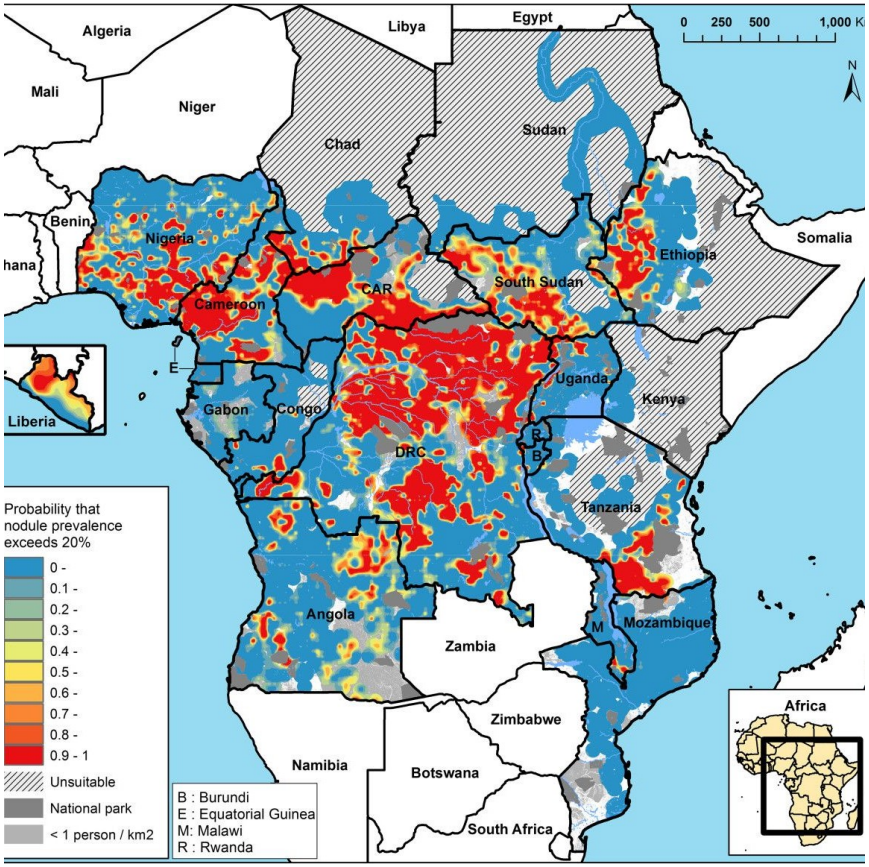
Geostatistical Methods for Disease Mapping

3-5 June 2019, Lancaster University

Peter J Diggle and Emanuele Giorgi

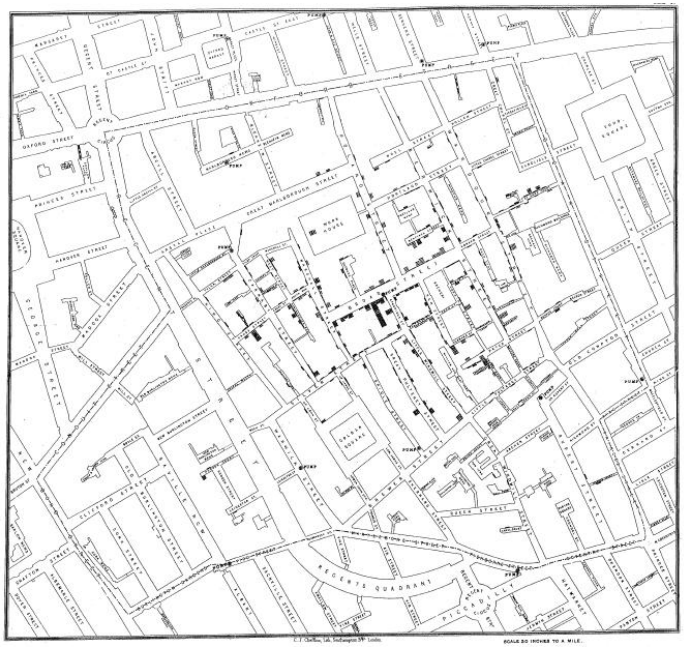
(CHICAS, Lancaster University)

1. **Introduction.** Some history, motivating examples, exploratory analysis, spatial residuals, the variogram
2. **Linear models.** Spatial corrrelation, Gaussian processes, parametric specifications of spatial correlation structure, parameter estimation, spatial prediction, application to lead poillution in Galicia
3. **Generalized linear models.** Prevalence surveys, binomial models, parameter estimation, spatial prediction, application to river blindness in Liberia
4. **Extensions.** ... as time permits!



Diggle, P.J. and Giorgi, E. (2019). Model-based Geostatistics: Methods and Applications in Global Public Health. (CRC Press)

Origins of spatial epidemiology: Cholera in Victorian London, 1854



The physician John Snow famously removed the handle of the Broad Street water-pump, having concluded (correctly) that infected water was the source of the disease contrary to conventional wisdom at the time.

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Epidemic vs endemic patterns of incidence

- **Foot-and-mouth in Cumbria (the 2001 epidemic)**

  Diggle (2006)

- **Gastro-enteric disease in Hampshire (AEGISS)**

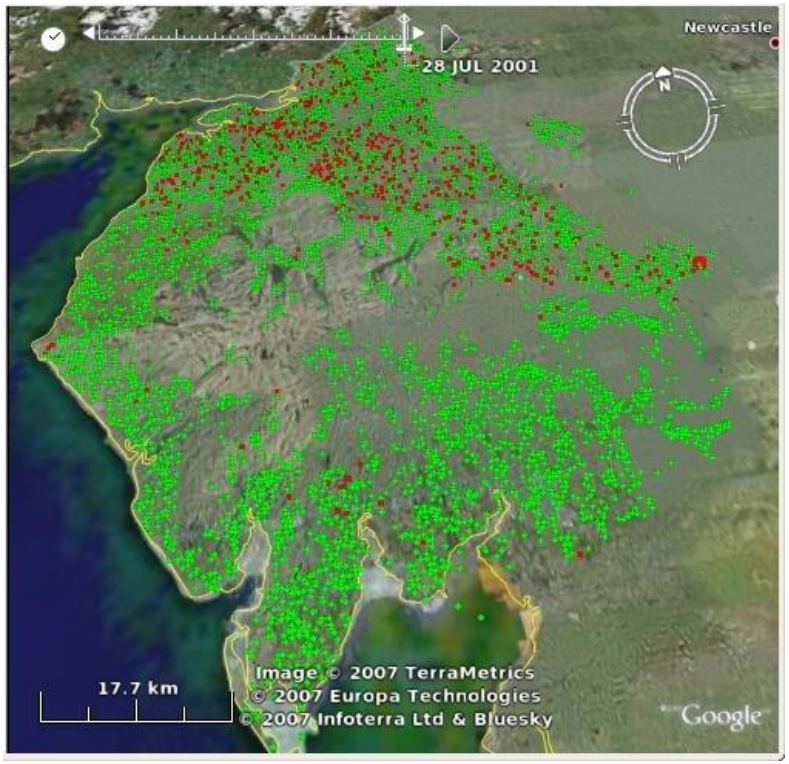
  Diggle, Rowlingson and Su (2005)

Animations at: `http://www.lancaster.ac.uk/staff/diggle/`

What are the similarities and differences between the two phenomena?

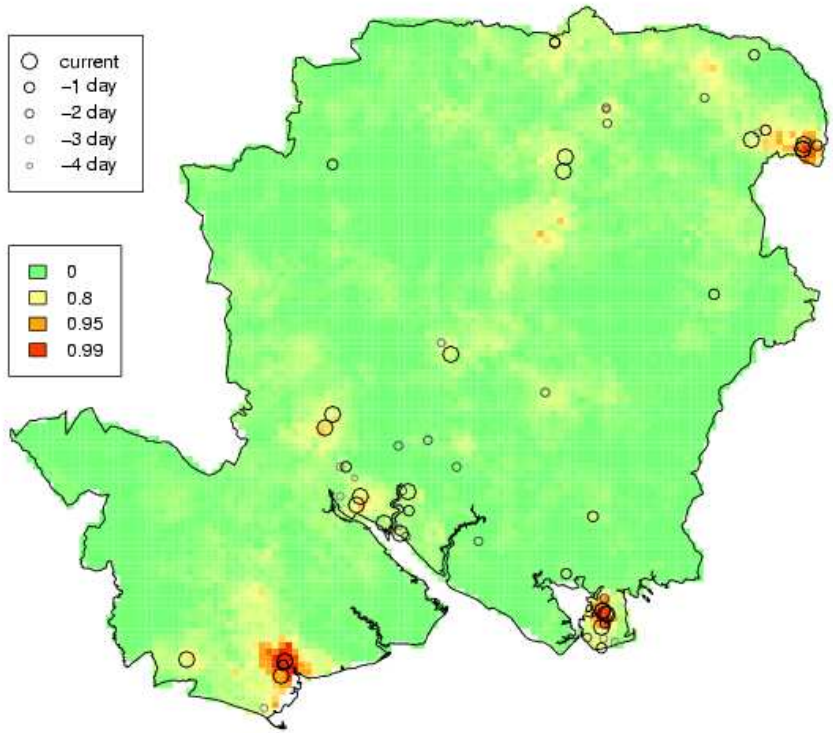
Mechanistic modelling: the 2001 UK FMD epidemic (Diggle, 2006)

- **Predominantly a classic epidemic pattern of spread from an initial source**
- **Occasional apparently spontaneous outbreaks remote from prevalent cases**
- **$\lambda(x, t|\mathcal{H}_t)$ =conditional intensity, given history $\mathcal{H}_t$**



Empirical modelling: The AEGISS project (Diggle, Rowlingson and Su, 2005)

- **early detection of anomalies in local incidence**
- **data on 3374 consecutive reports of non-specific gastro-intestinal illness**
- **log-Gaussian Cox process, space-time correlation $\rho(u, v)$**



A hierarchical modelling framework

Need to distinguish between:

- **(scientific) modelling of a process whose behaviour we wish to understand;**
- **(statistical) modelling of data that tell us something about the process**

A general framework:

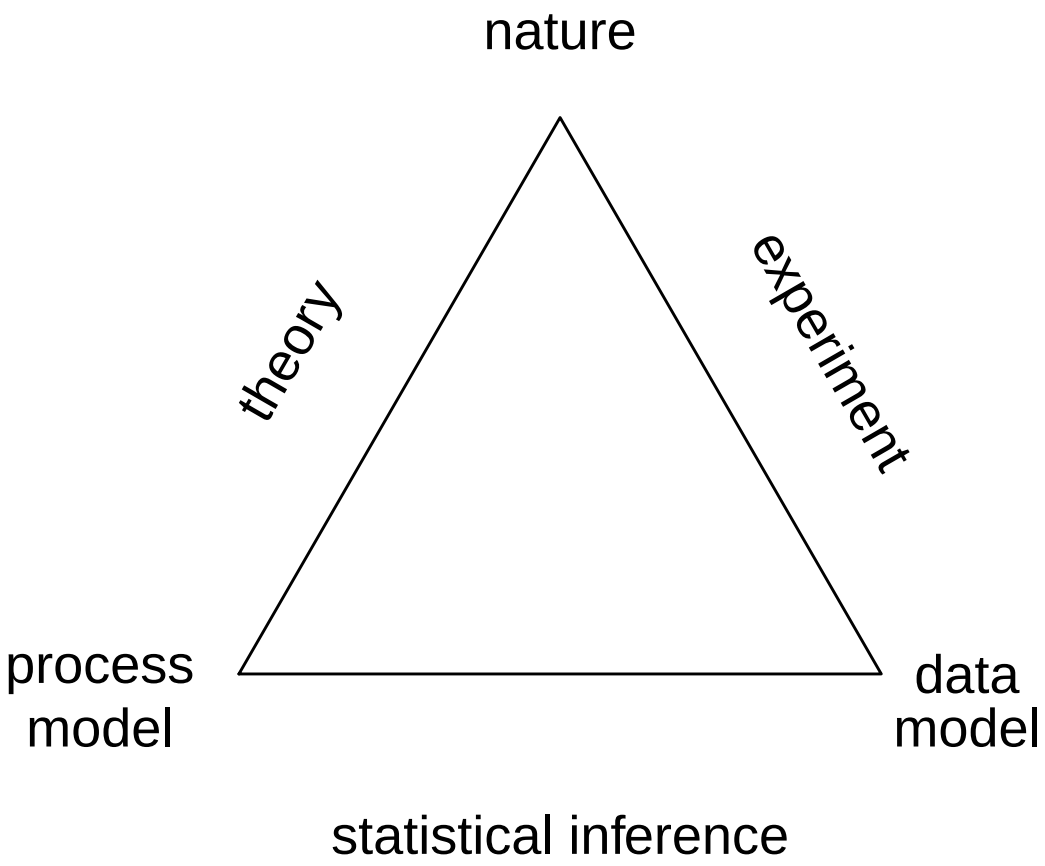
- **$[\cdot]$ means** *the distribution of*

  S : **the scientific process we wish to understand** **(signal)**
  Y : **data that can help us understand the process** **(noise)**

- **hierarchical formulation:**

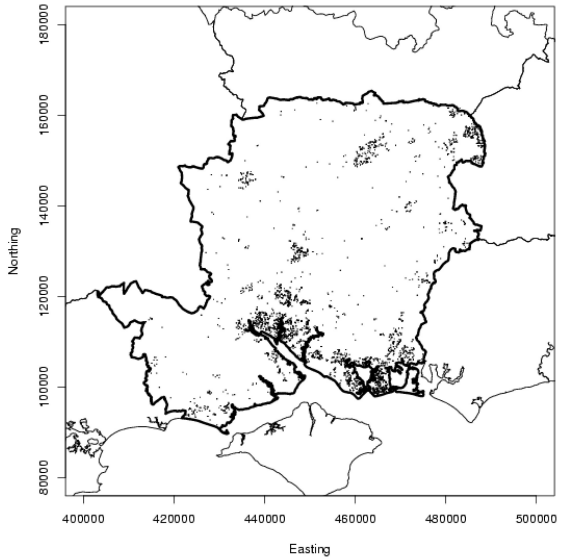
$$[S, Y] = [S][Y|S]$$

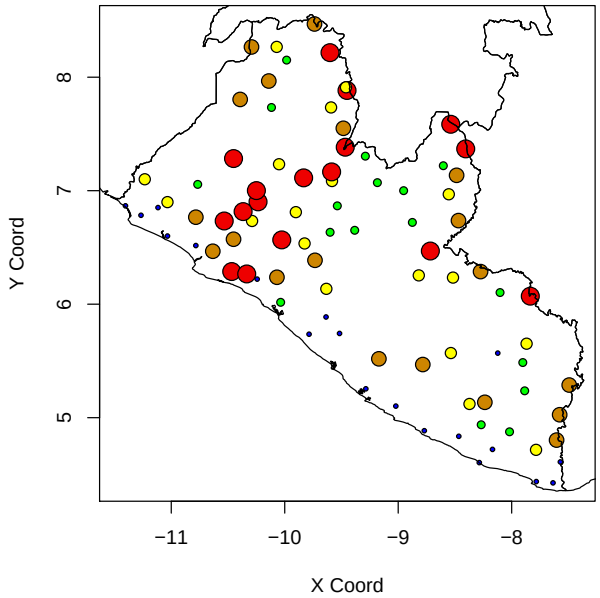
Statistics and scientific method



Diggle and Chetwynd (2011)

Three data-sets





Cancer rates in administrative areas

Calls to NHS Direct in Hampshire, UK

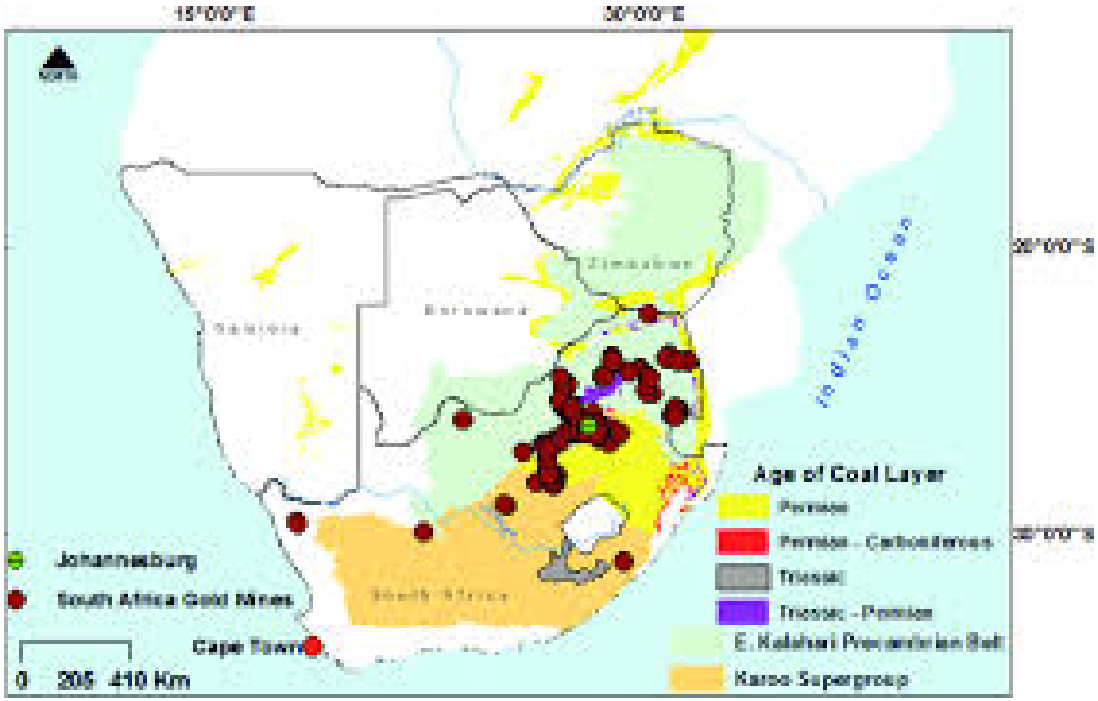
Prevalence surveys in Liberia

Are the three underlying processes fundamentally different?

