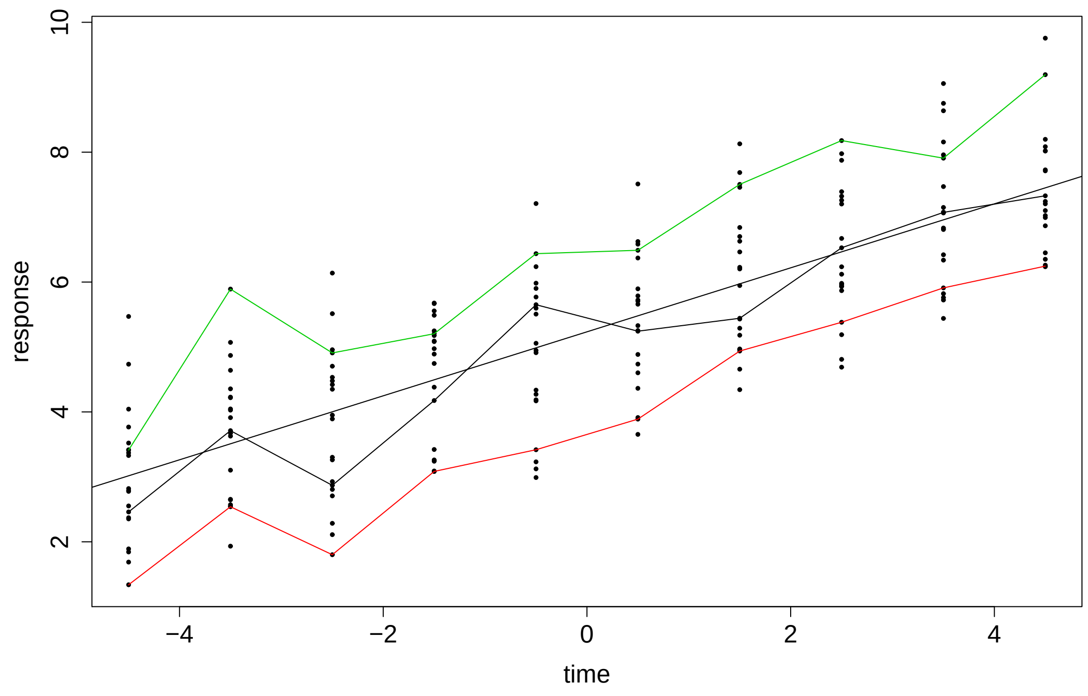
Correlation may or may not hurt you

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$



Correlation may or may not hurt you

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$



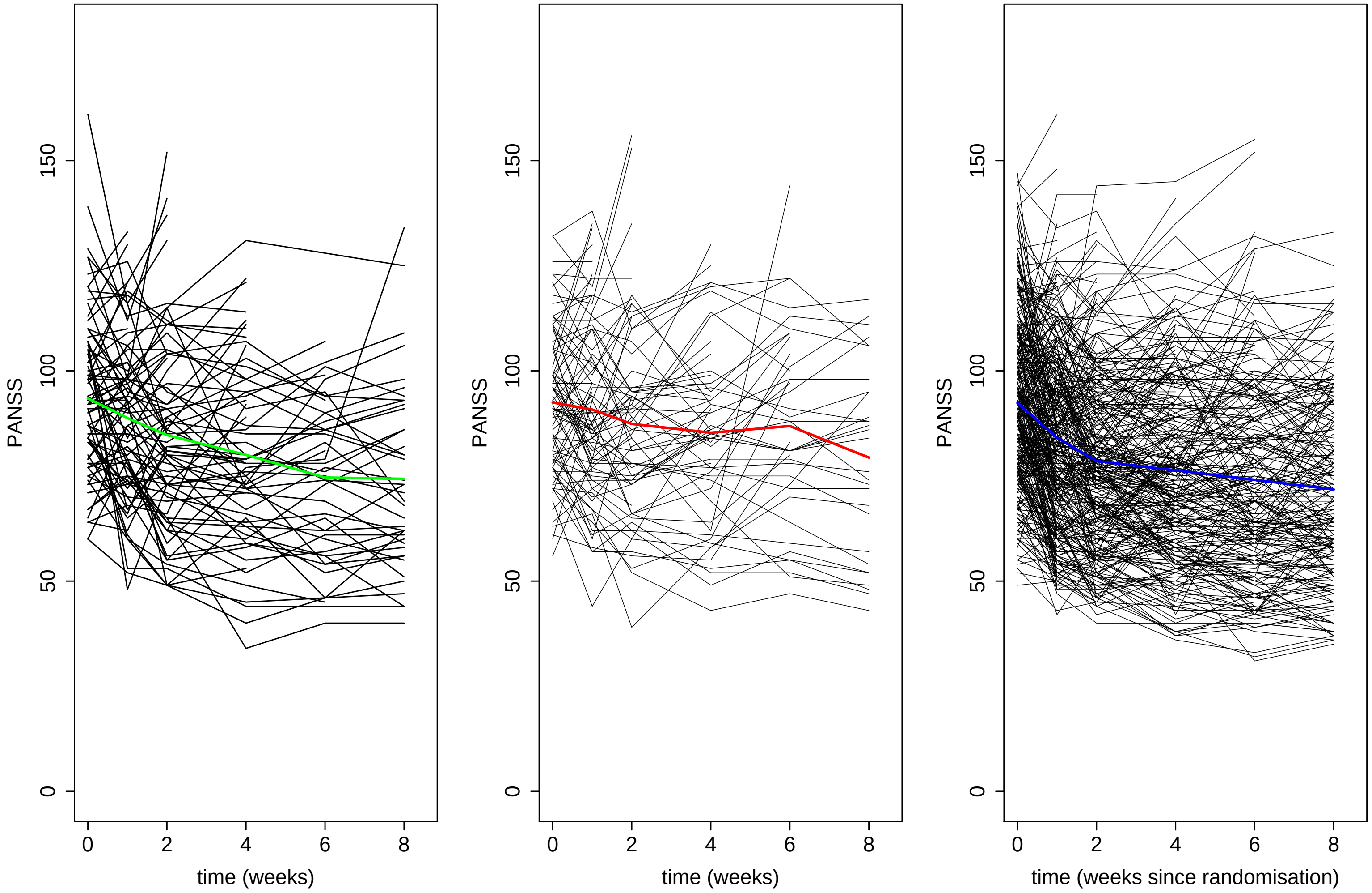
Correlation may or may not hurt you

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$

Parameter estimates and standard errors:

	ignoring correlation		recognising correlation	
	estimate	standard error	estimate	standard error
α	5.234	0.074	5.234	0.202
β	0.493	0.026	0.493	0.011

A spaghetti plot of the PANSS data



The variogram of a stochastic process $Y(t)$ is

$$V(u) = \frac{1}{2}\text{Var}\{Y(t) - Y(t-u)\}$$

- **well-defined for stationary and some non-stationary processes**
- **for stationary processes,**

$$V(u) = \sigma^2\{1 - \rho(u)\}$$

- **easier to estimate $V(u)$ than $\rho(u)$ when data are unbalanced**

Estimating the variogram

Data: $(Y_{ij}, t_{ij}) : i = 1, ..., m; \; j = 1, ..., n_i$

r_{ij} = residual from preliminary model for mean response

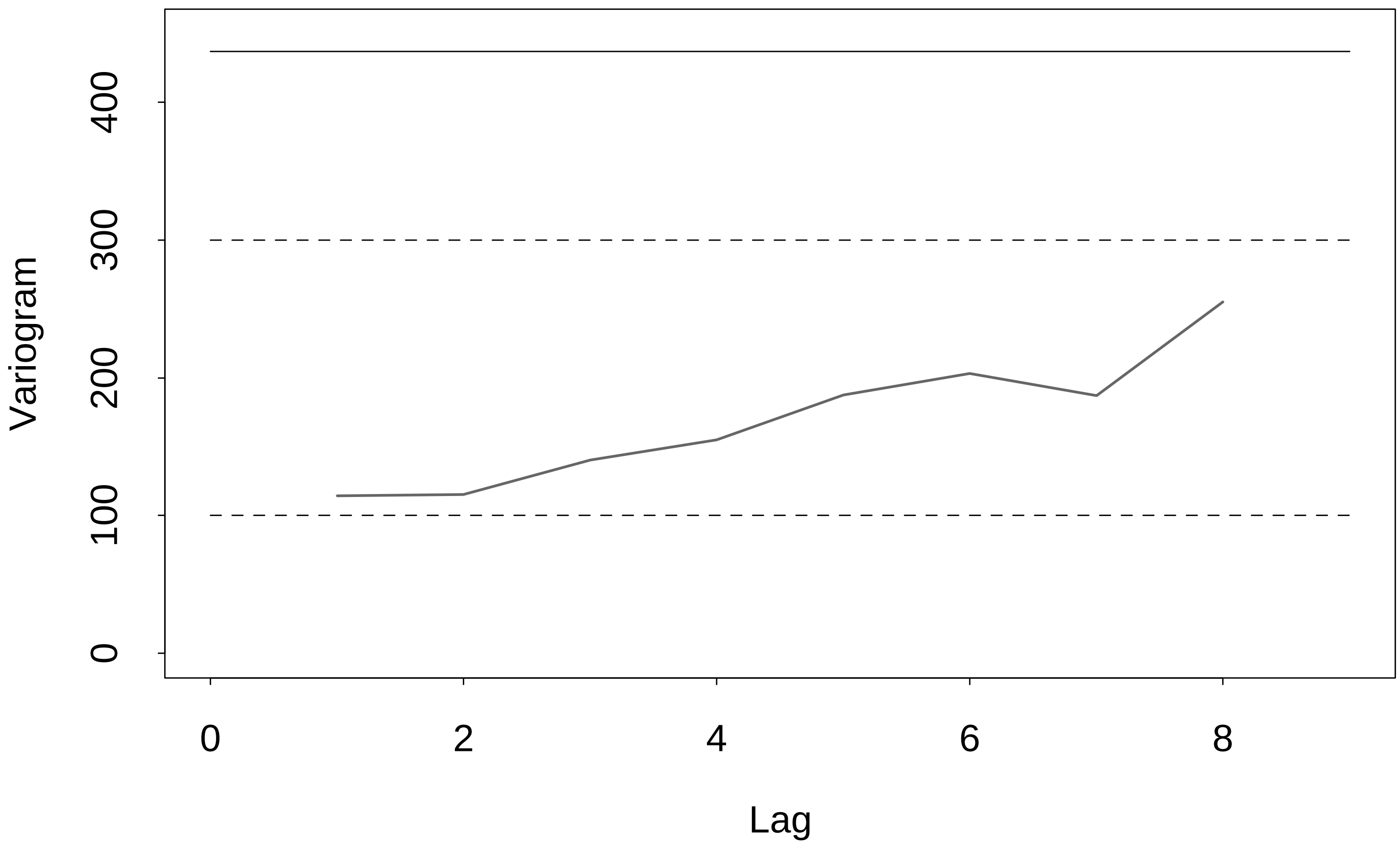
- **Define**

$$v_{ijk\ell} = \frac{1}{2}(r_{ij} - r_{k\ell})^2$$

- **Estimate**

$$\begin{aligned} \hat{V}(u) &= \text{average of all } v_{iji\ell} \text{ such that } |t_{ij} - t_{i\ell}| \simeq u \\ \hat{\sigma}^2 &= \text{average of all } v_{ijk\ell} \text{ such that } i \neq k. \end{aligned}$$

Example 3. schizophrenia trial



Where does the correlation come from?

- differences between subjects
- variation over time within subjects
- measurement error

General linear model, correlated residuals

i = subjects $\qquad$ j = measurements within subjects

$$\begin{aligned} E(Y_{ij}) &= x_{ij1}\beta_1 + ... + x_{ijp}\beta_p \\ Y_i &= X_i\beta + \epsilon_i \\ Y &= X\beta + \epsilon \end{aligned}$$

- measurements from different subjects independent
- measurements from same subject typically correlated.

Parametric models for covariance structure

Three sources of random variation in a typical set of longitudinal data:

- **Random effects (variation between subjects)**
  - **characteristics of individual subjects**
  - **for example, intrinsically high or low responders**
  - **influence extends to all measurements on the subject in question.**

Parametric models for covariance structure

Three sources of random variation in a typical set of longitudinal data:

- **Random effects**
- **Serial correlation (variation over time within subjects)**
  - **measurements taken close together in time typically more strongly correlated than those taken further apart in time**
  - **on a sufficiently small time-scale, this kind of structure is almost inevitable**

Parametric models for covariance structure

Three sources of random variation in a typical set of longitudinal data:

- **Random effects**
- **Serial correlation**
- **Measurement error**
  - **when measurements involve delicate determinations, duplicate measurements at same time on same subject may show substantial variation**

Diggle, Heagerty, Liang and Zeger (2002, Chapter 5)

Some simple models

- **Compound symmetry**

$$Y_{ij} - \mu_{ij} = U_i + Z_{ij}$$

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

$Z_{ij} \sim \mathrm{N}(0, \tau^2)$

Implies that $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \nu^2/(\nu^2 + \tau^2)$, for all $j \neq k$

- **Random intercept and slope**

$$Y_{ij} - \mu_{ij} = U_i + W_i t_{ij} + Z_{ij}$$

$(U_i, W_i) \sim \mathrm{BVN}(0, \Sigma)$

$Z_{ij} \sim \mathrm{N}(0, \tau^2)$

Often fits short sequences well, but extrapolation dubious, for example Var(Y_{ij}) quadratic in t_{ij}

- **Autoregressive**

$$Y_{ij} - \mu_{ij} = \alpha(Y_{i,j-1} - \mu_{i,j-1}) + Z_{ij}$$

$$Y_{i1} - \mu_{i1} \sim \mathrm{N}\{0, \tau^2/(1-\alpha^2)\}$$

$$Z_{ij} \sim \mathrm{N}(0, \tau^2), \quad j = 2, 3, \ldots$$

Not a natural choice for underlying continuous-time processes

- **Stationary Gaussian process**

$$Y_{ij} - \mu_{ij} = W_i(t_{ij})$$

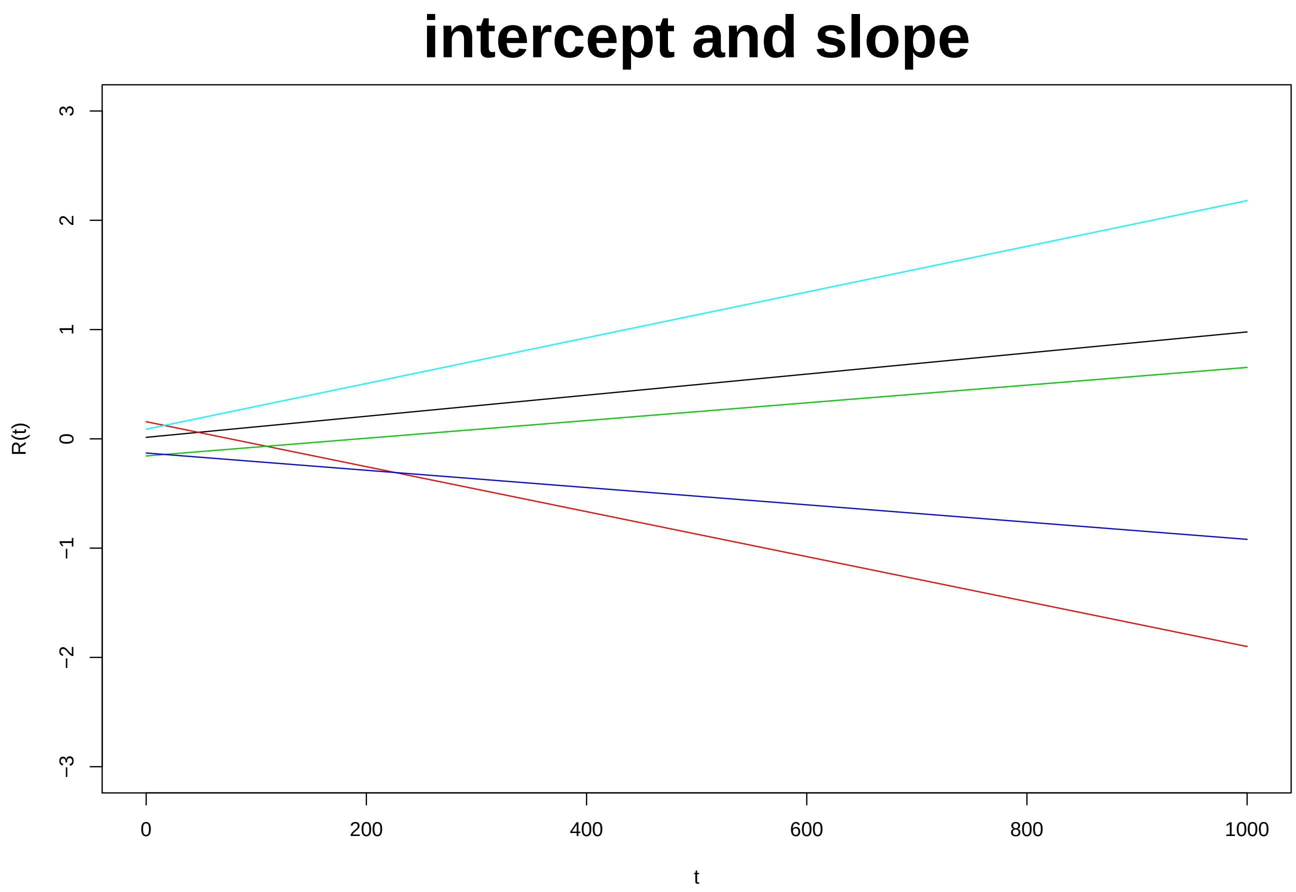
$W_i(t)$ a continuous-time Gaussian process

$\mathrm{E}[W(t)] = 0 \quad \mathrm{Var}\{W(t)\} = \sigma^2$

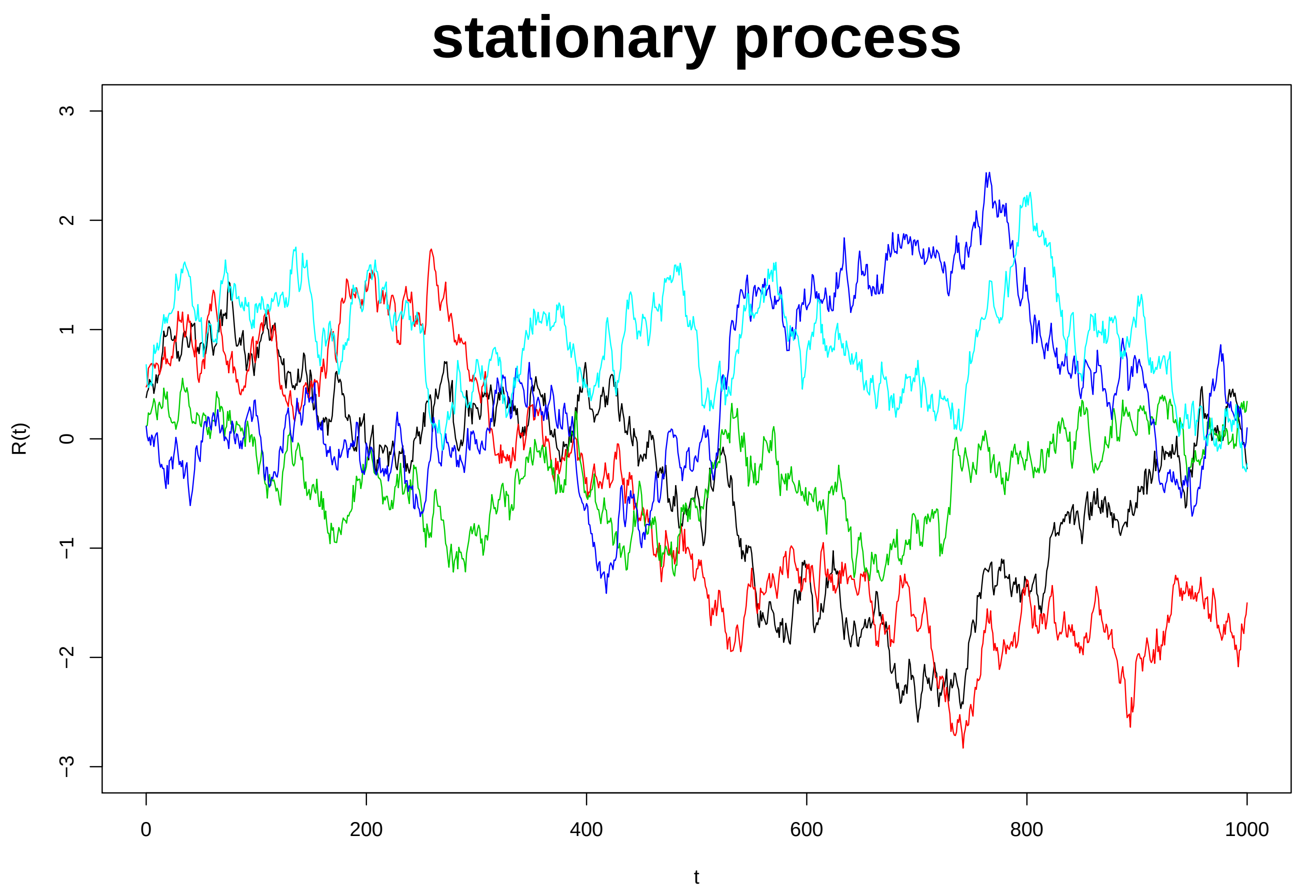
$\mathrm{Corr}\{W(t), W(t-u)\} = \rho(u)$

$\rho(u) = \exp(-u/\phi)$ gives continuous-time version of the autoregressive model

Time-varying random effects



Time-varying random effects: continued



- **A general model**

$$Y_{ij} - \mu_{ij} = d_{ij}' U_i + W_i(t_{ij}) + Z_{ij}$$

$U_i \sim \mathrm{MVN}(0, \Sigma)$
(random effects)

d_{ij} = vector of explanatory variables for random effects

$W_i(t)$ = continuous-time Gaussian process
(serial correlation)

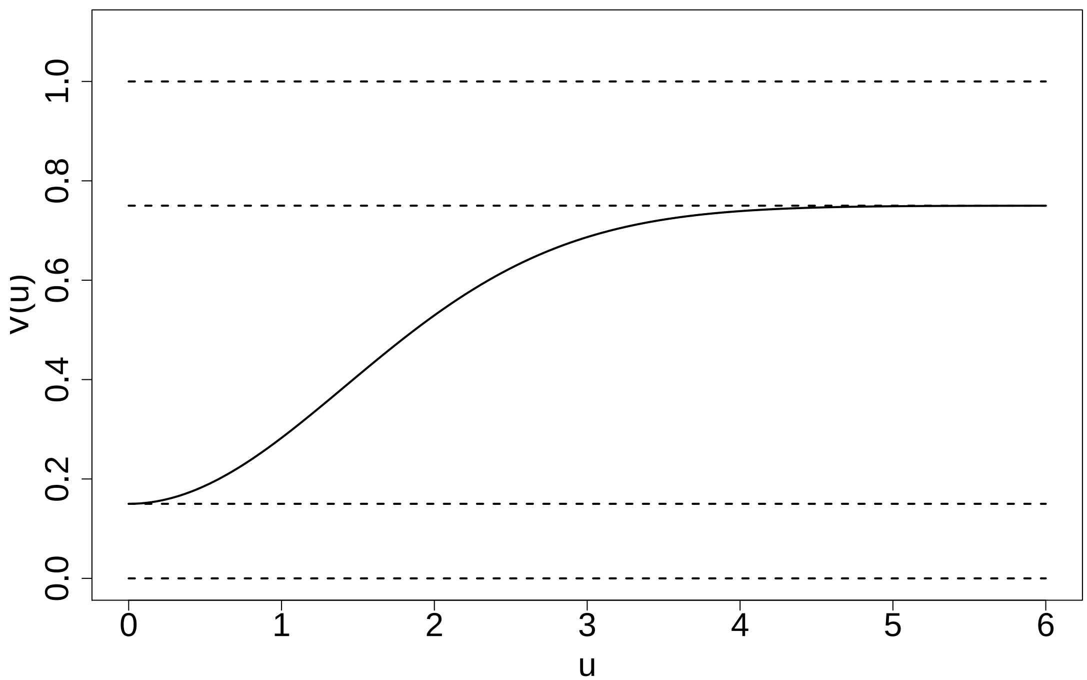
$Z_{ij} \sim \mathrm{N}(0, \tau^2)$
(measurement errors)

Even when all three components of variation are needed in principle, one or two may dominate in practice

The variogram of the general model

$$Y_{ij} - \mu_{ij} = d'_{ij} U_i + W_i(t_{ij}) + Z_{ij}$$

$$V(u) = \tau^2 + \sigma^2\{1 - \rho(u)\} \quad \mathrm{Var}(Y_{ij}) = \nu^2 + \sigma^2 + \tau^2$$



Fitting the model: non-technical summary

- **Ad hoc methods won't do**
- **Likelihood-based inference is the statistical gold standard**
- **But be sure you know what you are estimating when there are missing values**

Maximum likelihood estimation (V_0 known)

Log-likelihood for observed data y is

$$\begin{aligned} L(\beta, \sigma^2, V_0) &= -0.5\{nm \log \sigma^2 + m \log |V_0| \\ &\quad + \sigma^{-2}(y - X\beta)'(I \otimes V_0)^{-1}(y - X\beta)\} \end{aligned} \tag{1}$$

Given V_0, estimator for β is

$$\hat{\beta}(V_0) = (X'(I \otimes V_0)^{-1}X)^{-1}X'(I \otimes V_0)^{-1}y, \tag{2}$$

the weighted least squares estimates with $W = (I \otimes V_0)^{-1}$.

Explicit estimator for σ^2 also available as

$$\hat{\sigma}^2(V_0) = RSS(V_0)/(nm) \tag{3}$$

$$RSS(V_0) = \{y - X\hat{\beta}(V_0)\}'(I \otimes V_0)^{-1}\{y - X\hat{\beta}(V_0)\}.$$

Maximum likelihood estimation, V_0 unknown

Substitute (2) and (3) into (1) to give reduced log-likelihood

$$\mathcal{L}(V_0) = -0.5m[n \log\{RSS(V_0)\} + \log|V_0|]. \tag{4}$$

Numerical maximization of (4) then gives $\hat{V}_0$, hence $\hat{\beta} \equiv \hat{\beta}(\hat{V}_0)$ and $\hat{\sigma}^2 \equiv \hat{\sigma}^2(\hat{V}_0)$.

- **Dimensionality of optimisation is $\frac{1}{2}n(n+1) - 1$**
- **Each evaluation of $\mathcal{L}(V_0)$ requires inverse and determinant of an n by n matrix.**

A random effects model for CD4 cell counts

```
data<-read.table("CD4.data",header=T)
data[1:3,]
time<-data$time
CD4<-data$CD4
plot(time,CD4,pch=19,cex=0.25)
id<-data$id
uid<-unique(id)
for (i in 1:10) {
   take<-(id==uid[i])
   lines(time[take],CD4[take],col=i,lwd=2)
   }
```

```
# Simple linear model assuming uncorrelated residuals
#
fit1<-lm(CD4~time)
summary(fit1)
#
# random intercept and slope model
#
library(nlme)
?lme
fit2<-lme(CD4~time,random=~1|id)
summary(fit2)
```

```
# make fitted value constant before sero-conversion
#
timeplus<-time*(time>0)
fit3<-lme(CD4~timeplus,random=~1|id)
summary(fit3)
tfit<-0.1*(0:50)
Xfit<-cbind(rep(1,51),tfit)
fit<-c(Xfit%*%fit3$coef$fixed)
Vmat<-fit3$varFix
Vfit<-diag(Xfit%*%Vmat%*%t(Xfit))
upper<-fit+2*sqrt(Vfit)
lower<-fit-2*sqrt(Vfit)
#
# plot fit with 95% point-wise confidence intervals
#
plot(time,CD4,pch=19,cex=0.25)
lines(c(-3,tfit),c(upper[1],upper),col="red")
lines(c(-3,tfit),c(lower[1],lower),col="red")
```

Missing values and dropouts

Issues concerning missing values in longitudinal data can be addressed at two different levels:

- **technical: can the statistical method I am using cope with missing values?**

- **conceptual: *why* are the data missing? Does the fact that an observation is missing convey partial information about the value that would have been observed?**

These same questions also arise with cross-sectional data, but the correlation inherent to longitudinal data can sometimes be exploited to good effect.