Spatial stochastic processes

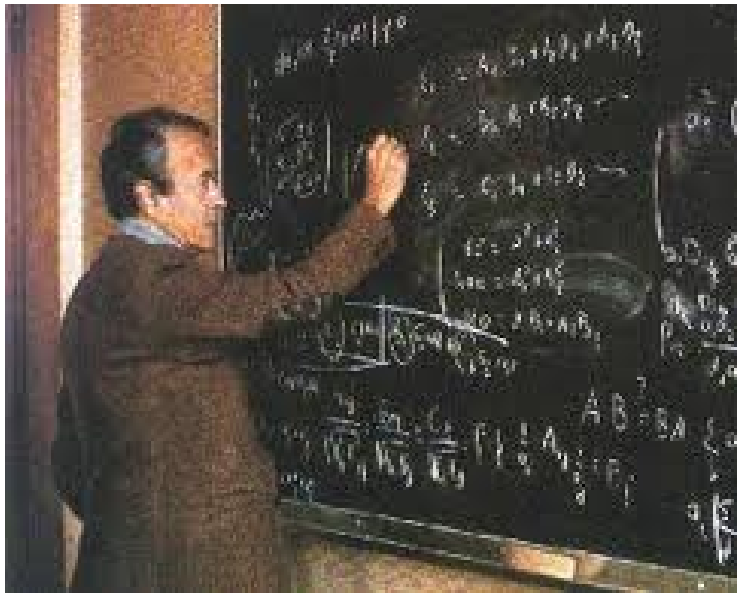
1. A **stochastic process** is a collection of random variables
2. A **spatial stochastic process** is a stochastic process in which each random variable is associated with a position in space
3. Three important types of spatial stochastic process:
   - **discrete spatial variation:** the random variables associate a real value with a particular, pre-specified, set of points in space, hence $\{(S_i, x_i) : i = 1, ..., n\}$
   - **point processes:** the random variables are the locations themselves, $\{x_i : i = 1, ..., n\}$
   - **continuous spatial variation:** the random variables associate a real value with every point in the space, hence $\{S(x) : x \in \mathbb{R}^2\}$

History: D G Krige and the South African mining industry



Krige (1951)

From South Africa to Fontainebleau...Georges Matheron and classical geostatistics

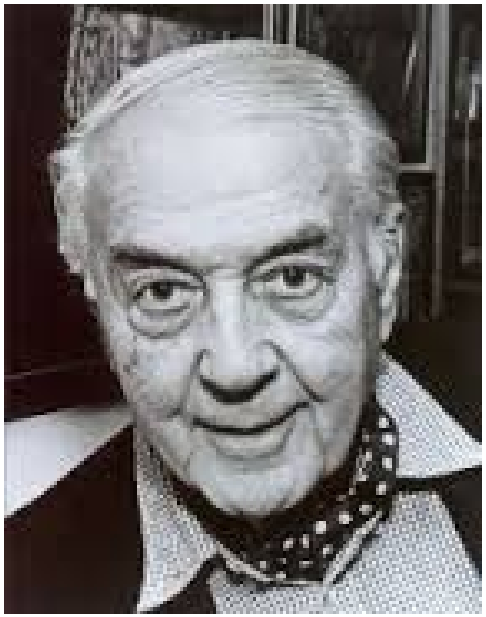


Matheron (1963)

Meanwhile in the forest...Bertil Matérn and the Royal College of Forestry

Matérn (1960)

Into the statistical mainstream...Watson, Ripley, Cressie



Watson (1972)

Ripley (1981)

Cressie (1991)

Geostatistics

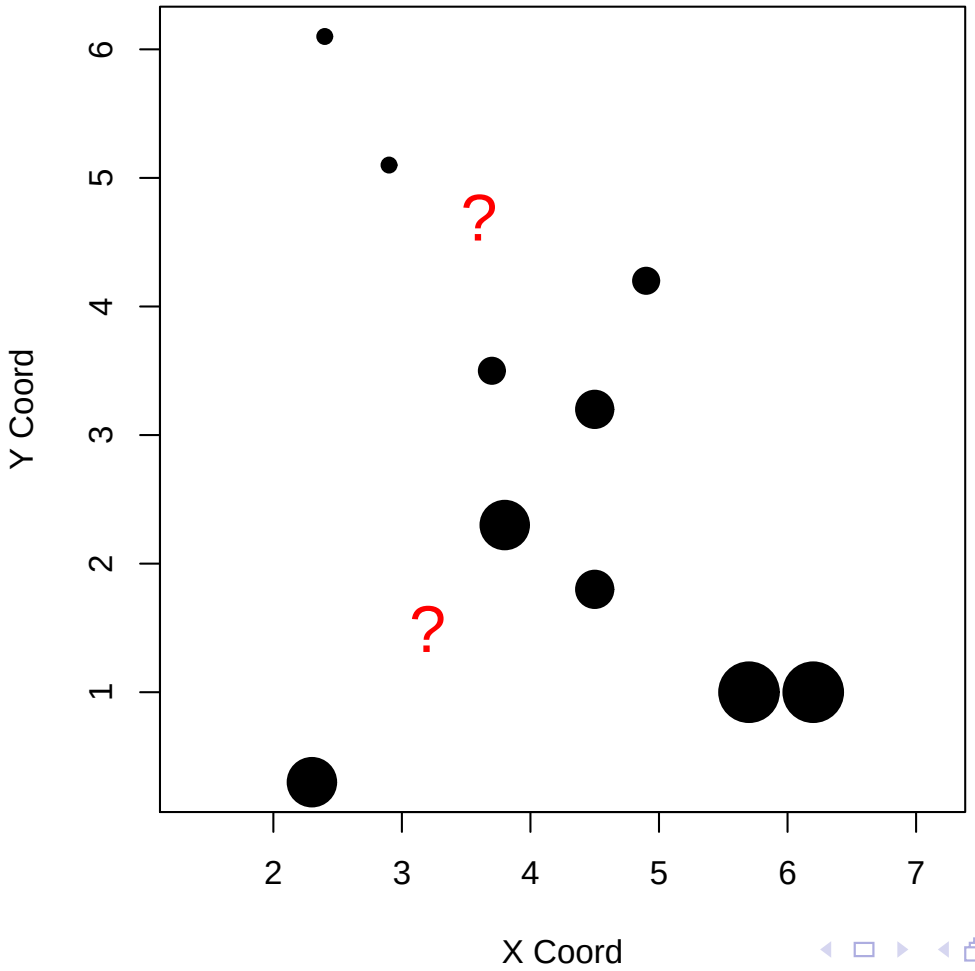
- **traditionally, a self-contained methodology for spatial prediction:**
  - **origins in the South African mining industry**
  - **subsequently developed at École des Mines, Fontainebleau, France**
- **nowadays, that part of spatial statistics which is concerned with data obtained by spatially discrete sampling of a spatially continuous process**

Model-based geostatistics: the application of general principles of statistical modelling and inference to geostatistical problems

Diggle, Moyeed and Tawn (1998); Diggle and Ribeiro (2007)

The canonical geostatistical problem

Given a set of measurements $Y_i : i = 1, \ldots, n$ at locations x_i in a spatial region A, presumed to be (noisy) measurements of a spatially continuous phenomenon $S(x_i)$, what can we say about the realisation of $S(x)$ throughout A?



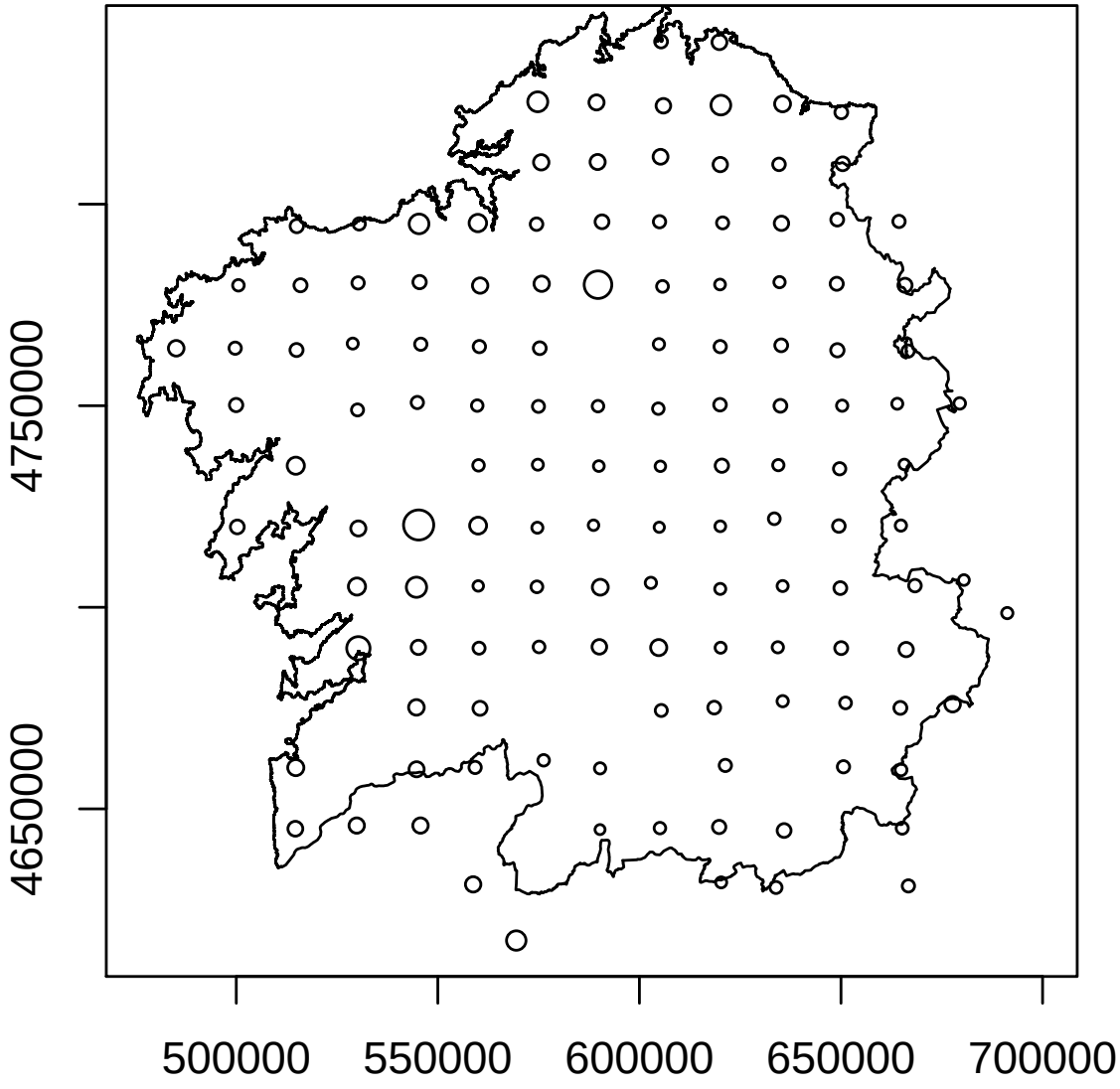
Geostatistical problems more generally

- **Design:** how to choose locations x_i at which to collect outcome data
- **Estimation:** how to investigate relationship between outcome and covariates when data may be spatially correlated
- **Prediction:** how to map (expected value of) outcome throughout the study-region

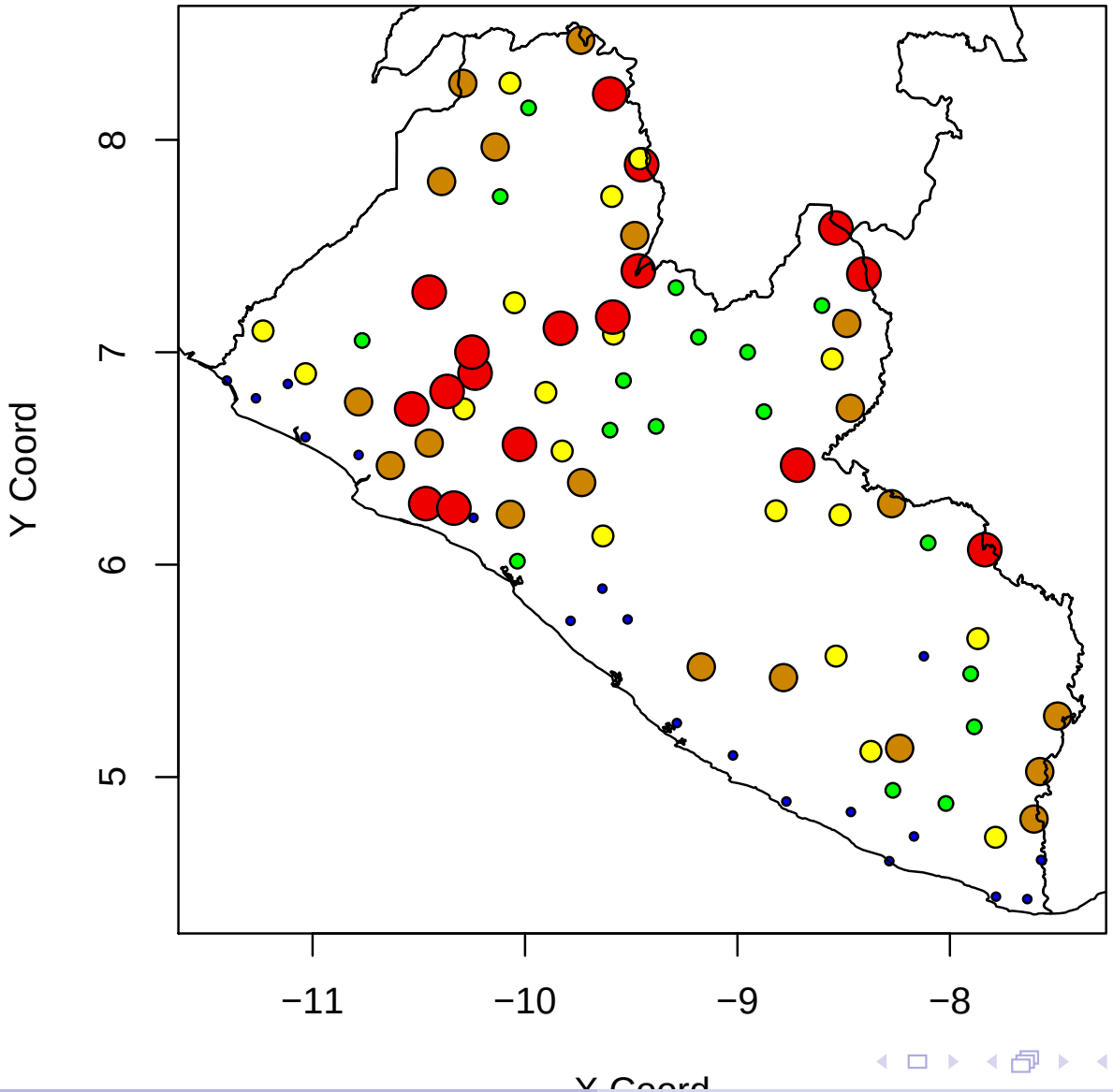
Practical point:

- **Estimation** only requires covariate information at locations x_i
- **Prediction** requires covariate information throughout the study-region.

Example 1. Heavy metal monitoring in Galicia



Example 2. River-blindness in Liberia



Plotting geostatistical data: the PrevMap package

```
library(PrevMap)
data<-read.csv(file.choose()) # galicia.csv
names(data)
table(data$survey)
lead2000<-data[data$survey==2000,1:3]
lead2000<-as.data.frame(lead2000)
point.map(lead2000,var.name=~lead,coords=~x+y)
point.map(lead2000,var.name=~lead,coords=~x+y,
pt.div="quintiles",cex.min=0.7,cex.max=1.3)
bound<-read.csv(file.choose()) # galicia_bndrs.csv
lines(bound)
```

The first law of geography

Everything is relatd to everyting else, are related, but close things are more related than distant things

Waldo Tobler

- **spatial correlation is the statistical version of Tobler's law**
- **correlation as a function of distance**
- **but for irregular spatial distributions, we can't line up data-points in pairs to calculate empirical correlations**
- **solution is the variogram**

The variogram

Geostatistical data

Locations $x_i : i = 1, \ldots, n$ with associated measurements y_i

Empirical variogram

$$u_{ij} = \|x_i - x_j\| \qquad v_{ij} = \frac{1}{2}(y_i - y_j)^2$$

- **variogram cloud is a scatterplot of the points (u_{ij}, v_{ij}) (not recommended)**
- **empirical variogram smooths the variogram cloud by averaging within bins: $u - h/2 \leq u_{ij} < u + h/2$**

An empirical variogram for the lead pollution data

```
vario1<-variogram(lead2000,var.name=~lead,coords=~x+y)
plot(vario1)
plot(vario1,pch=19,col="red")
?variogram
names(vario1)
vario1$u
u<-5000*(1:20)    # choose your own distance bins
vario2<-variogram(lead2000,var.name=~lead,coords=~x+y,uvec=u)
plot(vario2$u,vario2$v,type="l",xlim=c(0,100000),
   xlab="u",ylab="V(u)")    # too noisy?
u<-10000*(1:10)
vario3<-variogram(lead2000,var.name=~lead,coords=~x+y,uvec=u)
plot(vario3$u,vario3$v,type="l",xlim=c(0,100000),ylim=c(0,max(va
   xlab="u",ylab="V(u)")
```

What is the variogram estimating?

Suppose random variables Y_1 and Y_2 have expectation zero, variance σ^2 and correlation ρ

$$\begin{aligned} \mathrm{E}[\{Y_1 - Y_2\}^2] &= \mathrm{E}[(Y_1^2] + \mathrm{E}[Y_2^2] - 2\mathrm{E}[Y_1 Y_2] \\ &= \mathrm{Var}[Y_1] + \mathrm{Var}[Y_2] - 2\mathrm{Cov}[Y_1, Y_2] \\ &= 2\sigma^2(1-\rho) \end{aligned}$$

Spatial version: $Y(x), Y(x-u)$

$$\frac{1}{2}E[\{Y(x) - Y(x-u)\}^2] = V(u) = \sigma^2\{1-\rho(u)\}$$

What is the variogram estimating? (2)

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

The function $\rho(u)$ is called the spatial correlation function of the spatial process $Y(x)$

Key assumption is stationarity:

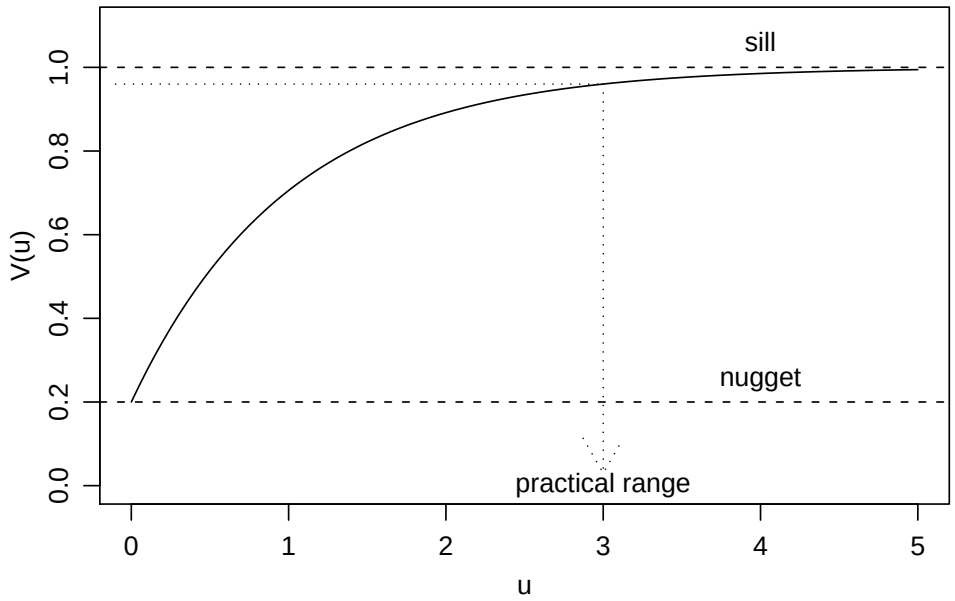
- **spatially constant expectation, $\mu(x) = \mu$**
- **spatially constant variance, $\sigma^2(x) = \sigma^2$**
- **correlation is a function of distance**

How to accommodate explanatory variables?

- **fit a provisional regression model**
- **calculate variogram of standardised residuals**

Structural parameters

$$(\tau^2, \sigma^2, \phi)$$



- **the nugget variance**: τ^2
- **the sill**: $\tau^2 + \sigma^2$, where $\sigma^2 = \mathrm{Var}\{S(x)\}$
- **the practical range** r: $\rho(r) = 0.05$

Example: $\rho(u) = \exp(-u/\phi) \Rightarrow r \approx 3\phi$

Linear models

Data

Measurements Y_i are made at locations x_i with associated explanatory variables $d(x_i)$

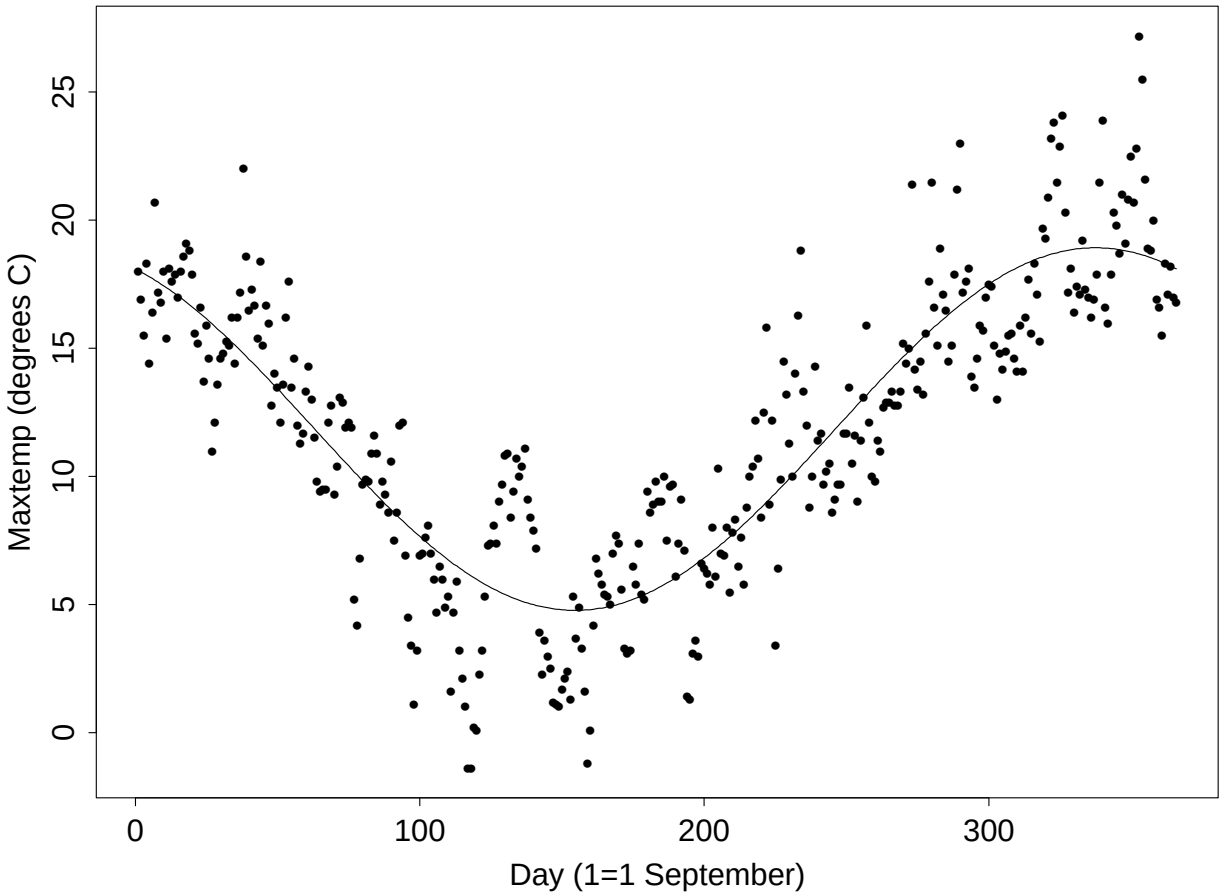
Classical (non-spatial) linear model

$$Y_i = d(x_i)^\top \beta + U_i : U_i \sim \mathrm{N}(0, \sigma^2)$$

- **Independence.** Y_i are mutually independent random variables.
- **Linearity.** $E[Y_i] = \mu_i = d(x_i)^\top \beta$
- **Homoscedasticity.** $\mathrm{Var}\,[Y_i] = \sigma^2, i = 1, \ldots, n$
- **Error-free explanatory variables?**
- **Normality.** $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$

Questioning independence: a digression into time series

- **maximum daily temperatures (degrees C) at Bailrigg (Lancaster) field-station, September 1995 to August 1996**
- **note that an unusually cold Christmas 1995 was followed by a mild period in January-February**



Bailrigg temperature data: points for discussion

- **what are the main features of the data?**
- **how did I fit the smooth curve to the data?**
- **what features are and are not explained by the fitted curve?**

A harmonic regression model

$$Y(t) \quad = \quad \mu + \alpha \cos(2\pi t/p + \phi) + \text{residualresidual}$$

- μ = overall mean value (of time series $Y(t)$)
- p = period
- α = amplitude
- ϕ = phase

Usually, the period is known, but the mean, amplitude and phase are not

Fitting the model

Using a standard trigonometric identify,

$$\cos(A + B) = \cos(A)\cos(B) - \sin(A)\sin(B)$$

we re-write the model as

$$Y(t) = \mu + \beta_1 \cos(2\pi t/p) + \beta_2 \sin(2\pi t/p) + \text{residual}$$

Note that if we know the period, p, we also know the values of

$$x_1(t) = \cos(2\pi t/p) \qquad x_2(t) = \sin(2\pi t/p)$$

Re-write the model as a linear regression model,

$$Y = \mu + \beta_1 x_1 + \beta_2 x_2 + \text{residual}$$

Using the lm() function to fit the model

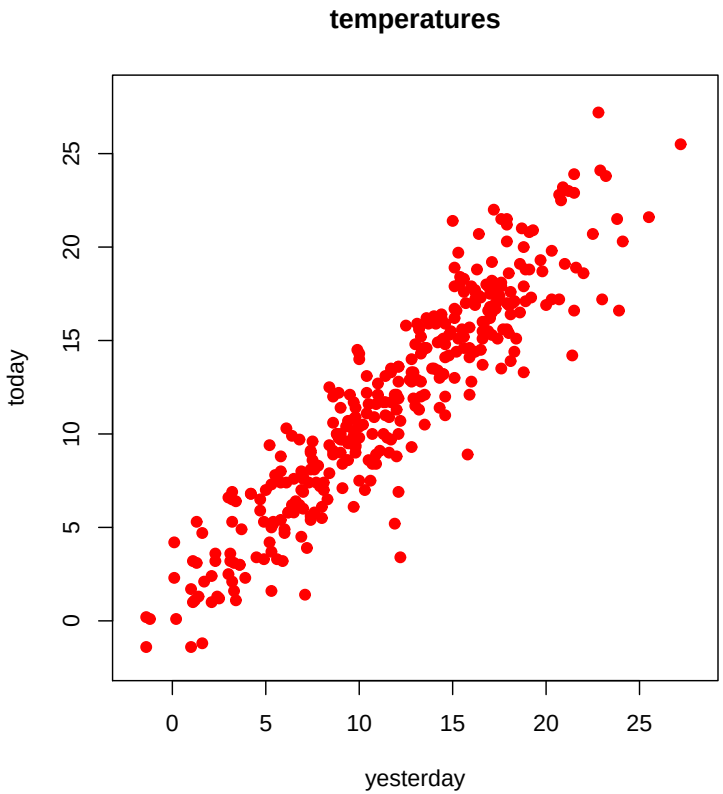
```
data<-read.csv(file.choose()) # maxtemp_data.csv
names(data)
y<-data[,4]
day<-1:366                          # data from 1/09/1995 to 31/08/19
x1<-cos(2*pi*day/366); x2<-sin(2*pi*day/366)
fit<-lm(y~x1+x2)
summary(fit)
```

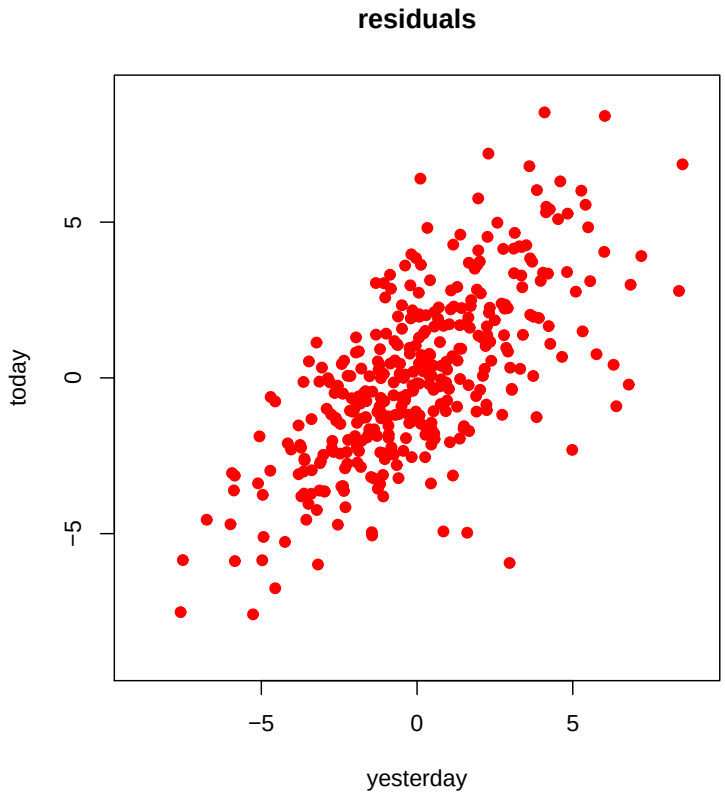
Can we trust the results?