Rubin's classification

- **MCAR (completely at random): P(missing) depends neither on observed nor unobserved measurements**
- **MAR (at random): P(missing) depends on observed measurements, but not on unobserved measurements**
- **MNAR (not at random): conditional on observed measurements, P(missing) depends on unobserved measurements.**

Rubin (1976)

Dropout

Once a subject goes missing, they never return

Example : Longitudinal clinical trial

- **completely at random: patient leaves the the study because they move house**

- **at random : patient leaves the study on their doctor's advice, based on observed measurement history**

- **not at random : patient misses their appointment because they are feeling unwell.**

Little (1995)

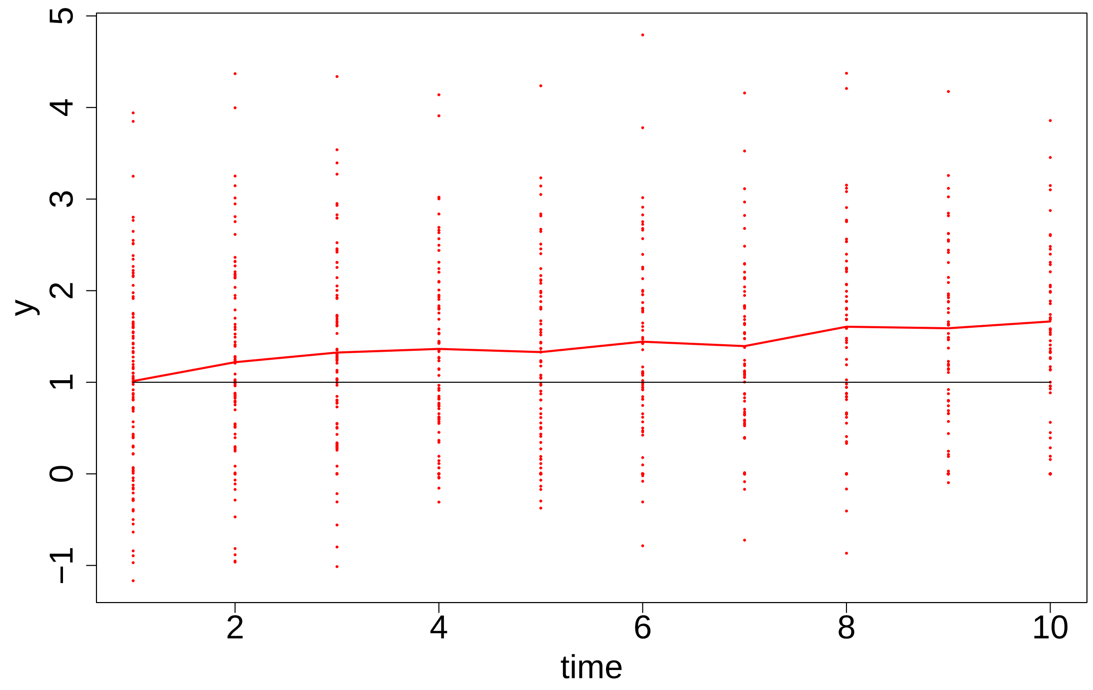
Conventional wisdom

- any sensible method of analysis valid if dropout is MCAR
- likelihood-based analysis valid if dropout is MAR

But: under MAR, target of likelihood-based inference is model for hypothetical dropout-free population

Example

- **Model is $Y_{ij} = \mu + U_i + Z_{ij}$ (random intercept)**
- **Dropout is MAR: logit$p_{ij} = -1 - 2 \times Y_{i,j-1}$**



PJD's take on ignorability

For correlated data, dropout mechanism can be ignored only if dropouts are completely random

In all other cases, need to:

- **think carefully what are the relevant practical questions,**
- **fit an appropriate model for both measurement process and dropout process**
- **use the model to answer the relevant questions.**

Diggle, Farewell and Henderson (2007)

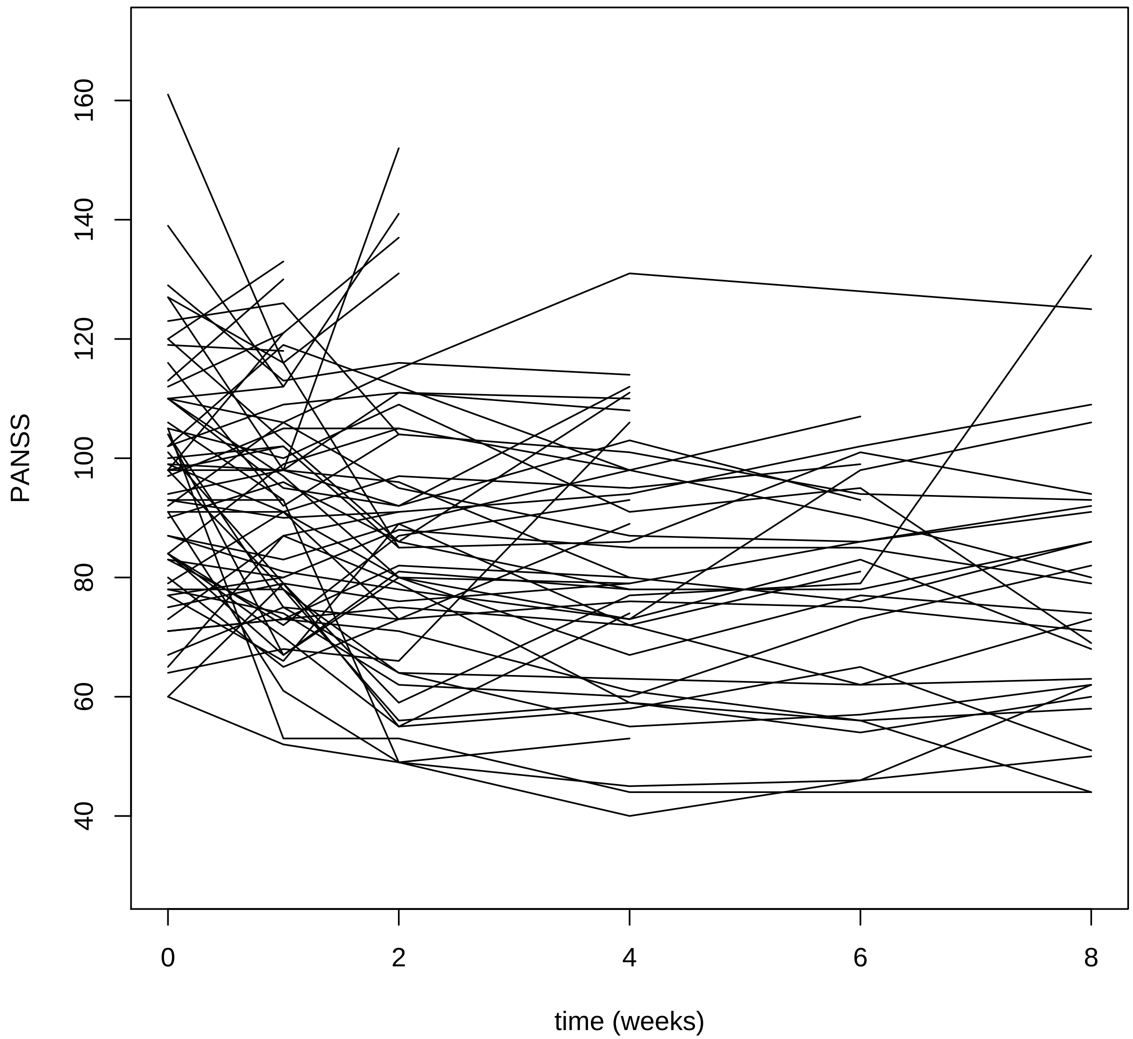
Schizophrenia trial data

- Data from placebo-controlled RCT of drug treatments for schizophrenia:
  - Placebo; Haloperidol (standard); Risperidone (novel)
- Y = sequence of weekly PANSS measurements
- F = dropout time
- Total $m = 516$ subjects, but high dropout rates:

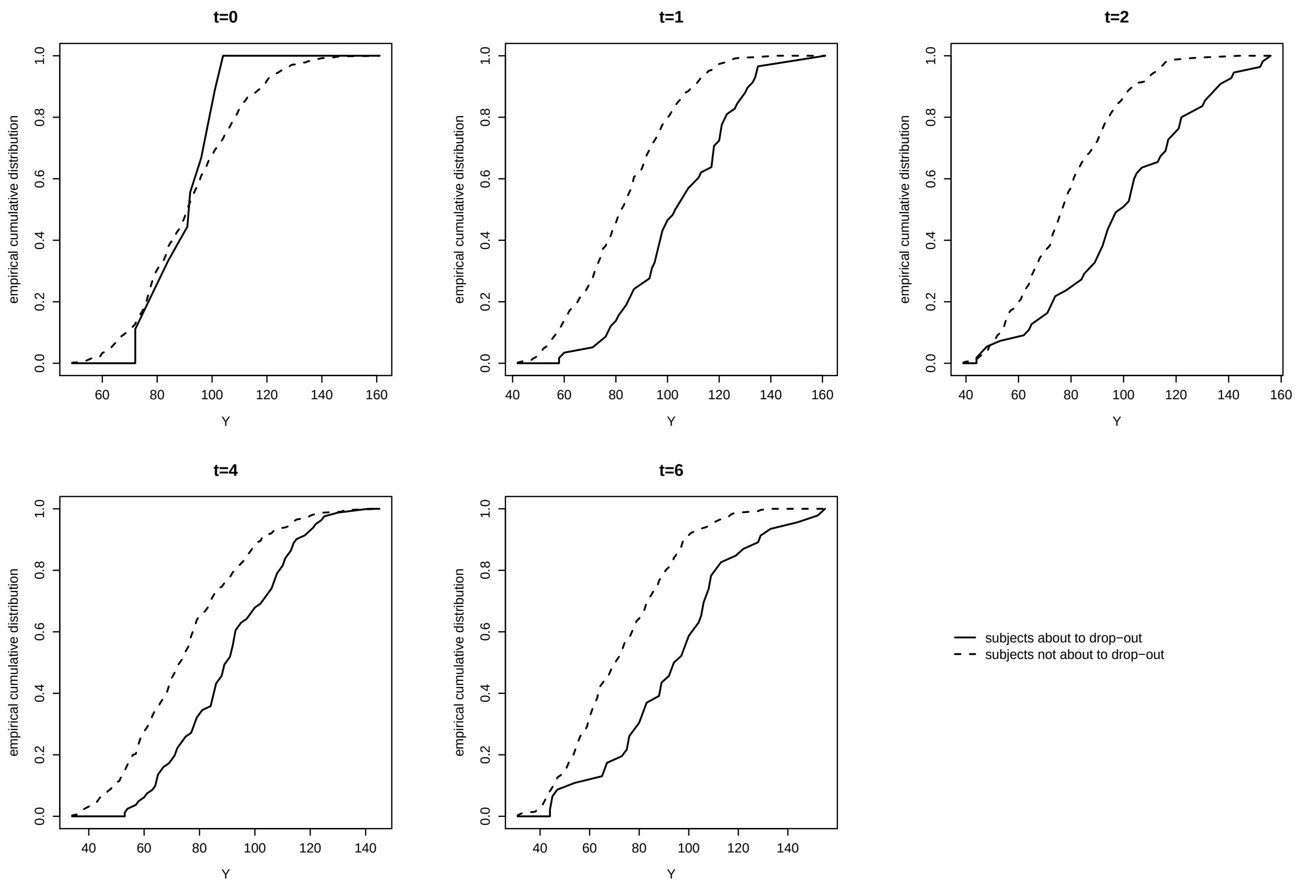
week	−1	0	1	2	4	6	8
missing	0	3	9	70	122	205	251
proportion	0.00	0.01	0.02	0.14	0.24	0.40	0.49

- Dropout rate also treatment-dependent ($P > H > R$)

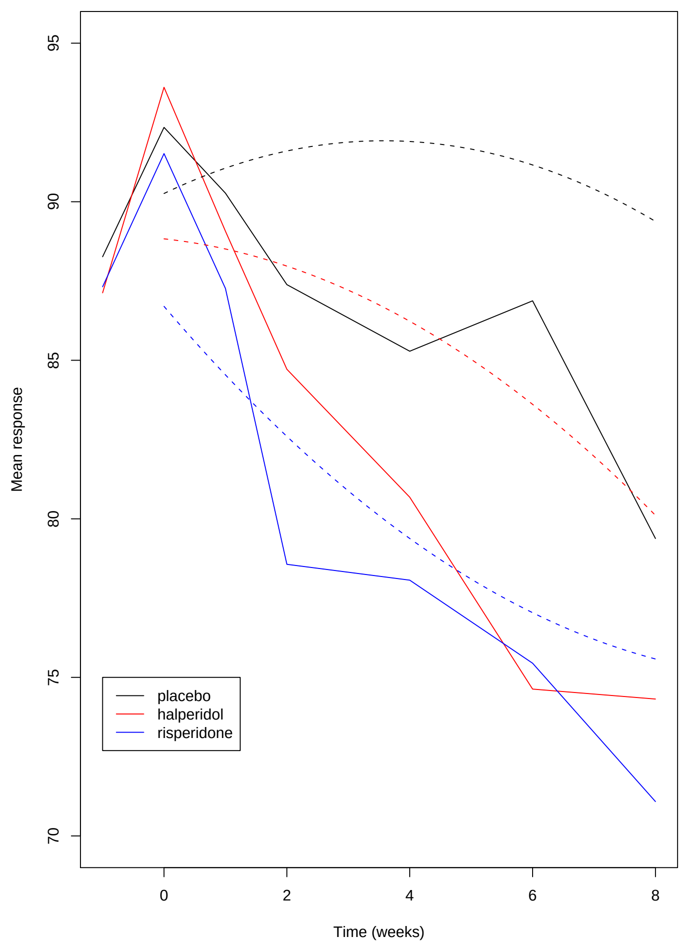
Schizophrenia data
PANSS responses from haloperidol arm



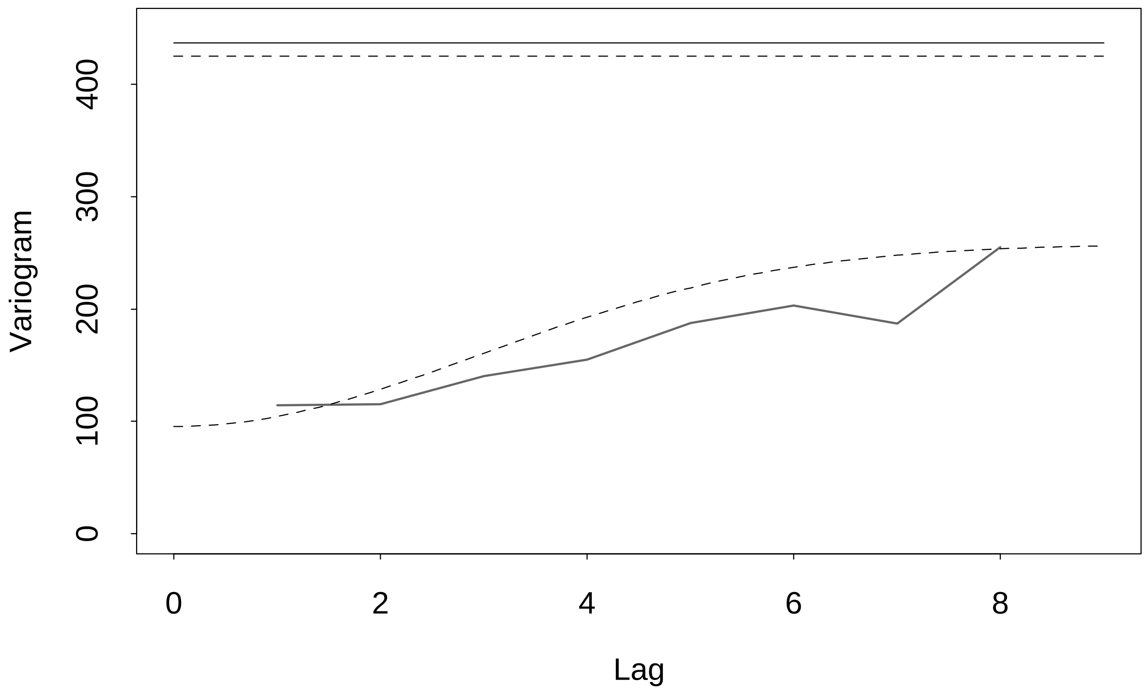
Dropout is not completely at random



Schizophrenia trial data
Mean response (random effects model)



Schizophrenia trial data
Empirical and fitted variograms



Generalized linear models for longitudinal data

- **random effects models**
- **transition models**
- **marginal models**

Diggle, Heagerty, Liang and Zeger (2002, Chapter 7)

Random effects GLM

Responses $Y_1, \ldots, Y_n$ on an individual subject conditionally independent, given unobserved vector of random effects U

$U \sim g(u)$ represents properties of individual subjects that vary randomly between subjects

- $E(Y_j|U) = \mu_j : h(\mu_i) = \mathrm{x}_j'\beta + U'\alpha$
- $\mathrm{Var}(Y_j|U) = \phi v(\mu_j)$
- **$(Y_1, \ldots, Y_n)$ are mutually independent conditional on U.**

Likelihood inference requires evaluation of

$$f(y) = \int \prod_{j=1}^{n} f(y_j|U)g(U)dU$$

Transition GLM

Each Y_j modelled conditionally on preceding $Y_1, Y_2, \ldots, Y_{i-1}$.

- $E(Y_j|\text{history}) = \mu_i$
- $h(\mu_j) = \mathrm{x}_j'\beta + \sum_{k=1}^{j-1} Y_{j-k}'\alpha_k$
- $\mathrm{Var}(Y_j|\text{history}) = \phi v(\mu_j)$

Construct likelihood as product of conditional distributions, usually assuming restricted form of dependence.
Example:

$$f_k(y_j|y_1, ..., y_{j-1}) = f_k(y_j|y_{j-1})$$

and condition on y_1 as model does not directly specify $f_1(y_1)$.

Marginal GLM

Let $h(\cdot)$ be a link function which operates component-wise,

- $E(y) = \mu : h(\mu) = X\beta$
- $\text{Var}(y_i) = \phi v(\mu_i)$
- $\text{Corr}(y) = R(\alpha)$.

Not a fully specified probability model

May require constraints on variance function $v(\cdot)$ and correlation matrix $R(\cdot)$ for valid specification

Inference for β uses generalized estimating equations

Liang and Zeger (1986)

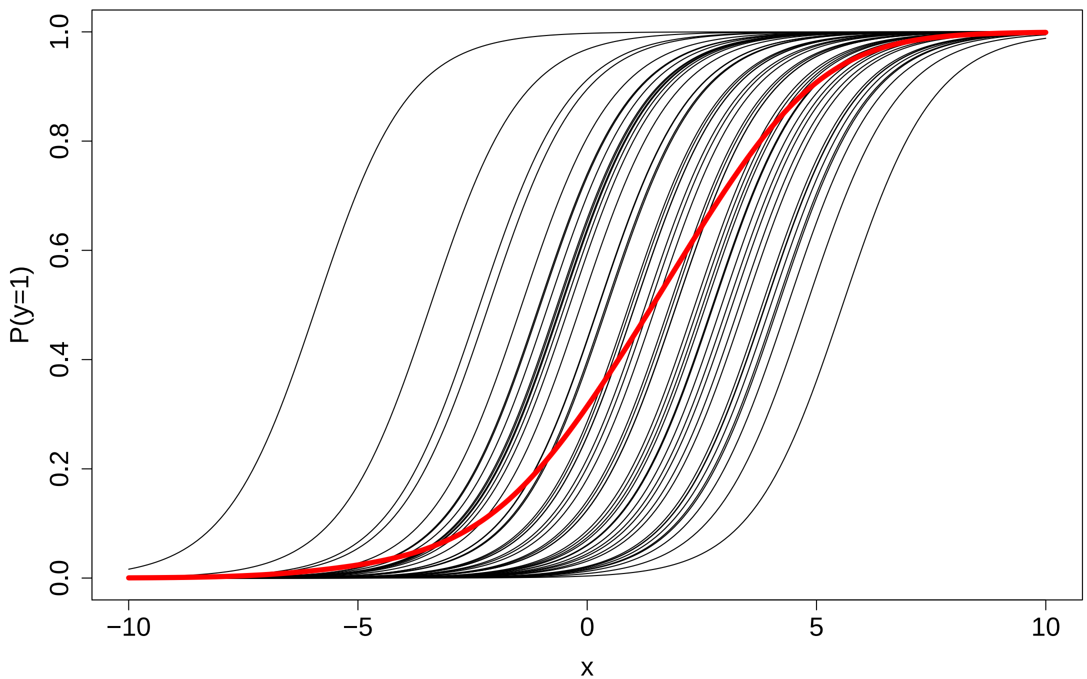
What are we estimating?

- in marginal modelling, β measures population-averaged effects of explanatory variables on mean response
- in transition or random effects modelling, β measures effects of explanatory variables on mean response of an individual subject, conditional on
  - subject's measurement history (transition model)
  - subject's own random characteristics U_i (random effects model)

Example: **Simulation of a logistic regression model, probability of positive response from subject i at time t is $p_i(t)$,**

$$\mathrm{logit}\{p_i(t)\} : \alpha + \beta x(t) + \gamma U_i,$$

$x(t)$ is a continuous covariate and U_i is a random effect



Example: Effect of mother's smoking on probability of intra-uterine growth retardation (IUGR).

Consider a binary response $Y = 1/0$ to indicate whether a baby experiences IUGR, and a covariate x to measure the mother's amount of smoking.

Two relevant questions:

1. **public health:** by how much might population incidence of IUGR be reduced by a reduction in smoking?

2. **clinical/biomedical:** by how much is a baby's risk of IUGR reduced by a reduction in their mother's smoking?

Question 1 is addressed by a marginal model, question 2 by a random effects model

4. Continuous spatial variation

- **stationary Gaussian processes;**
- **variogram estimation;**
- **likelihood-based estimation;**
- **spatial prediction.**

What is this thing called geostatistics?

biostatistics = bio–statistics

geostatistics ≠ geo–statistics

The core geostatistical problem: given a set of measured values Y_i at locations $x_i \in A$ of some spatial phenomenon $S(\cdot)$, what can you say about the complete surface $\{S(x) : x \in A\}$?

Krige, 1951; Matérn, 1960; Mathéron, 1963; Watson, 1972; Ripley, 1981

Recall from LDA lectures

- Stationary Gaussian process $Y_{ij} - \mu_{ij} = W_i(t_{ij})$

  $W_i(t)$ a continuous-time Gaussian process

  $\mathrm{E}[W(t)] = 0$ $\mathrm{Var}\{W(t)\} = \sigma^2$ $\mathrm{Corr}\{W(t), W(t-u)\} = \rho(u)$

- Variogram of a stochastic process $Y(t)$ is

$$V(u) = \frac{1}{2}\mathrm{Var}\{Y(t) - Y(t-u)\}$$

  For stationary processes,

$$V(u) = \sigma^2\{1 - \rho(u)\}$$

For geostatistics, simply substitute a spatial process $S(x)$ for the temporal process $W(t)$, and off you go

Model-based Geostatistics

- **the application of general principles of statistical modelling and inference to geostatistical problems**

- **Example: kriging as minimum mean square error prediction under Gaussian modelling assumptions**

Computation with geoR

```
library(geoR)
lead<-read.table("lead2000_data.txt",header=T)
lead<-as.geodata(lead)
summary(lead)
plot(lead)
?points.geodata
points(lead,cex.min=1,cex.max=4)
points(lead,cex.min=0.5,cex.max=2)
points(lead,cex.min=0.5,cex.max=2,pt.div="quint")
loglead<-lead
loglead$data<-log(loglead$data)
points(loglead,cex.min=0.5,cex.max=2,pt.div="quint")
```

Notation

- $Y = \{Y_i : i = 1, ..., n\}$ is the measurement data
- $\{x_i : i = 1, ..., n\}$ is the sampling design
- A is the region of interest
- $Y = \{Y(x) : x \in A\}$ is the measurement process
- $S^* = \{S(x) : x \in A\}$ is the signal process
- $T = \mathcal{F}(S^*)$ is the target for prediction
- $[S^*, Y] = [S^*][Y|S^*]$ is the geostatistical model

Gaussian model-based geostatistics

Model specification:

- **Stationary Gaussian process** $S(x) : x \in \mathbb{R}^2$
  - $\mathrm{E}[S(x)] = \mu$
  - $\mathrm{Cov}\{S(x), S(x')\} = \sigma^2 \rho(\|x - x'\|)$
- **Mutually independent** $Y_i | S(\cdot) \sim \mathrm{N}(S(x), \tau^2)$

Minimum mean square error prediction

$$[S, Y] = [S][Y|S]$$

- $\hat{T} = t(Y)$ is a point predictor
- $\text{MSE}(\hat{T}) = \text{E}[(\hat{T} - T)^2]$

Theorem: $MSE(\hat{T})$ takes its minimum value when $\hat{T} = \text{E}(T|Y)$.

Proof uses result that for any predictor $\tilde{T}$,

$$\text{E}[(T - \tilde{T})^2] = \text{E}_Y[\text{Var}_T(T|Y)] + \text{E}_Y\{[\text{E}_T(T|Y) - \tilde{T}]^2\}$$

Immediate corollary is that

$$\text{E}[(T - \hat{T})^2] = \text{E}_Y[\text{Var}(T|Y)] \approx \text{Var}(T|Y)$$

Simple and ordinary kriging

Recall Gaussian model:

- **Stationary Gaussian process** $S(x) : x \in \mathbb{R}^2$
  - $\mathrm{E}[S(x)] = \mu$
  - $\mathrm{Cov}\{S(x), S(x')\} = \sigma^2 \rho(\|x - x'\|)$
- **Mutually independent** $Y_i | S(\cdot) \sim \mathrm{N}(S(x), \tau^2)$

Gaussian model implies

$$Y \sim \text{MVN}(\mu 1, \sigma^2 V)$$

$$V = R + (\tau^2/\sigma^2)I \qquad R_{ij} = \rho(\|x_i - x_j\|)$$

Target for prediction is $T = S(x)$, write $r = (r_1, ..., r_n)$ where

$$r_i = \rho(\|x - x_i\|)$$

Standard results on multivariate Normal then give $[T|Y]$ as multivariate Gaussian with mean and variance

$$\hat{T} = \mu + r'V^{-1}(Y - \mu 1) \tag{5}$$

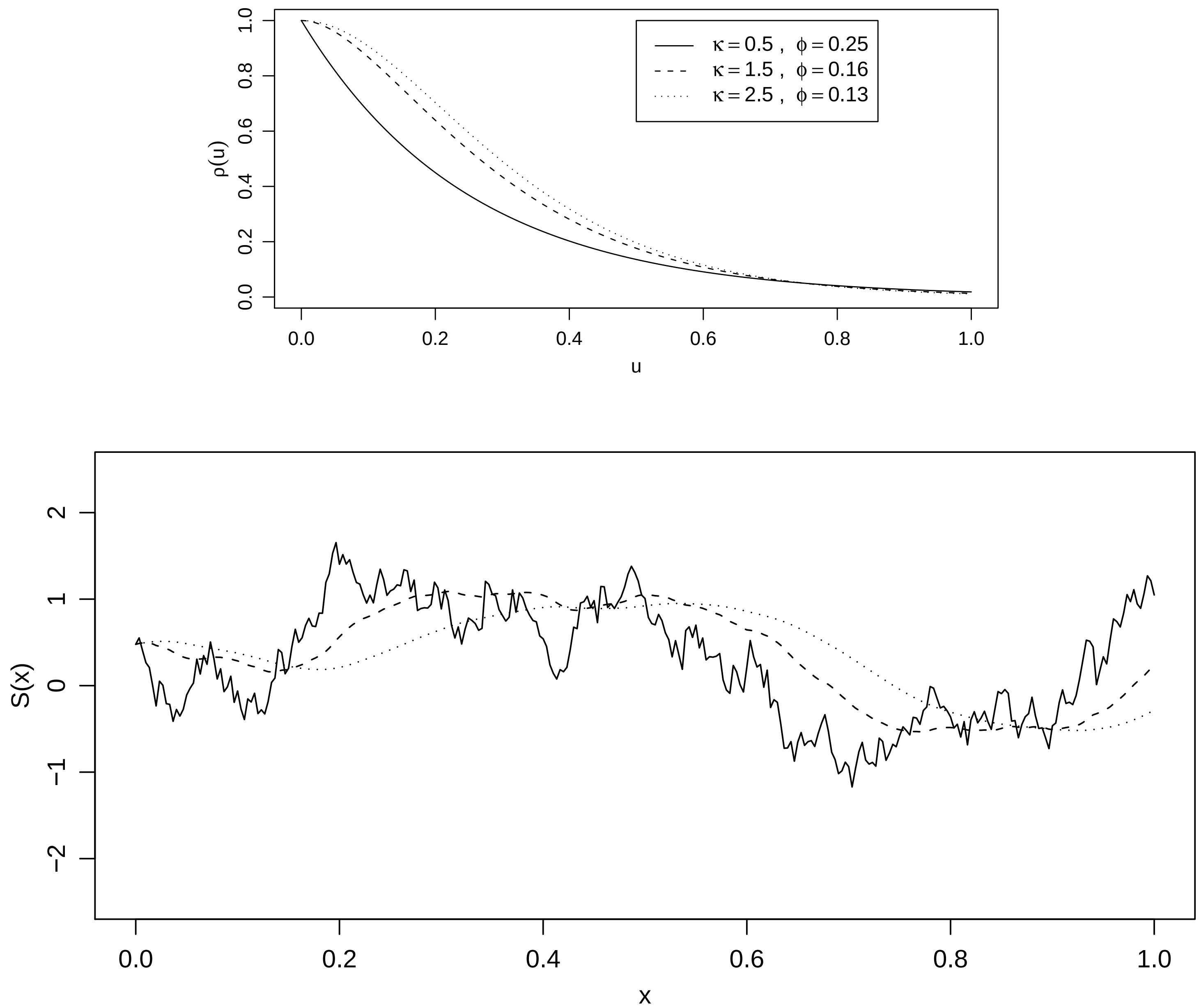
$$\text{Var}(T|Y) = \sigma^2(1 - r'V^{-1}r). \tag{6}$$

Simple kriging: $\hat{\mu} = \bar{Y}$ Ordinary kriging: $\hat{\mu} = (1'V^{-1}1)^{-1}1'V^{-1}Y$