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5921 -1.8240 -0.1475  1.7140  8.5232

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.8467     0.1441   82.22   <2e-16 ***
x1                6.2508     0.2038   30.68   <2e-16 ***
x2               -3.3177     0.2038  -16.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.756 on 363 degrees of freedom
Multiple R-Squared: 0.7687,     Adjusted R-squared: 0.7674
F-statistic: 603.1 on 2 and 363 DF,  p-value: < 2.2e-16
```

Autocorrelation

- relationship between today's and yesterday's temperature?
- relationship between today's and yesterday's residual?





- how and why are the two relationships different?

Conclusion: failure to correct for a time-varying mean results in misleading estimates of autocorrelation

Back to geostatistics: linear models

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

Classical linear model

$$Y_i = \mu(x_i) + U_i : i, ..., n \qquad \mu(x_i) = d(x_i)'\beta$$

Geopatial linear model

$$Y_i = \mu(x_i) + S(x_i) + Z_i : i, ..., n \qquad \mu(x_i) = d_i'\gamma + w(x_i)'\beta$$

- $S(x)$ spatially corrrelated stochastic process, mean zero, variance σ^2
- Z_i uncorrelated $\mathrm{N}(0, \tau^2)$ (measurement errors?)
- if $x_i = x_j$ then $S(x_i) = S(x_j)$ but $Z_i \neq Z_j$

A graphical diagnostic for residual spatial correlation

- calculate empirical variogram of standardised residuals,

$$\hat{V}(u_k) : k = 1, 2, ..., m$$

- re-calculate and plot empirical variogam after independent random permutations of residuals

Application to the lead pollution data

```
u<-10000*(1:10)
spat.corr.diagnostic(lead~1,coords=~x+y,data=lead2000,
                                    likelihood="Gaussian",uvec=u)
lead2000$loglead<-log(lead2000$lead)
diag<-spat.corr.diagnostic(loglead~1,coords=~x+y,data=lead2000,
                                   likelihood="Gaussian",uvec=u)
```

Which result should we believe?

Transforming the response variable

Reasons to transform the response variable

In decreasing order of importance:

- **to linearise the relationship with explanatory variables**
- **to achieve a constant variance**
- **to satisfy Gaussian distributional assumptions**

The first rule when analysing biological data is to take logs

Robert Elston

```
par(mfrow=c(1,2))
hist(lead2000$lead); hist(lead2000$loglead)
```

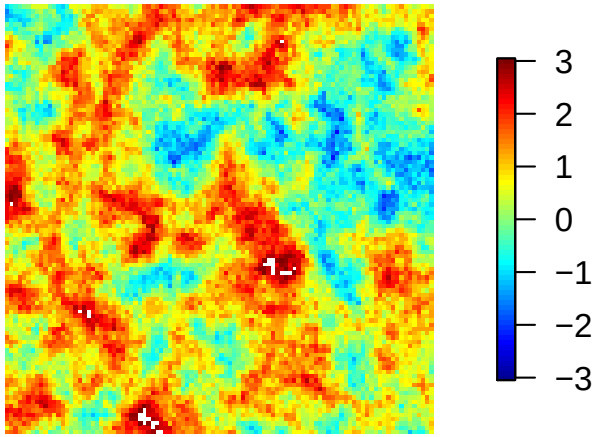
Fitting a geospatial linear model

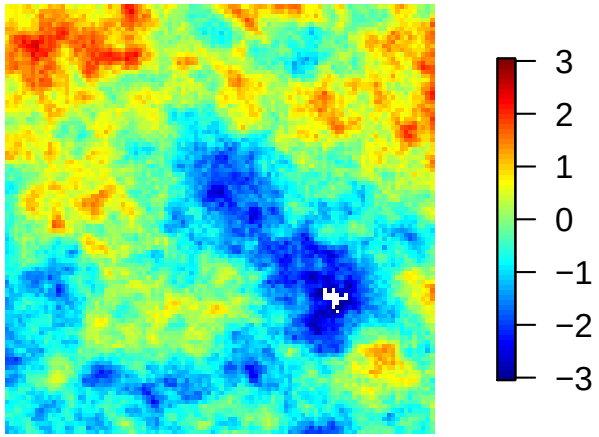
- **identify explanatory variables**
- **fit classical linear model by ordinary least squares**
- **calculate empirical variogram of residuals**
- **identify parametric model for spatial correlation**
- **re-fit using maximum likelihood, typically assuming Gaussian model for spatial process $S(x)$ and measurement errors Z_i**

Gaussian processes

Why Gaussian?

A spatial process $S(x)$ is Gaussian if all of its finite-dimensional distributions are multivariate Normal





The geospatial linear Gaussian model

- **Stationary** Gaussian process $S(x) : x \in \mathbb{R}^2$
  - $\mathrm{E}[S(x)] = 0$
  - $\mathrm{Cov}\{S(x), S(x')\} = \sigma^2 \rho(u) \qquad u = \|x - x'\|$
- Mutually independent $Y_i | S(\cdot) \sim \mathrm{N}(\mu(x_i), \tau^2)$
- $\mu(x) = d(x)'\beta$

Correlation function: how to parameterise $\rho(u)$?

The Matérn family of correlation functions

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

$K_{\kappa}(\cdot)$: modified Bessel function of order κ

Interpretation

- κ determines smoothness of underlying Gaussian process

$$\kappa > r \Rightarrow S(x) \text{ is } r \text{ times differentiable}$$

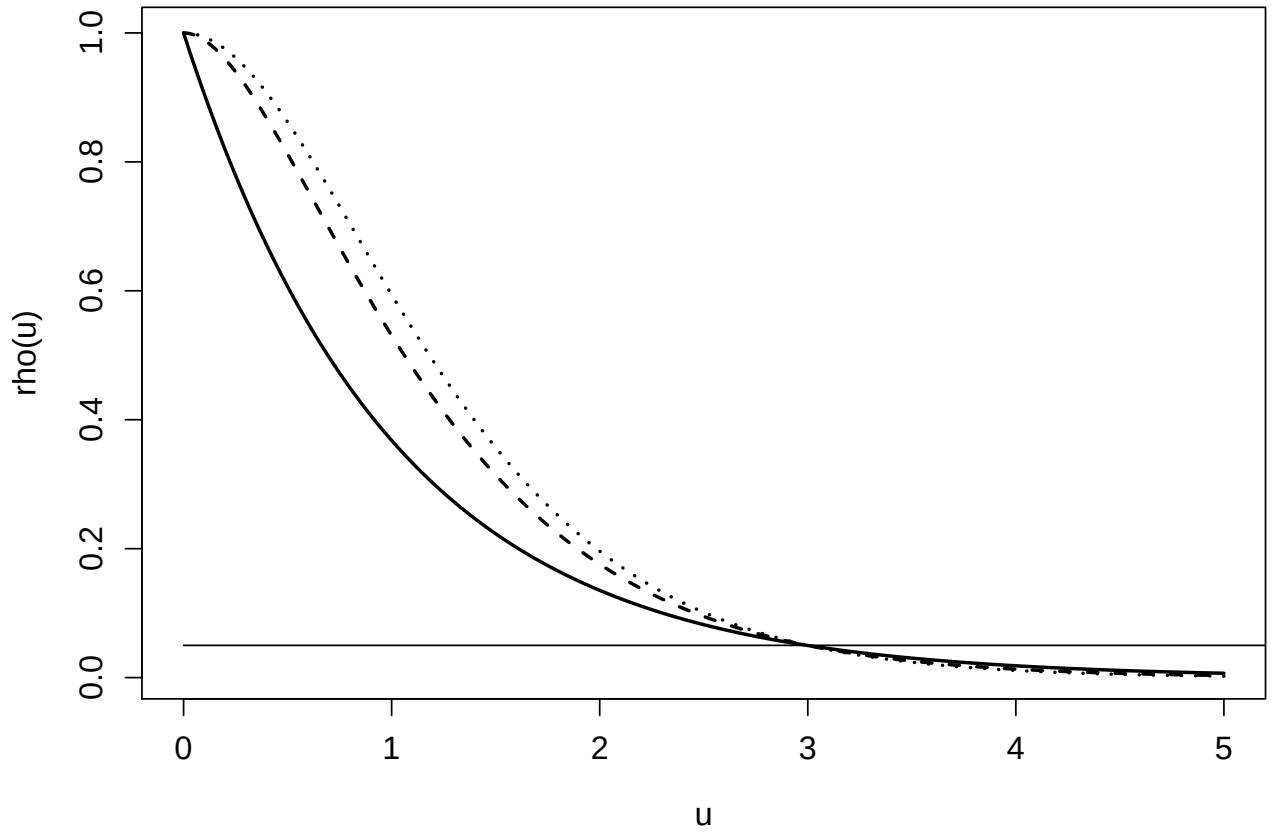
- ϕ determines scale of spatial correlation

Special cases

- $\kappa = 0.5$ gives $\rho(u) = \exp\{-u/\phi\}$
- $\kappa \to \infty$ gives $\rho(u) = \exp\{-(u/\phi)^2\}$
- $\kappa \to 0$ gives non-stationary limiting process

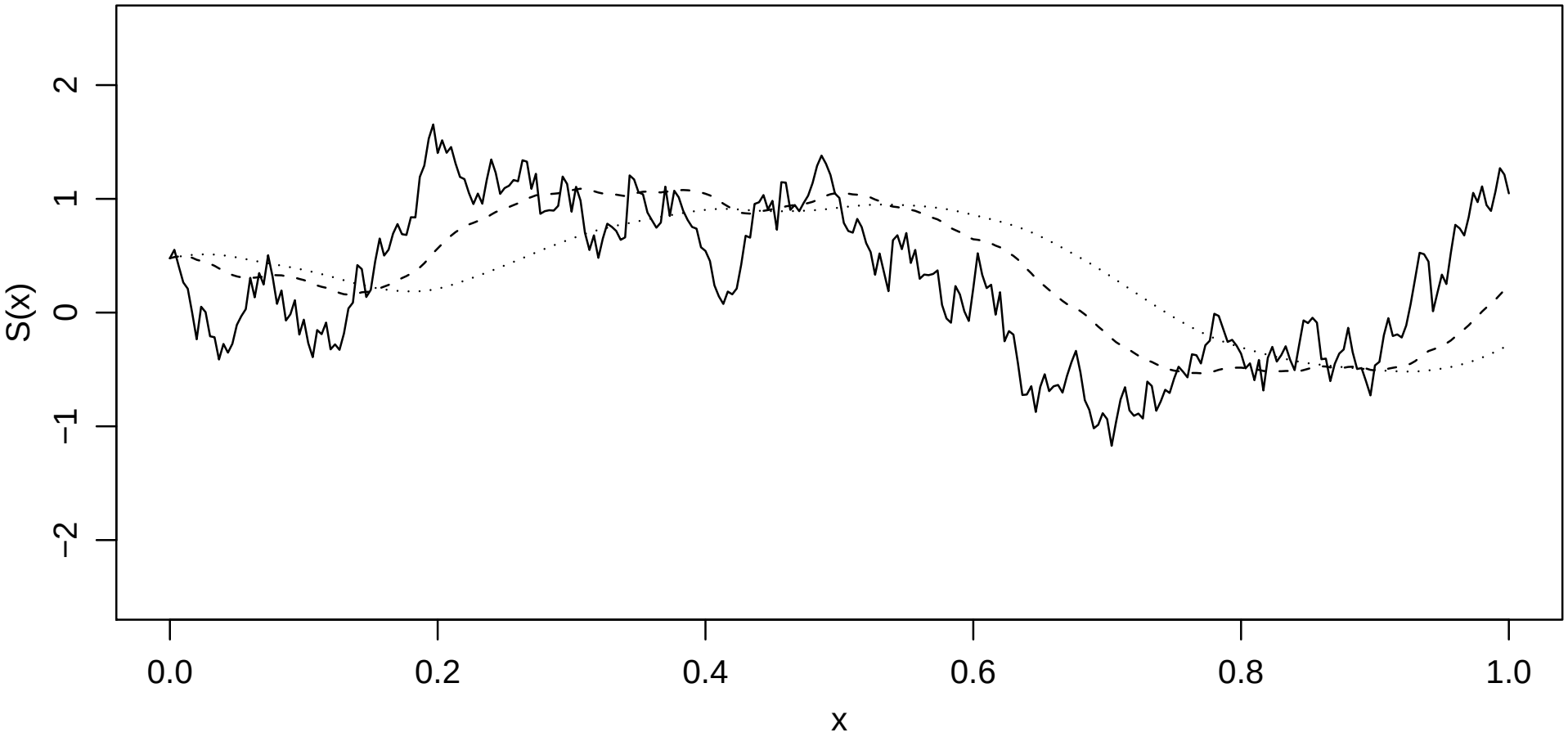
Often sufficient to choose amongst $\kappa = 0.5, 1.5, 2.5$

The Matérn correlation function



- $\kappa \leq 1 \Rightarrow S(x)$ **is continuous but non-differentiable**
- $\kappa > r \Rightarrow S(x)$ **is** r **times differentiable**

Matérn simulated realisations



Solid, dashed and dotted lines are for $\kappa = 0.5, 1.5, 2.5$, respectively

Initial parameter estimates: the variogram

Widely used, but not recommended except for initial analysis

- **weighted least squares criterion:**

$$W(\theta) = \sum_k n_k \{\hat{V}(u_k) - V(u_k;\theta)\}^2$$

  where θ denotes vector of covariance parameters

- **arbitrary upper limit for u_k**
- **standard errors not available**

Variogram estimation: lead pollution data

```
u<-10000*(1:10)
vario4<-variogram(lead2000,var.name=~loglead,
                  coords=~x+y,uvec=u)
fit<-eyefit(vario4)
fit
```

Parameter estimation: maximum likelihood

Linear Gaussian model: $\boldsymbol{Y} \sim \mathrm{MVN}(\boldsymbol{D}\beta, \sigma^2\boldsymbol{R} + \tau^2\boldsymbol{I})$

- $\boldsymbol{D}$ is $n \times p$ matrix, $(i,k)^{th}$ element $d_k(x_i)$
- $\boldsymbol{R}$ is $n \times n$ matrix, $(i,j)^{th}$ element $\rho(u_{ij})$
- $u_{ij} = ||x_i - x_j||$, Euclidean distance between x_i and x_j

Fitting process:

- initial estimates of regression parameters β by ordinary least squares
- initial estimates of covariance parameters θ from variogram of residuals
- numerical maximisation of

$$L(\beta, \theta) = \log f(y; \beta, \theta)$$

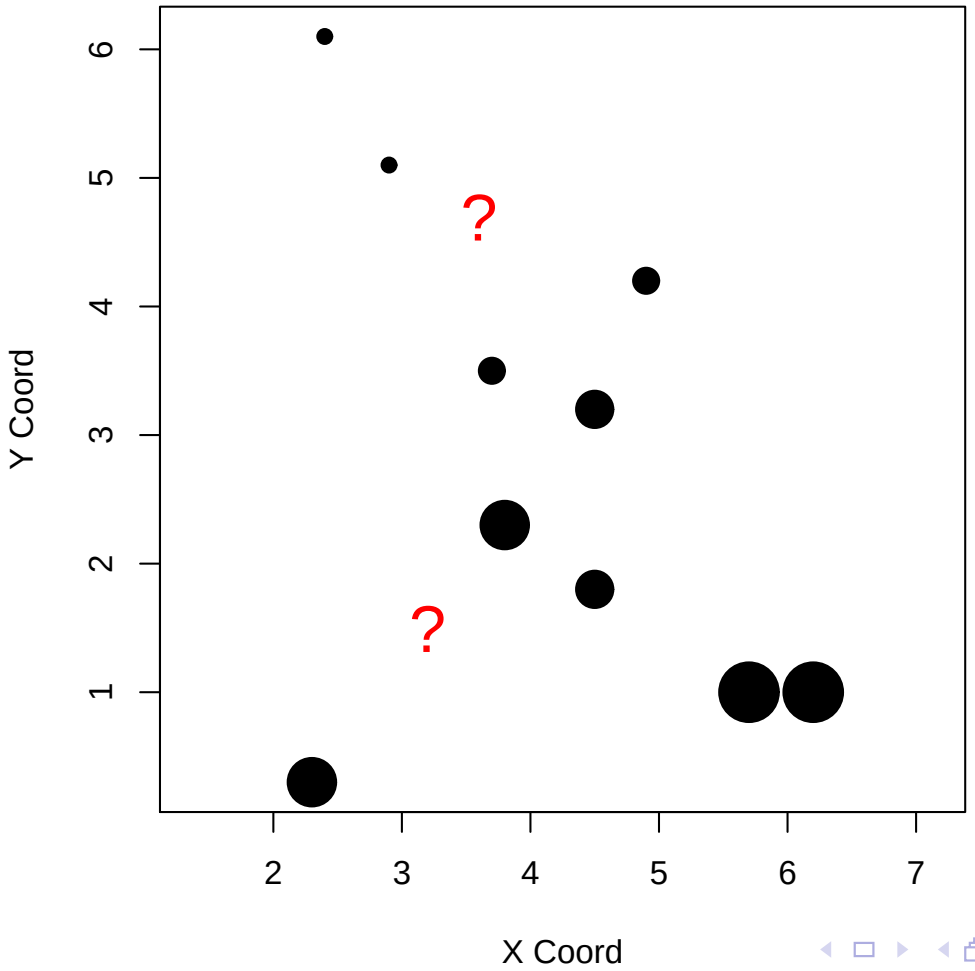
where $f(\cdot)$ denotes multivariate Normal density

Application: lead pollution data

```
fit.MLE <- linear.model.MLE(loglead~1,coords=~x+y,
                            start.cov.pars = c(30000,0.25),
                            data=lead2000,kappa=0.5)
summary(fit.MLE)
fit.MLE$estimate
theta<-exp(fit.MLE$estimate[2:4])
u<-1000*(0:100)
v<-theta[1]*(1-exp(-u/theta[2]))
lines(u,v,col="red")
diag.fit <- variog.diagnostic.lm(fit.MLE)
```

The canonical geostatistical problem

Given a set of measurements $Y_i : i = 1, \ldots, n$ at locations x_i in a spatial region A, presumed to be (noisy) measurements of a spatially continuous phenomenon $S(x_i)$, what can we say about the realisation of $S(x)$ throughout A?



Geostatistical prediction

- **Interpolation or smoothing?**
- **Point predictions or predictive distributions?**
- **Predicting linear or non-linear properties of a partially observed spatial surface.**
- **Bayesian inference: parameter estimation and spatial prediction as a single process.**
- **Parameter uncertainty and predictive uncertainty**

Prediction

The answer to any prediction problem is a probability distribution

Peter McCullagh

- T = any quantity of scientific interest
- Y = data that can tell us something about T.

The predictive distribution of T is the conditional probability distribution of T given Y

Geostatistical prediction: a general algorithm

$Y = \{Y_1, ..., Y_n\}$ at locations $x_1, ..., x_n$

$S^* = \{S(x_1^*), ..., S(x_M^*)\}$ for any set of locations $\{x_1^*, ..., x_M^*\}$

Algorithm

- model specifies $[Y, S^*] = [S^*][Y|S^*] \Rightarrow [S^*|Y]$ (Bayes'Theorem)
- simulate samples of S^* conditional on Y
- calculate corresponding $T^* = \mathcal{T}(S^*)$ from each sample of S^*
- resulting T^* are samples from predictive distribution of T

Minimum mean square error point prediction (kriging)

Model

- $[S^*]$ = distribution of latent signal process
- $[Y|S^*]$ = distribution of data conditional on latent process
- Bayes' theorem:

$$[S^*|Y] = [S^*][Y|S^*] / \int [S^*][Y|S^*] dS^*$$

Mean square error

- $\hat{T} = t(Y)$ is a point predictor
- $\mathrm{MSE}(\hat{T}) = \mathrm{E}[(\hat{T} - T)^2]$ is the mean square error

Theorem

1. $MSE(\hat{T})$ takes its minimum value when $\hat{T} = \mathrm{E}(T|Y)$.
2. $\mathrm{Var}(T|Y)$ estimates the achieved mean square error

Simple and ordinary kriging

$$Y \sim \mathrm{MVN}(\mu 1, \sigma^2 V) \quad V = R + (\tau^2/\sigma^2) \quad R_{ij} = \rho(\|x_i - x_j\|)$$

Target for prediction is $T = S(x)$

Write $r = (r_1, \ldots, r_n)$, where $r_i = \rho(\|x - x_i\|)$

T and Y jointly multivariate Normal $\Rightarrow [T|Y]$ univariate Normal:

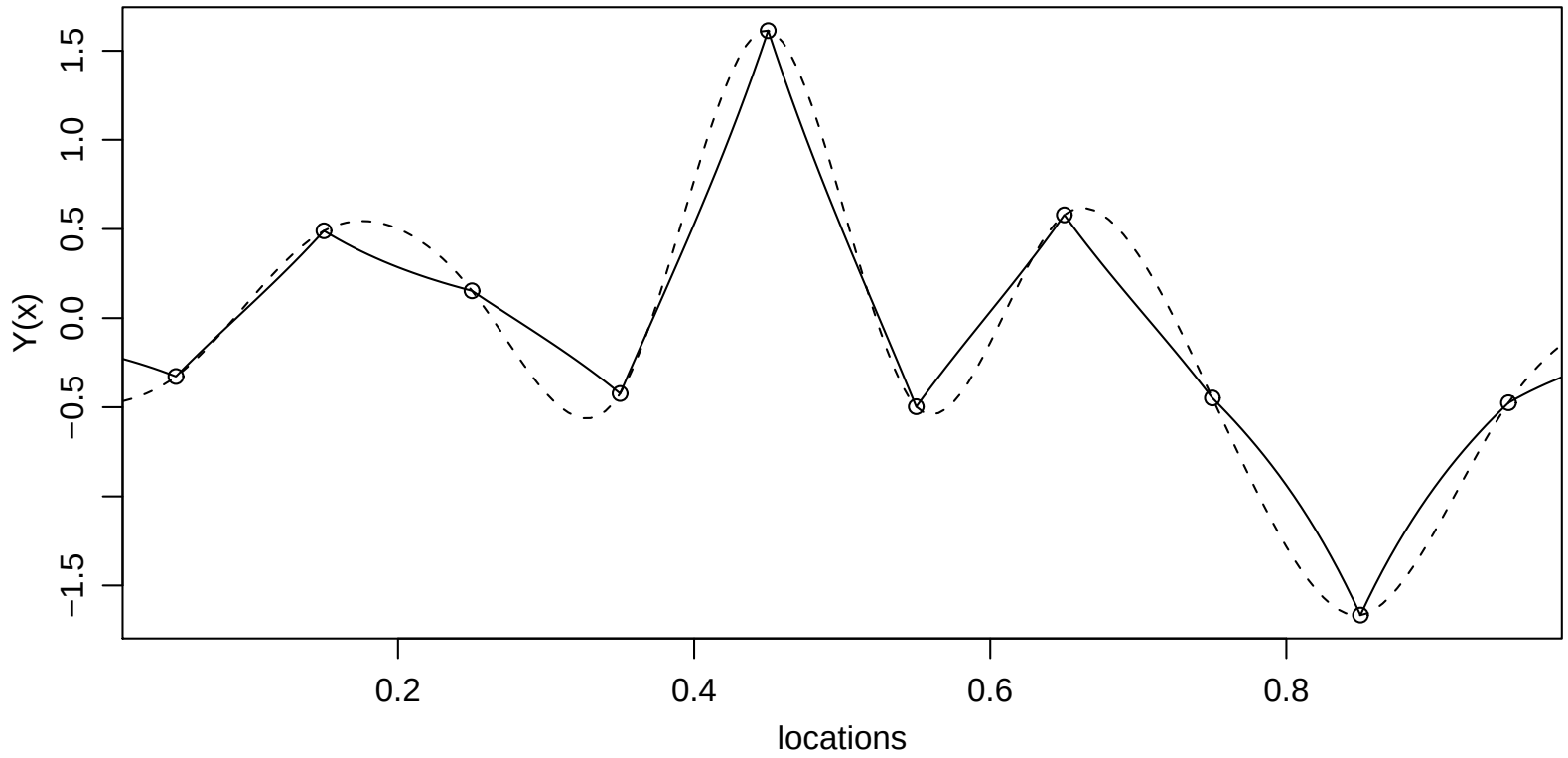
$$\hat{T} = \mathrm{E}[T|Y] = \mu + r' V^{-1}(Y - \mu 1)$$

$$\mathrm{Var}(T|Y) = \sigma^2(1 - r' V^{-1} r)$$

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

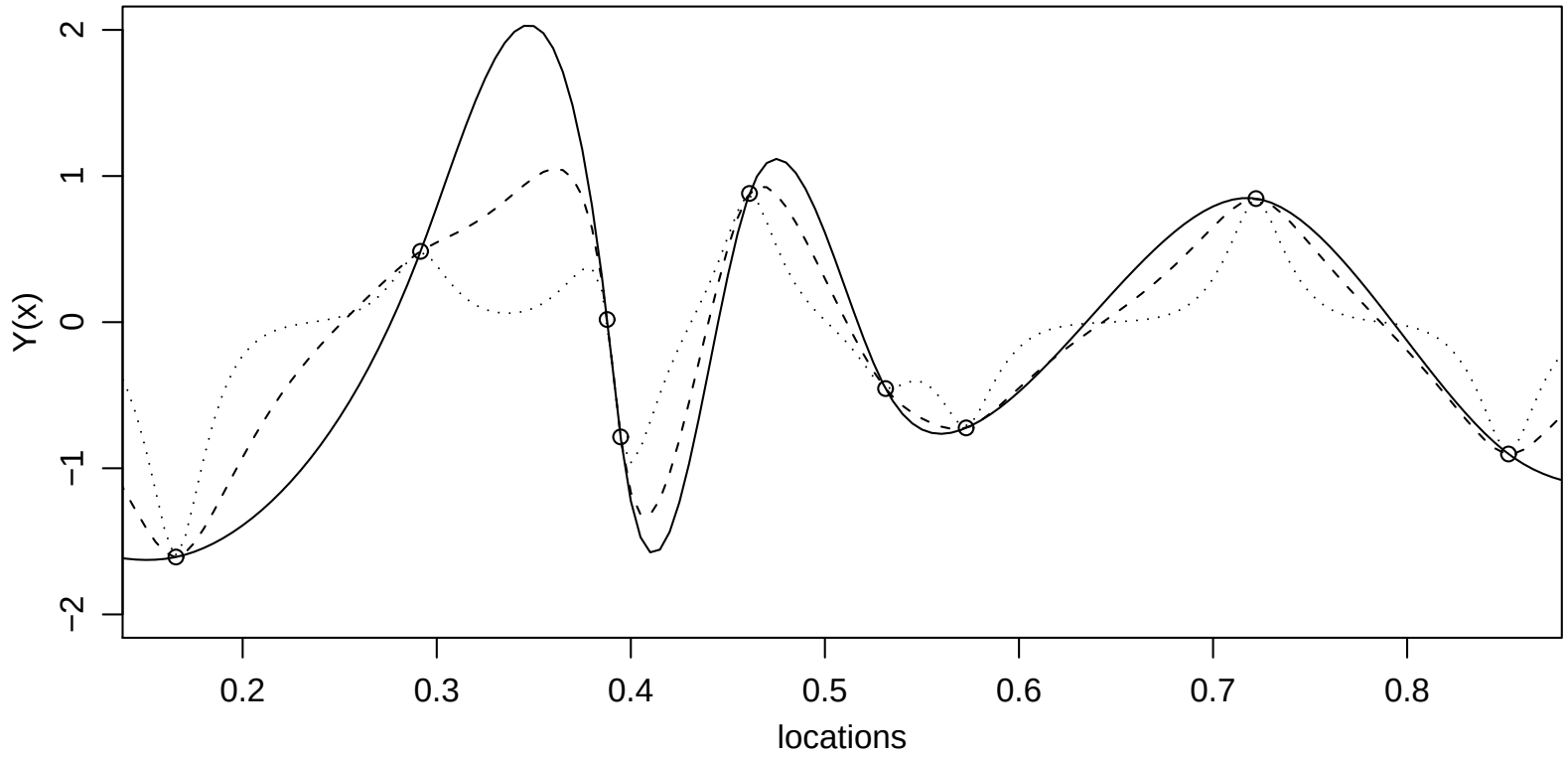
Simple kriging: three examples

1. Varying κ (smoothness of $S(x)$)



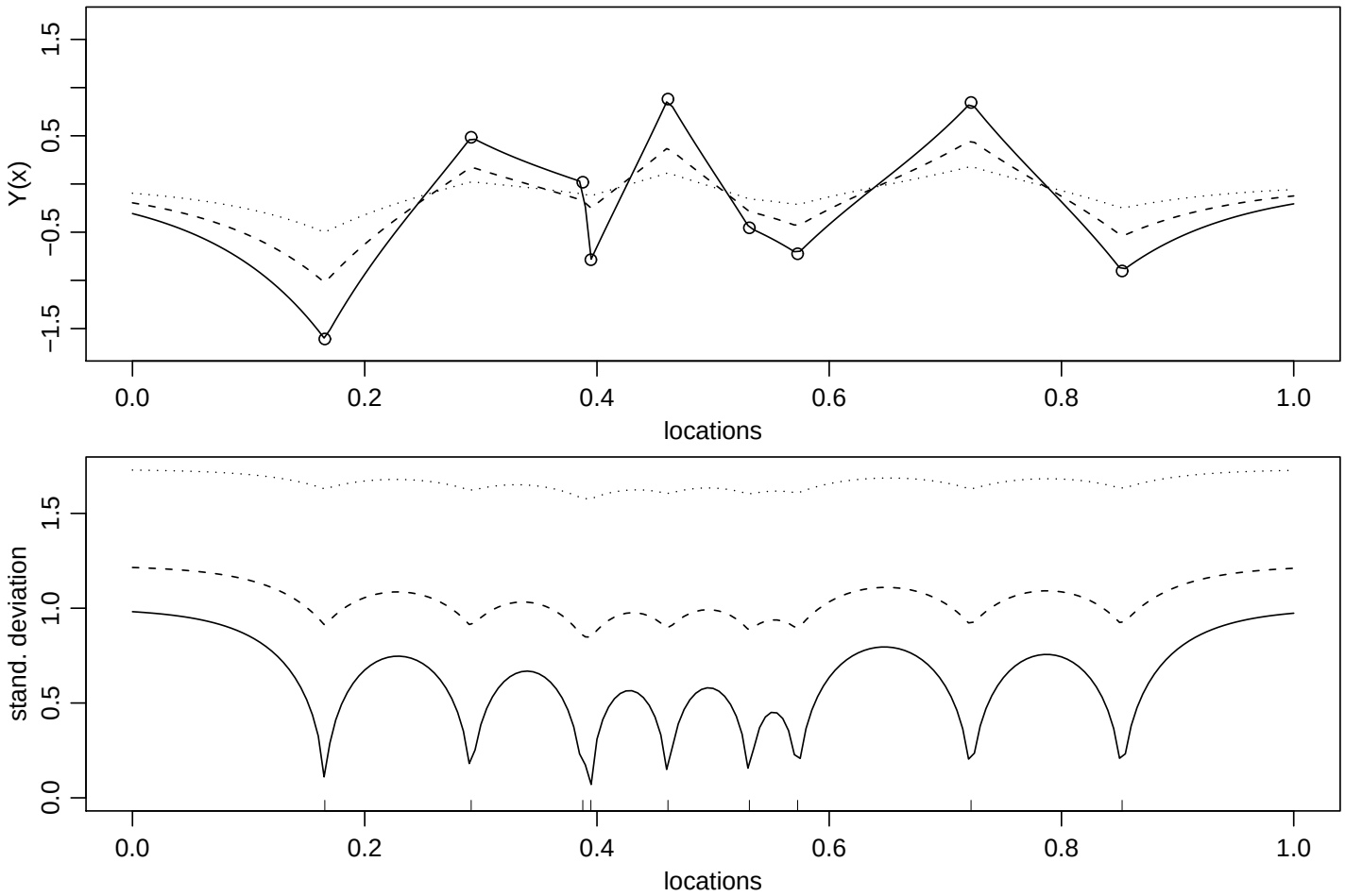
Solid, dashed and dotted lines are for $\kappa = 0.5, 1.5, 2.5$, respectively

2. Varying ϕ (range of spatial correlation



Solid, dashed and dotted lines are for large, medium and small ϕ, respectively

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



Interpolation when $\tau^2 = 0$ (solid line), smoothing when $\tau^2 > 0$ (dashed and dotted lines)

Trans-Gaussian models

Assume Gaussian model holds after point-wise transformation

Example: log-Gaussian kriging, $T(x) = \exp\{S(x)\}$

- **$S_1, ..., S_m$ are a sample from $[S|Y]$**
- **$T_i = \exp(S_i) \Rightarrow T_1, ..., T_m$ are a sample from $[T|Y]$**

Exercise: show that $\hat{T}(x) = \exp\{\hat{S}(x) + v(x)/2\}$, where $v(x) = \mathrm{Var}\{S(x)|Y\}$.

Reminder: 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 $[S^*|Y]$**
- **calculate corresponding $T^* = \mathcal{T}(S^*)$**

Application to lead pollution data