The Matérn family of correlation functions

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

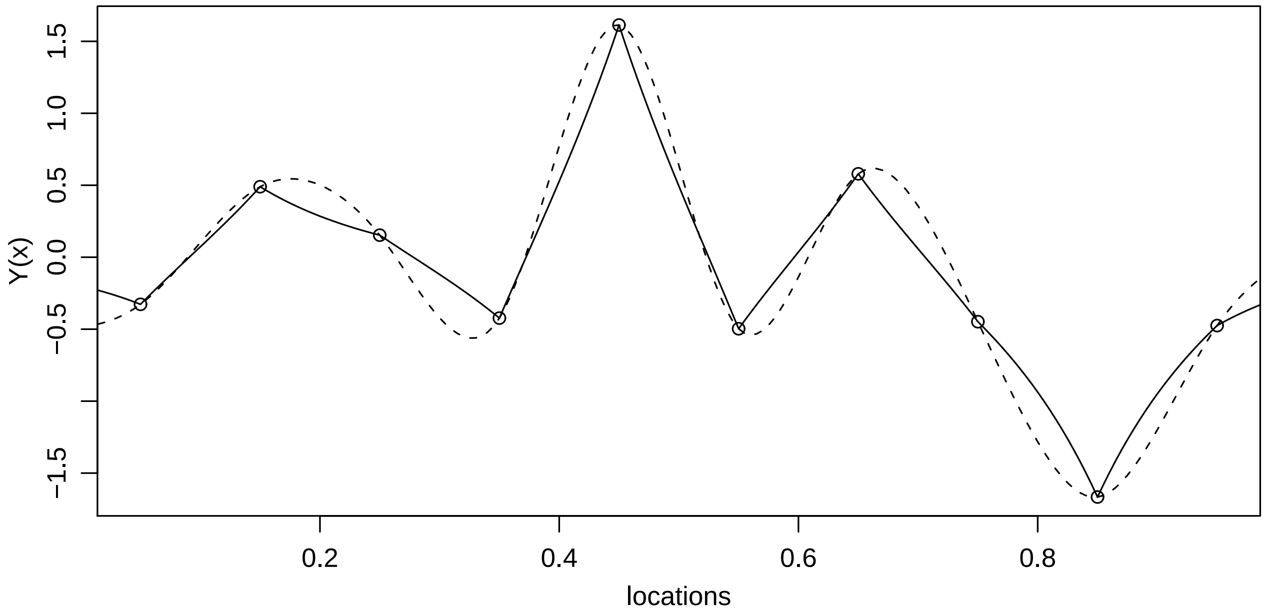
- parameters $\kappa > 0$ and $\phi > 0$
- $K_{\kappa}(\cdot)$: modified Bessel function of order κ
- $\kappa = 0.5$ gives $\rho(u) = \exp\{-u/\phi\}$
- $\kappa \rightarrow \infty$ gives $\rho(u) = \exp\{-(u/\phi)^2\}$
- κ and ϕ are not orthogonal:
  - helpful re-parametrisation: $\alpha = 2\phi\sqrt{\kappa}$
  - but estimation of κ is difficult



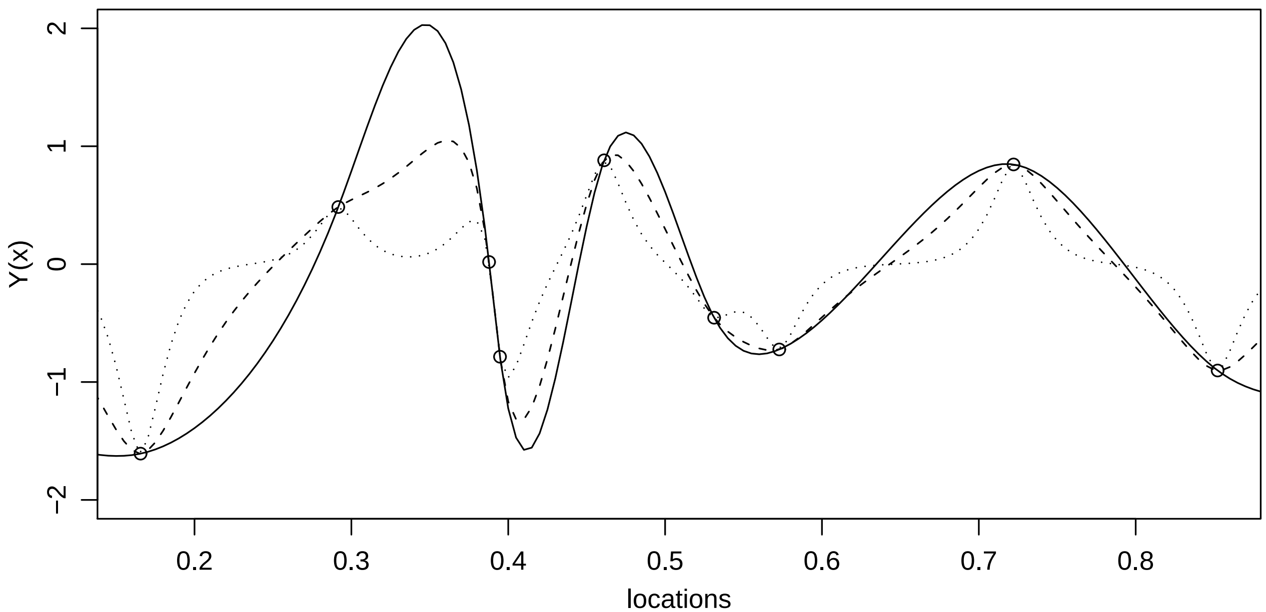
κ controls mean-square differentiability of $S(x)$

Simple kriging: three examples

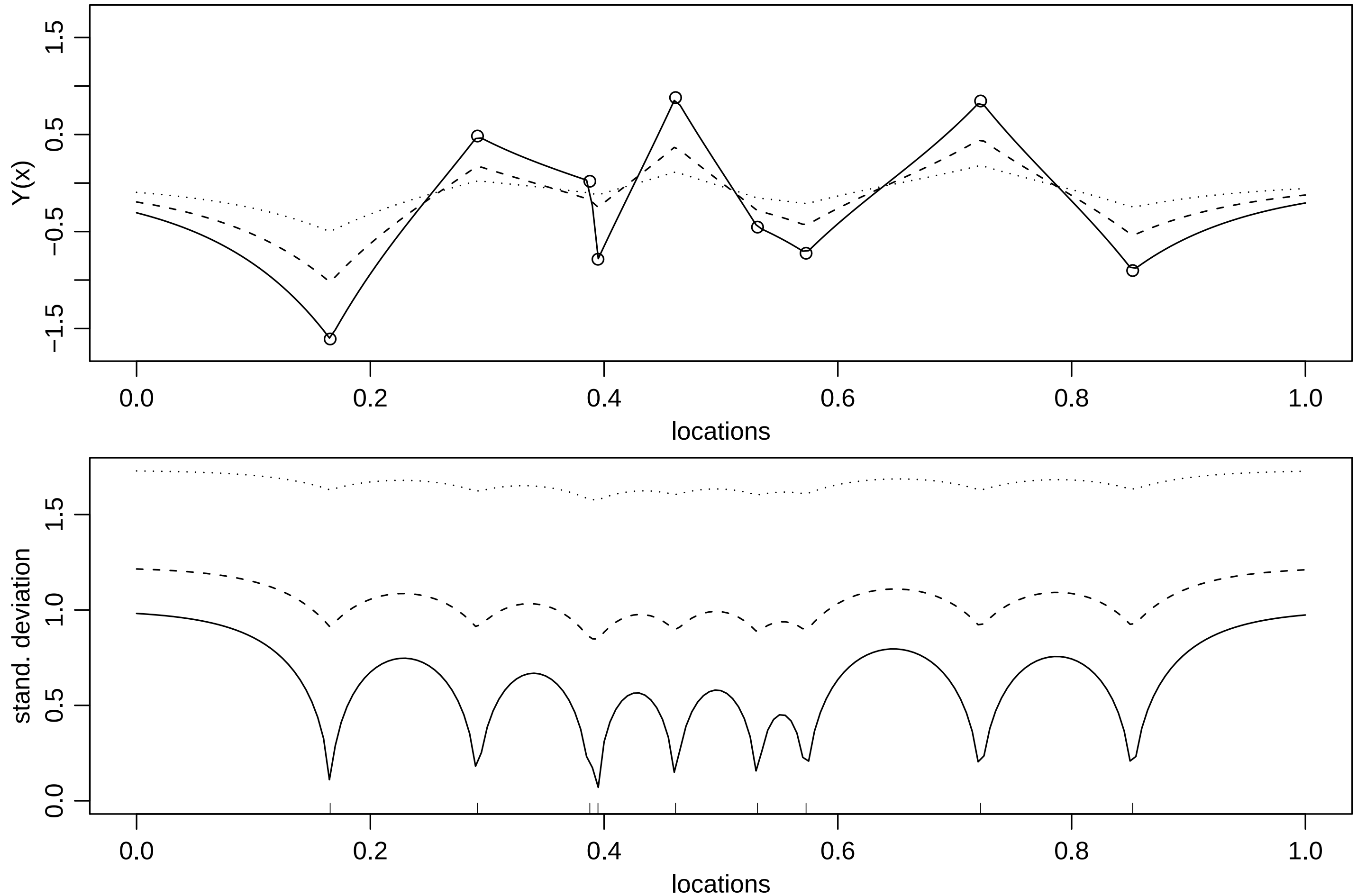
1. Varying κ (smoothness of $S(x)$)



2. Varying ϕ (range of spatial correlation



3. Varying τ^2/σ^2 (noise-to-dignal ratio)



Predicting non-linear functionals

- **minimum mean square error prediction is not invariant under non-linear transformation**

- **the complete answer to a prediction problem is the predictive distribution, $[T|Y]$**

- **Recommended strategy:**

  – **draw repeated samples from $[S^*|Y]$ (conditional simulation)**

  – **calculate required summaries (examples to follow)**

The variogram re-visited

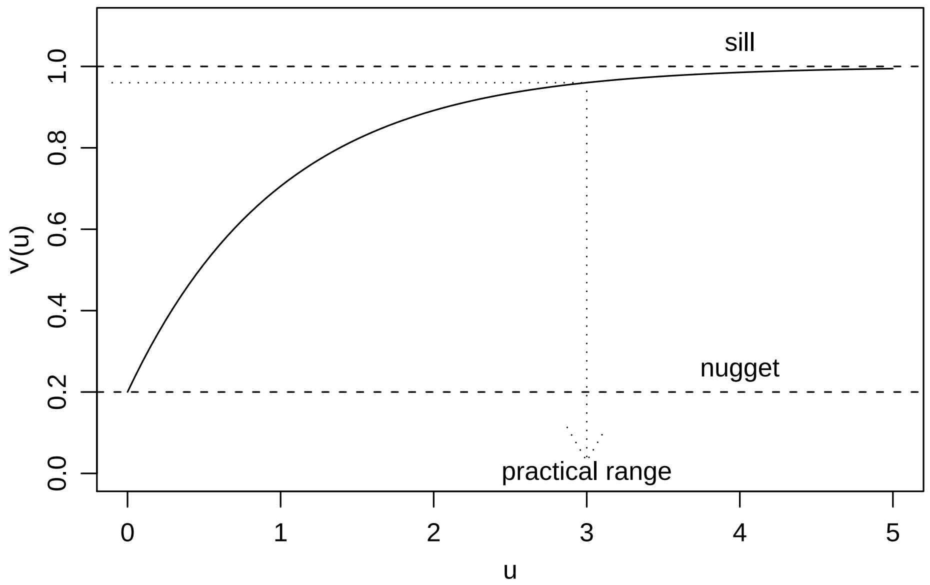
- the variogram of a process $Y(x)$ is the function

$$V(x, x') = \frac{1}{2}\text{Var}\{Y(x) - Y(x')\}$$

- for the spatial Gaussian model, with $u = ||x - x'||$,

$$V(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}$$

- provides a summary of the basic structural parameters of the spatial Gaussian process



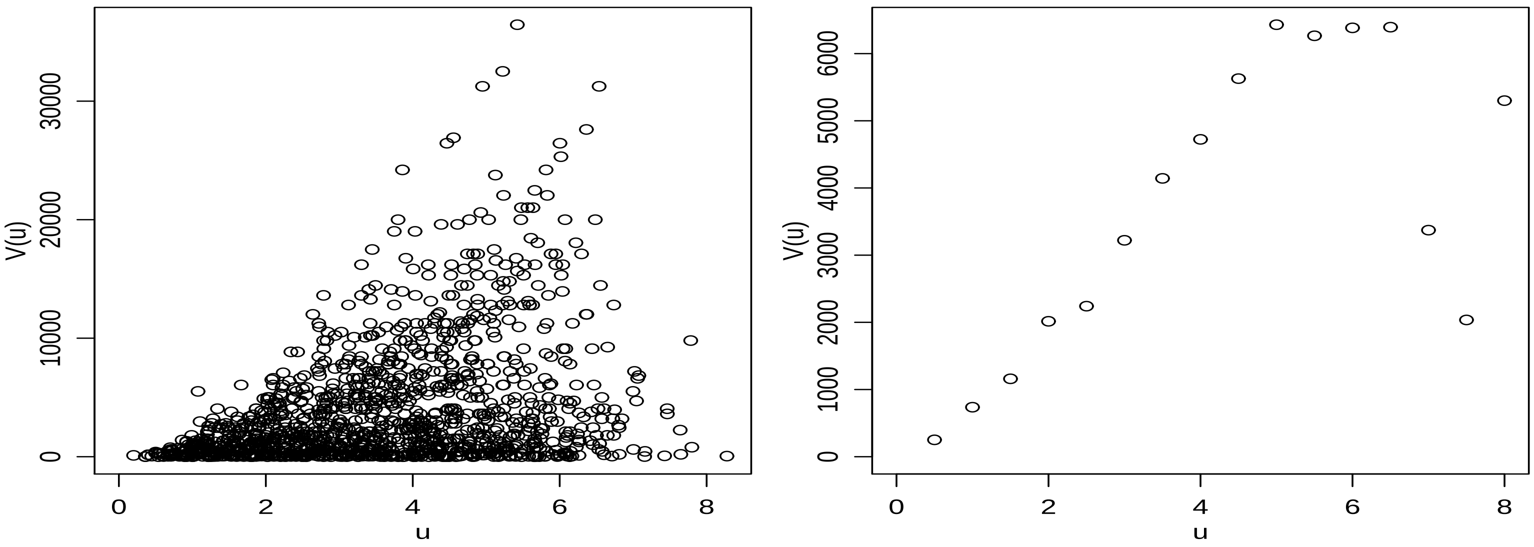
- **the nugget variance: τ^2**
- **the sill: $\sigma^2 = \mathrm{Var}\{S(x)\}$**
- **the practical range: ϕ, such $\rho(u) = \rho(u/\phi)$**

Empirical variograms

$$u_{ij} = \|x_i - x)j\| \qquad v_{ij} = 0.5[y(x_i) - y(x_j)]^2$$

- the variogram cloud is a scatterplot of the points (u_{ij}, v_{ij})
- the empirical variogram smooths the variogram cloud by averaging within bins: $u - h/2 \le u_{ij} < u + h/2$
- for a process with non-constant mean (covariates), use residuals $r(x_i) = y(x_i) - \hat{\mu}(x_i)$ to compute v_{ij}

Limitations of $\hat{V}(u)$



1. $v_{ij} \sim V(u_{ij})\chi_1^2$

2. the v_{ij} are correlated

Consequences:

- variogram cloud is unstable, pointwise and in overall shape

- binning addresses point 1, but not point 2

Parameter estimation using the variogram

What not to do and how to do it

- **weighted least squares criterion:**

$$W(\theta) = \sum_k n_k \{[\bar{V}_k - V(u_k;\theta)]\}^2$$

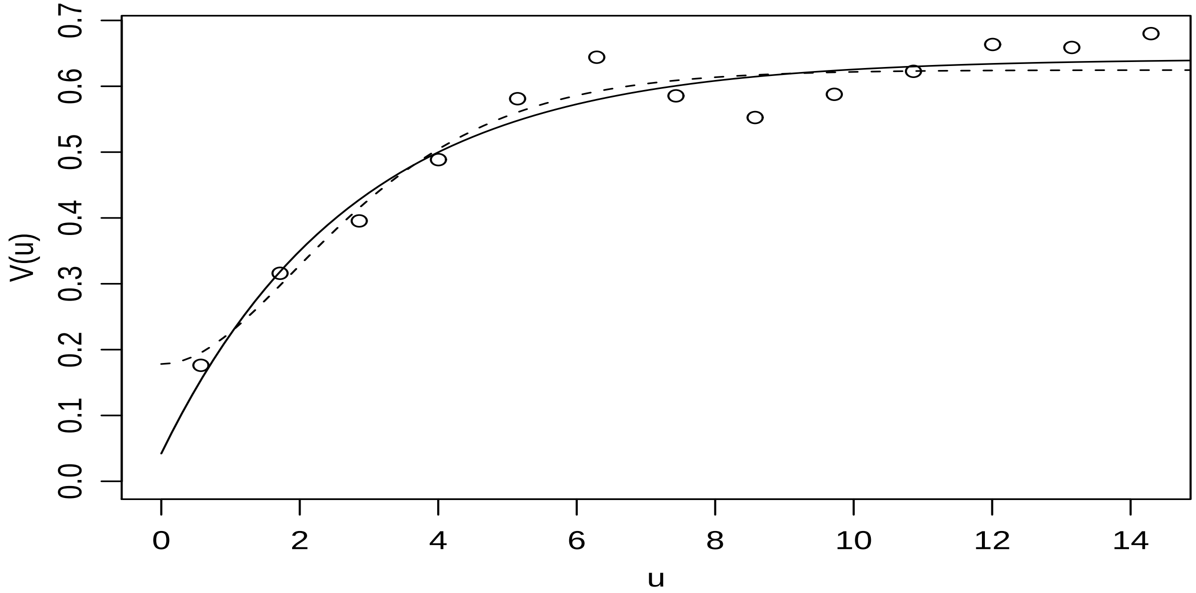
  where θ denotes vector of covariance parameters and $\bar{V}_k$ is average of n_k variogram ordinates v_{ij}.

- **need to choose upper limit for u (arbitrary?)**

- **variations include:**
  - **fitting models to the variogram cloud**
  - **other estimators for the empirical variogram**
  - **different proposals for weights**

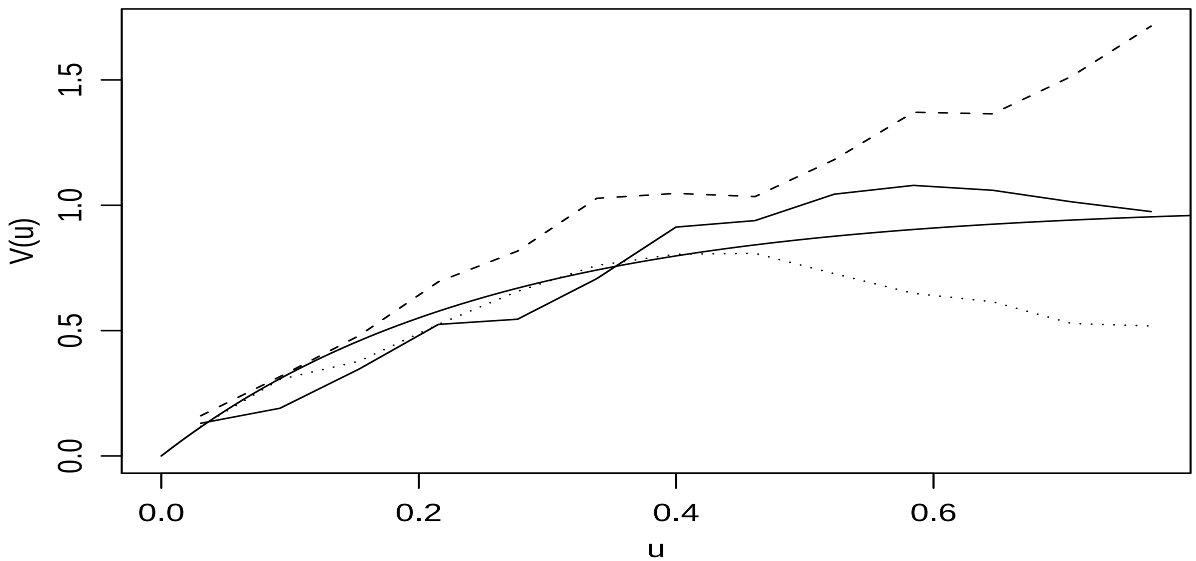
Comments on variogram fitting

1. **Can give equally good fits for different extrapolations at origin.**



2. Correlation between variogram points induces smoothness.

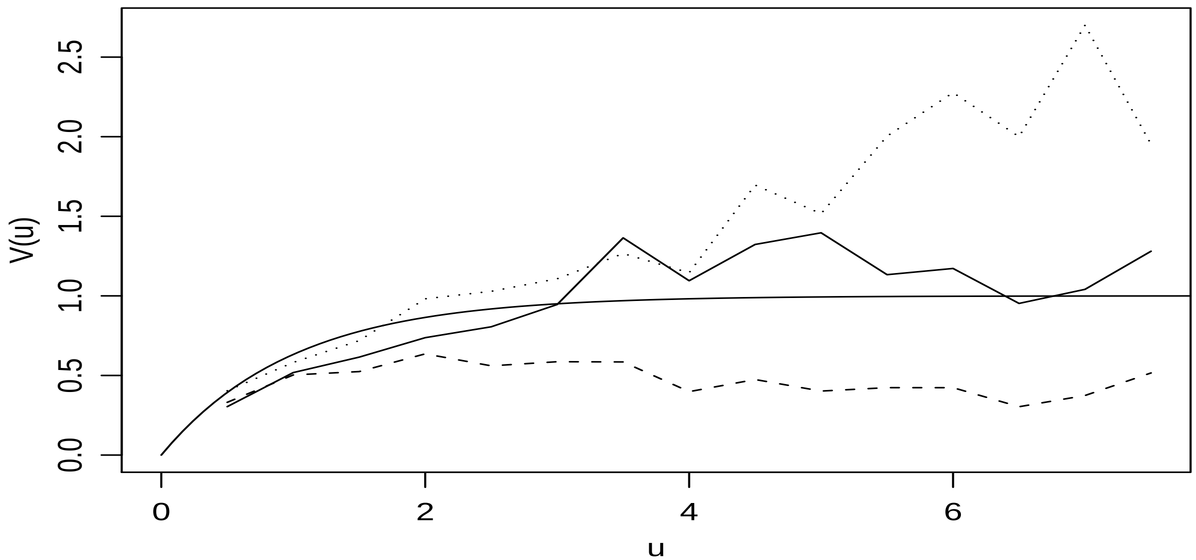
Empirical variograms for three simulations from the same model.



3. Fit is highly sensitive to specification of the mean.

Illustration with linear trend surface:

- **solid smooth line: theoretical variogram;**
- **dotted line: from data;**
- **solid line: from true residuals;**
- **dashed line : from estimated residuals.**



Parameter estimation: maximum likelihood

$$Y \sim \text{MVN}(\mu 1, \sigma^2 R + \tau^2 I)$$

R is the $n \times n$ matrix with $(i,j)^{th}$ element $\rho(u_{ij})$ where $u_{ij} = ||x_i - x_j||$, Euclidean distance between x_i and x_j.

Or more generally:

$$\mu(x_i) = \sum_{j=1}^{k} f_k(x_i)\beta_k$$

where $d_k(x_i)$ is a vector of covariates at location x_i, hence

$$Y \sim \text{MVN}(D\beta, \sigma^2 R + \tau^2 I)$$

Gaussian log-likelihood function:

$$L(\beta, \tau, \sigma, \phi, \kappa) \propto \quad -0.5\{\log|(\sigma^2 R + \tau^2 I)| + (y - D\beta)'(\sigma^2 R + \tau^2 I)^{-1}(y - D\beta)\}.$$

- write $\nu^2 = \tau^2/\sigma^2$, hence $\sigma^2 V = \sigma^2(R + \nu^2 I)$
- log-likelihood function is maximised for

$$\hat{\beta}(V) = (D'V^{-1}D)^{-1}D'V^{-1}y$$
$$\hat{\sigma}^2 = n^{-1}(y - D\hat{\beta})'V^{-1}(y - D\hat{\beta})$$

- substitute $(\hat{\beta}, \hat{\sigma^2})$ to give reduced maximisation problem

$$L^*(\nu^2, \phi, \kappa) \propto -0.5\{n \log|\hat{\sigma^2}| + \log|(R + \nu^2 I)|\}$$

- usually just consider κ in a discrete set $\{0.5, 1, 2, 3, ..., N\}$

Comments on maximum likelihood

- likelihood-based methods preferable to variogram-based methods
- restricted maximum likelihood is widely recommended but in our experience is sensitive to mis-specification of the mean model.
- in spatial models, distinction between $\mu(x)$ and $S(x)$ is not sharp.
- composite likelihood treats contributions from pairs (Y_i, Y_j) as if independent
- approximate likelihoods useful for handling large data-sets
- examining profile likelihoods is advisable, to check for poorly identified parameters

Trans-Gaussian models

- assume Gaussian model holds after point-wise transformation

- Box-Cox family is widely used

$$Y_i^* = h_\lambda(Y_i) = \begin{cases} (Y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(Y_i) & \text{if } \lambda = 0 \end{cases}$$

- bias-correction? only if point prediction is required

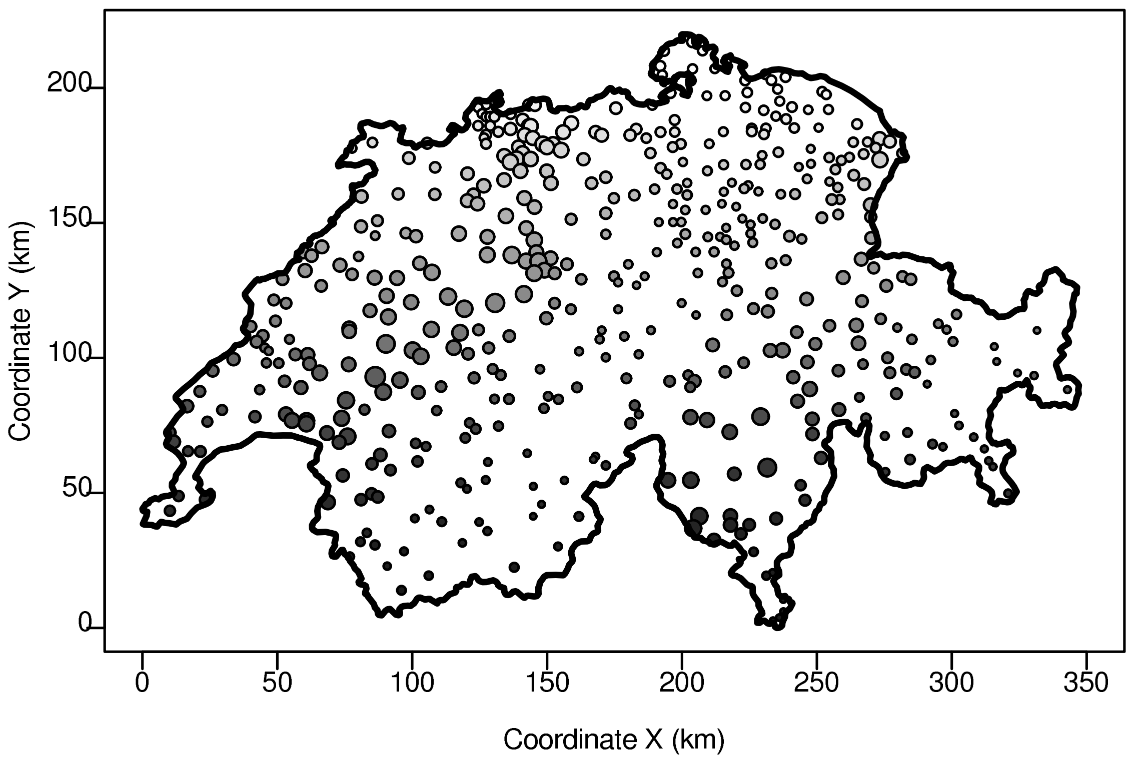
Example: log-Gaussian kriging

- $T(x) = \exp\{S(x)\} \quad \hat{T}(x) = \exp\{\hat{S}(x) + v(x)/2\}$

- $S_1, \ldots, S_m$ are a sample from $[S|Y]$

- $T_i = \exp(S_i) \Rightarrow T_1, \ldots, T_m$ are a sample from $[T|Y]$

Swiss rainfall data



Swiss rainfall: trans-Gaussian model

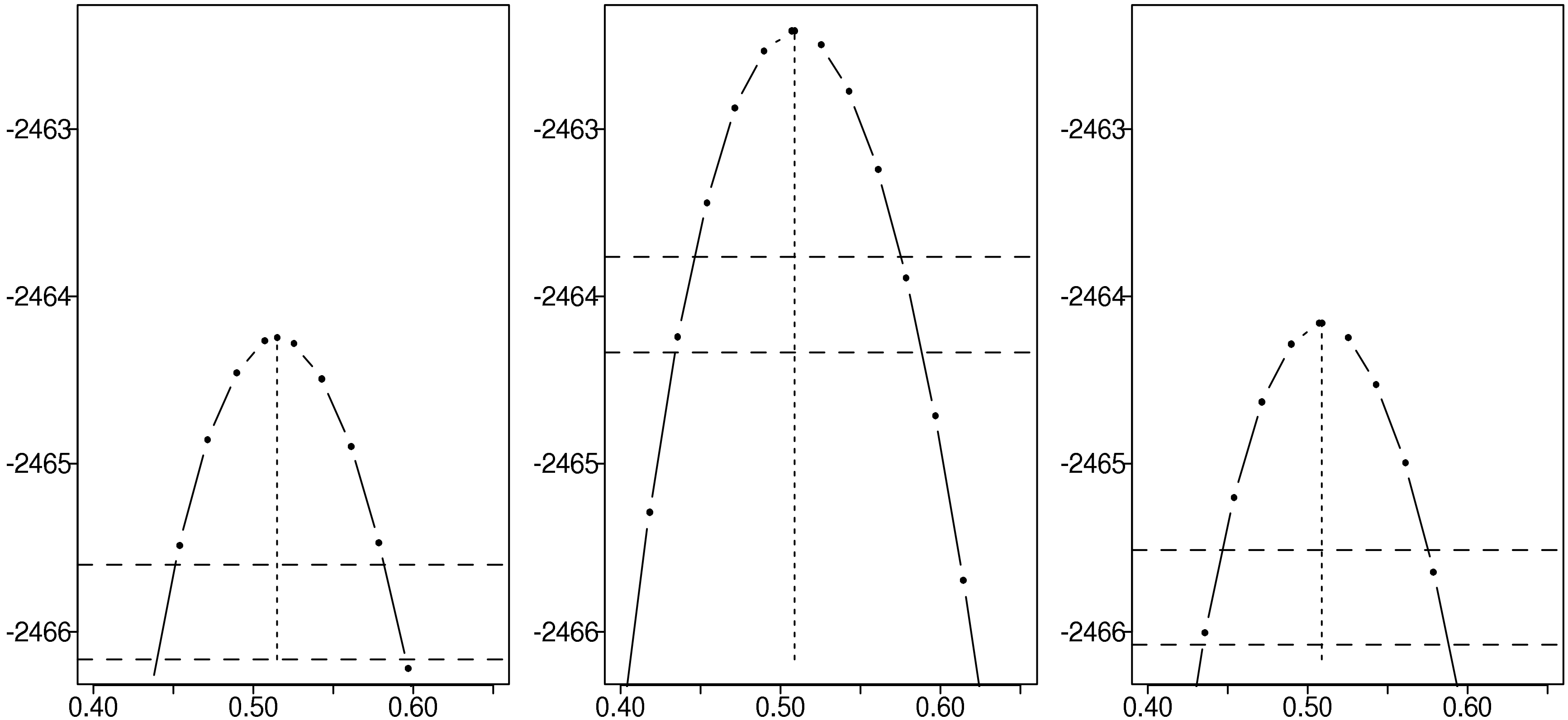
$$Y_i^* = h_\lambda(Y_i) = \begin{cases} (Y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(Y_i) & \text{if } \lambda = 0 \end{cases}$$

For log-likelihood, write $h_\lambda = h_\lambda(Y_1), ..., h_\lambda(Y_n)$,

$$\begin{aligned} \ell(\beta, \theta, \lambda) \quad = \quad & -\frac{1}{2}\{\log|\sigma^2 V| + (h_\lambda - D\beta)'\{\sigma^2 V\}^{-1}(h_\lambda - D\beta)\} \\ & + \sum_{i=1}^{n} \log\left((Y_i)^{\lambda-1}\right) \end{aligned}$$

Swiss rainfall: profile log-likelihoods for λ

Left panel: $\kappa = 0.5$ Centre panel: $\kappa = 1$ Right panel: $\kappa = 2$

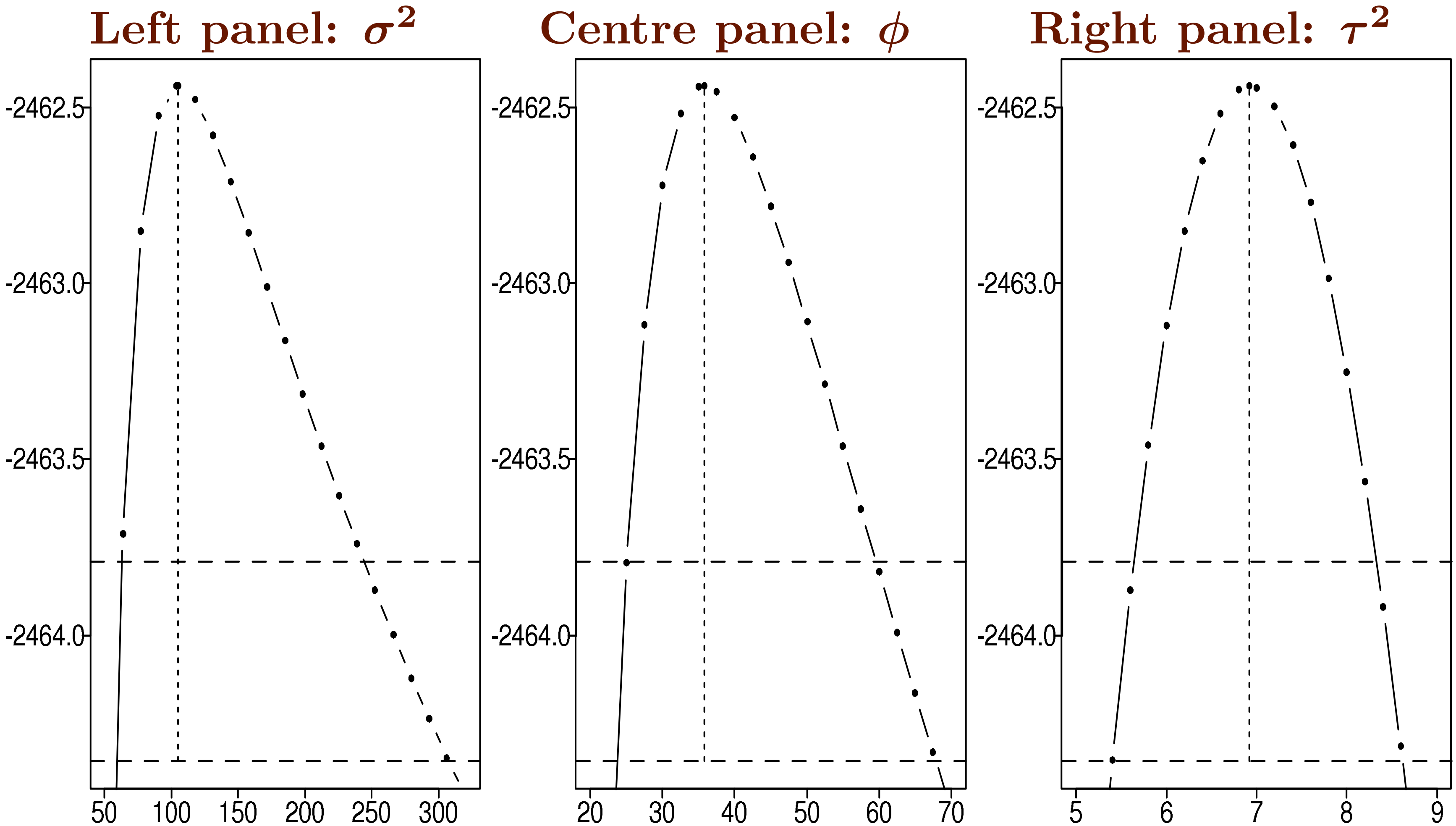


Swiss rainfall: MLE's ($\lambda = 0.5$)

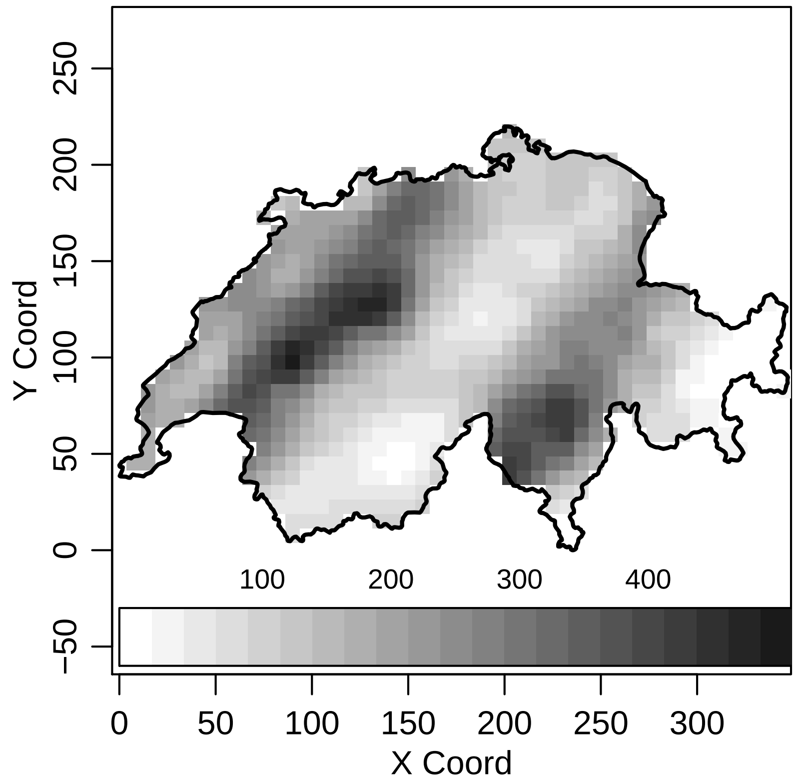
κ	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	$\log \hat{L}$
0.5	18.36	118.82	87.97	2.48	-2464.315
1	20.13	105.06	35.79	6.92	-2462.438
2	21.36	88.58	17.73	8.72	-2464.185

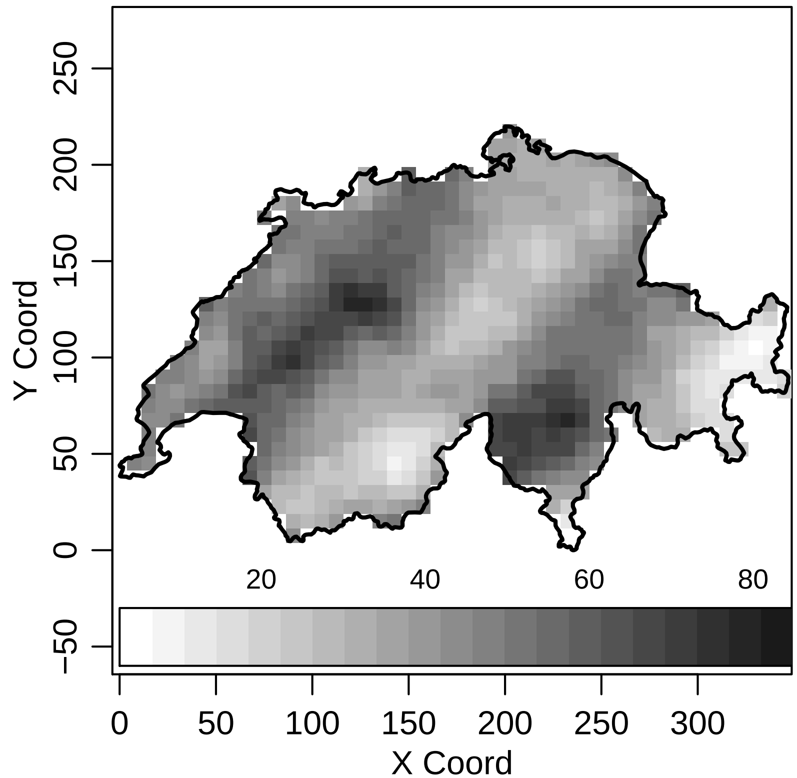
Likelihood criterion favours $\kappa = 1$

Swiss rainfall: profile log-likelihoods ($\lambda = 0.5$, $\kappa = 1$)

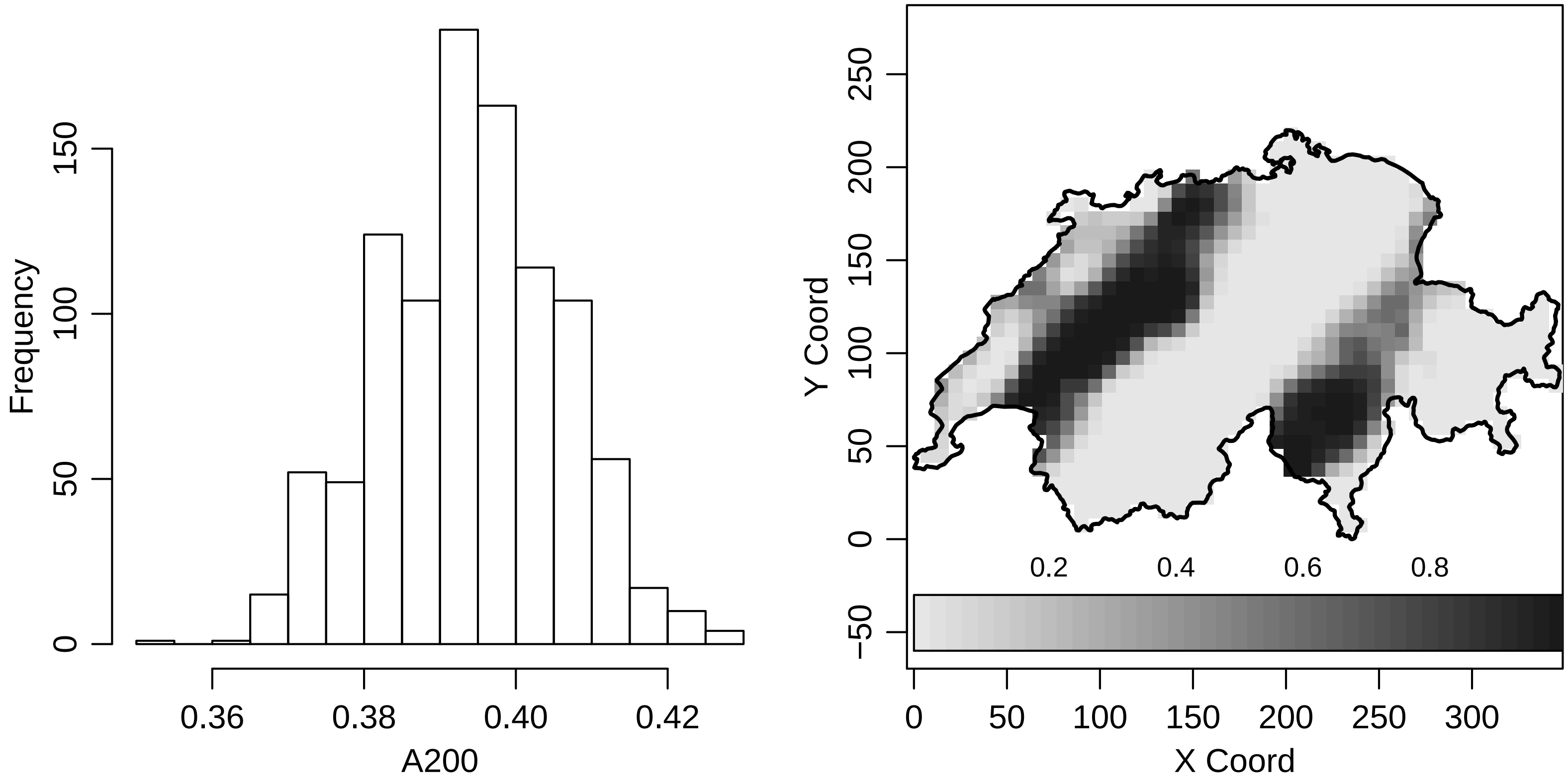


Swiss rainfall: plug-in predictions and prediction variances





Swiss rainfall: non-linear prediction



Left-panel: plug-in prediction for proportion of total area with rainfall exceeding 200 (= 20mm)

Right-panel: plug-in predictive map of $P(\text{rainfall} > 250|Y)$

Computation with geoR

```
vario1<-variog(loglead,uvec=5000*(0:30))
plot(vario1)
plot(vario1,pch=19,col="red")
?variog
vario2<-variog(loglead,uvec=5000*(0:30),trend="1st")
plot(vario2)
names(vario1)
plot(vario1$u,vario1$v,type="l",xlim=c(0,150000),ylim=c(0,0.25),
   xlab="u",ylab="V(u)")
lines(vario2$u,vario2$v,col="red")
```

```
loglead2<-loglead; loglead2$coords<-loglead$coords/100000
mlfit<-likfit(loglead2,ini.cov.pars=c(0.25,1),
   cov.model="matern",kappa=0.5)
region<-matrix(c(4.5,46.0,7.0,46.0,7.0,48.5,4.5,48.5),4,2,T)
grid<-pred_grid(region,by=0.1)
KC<-krige.control(obj.model=mlfit)
OC<-output.control(n.predictive=100)
set.seed(24367)
predictions<-krige.conv(geodata=loglead2,locations=grid,
   borders=region,krige=KC,output=OC)
image(predictions)
points(loglead2,add=T)
par(mfrow=c(1,2))
hist(loglead2$data,main="data")
predict.max<-NULL
for (sim in 1:100) {
   predict.max<-c(predict.max,max(predictions$simulations[,sim]))
   }
hist(predict.max,main="predicted maximum")
```

Bayesian inference: basics

Model specification

$$[\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\theta}] = [\boldsymbol{\theta}][\boldsymbol{S}|\boldsymbol{\theta}][\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\theta}]$$

Parameter estimation

- **integration gives**

$$[\boldsymbol{Y}, \boldsymbol{\theta}] = \int [\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\theta}] d\boldsymbol{S}$$

- **Bayes' Theorem gives posterior distribution**

$$[\boldsymbol{\theta}|\boldsymbol{Y}] = [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}]/[\boldsymbol{Y}]$$

- **where** $[\boldsymbol{Y}] = \int [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}] d\boldsymbol{\theta}$

Prediction: $S \rightarrow S^*$

- **expand model specification to**

$$[Y, S^*, \theta] = [\theta][S|\theta][Y|S,\theta][S^*|S,\theta]$$

- **plug-in predictive distribution is**

$$[S^*|Y, \hat{\theta}]$$

- **Bayesian predictive distribution is**

$$[S^*|Y] = \int [S^*|Y,\theta][\theta|Y]d\theta$$

- **for any target $T = t(S^*)$, required predictive distribution $[T|Y]$ follows**

Notes

- **likelihood function is central to both classical and Bayesian inference**
- **Bayesian prediction is a weighted average of plug-in predictions, with different plug-in values of θ weighted according to their conditional probabilities given the observed data.**
- **Bayesian prediction is usually more conservative than plug-in prediction**

Bayesian computation

1. **Evaluating the integral which defines $[S^*|Y]$ is often difficult**

2. **Markov Chain Monte Carlo methods are widely used**

3. **but for geostatistical problems, reliable implementation of MCMC is not straightforward (no natural Markovian structure)**

4. **for the Gaussian model, direct simulation is available**

Gaussian models: known (σ^2, ϕ)

$$Y \sim \mathrm{N}(D\beta, \sigma^2 R(\phi))$$

- choose conjugate prior $\beta \sim \mathrm{N}\left(m_\beta \; ; \; \sigma^2 V_\beta\right)$
- posterior for β is $\left[\beta | Y, \sigma^2, \phi\right] \sim \mathrm{N}\left(\hat{\beta}, \sigma^2 \, V_{\hat{\beta}}\right)$

$$\begin{aligned} \hat{\beta} &= (V_\beta^{-1} + D'R^{-1}D)^{-1}(V_\beta^{-1} m_\beta + D'R^{-1}y) \\ V_{\hat{\beta}} &= \sigma^2 \, (V_\beta^{-1} + D'R^{-1}D)^{-1}) \end{aligned}$$

- predictive distribution for S^* is

$$p(S^* | Y, \sigma^2, \phi) = \int p(S^* | Y, \beta, \sigma^2, \phi) \, p(\beta | Y, \sigma^2, \phi) \, d\beta.$$

Notes

- **mean and variance of predictive distribution can be written explicitly (but not given here)**
- **predictive mean compromises between prior and weighted average of Y**
- **predictive variance has three components:**
  - **a priori variance,**
  - **minus information in data**
  - **plus uncertainty in $\boldsymbol{\beta}$**
- **limiting case $V_{\boldsymbol{\beta}} \to \infty$ corresponds to ordinary kriging.**

Gaussian models: unknown (σ^2, ϕ)

Convenient choice of prior is:

$$[\beta|\sigma^2, \phi] \sim \mathrm{N}\left(m_b, \sigma^2 V_b\right) \quad [\sigma^2|\phi] \sim \chi^2_{ScI}\left(n_\sigma, S^2_\sigma\right) \quad [\phi] \sim \text{arbitrary}$$

- results in explicit expression for $[\beta, \sigma^2|Y, \phi]$ and computable expression for $[\phi|Y]$ whose form depends on choice of prior for ϕ
- in practice, use arbitrary discrete prior for ϕ and combine posteriors conditional on ϕ by weighted averaging

Algorithm 1:

1. choose lower and upper bounds for ϕ, assign a discrete uniform prior for ϕ over the chosen range

2. compute posterior $[\phi|Y]$ on this discrete support set

3. sample ϕ from posterior, $[\phi|Y]$

4. attach sampled value of ϕ to conditional posterior, $[\beta, \sigma^2|y, \phi]$, and sample (β, σ^2) from this distribution

5. repeat steps (3) and (4) as many times as required, to generate a sample from the joint posterior, $[\beta, \sigma^2, \phi|Y]$

Predictive distribution $[S^*|Y, \phi]$ is tractable, hence write

$$p(S^*|Y) = \int p(S^*|Y, \phi)\, p(\phi|y)\, d\phi = \mathbf{E}_{\phi|Y}[p(S^*|Y, \phi)]$$

Algorithm 2:

1. **discretise $[\phi|Y]$, as in Algorithm 1.**

2. **compute posterior $[\phi|Y]$**

3. **sample ϕ from posterior $[\phi|Y]$**

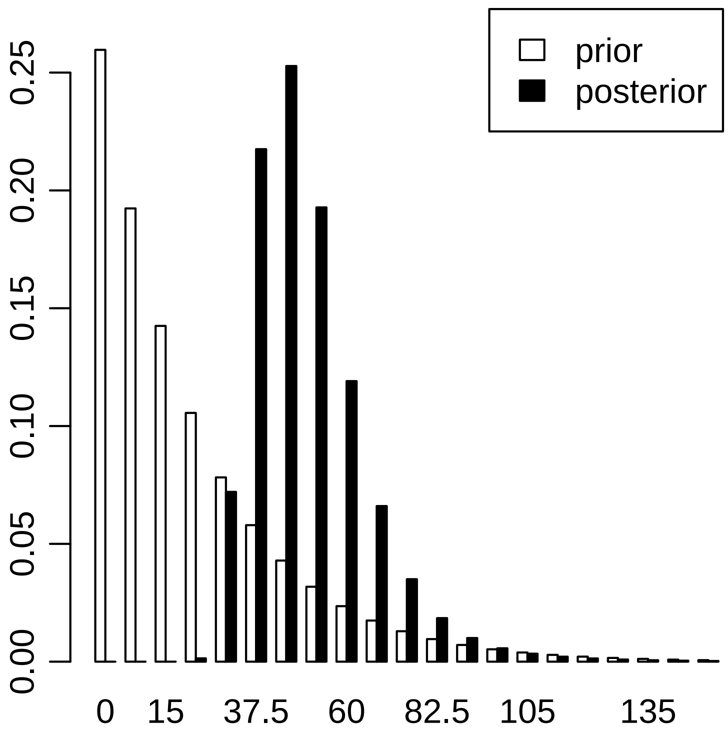
4. **attach sampled value of ϕ to $[S^*|y,\phi]$ and sample from this to obtain realisations from $[S^*|Y]$**

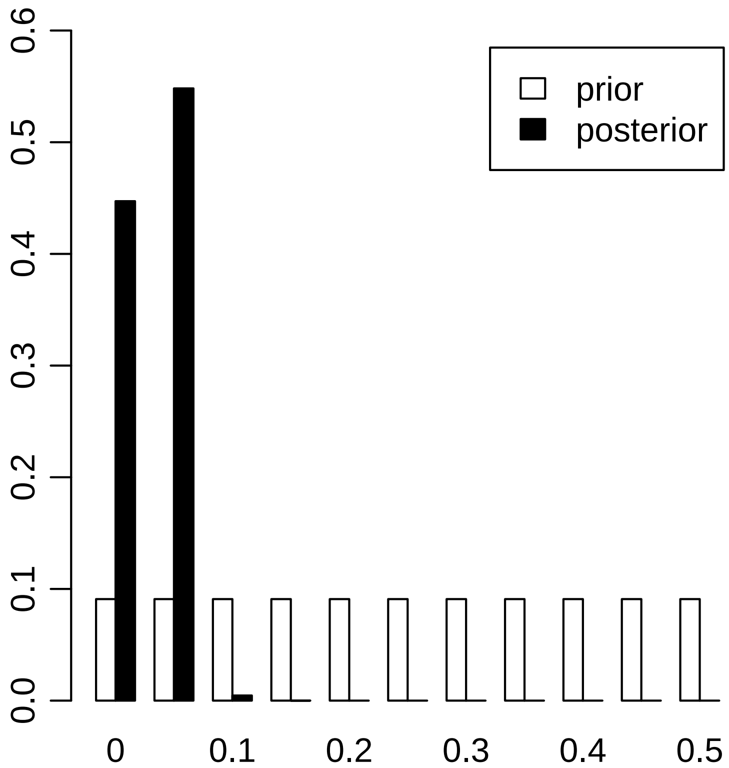
5. **repeat steps (3) and (4) as required**

Note: Extends immediately to multivariate ϕ (but may be computationally awkward)

Swiss rainfall

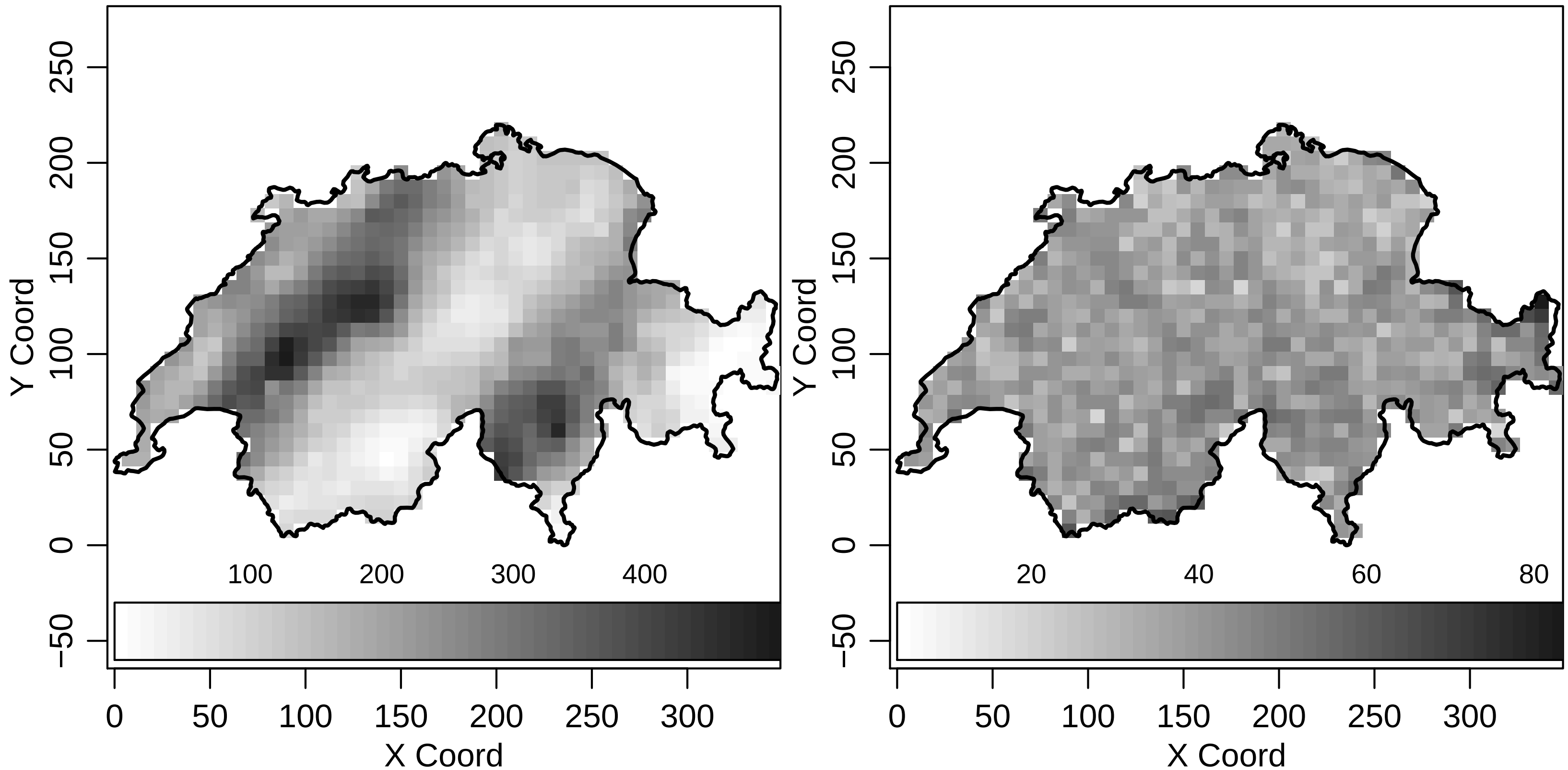
Priors/posteriors for ϕ (left) and ν^2 (right)





Swiss rainfall

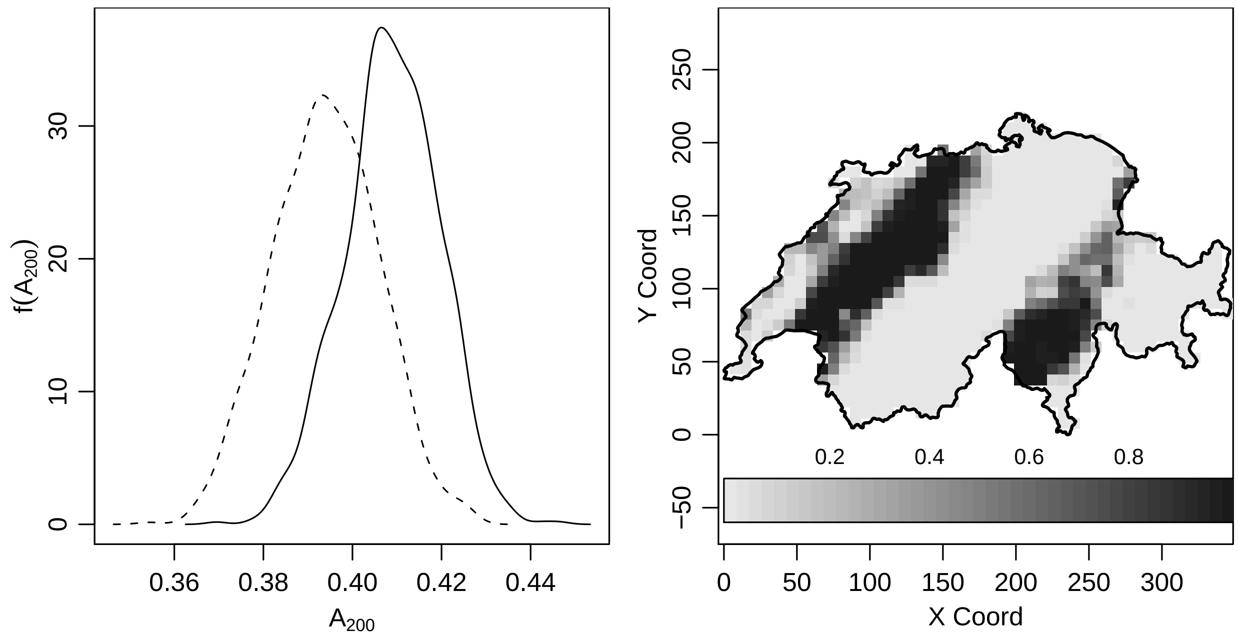
Mean (left-panel) and variance (right-panel) of predictive distribution