```
library(splancs)
galicia.grid <- gridpts(as.matrix(bound),xs=5000,ys=5000)
pred.MLE <- spatial.pred.linear.MLE(fit.MLE,
                                    grid.pred = galicia.grid,
                                    scale.predictions="logit",
                                    standard.errors = TRUE)
plot(pred.MLE,type="logit")
plot(pred.MLE,type="logit",summary="standard.errors")

pred.MLE <- spatial.pred.linear.MLE(fit.MLE,
                                    grid.pred = galicia.grid,
                                    scale.predictions="odds",
                                    standard.errors = TRUE)
plot(pred.MLE,type="odds")
```

Application to lead pollution data

```
pred.MLE <- spatial.pred.linear.MLE(fit.MLE,
                                    grid.pred = galicia.grid,
                                    scale.predictions="odds",
                                    n.sim.prev=1000)
plot(pred.MLE,type="odds")
range(pred.MLE$samples)

predict.max<-NULL
for (sim in 1:100) {
    predict.max<-c(predict.max,max(exp(pred.MLE$samples[,sim])))
    }
hist(predict.max,main="predicted maximum")
```

Generalised linear models

Data

Measurements Y_i are made at locations x_i with associated explanatory variables $d(x_i)$

Model

- **Independence.** Y_i are mutually independent random variables.
- **Link function.** $h(\mu_i) = d(x_i)^\top \beta$
- **Variance function.** $\mathrm{Var}[Y_i] = v(\mu_i)$
- **Error-free explanatory variables.**
- **Distribution.** $Y_i \sim f(y; \mu_i, \phi)$

Disease mapping 1

Single prevalence survey

Sample n individuals, observe Y positives

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

Multiple prevalence surveys

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

$$Y_i \sim \mathrm{Bin}(n_i, p_i) \quad \log\{p_i/(1-p_i)\} = d(x_i)^\top \beta$$

Disease mapping 2

Extra-binomial variation

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

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

This lecture

What to do if the U_i are spatially structured

Binomial logistic geostatistical model

- **Latent spatial process**

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

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

- **Linear predictor**

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

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

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

$$P(x) = \exp\{\eta(x)\}/[1 + \exp\{\eta(x)\}]$$

- **Error distribution**

$$Y_i|S(\cdot) \sim \mathrm{Bin}\{m_i, p(P_i)\}$$

Parameter estimation

Linear predictor: $LP = (LP_1, ..., LP_n)$, where

$$LP_i = d(x_i)'\beta + S(x_i)$$

Data: $(m_i, y_i) : i = 1, ..., n$

Model:

- $LP \sim \mathrm{MVN}(\theta)$
- $Y_i | LP_i \sim$ independent, binomially distributed

Likelihood:

$$L(\theta) = \int_T [T; \theta] \times [Y|T] dT$$

Evaluation needs Monte Carlo integration...handle with care!

Prediction in the binomial model

S = signal at data-locations
S^* = signal at all prediction locations of interest
(typically a finely spaced grid over the region of interest, A)

Plug-in prediction

- 1. Estimate $\hat{\theta}$ by Monte Carlo maximum likelihood
- 2. Sample from $[S|Y;\hat{\theta}]$ by MCMC
- 3. Sample from multivariate Normal distribution $[S^*|S]$

Allowing for parameter uncertainty

Use general result: $\hat{\theta} \sim \mathrm{MVN}(\theta, V)$

Replace step 2 above by:

- 2a. Sample $\theta^* \sim \mathrm{MVN}(\hat{\theta}, \hat{V})$
- 2b. Sample from $[S|Y;\theta^*]$ by MCMC

Extra-binomial variation: person or place?

Extend logistic geostatistical model to

$$\log\{P_i/(1-P_i)\} = \alpha + \{d(x_i)'\beta + S(x_i)\} + \{e_i'\gamma + U_i\}$$

- $d(x)$: measured properties of location x
- $S(x)$: stochastic process, proxy for unmeasured properties of x
- e_i : measured properties of people at location x_i
- U_i : independent random variables, proxy for unmeasured properties of community at x_i

Further extension to include individual-level explanatory variables

$$\log\{P_{ij}/(1-P_{ij})\} = \alpha + \{d(x_i)'\beta + S(x_i)\} + \{e_{ij}'\gamma + U_i\}$$

Exploratory analysis: empirical logits

Fitting the binomial logistic model is computationally demanding and requires judgement:

- **convergence of iterative algorithms?**
- **accuracy of approximations?**

Empirical logit transform

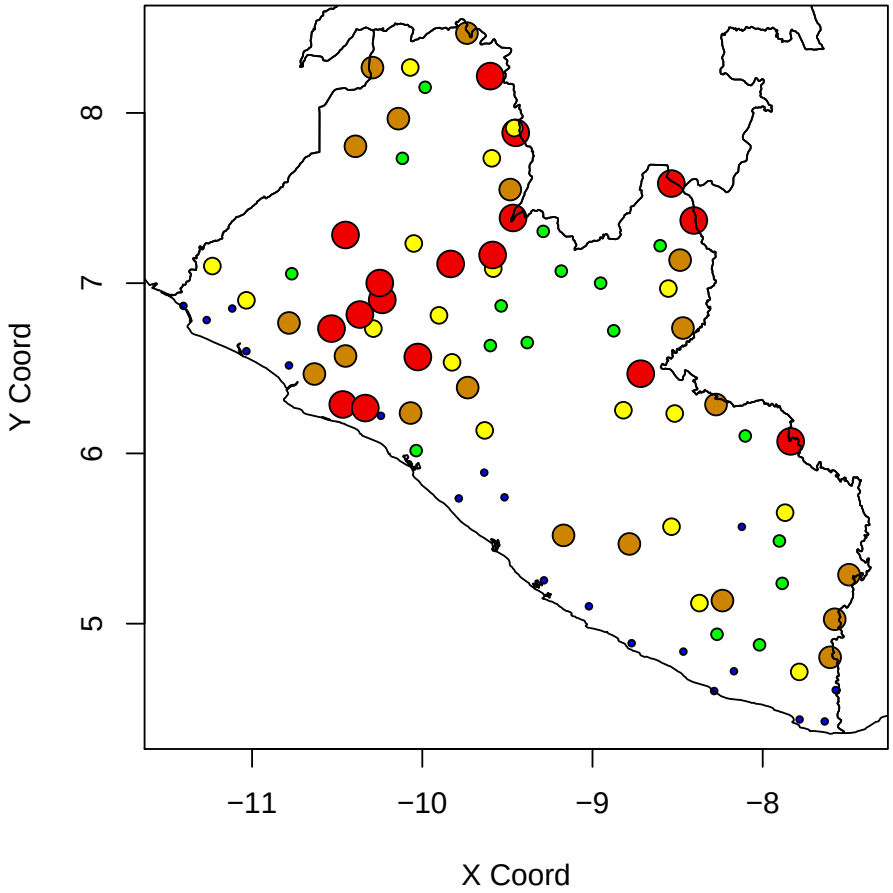
$$Z_i = \log\{(Y_i + 0.5)/(n_i - Y_i + 0.5)\}$$

Fitting a linear model with Z_i as response is comparatively straightforward

Case-study: Onchocerciasis (river blindness) in Liberia

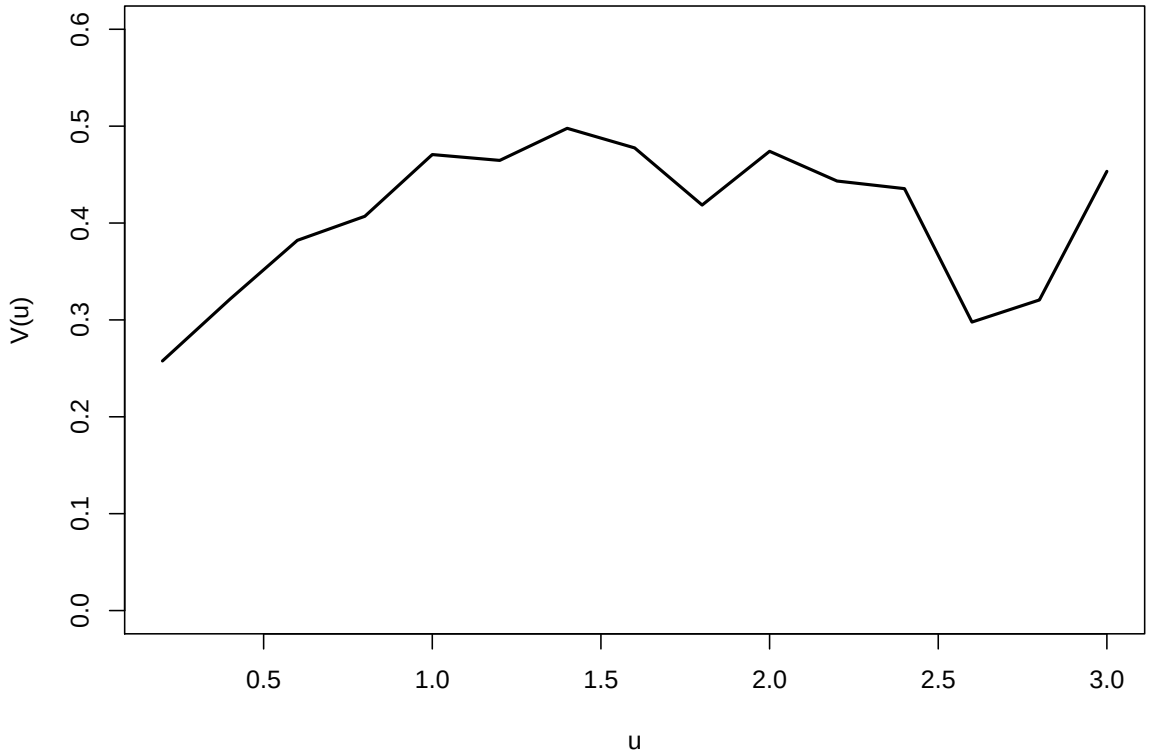
- **prevalence data from 90 villages in Liberia**
- **sample sizes 40 to 50**
- **empirical prevalences 0% to 35%**
- **use empirical logit transformation for exploratory analysis**

Exploratory analysis of onchocerciasis data



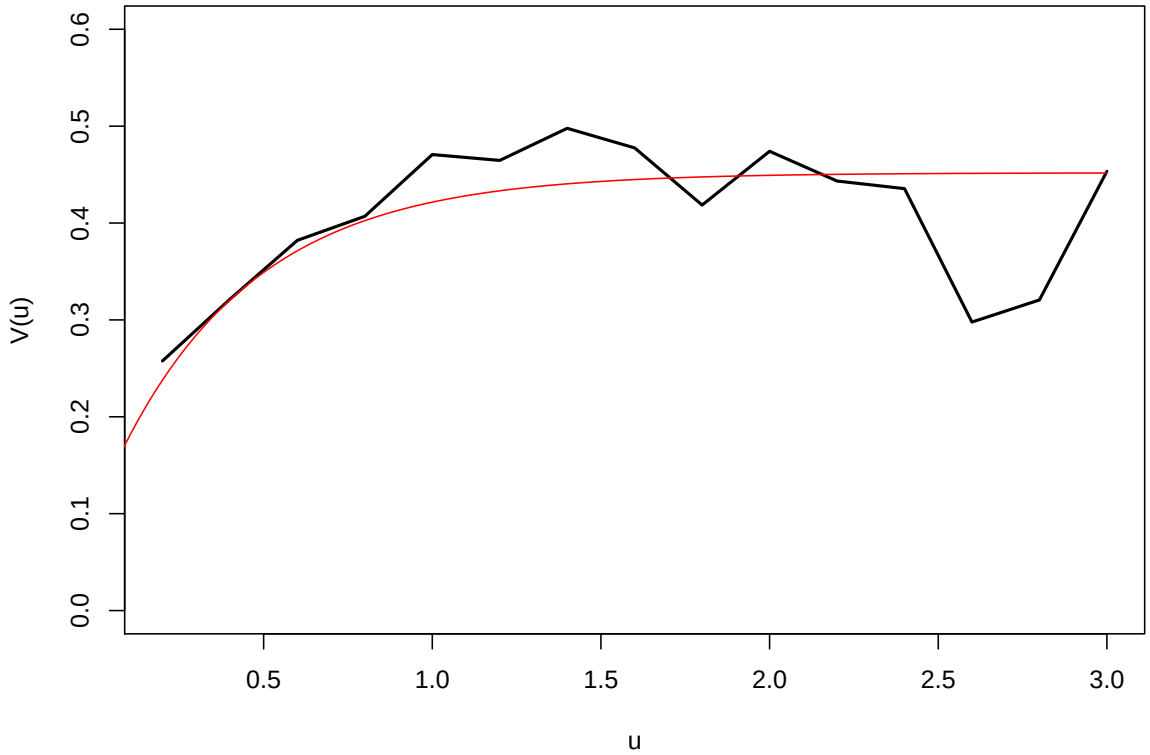
- **patches of high and low prevalence**
- **increasing trend away from coast?**

Exploratory analysis of onchocerciasis data (2)



- **residual variogram after fitting linear trend surface to empirical logits by ordinary least squares**

Exploratory analysis of onchocerciasis data (3)



- **fitted variogram from likelihood-based analysis of geostatistical model with Matérn correlation, $\kappa = 0.5$**

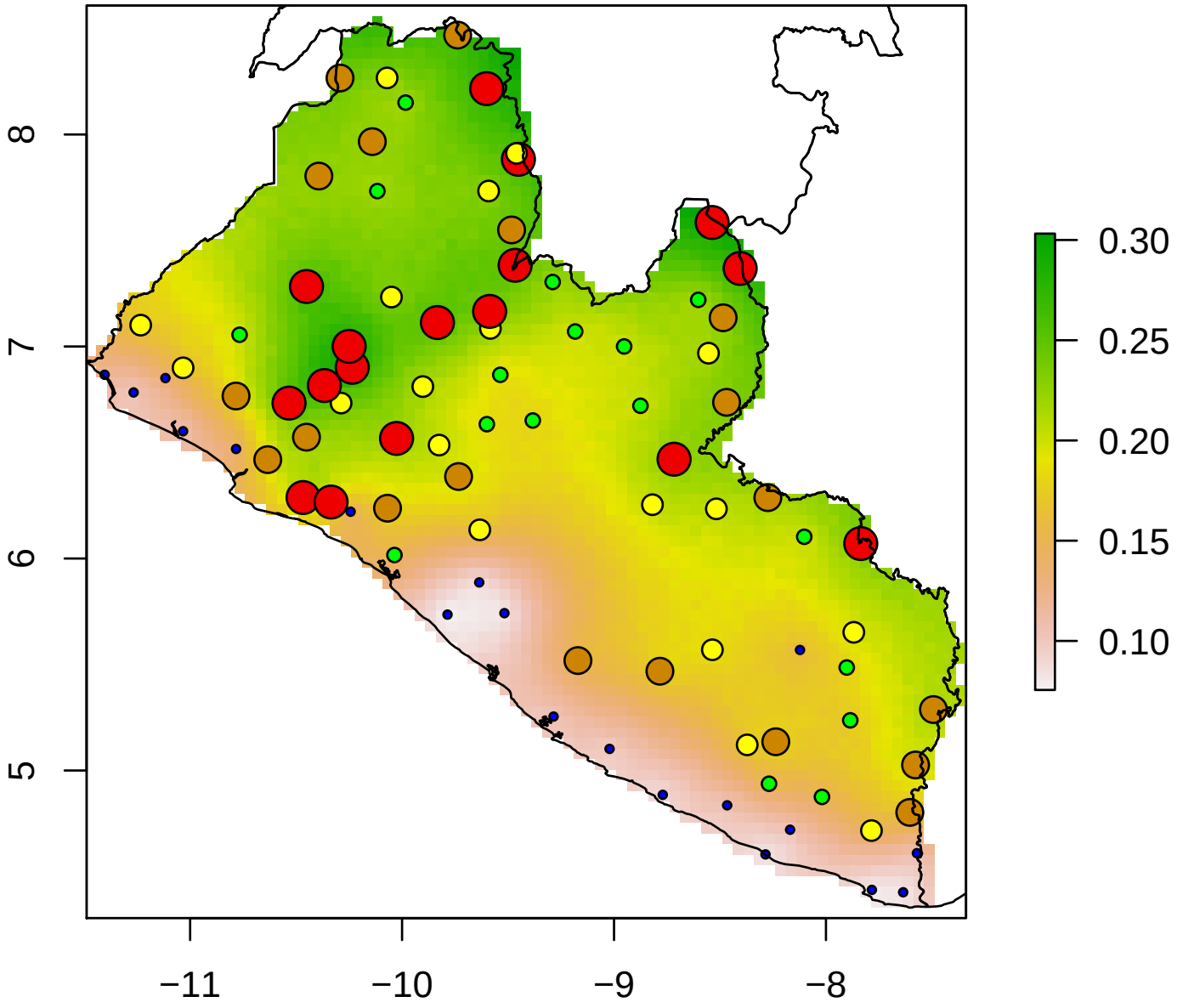
Liberia onchocerciasis: fitting the binomial logistic model