Swiss rainfall: posterior means and 95% credible intervals

parameter	estimate	95% interval
β	144.35	[53.08, 224.28]
σ^2	13662.15	[8713.18, 27116.35]
ϕ	49.97	[30, 82.5]
ν^2	0.03	[0, 0.05]

Swiss rainfall: non-linear prediction



Left-panel: **Bayesian (solid) and plug-in (dashed) prediction for proportion of total area with rainfall exceeding 200 (= 20mm)**

Right-panel: **Bayesian predictive map of $P(\text{rainfall} > 250|Y)$**

Computation with geoR

```
MC<-model.control()
?model.control
PC<-prior.control(beta.prior="flat",sigmasq.prior="sc.inv.chisq",
   sigmasq=0.2,df.sigmasq=4,phi.discrete=0.1*(1:5),
   tausq.rel.prior="uniform",tausq.rel.discrete=0.1*(0:3))
OC<-output.control(n.posterior=100,n.predictive=100,
   simulations.predictive=T,signal=T,moments=F)
set.seed(24367)
results.bayes<-krige.bayes(geodata=loglead2,locations=grid,
   borders=region,model=MC,prior=PC,output=OC)
```

```
plot(results.bayes)
posterior.bayes<-results.bayes$posterior
posterior.sample<-posterior.bayes$sample
par.names<-names(posterior.sample)
par(mfrow=c(2,2))
for (i in 1:4) {
   hist(posterior.sample[,i],xlab=par.names[i],main=" ")
   }
par(mfrow=c(1,1))
plot(posterior.sample[,2],posterior.sample[,3],
   xlab=par.names[2],ylab=par.names[3])
```

```
par(mfrow=c(1,1),pty="s")
predictions.bayes<-results.bayes$predictive
image(unique(grid[,1]),unique(grid[,2]),
   matrix(predictions.bayes$mean.simulations,26,26))
points(loglead2,add=T)
par(mfrow=c(1,2))
predict.max<-NULL
for (sim in 1:100) {
   predict.max<-c(predict.max,max(predictions$simulations[,sim]))
   }
hist(predict.max,xlab="predictive distribution of maximum",
   main="plug-in",breaks=0.1*(16:28))
predict.bayes.max<-NULL
for (sim in 1:100) {
   predict.bayes.max<-c(predict.bayes.max,
      max(predictions.bayes$simulations[,sim]))
   }
hist(predict.bayes.max,xlab="predictive distribution of maximum",
   main="Bayesian",breaks=0.1*(16:28))
```

Generalized linear geostatistical model (GLGM)

- **Latent spatial process**

  $S(x) \sim \mathrm{SGP}\{0, \sigma^2, \rho(u))\}$

  $\rho(u) = \exp(-|u|/\phi)$

- **Linear predictor**

  $\eta(x) = d(x)'\beta + S(x)$

- **Link function**

  $\mathrm{E}[Y_i] = \mu_i = h\{\eta(x_i)\}$

- **Conditional distribution for $Y_i : i = 1, ..., n$**

  $Y_i|S(\cdot) \sim f(y; \eta)$ mutually independent

GLGM

- **usually just a single realisation is available, in contrast with GLMM for longitudinal data analysis**
- **GLGM approach is most appealing when there is a natural sampling mechanism, for example Poisson model for counts or logistic-linear models for proportions**
- **transformed Gaussian models may be more useful for non-Gaussian continuous responses**
- **theoretical variograms can be derived but are less natural as summary statistics than in Gaussian case**
- **but empirical variograms of GLM residuals can still be useful for exploratory analysis**

The Loa loa prediction problem

Ground-truth survey data

- **random sample of subjects in each of a number of villages**
- **blood-samples test positive/negative for** *Loa loa*

Environmental data (satellite images)

- **measured on regular grid to cover region of interest**
- **elevation, green-ness of vegetation**

Objectives

- **predict local prevalence throughout study-region (Cameroon)**
- **compute local exceedance probabilities,**

$$\mathrm{P}(\text{prevalence} > 0.2 | \text{data})$$

Loa loa: a generalised linear model

- **Latent spatial process**

  $S(x) \sim \mathrm{SGP}\{0, \sigma^2, \rho(u))\}$

  $\rho(u) = \exp(-|u|/\phi)$

- **Linear predictor**

  $d(x) = \text{environmental variables at location } x$

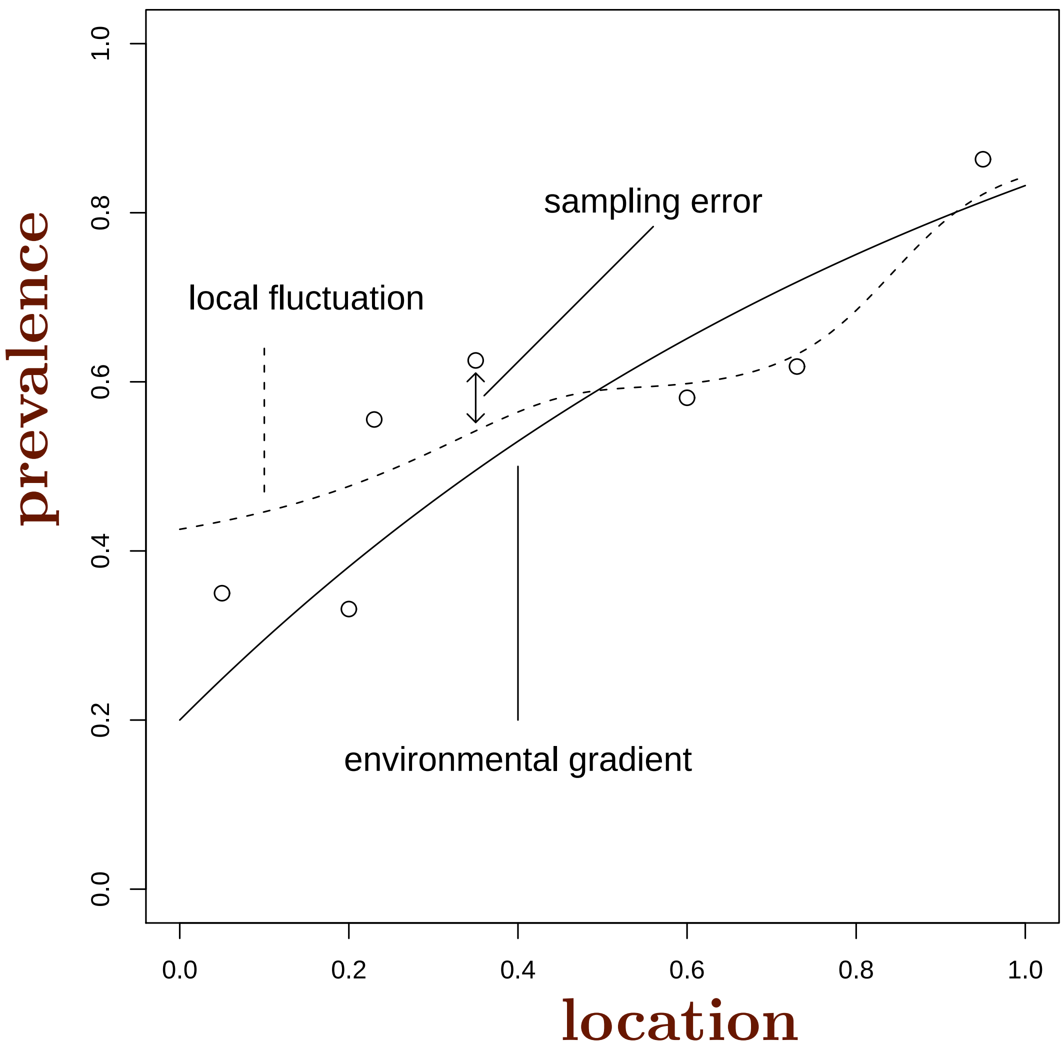
  $\eta(x) = d(x)'\beta + S(x)$

  $p(x) = \log[\eta(x)/\{1 - \eta(x)\}]$

- **Error distribution**

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

Schematic representation of Loa loa model



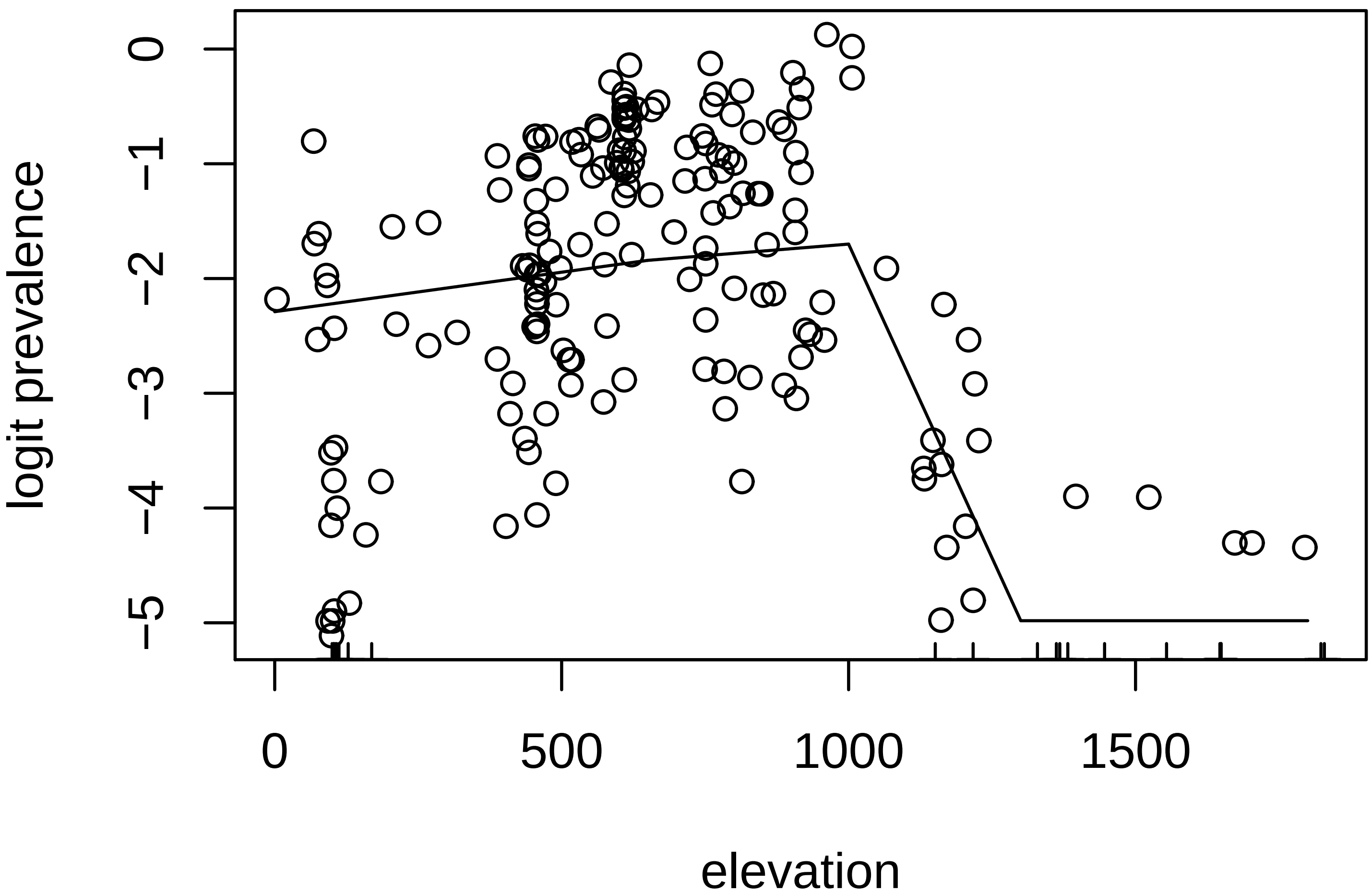
The modelling strategy

- **use relationship between environmental variables and ground-truth prevalence to construct preliminary predictions via logistic regression**

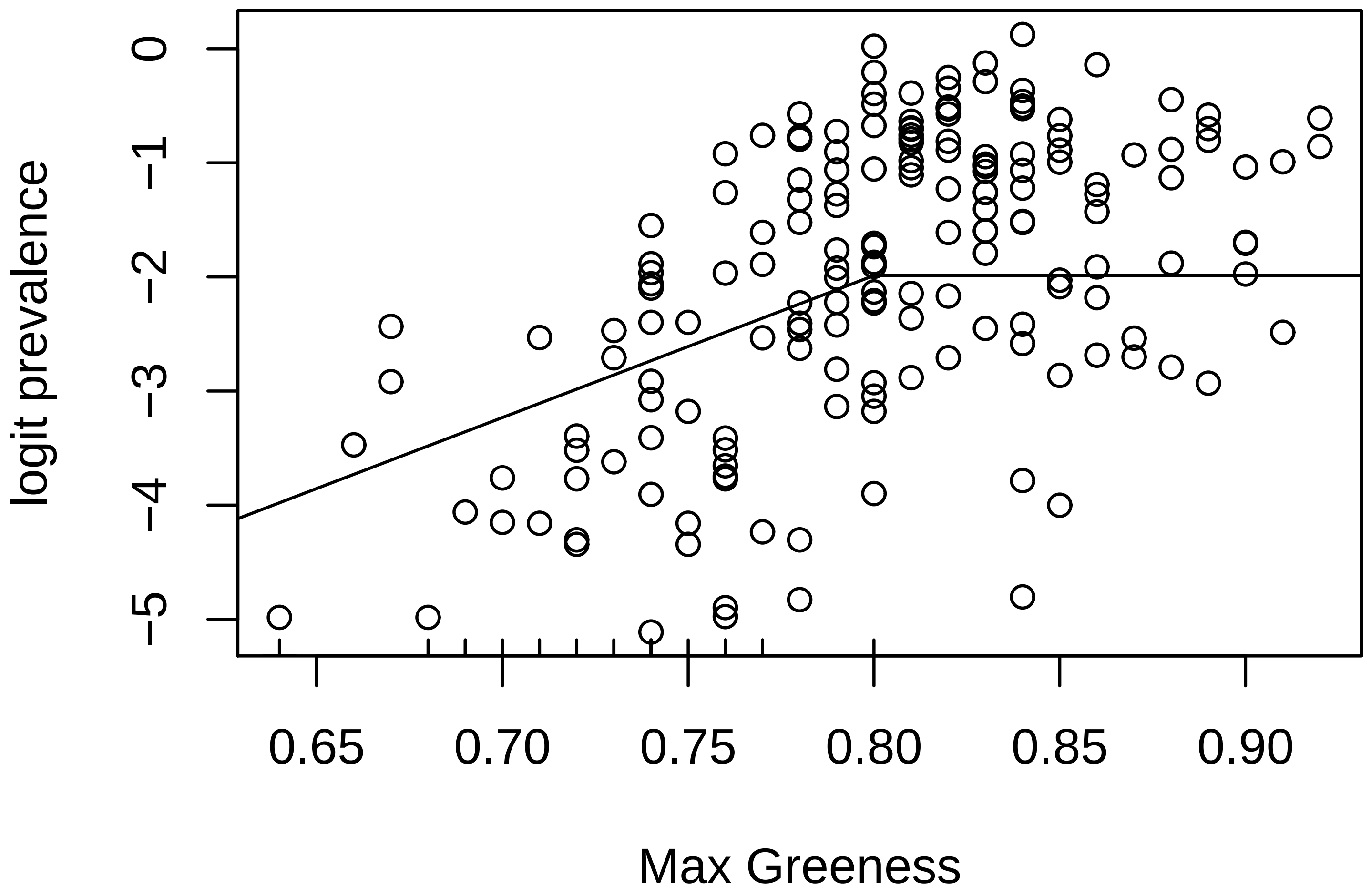
- **use local deviations from regression model to estimate smooth residual spatial variation**

- **Bayesian paradigm for quantification of uncertainty in resulting model-based predictions**

logit prevalence vs elevation

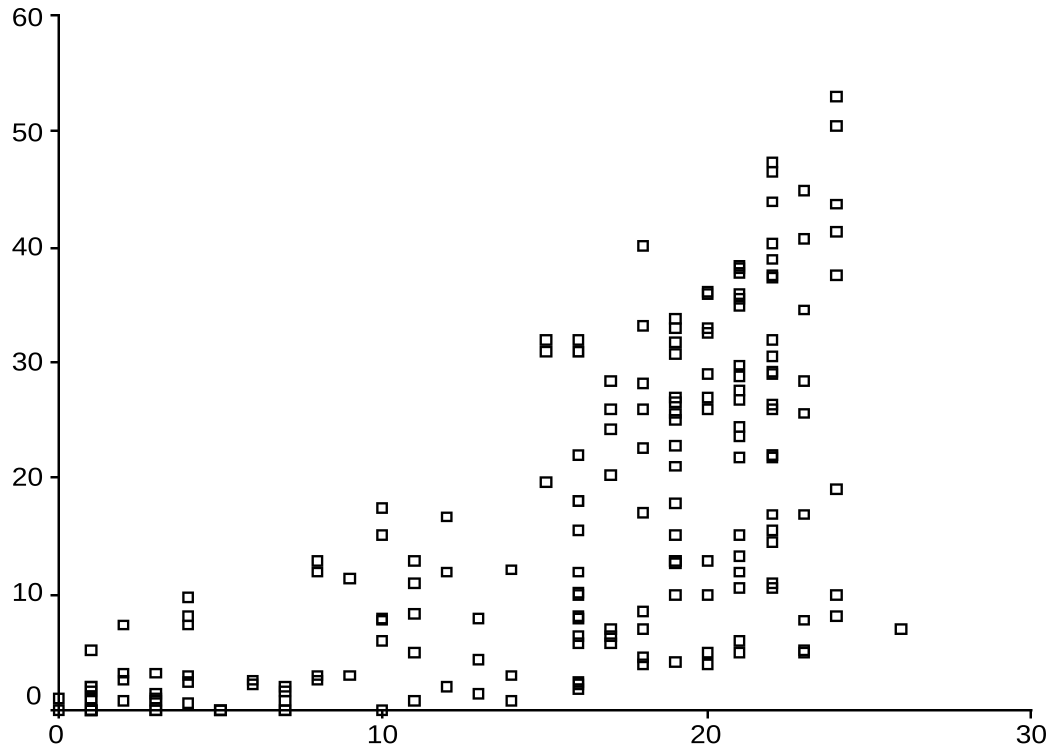


logit prevalence vs MAX = max NDVI

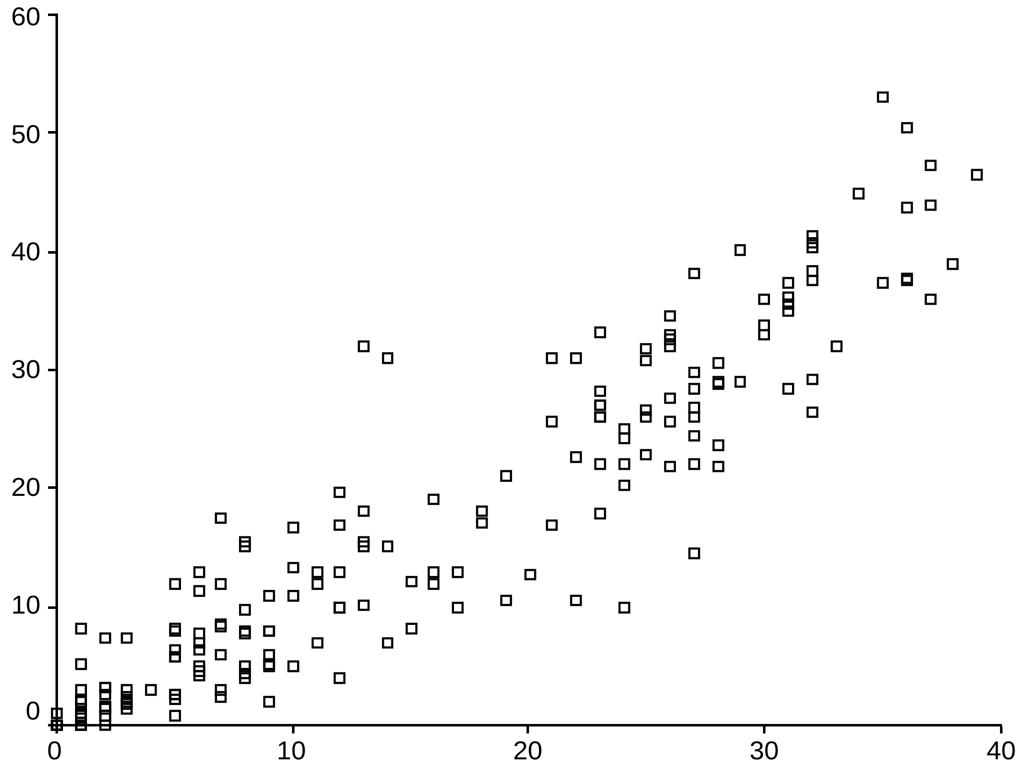


Comparing non-spatial and spatial predictions in Cameroon

Non-spatial



Spatial



Probabilistic prediction in Cameroon

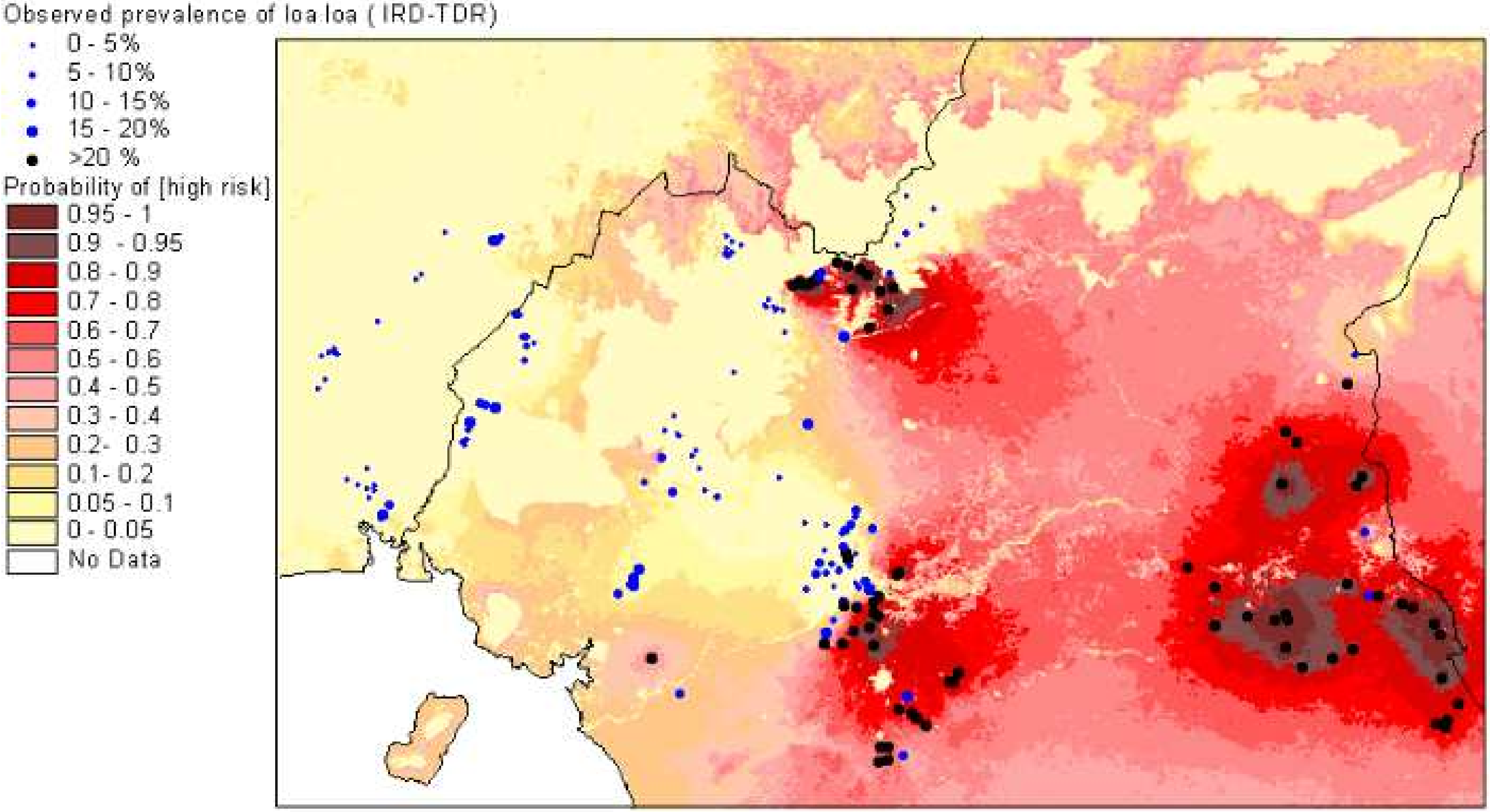


Figure 6: PCM for [high risk] in Cameroon based on ERMr with ground truth data.

5. Discrete spatial variation

- **Joint vs conditional specification**
- **Markov random field models**

Conditional specification of joint distributions

Theorem

$$\frac{f(y)}{f(z)} = \prod_{i=1}^{n} \frac{f_i(y_i | y_1, \ldots, y_{i-1}, z_{i+1}, \ldots, z_n)}{f_i(z_i | y_1, \ldots, y_{i-1}, z_{i+1}, \ldots, z_n)}$$

Outline of proof

Case $n = 3$ sufficient to show the idea, as follows

$$\begin{aligned} f(y_1, y_2, y_3) &= f(y_3 | y_1, y_2) \times f(y_2, y_1) \\ &= \frac{f(y_3 | y_1, y_2)}{f(z_3 | y_1, y_2)} \times f(z_3 | y_1, y_2) \times f(y_1, y_2) \\ &= \frac{f(y_3 | y_1, y_2)}{f(z_3 | y_1, y_2)} \times f(y_1, y_2, z_3) \end{aligned}$$

Same argument gives

$$\begin{aligned} f(y_1, y_2, z_3) &= f(y_2 | y_1, z_3) \times f(y_1, z_3) \\ &= \frac{f(y_2 | y_1, z_3)}{f(z_2 | y_1, z_3)} \times f(y_1, z_2, z_3) \end{aligned}$$

and so on, to give required result.

Exercise 2.2.2 re-visited

For the model

$$Y_i = \alpha(Y_{i-1} + Y_{i+1}) + Z_t : Z_t \sim \mathrm{N}(0, \tau^2),$$

the full conditional of Y_i depends on Y_{i-2}, Y_{i-1}, Y_{i+1} and Y_{i+2}.

- Re-write model in vector-matrix notation as

$$Y = AY + Z \Leftrightarrow Y = (I - A)^{-1}Z$$

where (using $n = 5$ for illustration)

$$A = \begin{bmatrix} 0 & \alpha & 0 & 0 & 0 \\ \alpha & 0 & \alpha & 0 & 0 \\ 0 & \alpha & 0 & \alpha & 0 \\ 0 & 0 & \alpha & 0 & \alpha \\ 0 & 0 & 0 & \alpha & 0 \end{bmatrix}$$

- Then, $Y \sim \mathrm{MVN}(0, (I - A)^{-2})$

- **Standard result from graphical modelling is that non-zero elements in $\mathrm{Var}(Y)^{-1}$ identify conditional dependencies (eg Whittaker, 1990, Proposition 5.7.3)**

- **Straightforward matrix algebra gives**

$$(I-A)^2 = \begin{bmatrix} 1+\alpha^2 & -2\alpha & \alpha^2 & 0 & 0 \\ -2\alpha & 1+2\alpha^2 & -2\alpha & \alpha^2 & 0 \\ \alpha^2 & -2\alpha & 1+2\alpha^2 & -2\alpha & \alpha^2 \\ 0 & \alpha^2 & -2\alpha & 1+2\alpha^2 & -2\alpha \\ 0 & 0 & \alpha^2 & -2\alpha & 1+\alpha^2 \end{bmatrix}$$

Hammersley-Clifford

Previous result says joint distribution of Y is determined by full conditionals provided full conditionals are self-consistent

General result: for any $A \subset \{1, 2, ..., n\}$, write $\mathcal{Y}_A = \{y_i : i \in A\}$, then

$$f(y) = \exp\left\{ \sum_{A \subset \{1,2,\dots,n\}} h(\mathcal{Y}_A) \right\} \qquad (1)$$

Definitions:

1) for any set of full conditionals $f_i(y_i|\{y_j : j \neq i\})$, index j is a neighbour of i if $f_i(\cdot)$ depends on y_j

2) a clique is a set of mutual neighbours.

Theorem (Hammersley-Clifford)

Expression (1) gives valid specification of $f(y)$ if and only if:

1. $h(\mathcal{Y}_A) = 0$ for all non-cliques A
2. $f(y)$ integrable (so can scale to $\int f(y) = 1$)
3. if $f(y_j) > 0$ for all $j \in A$, then $f(\mathcal{Y}_A) > 0$

Besag, 1974

Markov random field models

- Random vector $Y = (Y_1, ..., Y_n)$
- joint distribution $[Y]$ fully specified by full conditionals,

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

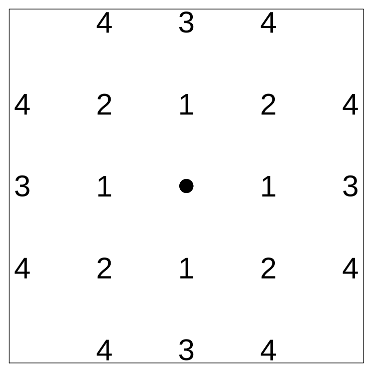
- Neighbourhood of i is $\mathcal{N}(i) \subset \{1, 2, ..., n\}$
- MRF: $[Y_i | \{Y_j : j \neq i\}] = [Y_i | Y_j : j \in \mathcal{N}(i)] : i = 1, ..., n$

Examples of MRF models

1. **Binary Y_i: auto-logistic model**

$$p_i = \mathrm{P}(Y_i = 1 | \{Y_j : j \neq i\}) \quad \mathrm{logit} p_i = \alpha + \beta \sum_{j \in \mathcal{N}_i} Y_j$$

Higher-order models defined naturally on regular lattices:



$$\mathrm{logit} p_i = \alpha + \sum_{k=1}^{m} \beta_k \sum_{j \in \mathcal{N}_k(i)} Y_j$$

2. Count Y_i: auto-Poisson model

$$\mu_i = \mathbf{E}[Y_i = 1 | \{Y_j : j \neq i\}] \qquad \log \mu_i = \alpha + \beta \sum_{j \in \mathcal{N}_i} Y_j$$

Restriction: the auto-Poisson model only defines a proper distribution when $\beta \leq 0$

3. Hierarchical model with latent Gaussian MRF

A better way to model spatial count data:

- **latent Gaussian MRF** $S = (S_1, ..., S_n)$
- **conditionally independent** $Y_i | S \sim \text{Poiss}(\alpha + \beta S_i)$

Even better if α **is replaced by** $\alpha_i = d_i' \theta$ **for vector of spatial explanatory variables** d_i

Besag, York and Mollié, 1991

Computational appeal of MRF models

- Gaussian MRF, mean μ, precision matrix $\Omega = \{\text{Var}(Y)\}^{-1}$, log-likelihood is

$$L = 0.5n \log |\Omega| - 0.5(Y - \mu)'\Omega(Y - \mu)$$

  Markov structure implies that Ω is sparse

- Gaussian or non-Gaussian MRF, Gibbs sampler for MCMC follows directly from model specification through full conditionals,

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

Limitations of MRF models for spatial data

- models are just multivariate probability distributions
  - parameterised in a way that has a spatial interpretation
  - but specific to a fixed set of locations $x_1, ..., x_n$
- 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,

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

6. Spatial point patterns

- exploratory analysis;
- Cox processes and the link to continuous spatial variation;
- pairwise interaction processes and the link to discrete spatial variation.

Notation

- spatial point process: countable set of events $x_i \in \mathbb{R}^2$
- $N(A) = \#(x_i \in A)$ for spatial region $A \subset \mathbb{R}^2$
- stationary if properties invariant under translation
- isotropic if properties invariant under rotation
- orderly if no multiple coincident events

The Poisson Process

1. $N(A) \sim \text{Poiss}(\mu(A))$, where

$$\mu(A) = \int_A \lambda(x)dx$$

2. given $N(A) = n$, events $x_i \in A$ iid, pdf $\propto \lambda(x)$

Complete spatial randomness: $\lambda(x) = \lambda$

Properties

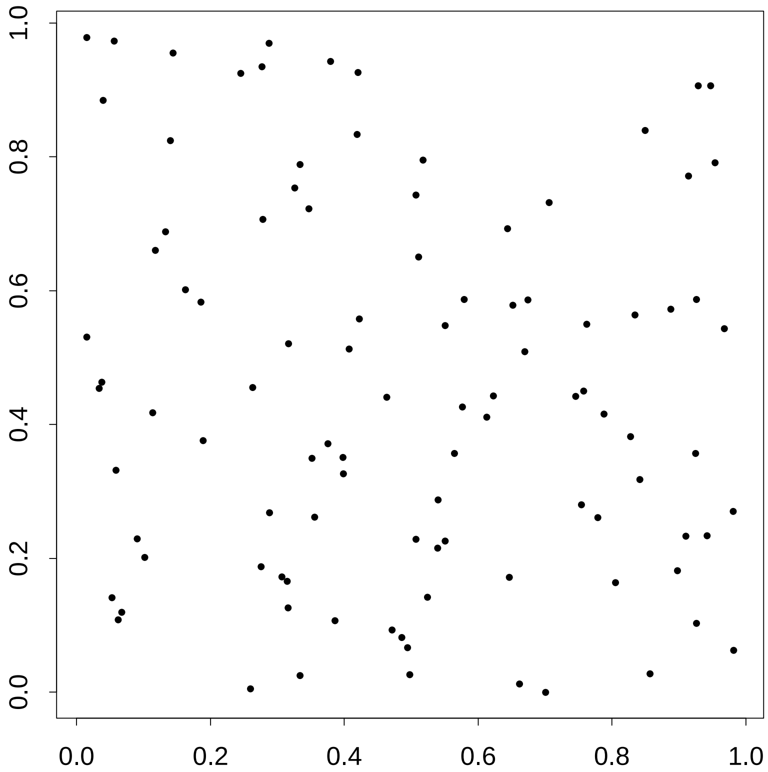
1. $N(A)$ and $N(B)$ independent when A and B disjoint

2. $\text{Var}\{N(A)\}/\text{E}[N(A)] = 1$, for all A

3. distance from an arbitrary point to the nearest event:

$$F(x) = 1 - \exp(-\pi\lambda x^2) : x > 0$$

Partial realisation of a Poisson process



Point process intensities

Def 1. **The (first-order) intensity function of a spatial point process is**

$$\lambda(x) = \lim_{|dx| \to 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\}$$

Def 2. **The second-order intensity function of a spatial point process is**

$$\lambda_2(x, y) = \lim_{\substack{|dx| \to 0 \\ |dy| \to 0}} \left\{ \frac{E[N(dx)N(dy)]}{|dx||dy|} \right\}$$

Def 3. **The covariance density of a spatial point process is**

$$\gamma(x, y) = \lambda_2(x, y) - \lambda(x)\lambda(y).$$

What if process is stationary and isotropic?

(i) $\lambda(x) \equiv \lambda = E[N(A)]/|A|$, (constant, for all A).

(ii) $\lambda_2(x, y) \equiv \lambda_2(\|x - y\|)$ (depends only on distance)

(iii) $\gamma(u) = \lambda_2(u) - \lambda^2$.

The K-function

Def 4 **The reduced second moment function of a stationary, isotropic spatial point process is**

$$K(s) = 2\pi\lambda^{-2} \int_0^s \lambda_2(r) r dr.$$

Theorem 1. For a stationary, isotropic, orderly process:

$K(s) = \lambda^{-1}$**E[number of further events within distance s of an arbitrary event]**

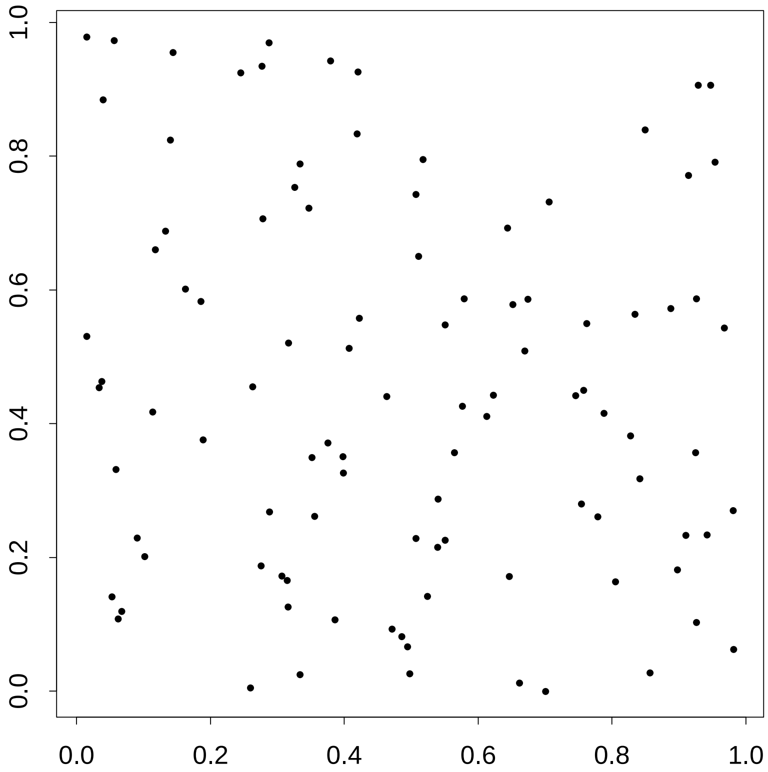
- **gives a tangible interpretation of $K(s)$**
- **suggests a method of estimating $K(s)$ from data**

Theorem 2. For a homogeneous, planar Poisson process,

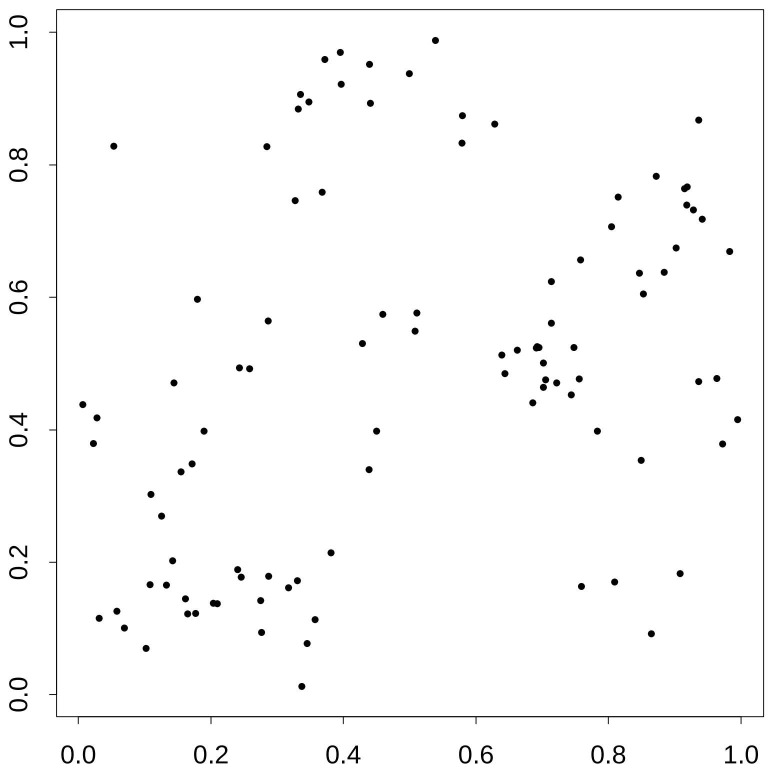
$$K(s) = \pi s^2$$

Three pictures

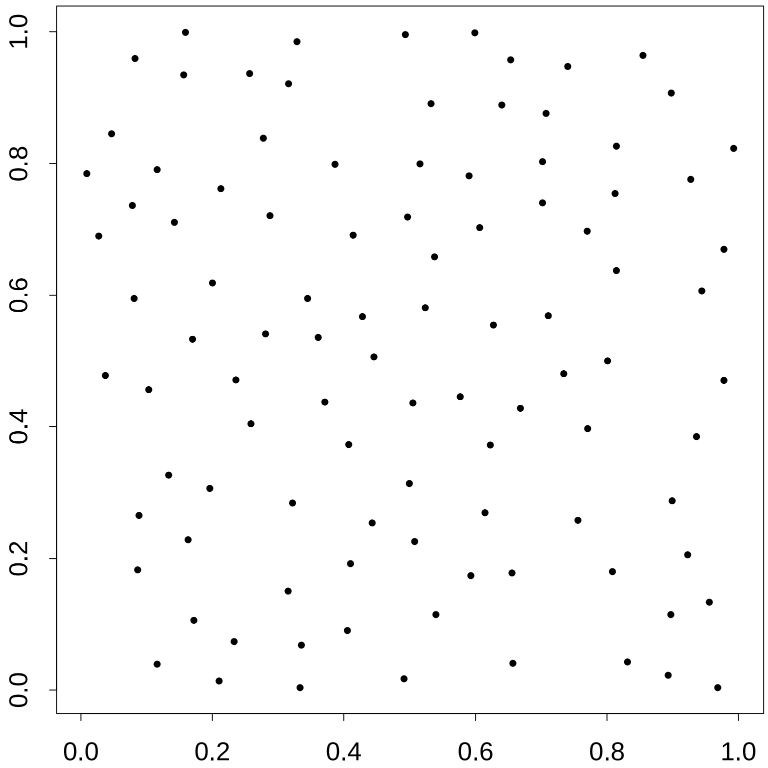
Completely random



Clustered



Regular



A useful property of the K-function

Def 5. **A random thinning, P', of a point process P, is a point process whose events are a sub-set of the events of P generated by retaining or deleting the events of P in a series of mutually independent Bernoulli trials.**

Theorem 3. $K(s)$ is invariant to random thinning.

Implication: the interpretation of an estimated K-function is robust to incomplete ascertainment of cases, provided the incompleteness is spatially neutral.

Estimating the K-function

Data: $x_i \in A : i = 1, \ldots, n$

Estimation of λ

$$\hat{\lambda} = n/|A|$$

Estimation of $K(s)$

$$\lambda K(s = \text{E[number of further events within distance } s \text{ of an arbitrary event]}$$

1. Define $E(s) = \lambda K(s)$.

2. Let d_{ij} be the distance between the events x_i and x_j.

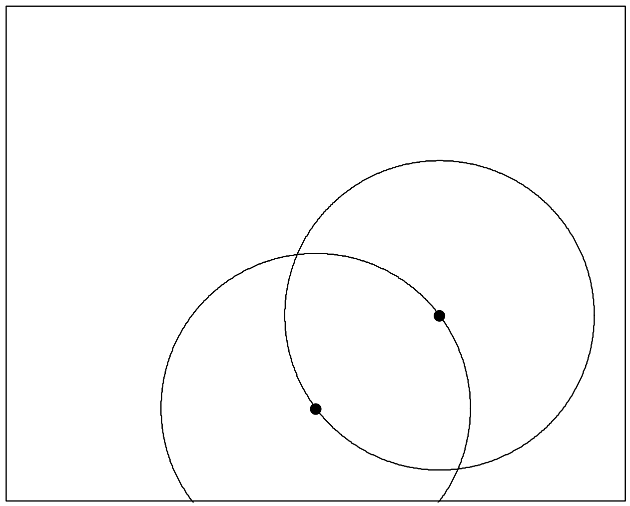
3. Define

$$\tilde{E}(s) = n^{-1} \sum_{i=1}^{n} \sum_{j \neq i} I(d_{ij} \leq s), \tag{7}$$

4. The estimator $\tilde{E}(s)$ is negatively biased because we do not observe events outside A

5. Introduce weights,

   w_{ij} = reciprocal of proportion of circumference of circle, centre x_i and radius d_{ij}, which is contained in A.



6. An edge-corrected estimator for $E(s)$ is

$$\hat{E}(s) = n^{-1} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(d_{ij} \leq s).$$

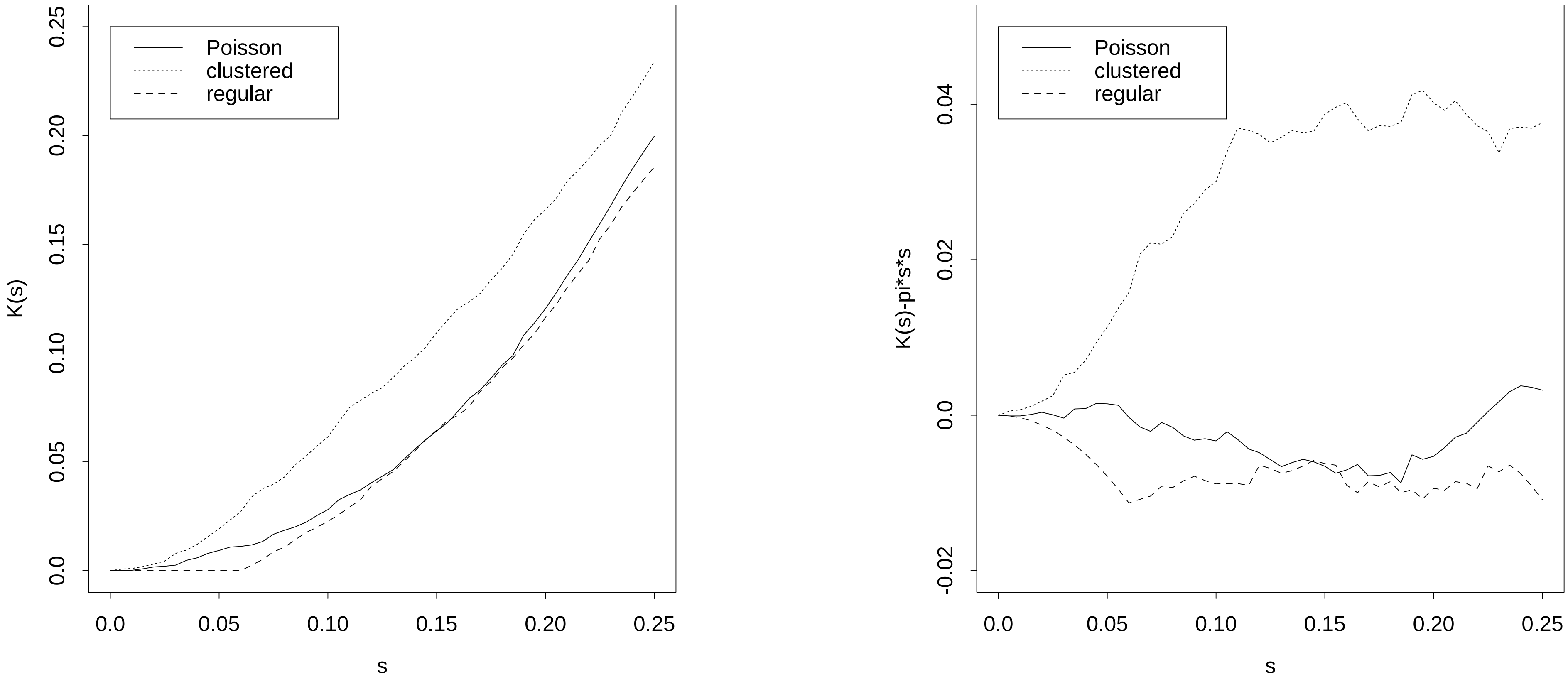
where $I(\cdot)$ is the indicator function.

7. Since $K(s) = E(s)/\lambda$, define

$$\hat{K}(s) = \hat{E}(s)/\hat{\lambda} \tag{8}$$

$$= n^{-2}|A| \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(d_{ij} \leq s) \tag{9}$$

Estimates $\hat{K}(s)$ for three simulated patterns



Bivariate K-functions

$\lambda_j : j = 1, 2$ denotes intensity of type j events.

$$\lambda_j K_{ij}(s) = \text{expected number of further type } j \text{ events within distance } s \text{ of an arbitrary type } i \text{ event}$$

- if type j events form a homogeneous Poisson process, then

$$K_{jj}(s) = \pi s^2$$

- if type 1 and type 2 events form independent processes, then

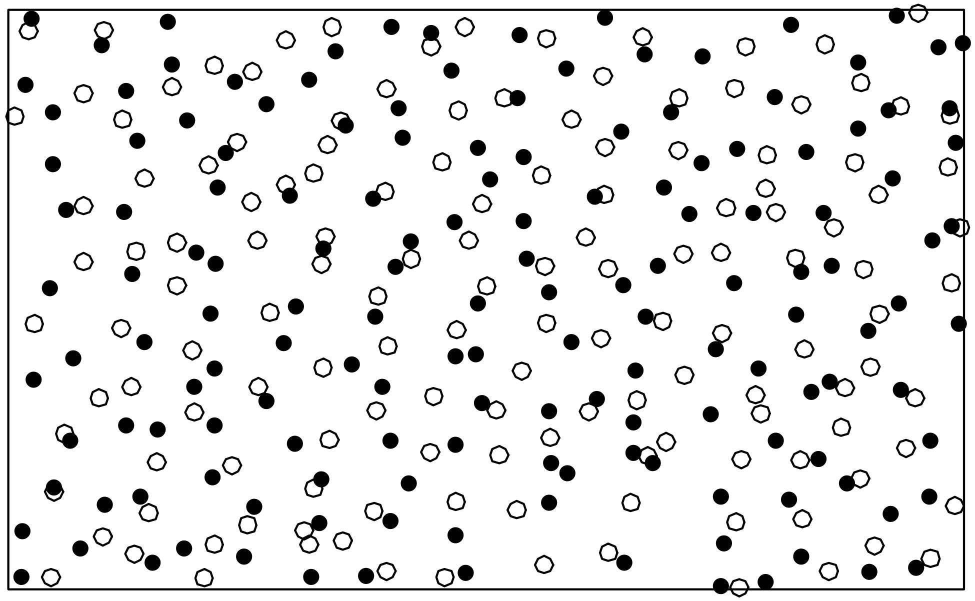
$$K_{12}(s) = \pi s^2$$

- if type 1 and type 2 events form a random labelling of a univariate process with K-function $K(s)$, then

$$K_{11}(s) = K_{12}(s) = K_{22}(s) = K(s)$$

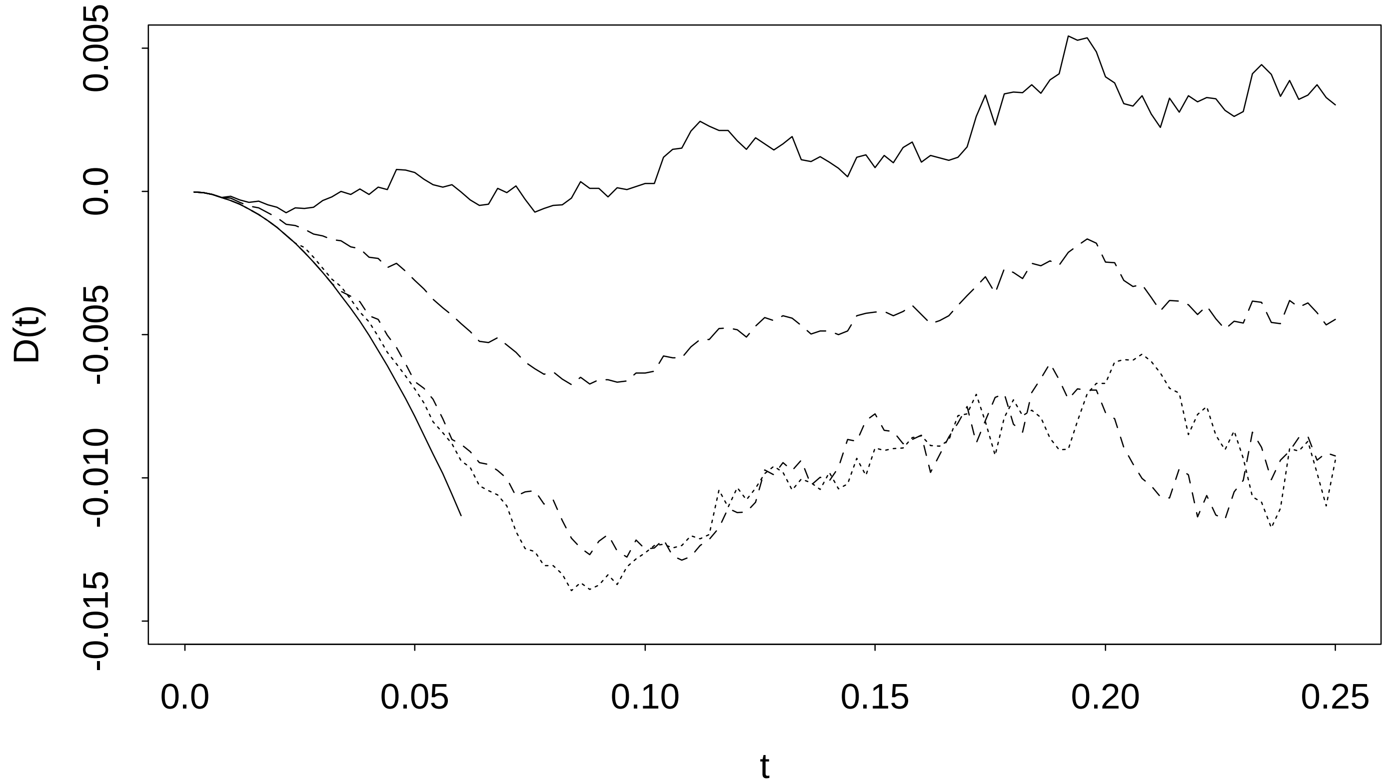
An example: displaced amacrine cells in rabbit retina

- **type 1 events transmit information to the brain when a light goes on**
- **type 2 events transmit information to the brain when a light goes off**
- **interest is in discriminating between two developmental hypotheses:**
  1. **on and off cells are initially generated in separate layers which later fuse to form the mature retina**
  2. **on and off cells are initially undifferentiated in a single layer and acquire their distinct functionality at a later stage**



Solid/open circles respectively identify ***on/off*** **cells**

Second-order properties:



Functions plotted are $\hat{D}(t) = \hat{K}(t) - \pi t^2$ as follows:

– – – : on cells; · · · · · : off cells; —— —— —— : all cells; ———— : bivariate.

The parabola $-\pi t^2$ is also shown as a solid line.

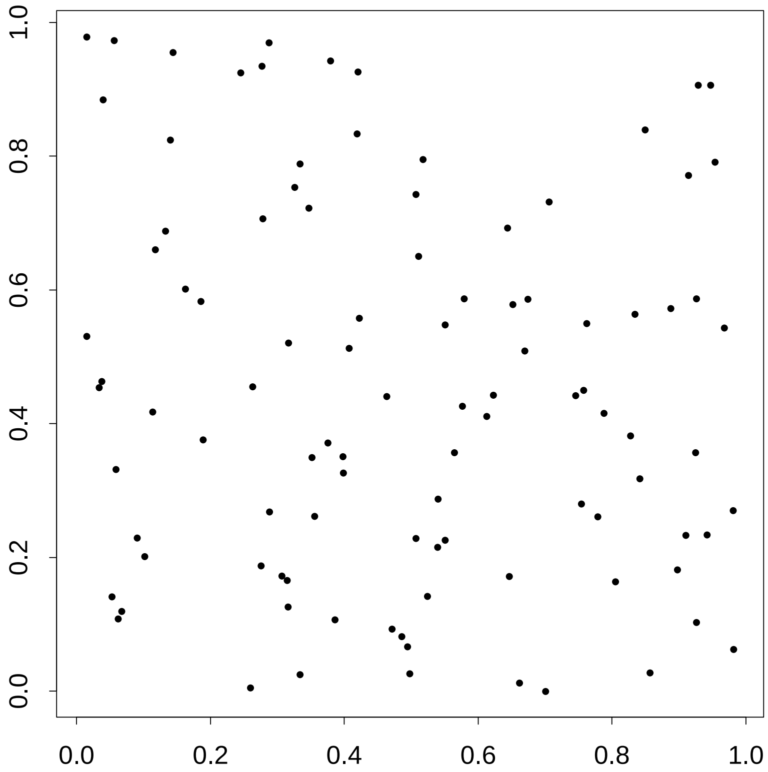
Computation with splancs

```
#
# Exploratory analysis of amacrine cell data
#
library(splancs)
on<-scan("amacrines_on.txt")
length(on)
on<-matrix(on,152,2,T)
off<-scan("amacrines_off.txt")
length(off)
off<-matrix(off,142,2,T)
a<-1060/662
poly<-matrix(c(0,0,a,0,a,1,0,1),4,2,T)
par(pty="s")
polymap(poly)
pointmap(on,add=T,pch=19,col="red")
pointmap(off,add=T,pch=19,col="blue")
```

```
?khat
s<-0.005*(0:51)
k.on<-khat(on,poly,s)
k.off<-khat(off,poly,s)
plot(s,k.on-pi*s*s,type="l",col="red",ylim=c(-0.015,0.005))
lines(s,k.off-pi*s*s,col="blue")
k.cross<-k12hat(on,off,poly,s)
lines(s,k.cross-pi*s*s)
```

Three pictures re-visited

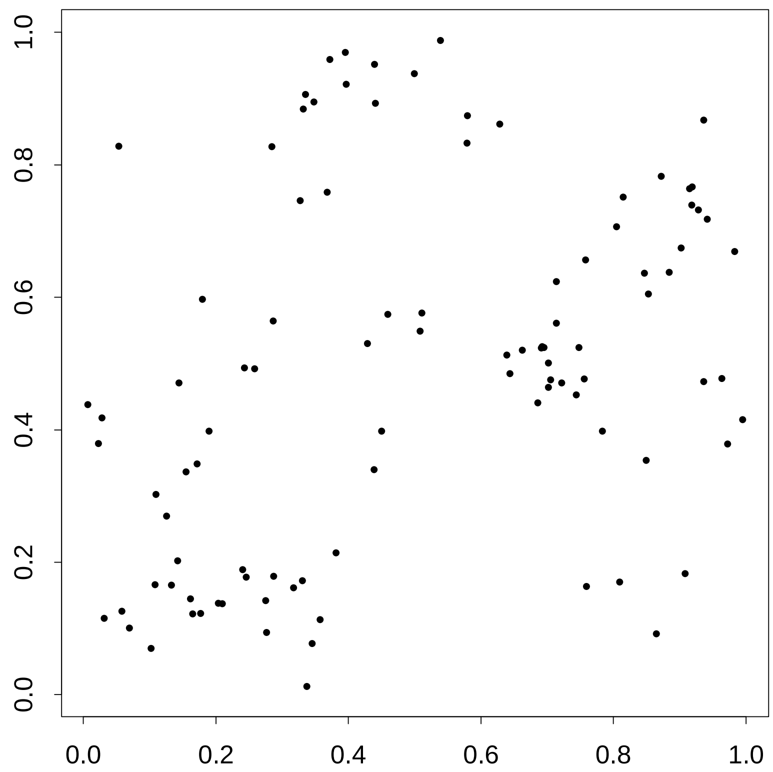
Completely random



A Poisson process

- $N(A) \sim \text{Poiss}(\lambda|A|)$
- conditional on $N(A) = n$, events $x_i : i = 1, ..., n$ are independent random sample from uniform distribution on A

Clustered



A Cox process

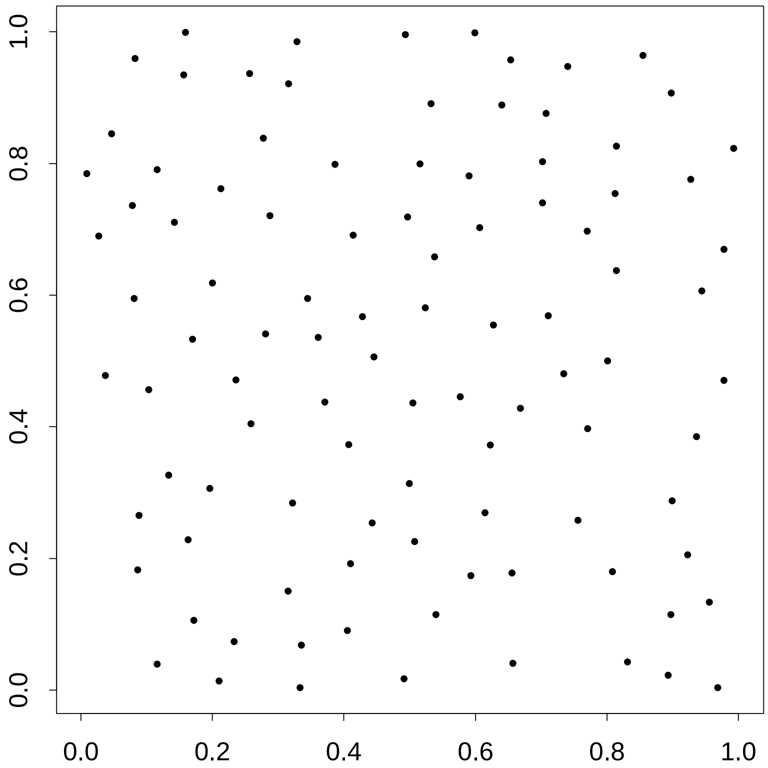
- $\Lambda(x)$ a non-negative-values spatial stochastic process
- conditional on $\Lambda(x) = \lambda(x)$, process is inhomogeneous Poisson

Cox, 1955

Picture: $\Lambda(x) = \sum g(x - X_i)$

- $X_i : i = 1, 2, \ldots$ homogenous Poisson process
- $g(\cdot) =$ bivariate Gaussian density, $N(0, \sigma^2 I)$

Regular



An inhibitory process

- events $\mathcal{X} = \{x_1, ..., x_n\}$ in spatial region A
- $LR(\mathcal{X})$ = likelihood ratio for $\mathcal{X}$ wrt Poisson process of unit intensity
- non-negative-valued interaction function $h(u) : u \geq 0$

$$LR(\mathcal{X}) \propto \beta^n \prod_{j \neq i} h(|x_i - x_j||)$$

Picture:

$$h(u) = \begin{cases} 0 & : \quad u < \delta \\ 1 & : \quad u \geq \delta \end{cases}$$

Poisson processes

- completely defined by their intensity function $\lambda(x)$
  - $N(A) \sim \text{Poiss}\left(\int_A \lambda(x)dx\right)$
  - conditional on $N(A) = n$, events $x_i : i = 1, \ldots, n$ are independent random sample from distribution with pdf $f(x) \propto \lambda(x)$
- Log-likelihood function,

$$L(\theta) = \sum_{i=1}^{n} \log \lambda(x_i; \theta) - \int_A \lambda(x; \theta)dx$$

- independence property often unrealistic, but may be useful approximation

Cox processes

- a Cox process is an inhomogeneous Poisson process with stochastic intensity $\Lambda(x)$

- useful class of models for environmentally driven processes

- even more useful when environmental covariates can explain part of the variation in $\Lambda(x)$

Cox (1955)

Cox processes: moment properties

Assume $\Lambda(x)$ stationary with mean λ and covariance function $\gamma(u)$, then:

- $\lambda =$ intensity
- $\gamma(u) =$ covariance density

$$K(s) = \pi s^2 + 2\pi\lambda^{-2}\int_0^s \gamma(u)u\,du$$

Cox processes: model-fitting

- likelihood generally intractable (except by Monte Carlo)
- ad hoc estimation by matching theoretical and empirical second moments (not entirely satisfactory)

$$\int_0^s w(u)\{\hat{K}(u) - K(u;\theta)\}^2 du$$

Pairwise interaction point processes (PIPP's)

- defined by their likelihood ratio wrt Poisson process
- useful for modelling inhibitory interactions between events
- can be derived as continuous limit of Poisson MRF models
  on a regular lattice
- problematic for modelling attractive interactions
  (recall similar reservation wrt auto-Poisson model)

PIPP's: formulation

- events $\mathcal{X} = \{x_1, ..., x_n\}$ in spatial region A
- $LR(\mathcal{X})$ = likelihood ratio for $\mathcal{X}$ wrt Poisson process of unit intensity
- non-negative-valued interaction function $h(u) : u \geq 0$

$$LR(\mathcal{X}) \propto \beta^n \prod_{j \neq i} h(|x_i - x_j||)$$

- process well-defined if $h(u) \leq 1$ for all u
- $h(u) = 1$ for all u gives homogeneous Poisson process

PIPP's: model-fitting

Conditional intensity at x, given $\mathcal{X} = \{x_1, ..., x_n\}$ in $A - \{x\}$,

$$\lambda(x|\mathcal{X}) = \beta \prod_{i=1}^{n} h(|x_i - x||)$$

- MCMC scheme for simulating realisations operates by alternating between:
  - adding event according to pdf $f(x) \propto \lambda(x|\mathcal{X})$
  - deleting event at random
- likelihood evaluation requires Monte Carlo methods

- **pseudo-likelihood:**
  - **treats $\boldsymbol{\lambda_c(\cdot)}$ as if unconditional intensity, hence**

$$L(\theta) = \sum_{i=1}^{n} \log \lambda_c(x_i | \mathcal{X}\ \{x_i\}; \theta) - \int_A \lambda(x | \mathcal{X}; \theta) dx$$

  - **gives good starting values for Monte Carlo inference**

Computation using spatstat

```
#
# fitting a pairwise interaction point process to the
#amacrine "on" cells
#
library(spatstat)
library(splancs)
#
xy.on<-matrix(scan("amacrines_on.txt"),152,2,T)
xy<-xy.on
?ppp
xy.ppp<-ppp(xy[,1],xy[,2],xrange=c(0,1060),yrange=c(0,662))
```

```
?ppm
?quadscheme
Q<-quadscheme(xy.ppp,nd=c(80,56))
#
# 80 by 56 quadrature grid gives approximate convergence of
# non-parametric estimate
#
stuff<-ppm(Q,interaction=PairPiece(r=20*(1:10)),
                              correction="Ripley")
h.nonparam.on<-c(0,0.0589,0.2857,0.6922,0.9524,1.0087,
                   0.9468,0.9230,0.8553,0.8415)
u.nonparam<-20*(0:9)+10
plot(u.nonparam,h.nonparam.on,type="l",xlab="r",ylab="h(u)")
```

PIPP's: Monte Carlo likelihood

Likelihood function for PIPP with parameter θ and data $\mathcal{X}$ can always be written as

$$\ell(\theta) = a(\theta) LR(\mathcal{X}, \theta)$$

Circumvent intractability of normalising constant $a(\theta)$ as follows:

- $$\begin{aligned} a(\theta)^{-1} &= \int LR(\mathcal{X}, \theta) d\mathcal{X} \\ &= \int LR(\mathcal{X}, \theta) \times \frac{a(\theta_0)}{a(\theta_0)} \times \frac{LR(\mathcal{X}, \theta_0)}{LR(\mathcal{X}, \theta_0)} d\mathcal{X} \end{aligned}$$

- Define $r(\mathcal{X}, \theta, \theta_0) = LR(\mathcal{X}, \theta)/LR(\mathcal{X}, \theta_0)$, then

$$\begin{aligned} a(\theta)^{-1} &= a(\theta_0)^{-1} \int r(\mathcal{X}, \theta, \theta_0)\ell(\mathcal{X}, \theta_0)d\mathcal{X} \\ &= a(\theta_0)^{-1}\mathrm{E}_{\theta_0}[r(\mathcal{X}, \theta, \theta_0)] \end{aligned}$$

- Since θ_0 is arbitrary, it follows that for any value θ_0, the MLE $\hat{\theta}$ maximises

$$L(\theta) = \log LR(\mathcal{X}, \theta) - \log \mathrm{E}_{\theta_0}[r(\mathcal{X}, \theta, \theta_0)]$$

which in turn can be approximated by

$$L^*(\theta) = \log LR(\mathcal{X}, \theta) - \log \left\{ s^{-1} \sum_{j=1}^{s} r(\mathcal{X}_j, \theta, \theta_0)] \right\},$$

where $\mathcal{X}_j : j = 1, ..., s$ are simulated with $\theta = \theta_0$

Algorithm

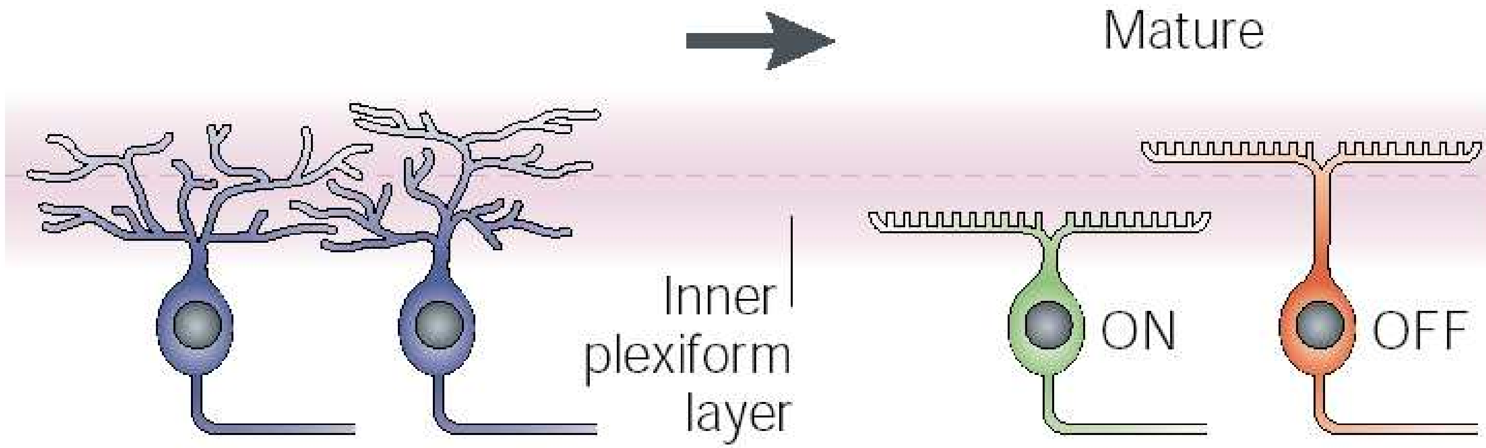
1. **Pick starting value θ_0 (eg maximum pseudo-likelihood estimate), and number of simulations s**

2. **Maximise resulting $L^*(\theta)$ to give $\theta = \tilde{\theta}$**

3. **Set $\theta_0 = \tilde{\theta}$, increase s and repeat**

Example: displaced amacrine cells

Biology (as of 2004)



Diggle, Eglen and Troy (2006)

Bivariate pairwise interaction point processes

Bivariate data

$$X_1 = \{x_{1i} : i = 1, ..., n_1\} \qquad X_2 = \{x_{2i} : i = 1, ..., n_2\}$$

Bivariate pairwise interaction model

$$f(X_1, X_2) \propto P_{11} P_{22} P_{12}$$

$$\begin{aligned}
P_{11} &= \prod_{i=2}^{n_1} \prod_{j=1}^{i-1} h_{11}(||x_{1i} - x_{1j}||) \\
P_{22} &= \prod_{i=2}^{n_2} \prod_{j=1}^{i-1} h_{22}(||x_{2i} - x_{2j}||) \\
P_{12} &= \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} h_{12}(||x_{1i} - x_{2j}||)
\end{aligned}$$

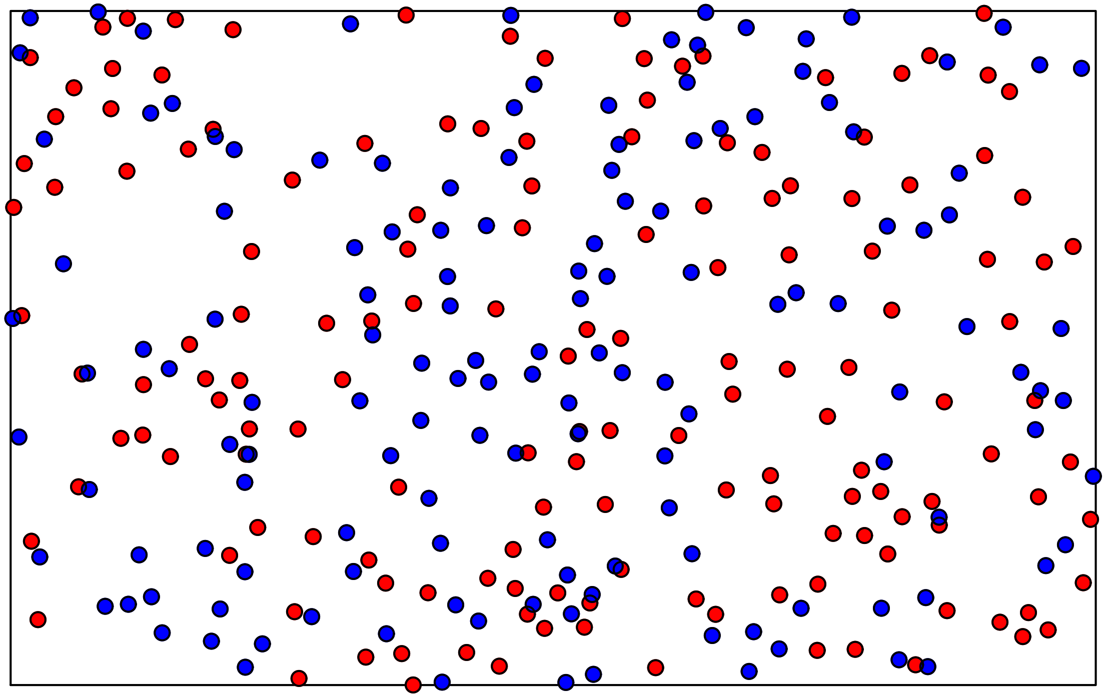
Parametric family of interaction functions

$$h(u;\theta) = \begin{cases} 0 & : \quad u \leq \delta \\ 1 - \exp[-\{(u-\delta)/\phi\}^{\alpha}] & : \quad u > \delta \end{cases}$$

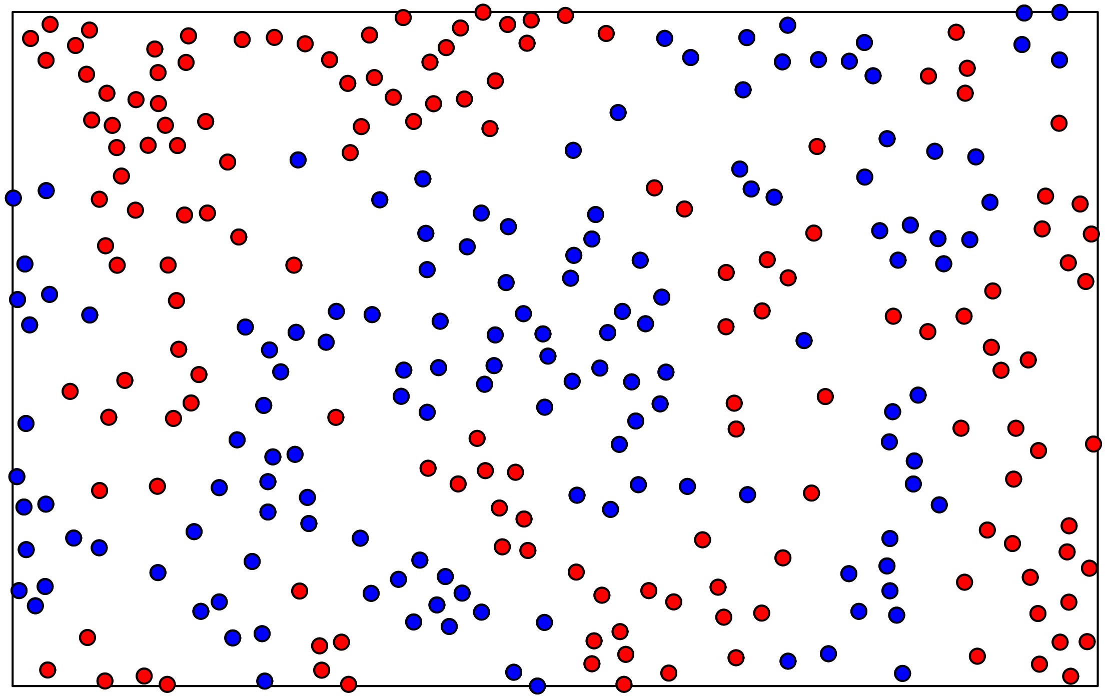
Special cases

- **Simple inhibition:** $\phi \to 0$
- **Independence:** $h_{12}(u) = 1$
- **Functional independence:** $h_{12}(\cdot)$ simple inhibitory

Marginal behaviour depends on $h_{12}(\cdot)$



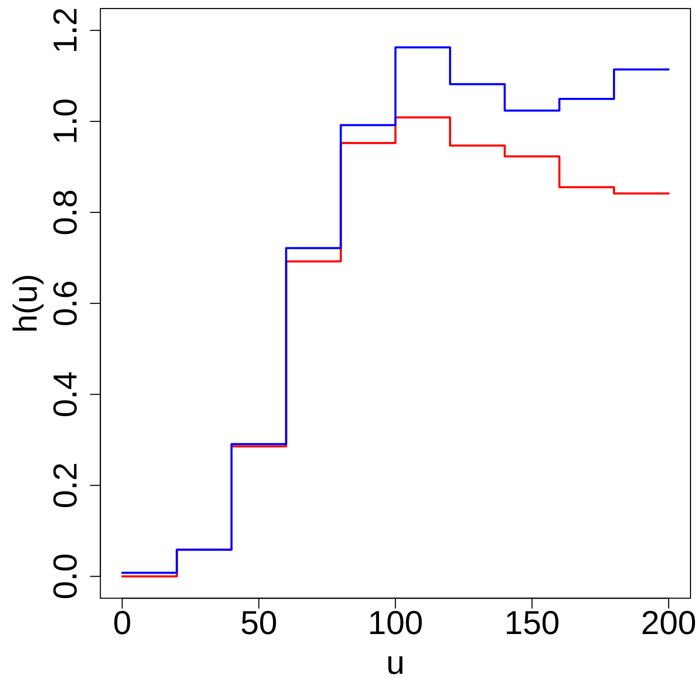
$\boldsymbol{\delta_{12} = 0}$ **(independence)**



$\delta_{12} = 50$ **(mutually inhibitory)**

Parametric analysis of the amacrine cells

Non-parametric estimates of $h(u)$ obtained by fitting step-function model using maximum pseudo-likelihood



on cells **off cells**

Fitted univariate models

$$h(u;\theta) = \begin{cases} 0 & : \quad u \le \delta \\ 1 - \exp[-\{(u-\delta)/\phi\}^{\alpha}] & : \quad u > \delta \end{cases}$$

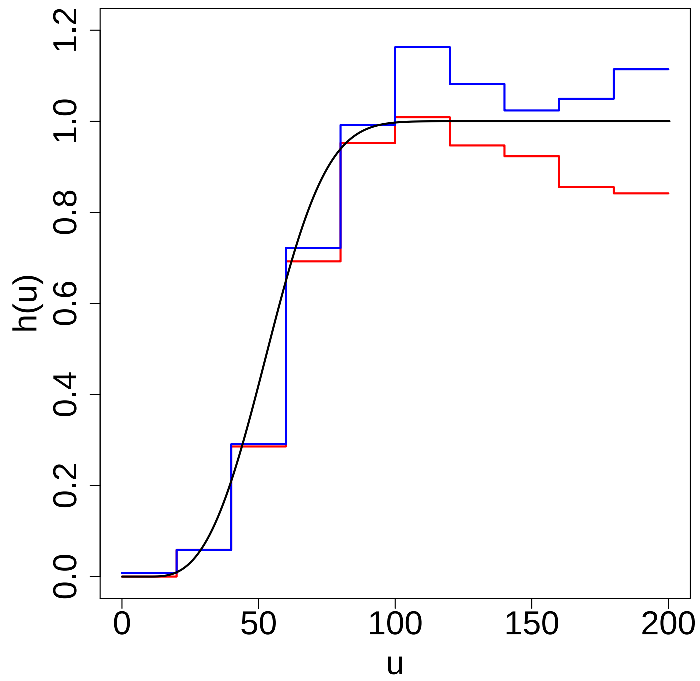
- Likelihood ratio statistic for common marginal parameters: $D = 1.36 \sim \chi^2_2$ $\quad p = 0.507$

- Pooled Monte Carlo MLE's

Parameter	Estimate	Std Error	Correlation
ϕ	49.08	2.51	
α	2.92	0.25	-0.06

Treat δ as known (physical size of cells)

Goodness-of-fit



A bivariate model for the amacrine cells

Likelihood ratio tests

- statistical independence vs functional independence

$$D = 5.30 \sim \chi^2_1 \qquad p = 0.021$$

- functional independence vs general bivariate

$$D = 0.30 \sim \chi^2_2 \qquad p = 0.861$$

- 95% confidence interval for δ_{12}

$$2.3 \leq \delta_{12} < 5.0$$

Goodness-of-fit

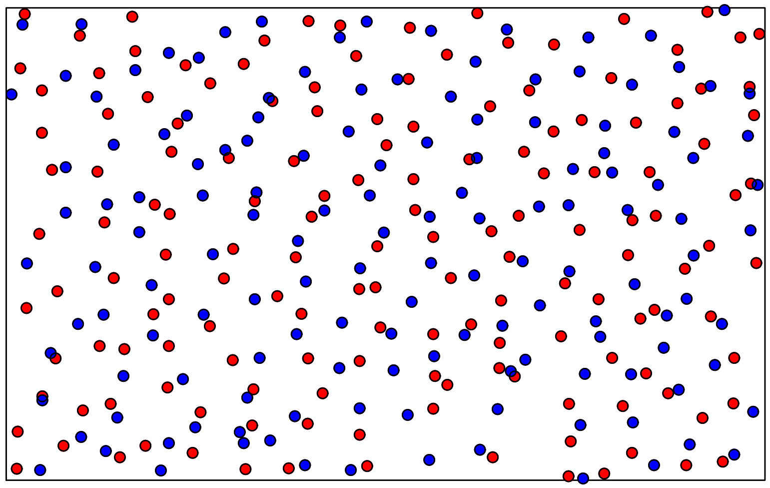
- $\hat{K}_{ij}(s)$ estimate from data
- $\bar{K}_{ij}(s)$ mean of estimates from 99 simulations of model
- three test statistics:

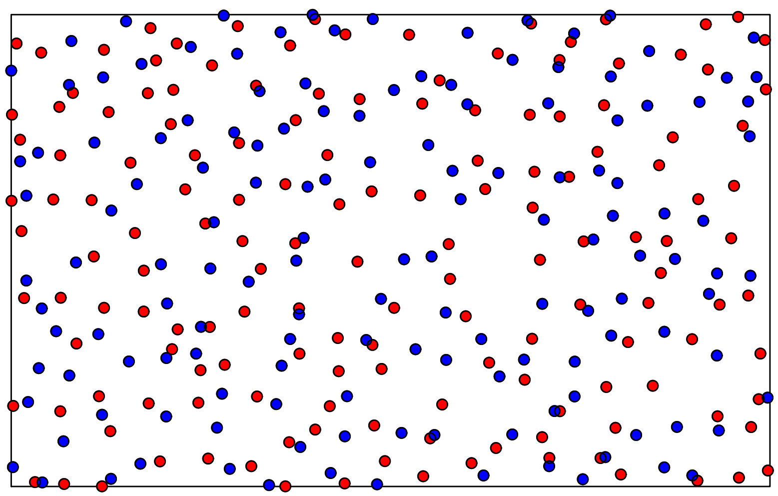
$$T_{ij} = \sum_{s=1}^{150} [\{\hat{K}_{ij}(s) - \bar{K}_i(s)\}/s]^2$$

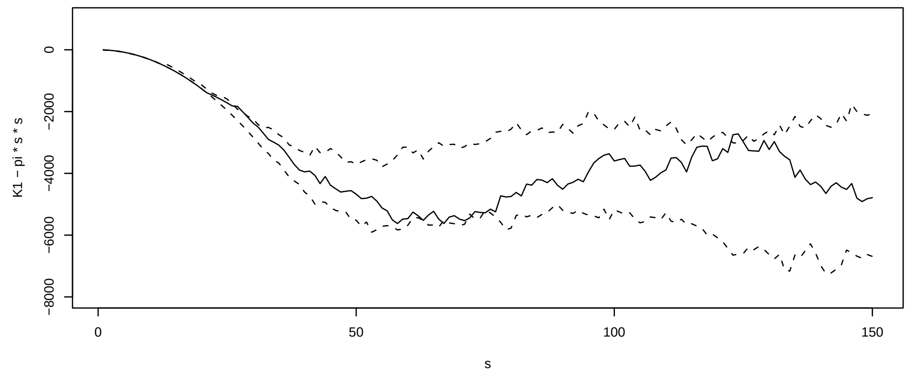
Results

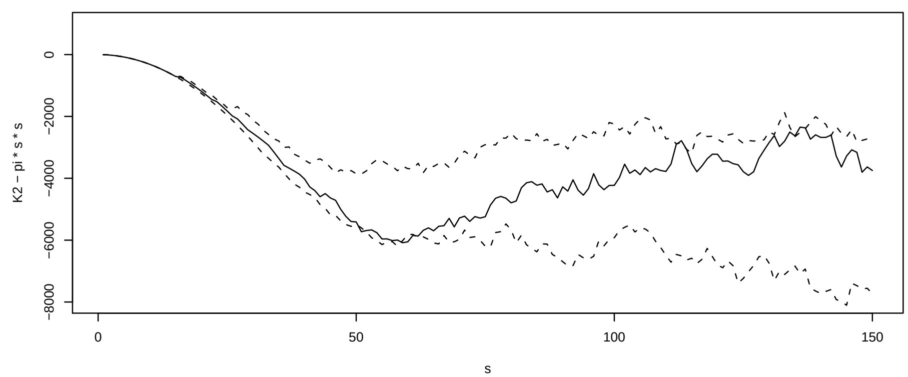
T_{11}, $p = 0.11$ (on cells)
T_{22}, $p = 0.05$ (off cells)
T_{12}, $p = 0.25$ (dependence)

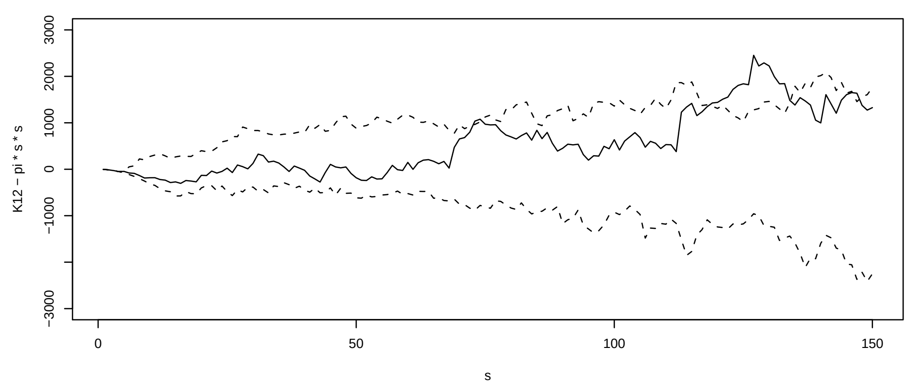
Bonferroni: $p \leq 0.15$






K1 − pi * s * s
0
−2000
−4000
−6000
−8000
0
50
100
150
s


K2 − pi * s * s
0
−2000
−4000
−6000
−8000
0
50
100
150
s


K12 − pi * s * s
3000
2000
1000
0
−1000
−3000
0
50
100
150
s

7. Spatio-temporal modelling

- spatial time series;
- spatio-temporal point processes.

Classification of spatio-temporal data?

Some possiblities:

- **geostatistical:** $(x_i, t_i, Y_i) : i = 1, ..., n;\ (x_i, t_i) \in \mathbb{R}^2 \times \mathbb{R}^+$
- **regular lattice:** $Y_{ijt} : i = 1, ..., n;\ j = 1, ..., m;\ t = 1, ..., T$ **(spatially discrete)**
- **spatial time series:** $(x_i, Y_{it}) : i = 1, ..., n;\ t = 1, ..., T$ **(spatially discrete or spatially continuous)**
- **point process:** $(x_i, t_i) : i = 1, ..., n$
- **various hybrids**

Spatial time series

$$(Y_{it}, x_i) : i = 1, ..., n;\ t = 1, ..., T$$

- spatially discrete sample from a spatially continuous phenomenon
- a common situation in practice, e.g. environmental monitoring networks
- implicit assumption that data are spatially sparse but temporally dense

Spatial time series: model specification

1. Direct specification: $\text{Cov}\{Y)x, t), Y(x', t')\} = \sigma^2 \rho(u, v)$, $u = ||x - x'||$, $v = |t - t'|$
   (a) separable: $\rho(u, v) = \rho_s(u)\rho_t(v)$
   (b) non-separable: $\rho(u, v) \neq \rho_s(u)\rho_t(v)$
2. Conditioning on the past:
   - $Y_t = \{Y_t(x) : x \in \mathbb{R}^2\}$
   - model Y_t conditional on $\{Y_s : s < t\}$

   Natural approach for a discrete-time, spatially continuous Markov process,

$$[Y_t|\{Y_s : s < t\}] = [Y_t|Y_{t-1}]$$

Separabilty then implies $[Y_t(x)|Y_{t-1}] = [Y_t(x)|Y_{t-1}(x)]$.

Real-time disease surveillance

Data: daily calls to NHS direct

Model: log-Gaussian Cox process

$$\begin{aligned} \Lambda(x,t) &= \lambda_0(x)\mu_0(t)\exp\{S(x,t)\} \\ S(x,t) &\sim \mathrm{SGP}\{-0.5\sigma^2, \sigma^2, \rho(u,v)\} \end{aligned}$$

Goal: real-time mapping of $\mathrm{P}\{S(x,t) > c\}$ for pre-specified c

Diggle, Rowlingson and Su (2005)

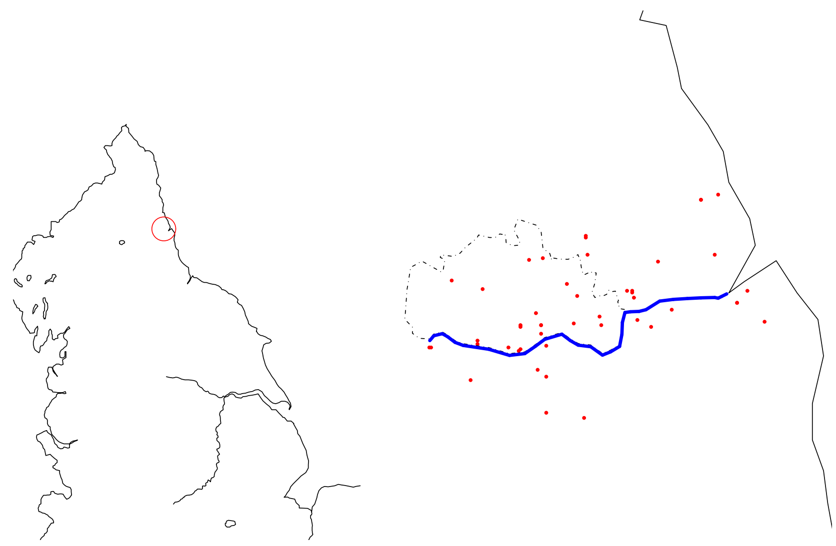
Animation at `www.lancaster.ac.uk/staff/diggle`

The PAMPER study

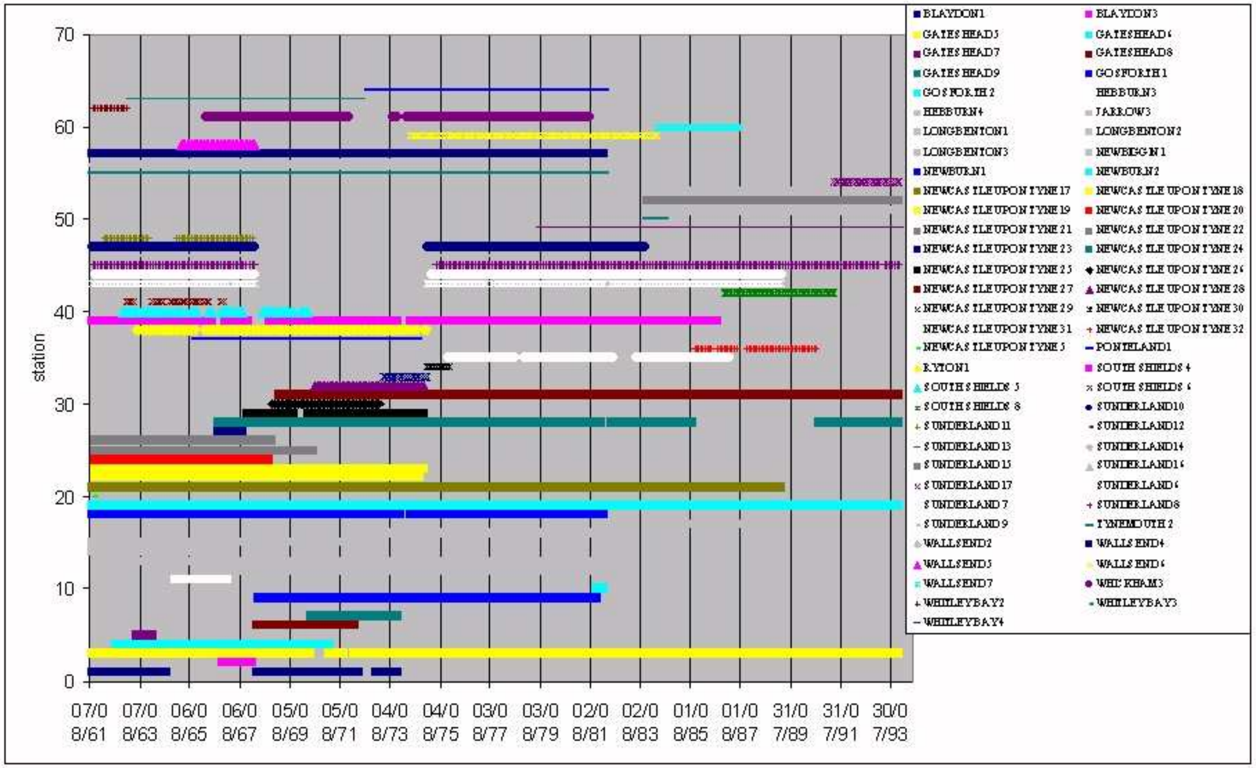
Goal: Construct predictions of black smoke levels, $S(x, t)$, over thirty-year period

Available data:

- monitored black smoke levels from spatially discrete monitoring network



- **monitors are only active intermittently**



Modelling strategy

Two-stage approach:

1. **model temporal variation in spatially averaged black smoke levels**

2. **model residual spatio-temporal variation about temporal average**

Model for temporal variation in spatially averaged black smoke

Y_t = spatially averaged black smoke at time t

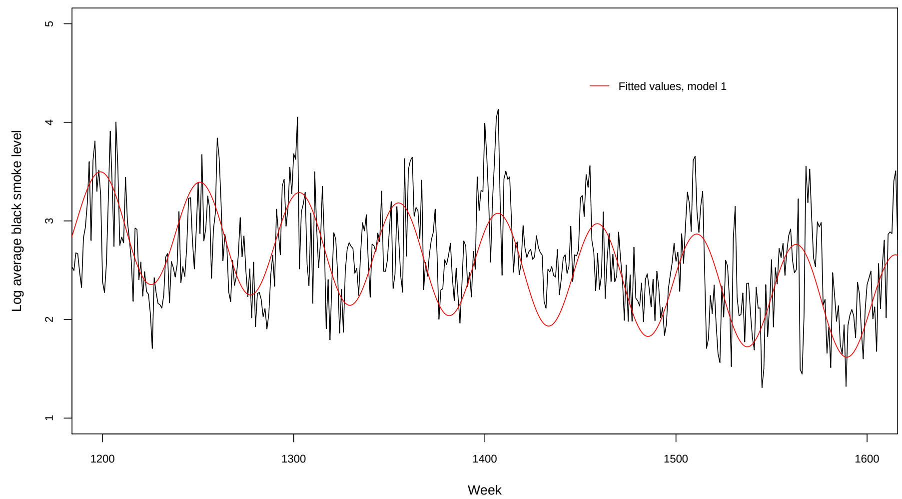
Model needs to take account of:

- long-term (decreasing) trend
- seasonal variation

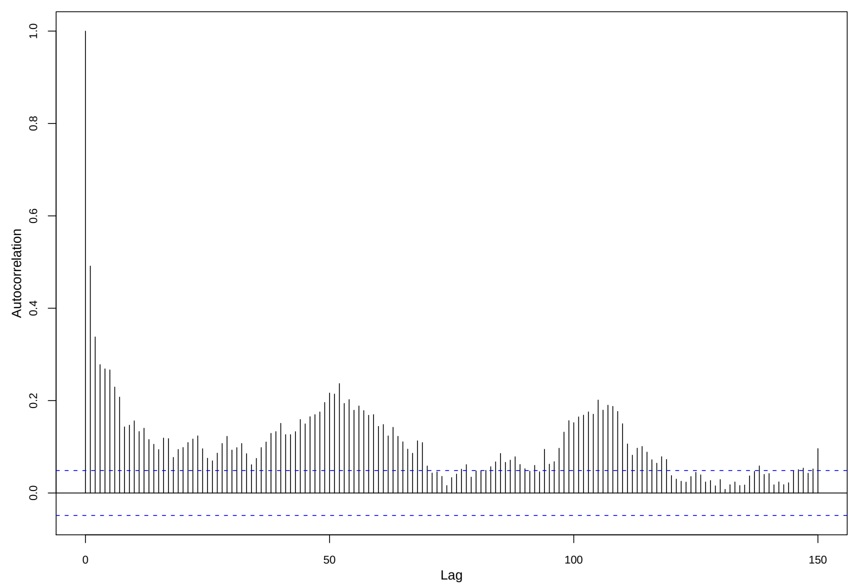
Classical regression model for Y_t is

$$\log P_t = \alpha + \beta t + \sum_{k=1}^{r} \{A_k \cos(k\omega t) + B_k \sin(k\omega t)\} + Z_t$$

Case $r = 1$ gives pure sinusoid, $r = 22, 3, ...$ allows non-sinusoidal seasonal patterns



Fitted values, model 1
Log average black smoke level
Week
1
2
3
4
5
1200
1300
1400
1500
1600



Autocorrelation
Lag
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150

Model for temporal variation in spatially averaged black smoke (continued)

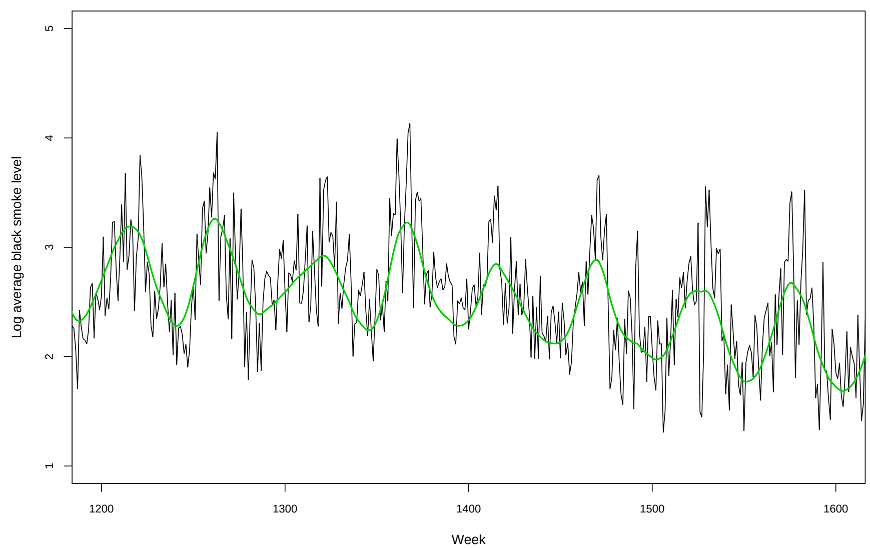
Classical model fails because seasonal pattern is stochastic.

Dynamic model:

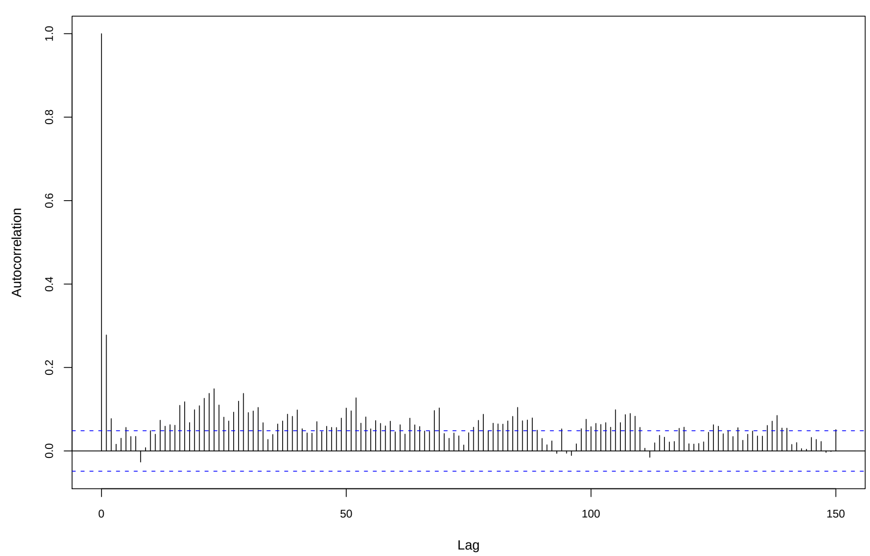
$$\log P_t = \alpha + \beta t + \{A_t \cos(\omega t) + B_t \sin(\omega t)\} + Z_t$$

$$\begin{aligned} A_t &= A_{t-1} + \epsilon_t \\ B_t &= B_{t-1} + \delta_t \end{aligned}$$

Allows locations and magnitudes of seasonal peaks and troughs to vary between years


Log average black smoke level
Week
1
2
3
4
5
1200
1300
1400
1500
1600

Model (2)


Autocorrelation
Lag
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150

Model for spatio-temporal variation in residuals

$$Y_t(x) = \log \hat{P}_t + S(x,t) + Z_t(x)$$

- $S(x,t)$ = spatio-temporally correlated (?) random field
- $Z_t(x)$ = mutually independent measurement errors

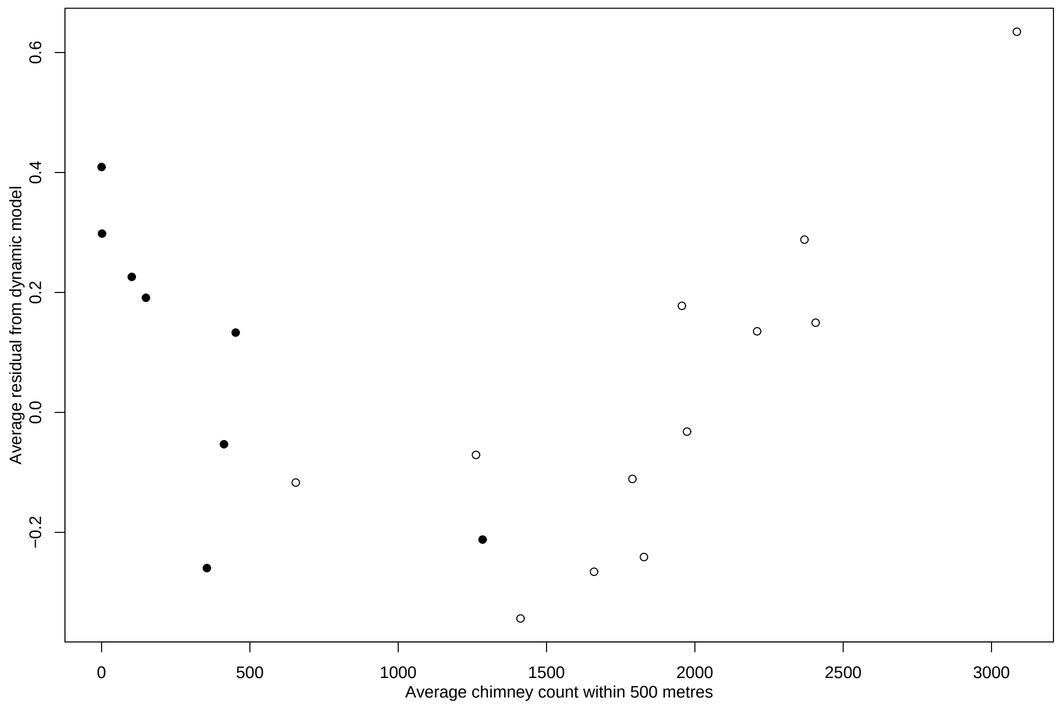
Constructed covariates

- **where does the spatio-temporal correlation come from?**
- **look for possible surrogate measures which:**
  - **are available at all locations and times**
  - **correlate well with measured black smoke concentrations at monitored locations**

Monitored black smoke vs domestic chimney density

Important interactions with:

- **non-residential/residential land-use (solid/open circles)**
- **clean-air act (staggered implementation)**



PAMPER analysis: discussion points

1. temporal takes precedence over spatial

2. construction of spatially continuous explanatory variables assists prediction of spatio-temporally continuous exposure surface

3. and may eliminate residual spatio-temporal correlation

Spatio-temporal point processes

$\mathcal{H}_t =$ complete history (locations and times of events)

$\lambda(x, t|\mathcal{H}_t) =$ conditional intensity (hazard) for new event at location x, time t, given history $\mathcal{H}_t$

Likelihood analysis

Log-likelihood for data $(x_i, t_i) \in A \times [0, T] : i = 1, ..., n$, with $t_1 < t_2 < ... < t_n$, is

$$L(\theta) = \sum_{i=1}^{n} \log \lambda(x_i, t_i | \mathcal{H}_{t_i}) - \int_0^T \int_A \lambda(x, t | \mathcal{H}_t) dx dt$$

Rarely tractable, but Monte Carlo methods available in special cases (eg log-Gaussian Cox processes)

Partial likelihood analysis

Data $(x_i, t_i) \in A \times [0, T] : i = 1, ..., n; \quad t_1 < t_2 < ... < t_n$

Condition on locations x_i and times t_i
Derive log-likelihood for observed ordering $1, 2, ..., n$

Need to distinguish between:

- **Spatially discrete set of potential points**
- **Spatially continuous set of potential points**

Partial Likelihood Formulation

- Condition on the locations x_i and times t_i

- $\mathcal{R}_i$: the risk set at time t_i

- Partial log-likelihood $L_p(\theta) = \sum_{i=1}^{n} \log p_i$
- Spatially discrete $\longrightarrow \mathcal{R}_i = \{i, i+1, ..., n\}$

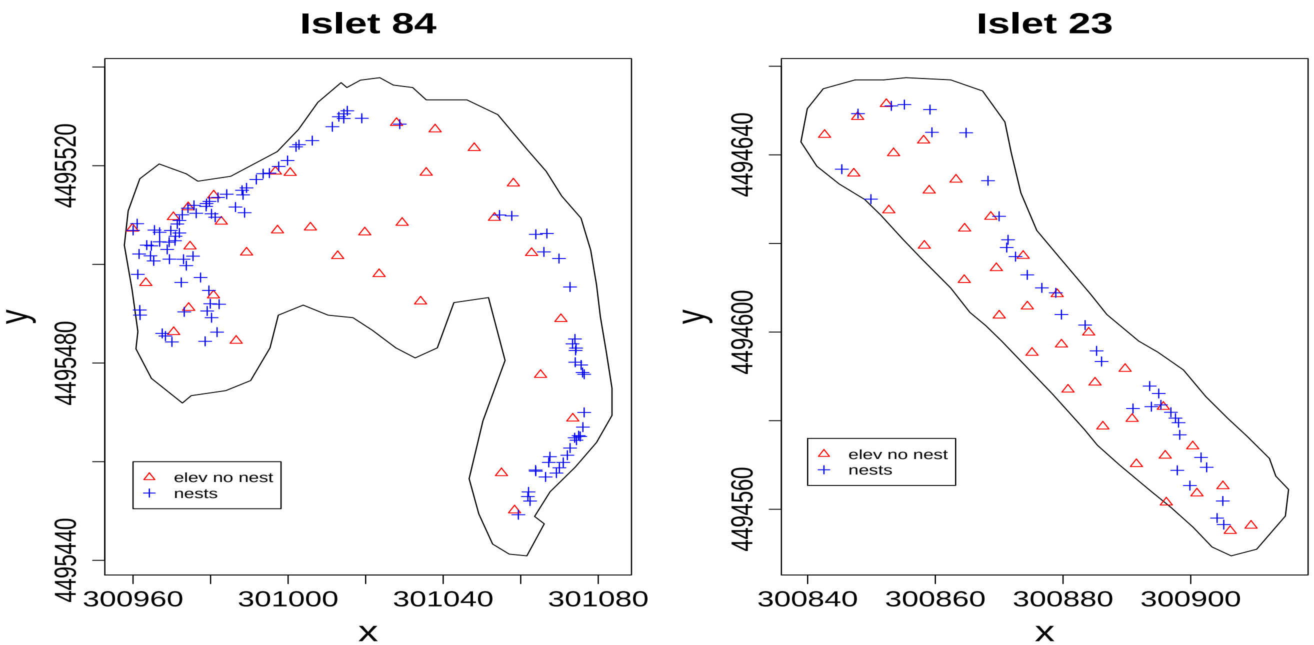
$$p_i = \frac{\lambda(x_i, t_i|\mathcal{H}_{t_i})}{\sum_{j \geq i} \lambda(x_j, t_i|\mathcal{H}_{t_i})}$$

- Spatially continuous $\longrightarrow \mathcal{R}_i \equiv A$

$$p_i = \frac{\lambda(x_i, t_i|\mathcal{H}_{t_i})}{\int_A \lambda(x, t_i|\mathcal{H}_{t_i}) dx}$$

Nesting colonies of common terns

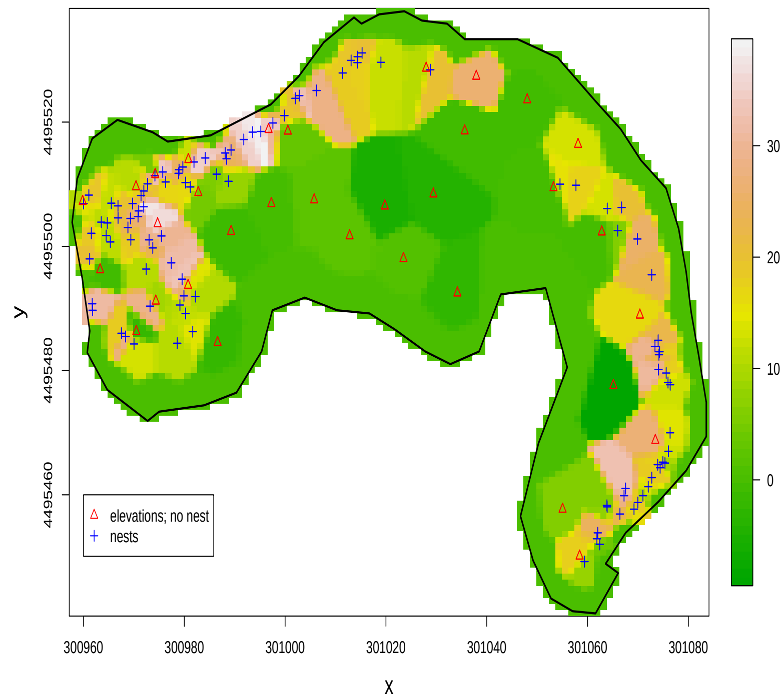
Islets 23 and 84



Coast boundaries (−), spatial locations of the nests (+), and other locations for which elevation is recorded (△) for islets 84 (left panel) and 23 (right panel)

Approximation of elevation surface

Approximate elevation surface $z(x)$ for islet 84 based on all available elevations and assuming piece-wise constant $z(x)$ within Voronoi tiles



Conditional intensity

$$\lambda(\mathrm{x}, t|\mathcal{H}_t) = \lambda_0(t) \exp\{\beta z(\mathrm{x})\} g(\mathrm{x}, t_i|\mathcal{H}_t)$$

- $g(\mathrm{x}, t|\mathcal{H}_t)$ models dependence on locations of earlier nests
- $\beta z(\mathrm{x})$ models log-linear effect of elevation

Two models for $g(\cdot)$

- $\mathcal{M}_1$:

$$g(\mathrm{x}, t|\mathcal{H}_t) = h\left(\min_{j:t_j<t}(||\mathrm{x}_j - \mathrm{x}||)\right)$$

- $\mathcal{M}_2$:

$$g(\mathrm{x}, t|\mathcal{H}_t) = \prod_{j:t_j<t} h(||\mathrm{x} - \mathrm{x}_j||)$$

$$h(u) = \begin{cases} 0, & u \leq d_0 \\ 1 + \theta \exp\left\{-\frac{(u-d_0)^c}{\phi}\right\}, & u > d_0 \end{cases}$$

- $c = 1 \longrightarrow$ exponential kernel
- $c = 2 \longrightarrow$ Gaussian kernel

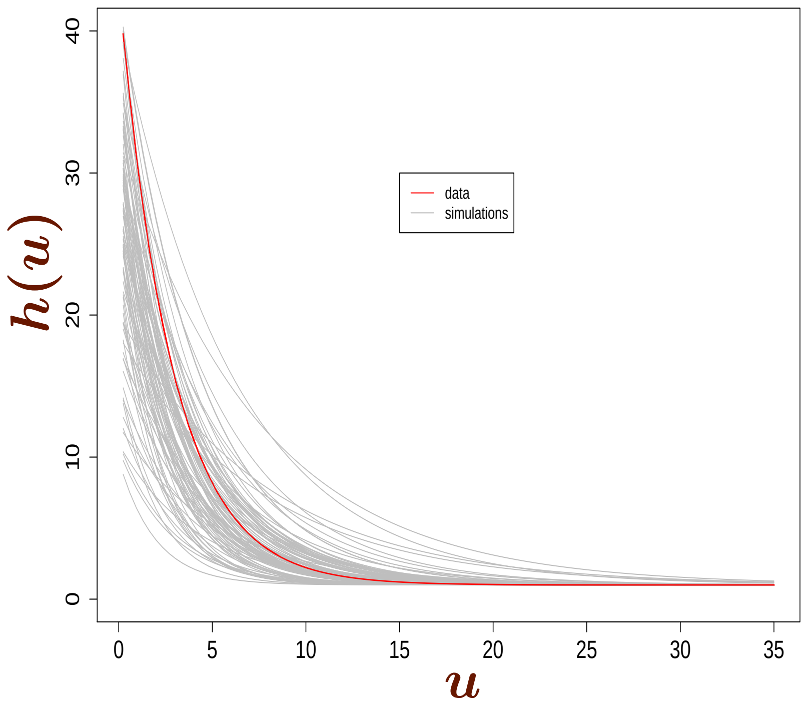
Results

- **assumption of common set of parameters in islets 23 and 84 is not supported by the data**
- **but dataset for islet 23 is uncomfortably small for formal inference (36 events)**
- **likelihood ratio tests favour model $\mathcal{M}_1$ (nearest neighbour distance only) with $c = 1$ (exponential kernel)**
- **highly significant effect of elevation**

  $\hat{\beta} = 0.05, SE = 0.0006, p << 0.001$

Monte Carlo interval estimation

Envelope of estimates $\hat{h}(u)$ from 99 simulations of fitted model



Conclusions

- **spatio-temporal data-sets becoming widely available**
- **different problems require different modelling strategies**
- **temporal should often take precedence over spatial**
- **routine implementation is an important consideration when exploring many different models**

Any questions?

And I leave you with...

- **the role of modelling**

  "We buy information with assumptions"

  Coombs (1964)

- **choice of model/method should relate to scientific purpose.**

  "Analyse problems, not data"

  PJD

- **spatial and longitudinal data analysis are challenging, but rewarding tasks**

> **“La peinture de l’huile,**
> **c’est tres difficile**
> **Mais c’est beaucoup plus beau,**
> **que la peinture de l’eau”**

Winston Churchill

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society* B **36**, 192–225.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with Discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Coombs, C.H. (1964). *A Theory of Data*. New York : Wiley

Cox, D.R. (1955). Some statistical methods related with series of events (with Discussion). *Journal of the Royal Statistical Society*, B **17**, 129–57.

Diggle, P.J., Eglen, S.J. and Troy, J.B. (2006). Modelling the bivariate spatial distribution of amacrine cells. In *Case Studies in Spatial Point Processes*, ed A Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, 215–233. New York: Springer.

Diggle, P.J. , Farewell, D. and Henderson, R. (2007). Longitudinal data with dropout: objectives, assumptions and a proposal (with Discussion). *Applied Statistics*, **56**, 499–550.

Diggle, P.J., Heagerty, P., Liang, K-Y and Zeger, S.L. (2002). *Analysis of Longitudinal Data (second edition)*. Oxford : Oxford University Press.

Diggle, P.J., Menenzes, R. and Su, T-L. (2008). Geostatistical analysis under preferential sampling. `http://www.bepress.com/jhubiostat/paper162/`

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

Eglin, S.J. and Wong, J.C.T. (2008). Spatial constraints underlying the retinal mosaics of two types of horizontal cells in cat and macaque. *Visual Neuroscience*, **25**, 209-213.

Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**, 119–39.

Liang, K-Y and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–21.

Matérn, B. (1960). *Spatial variation.* Technical Report, Statens Skogsforsningsinstitut, Stockholm (reprinted, 1986, Berlin : Springer-Verlag).

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, **58**, 1246–66.

Mercer, W.B. and Hall, A.D. (1911). The experimental error of field trials. *Journal of Agricultural Science*, **4**, 107–132.

Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society* B **39**, 172–212.

Ripley, B.D. (1981). *Spatial Statistics.* New York : Wiley.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–92.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* London: Chapman and Hall.

Watson, G.S. (1972). Trend surface analysis and spatial correlation. *Geological Society of America Special Paper*, **146**, 39–46.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, **8**, 158–183.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester: Wiley