```
library(PrevMap)
response<-cbind(npositive,ntested-npositive)
fit.glm<-glm(response~longitude+latitude,
       data=data,family=binomial)
beta<-fit.glm$coef; par0<-c(beta,fit$cov.pars,fit$nugget)
mcmc<-control.mcmc.MCML(n.sim=10000,burnin=2000,thin=8,
       h=1.65/(nrow(data)**(1/6)))
fit.bl<-binomial.logistic.MCML(npositive~longitude+latitude,
       units.m= ntested,coords=~longitude+latitude,
       data=data,par0=par0,control.mcmc=mcmc,kappa=0.5,
       start.cov.pars=c(par0[5],par0[6]/par0[4]))
```

Liberia onchocerciasis: mapping the results

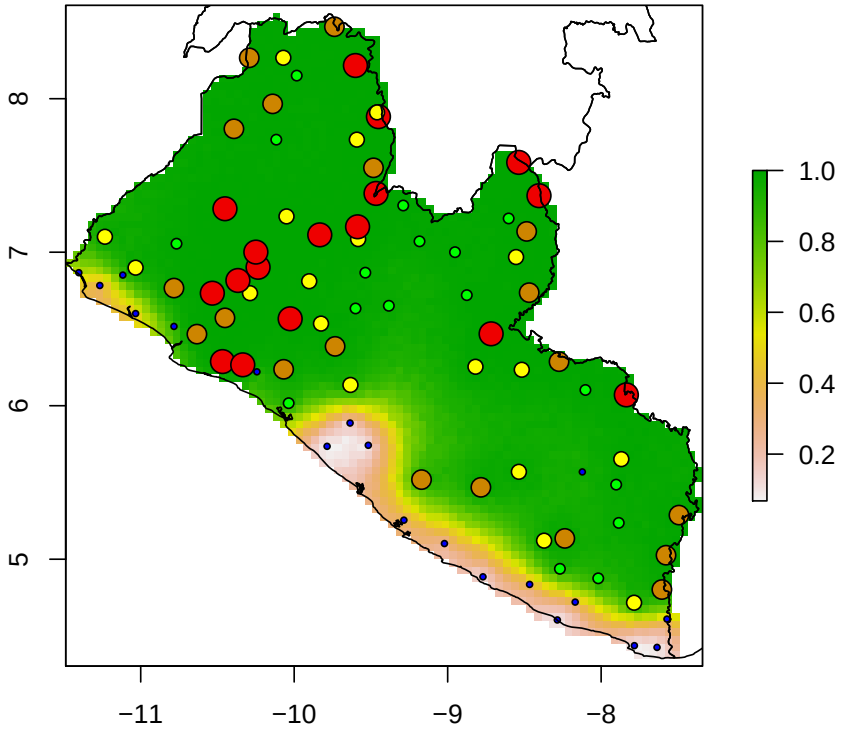
```
library(splancs)
xy<-data.frame(longitude=data$longitude,latitude=data$latitude)
par(pty="s"); pointmap(xy)
poly<-getpoly()
grid.predict<-as.data.frame(gridpts(poly,xs=0.1,ys=0.1))
names(grid.predict)<-c("longitude","latitude")
predict.MCML<spatial.pred.binomial.MCML(fit.bl,
       grid.pred=grid.predict,predictors=grid.predict,
       control.mcmc=mcmc,scale.predictions="prevalence",
       standard.errors=TRUE,thresholds=0.2,
       scale.thresholds="prevalence")
plot(predict.MCML,type="prevalence")
points(gd,add=TRUE)
polymap(poly,add=TRUE)
```

Liberia: onchocerciasis prevalence map

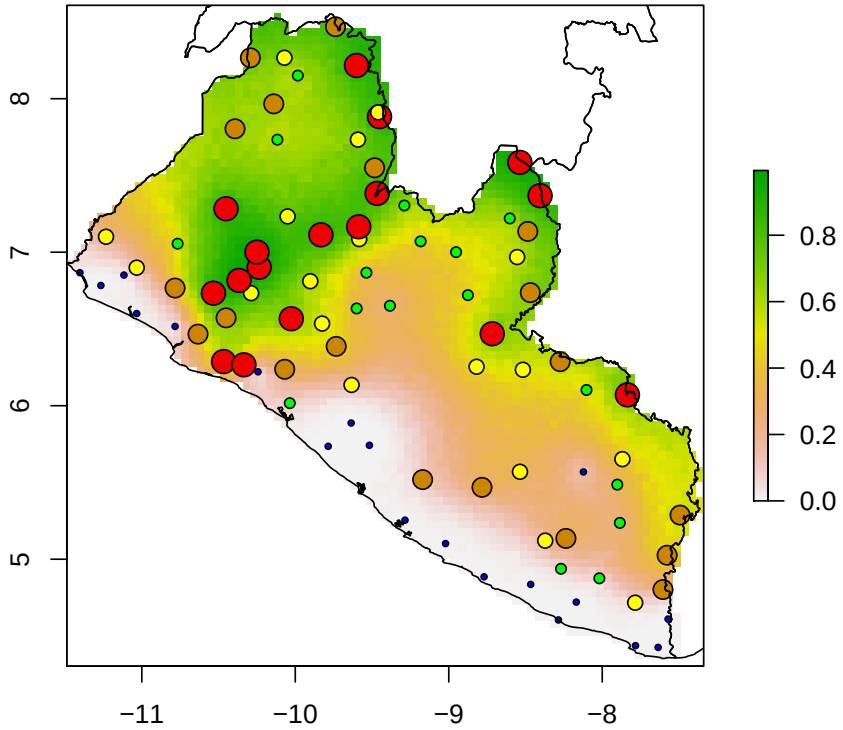


Liberia: onchocerciasis predictive probability maps

P(prevalence>10%)



P(prevalence>20%)

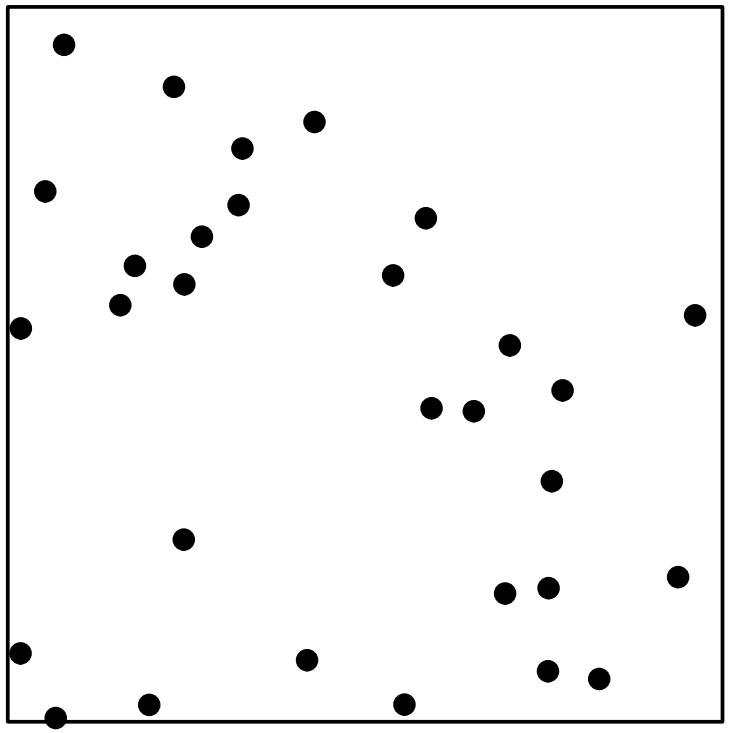


Additional topics

- **Geostatistical design**
- **Bayesian inference**
- **Combining data from:**
  - **randomised and non-randomised surveys**
  - **multiple diagnostics**
  - **multiple sampling occasions**
- **Preferential sampling**
- **Hybrid geospatial/mechanistic models**
- **Zero-inflation**
- **spatio-temporal mapping**

Lattice-free spatially regular sampling designs

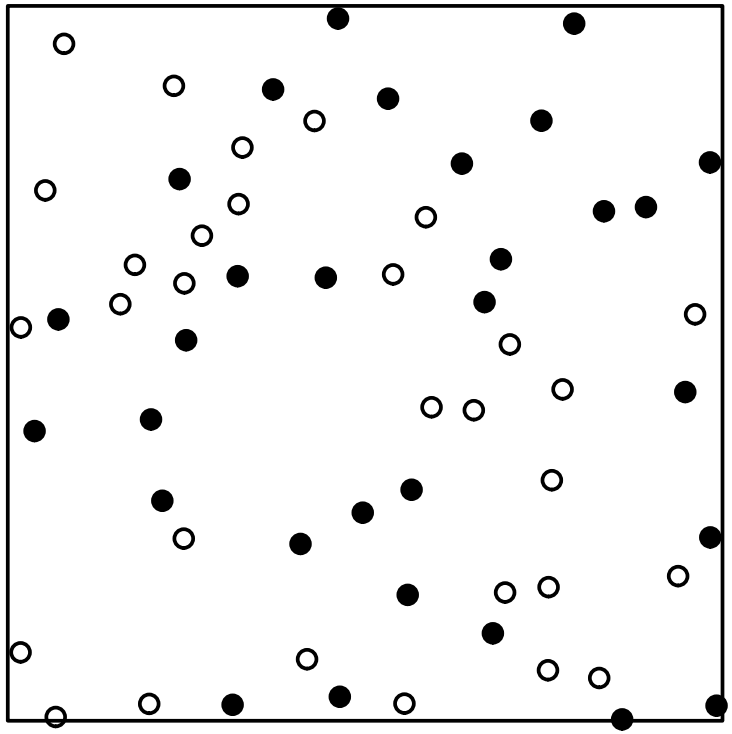
Sample at random subject to a minimum distance constraint



Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2017). Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, **28, DOI: 10.1002/env.2425.**

Lattice-free spatially regular sampling designs

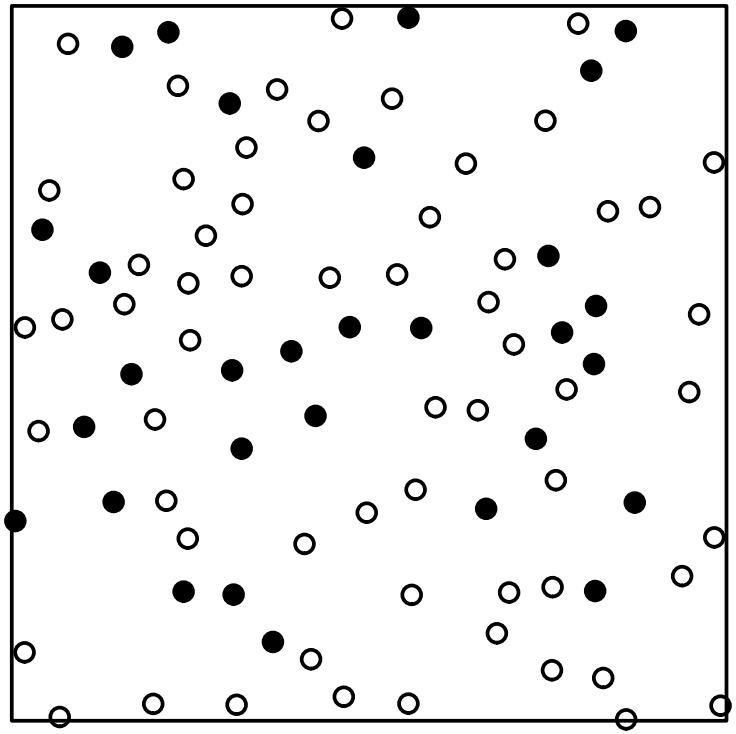
Sample at random subject to a minimum distance constraint



Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2017). Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, **28, DOI: 10.1002/env.2425.**

Lattice-free spatially regular sampling designs

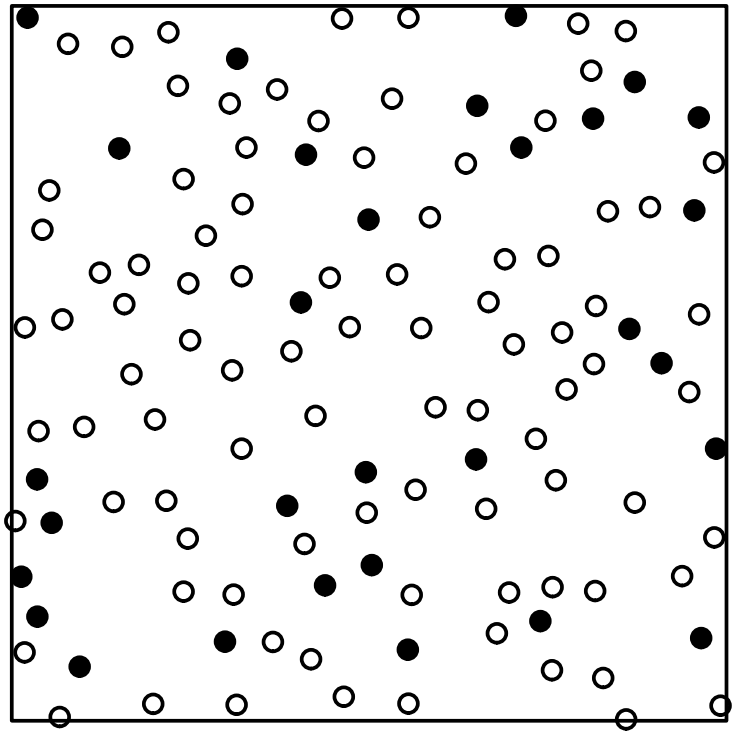
Sample at random subject to a minimum distance constraint



Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2017). Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, **28, DOI: 10.1002/env.2425.**

Lattice-free spatially regular sampling designs

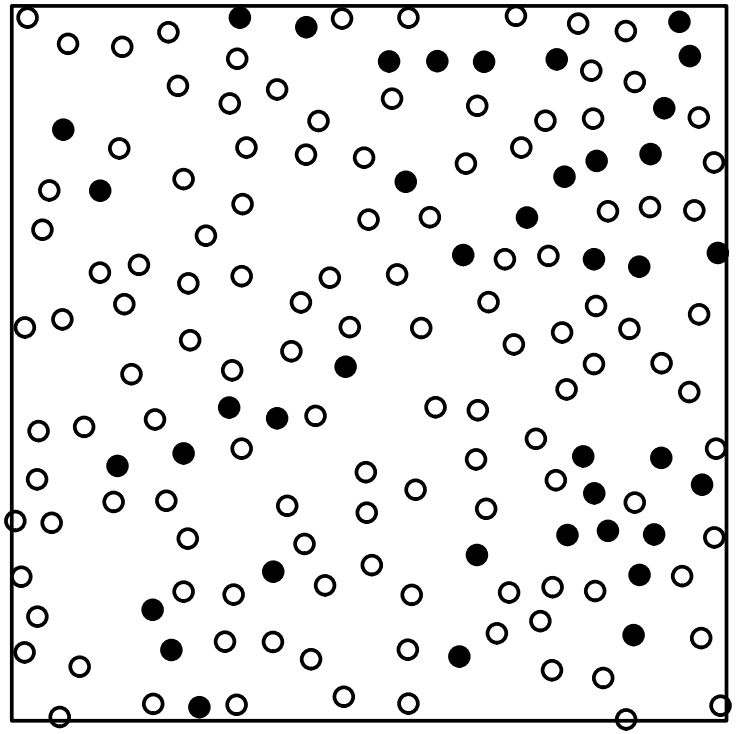
Sample at random subject to a minimum distance constraint



Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2017). Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, **28, DOI: 10.1002/env.2425.**

Lattice-free spatially regular sampling designs

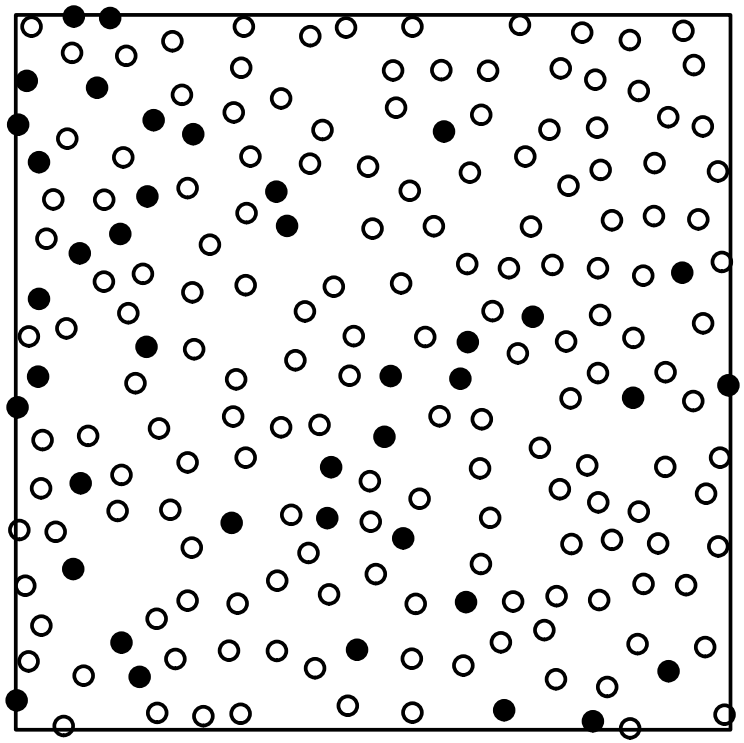
Sample at random subject to a minimum distance constraint



Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2017). Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, **28, DOI: 10.1002/env.2425.**

Lattice-free spatially regular sampling designs

Sample at random subject to a minimum distance constraint



Chipeta, M.G., Terlouw, D.J., Phiri, K.S. and Diggle, P.J. (2017). Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, **28, DOI: 10.1002/env.2425.**

Bayesian inference

Classical model specification

$$[S;\theta][Y|S;\theta]$$

Bayesian model specification replaces **unknown constant** θ by **unobserved random variable** θ to give

$$[\theta][S|\theta][Y|S,\theta]$$

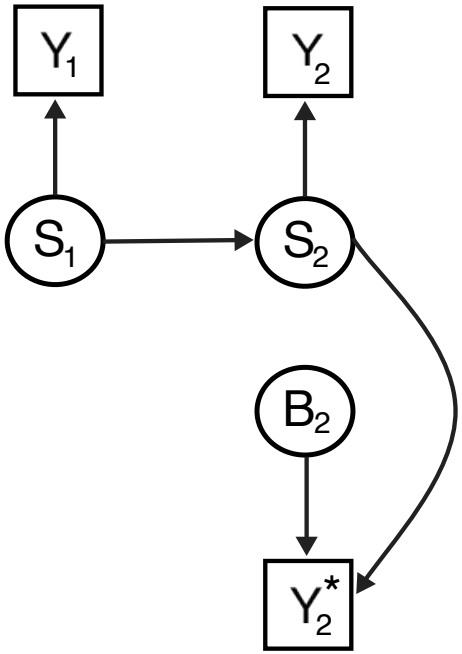
where $[\theta]$ is the **prior distribution** for θ

Bayes' Theorem

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

where $[\theta|Y]$ is the **posterior distribution** for θ

Combining randomised and convenience samples



- **Randomised: Rolling malaria indicator surveys**
- **Convenience: Presentations at general health clinics**

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

Preferential sampling

locations X **signal S** **measurements Y**

- **Conventional model:**

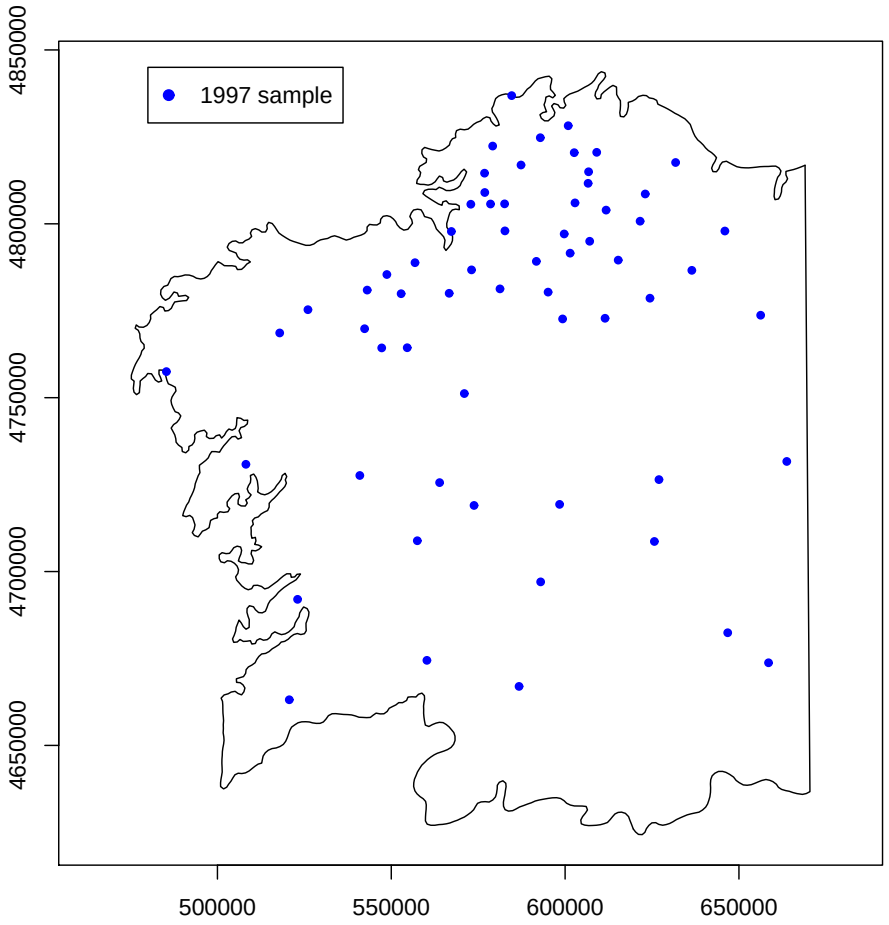
$$[X, S, Y] = [S][X][Y|S] \quad (1)$$

- **Preferential sampling model:**

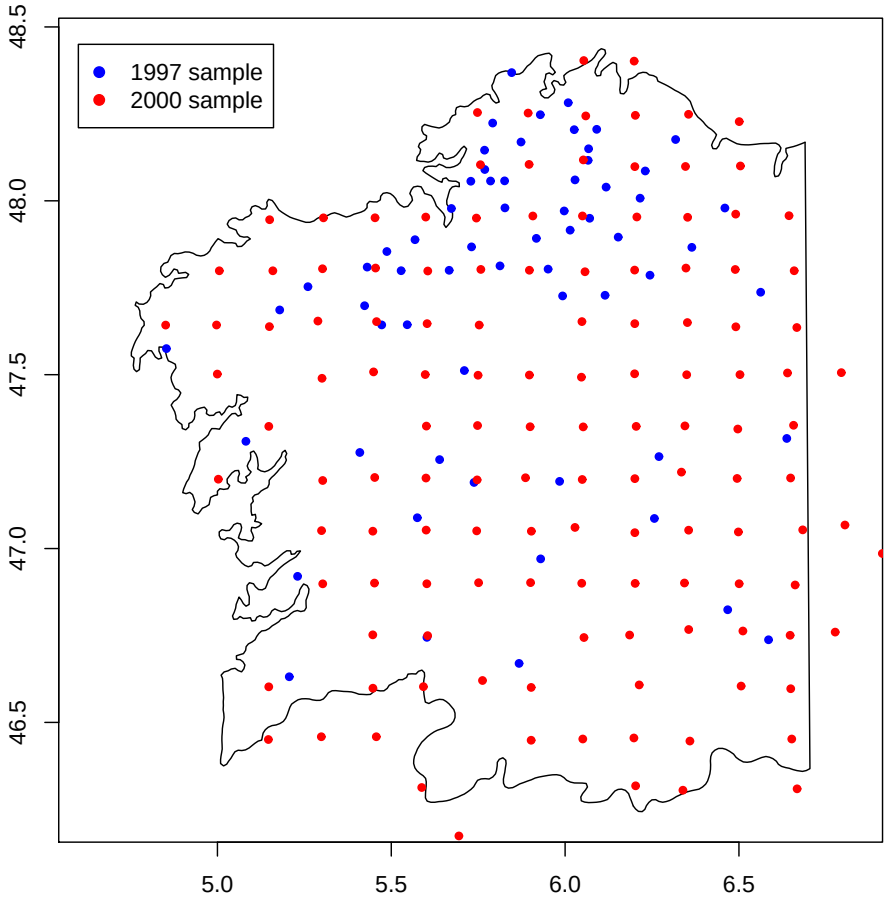
$$[X, S, Y] = [S][X|S][Y|S, X] \quad (2)$$

If X and S are stochastically dependent, the term $[X|S]$ in (1) needs to be included for correct inference.

Heavy metal monitoring in Galicia



Heavy metal monitoring in Galicia



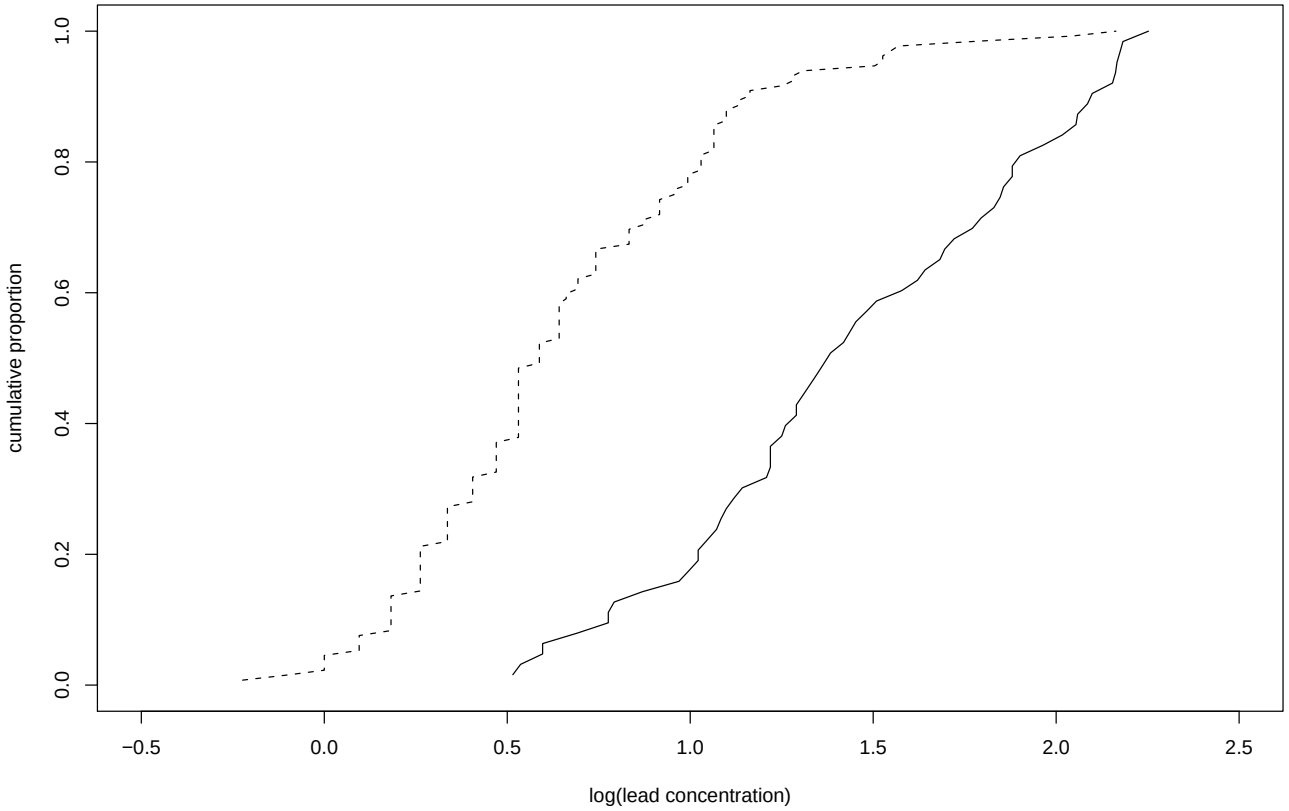
Heavy metal monitoring in Galicia

- **1997 sampling design highly non-uniform**
- **may lead to biased estimates of spatial variation**
- **2000 sampling design spatially uniform, so assume unbiased**
- **possible modelling framework is:**
  - **2000 sampling is non-preferential**
  - **1997 sampling may be preferential**
  - **some parameters in common between 1997 and 2000?**

Summary statistics

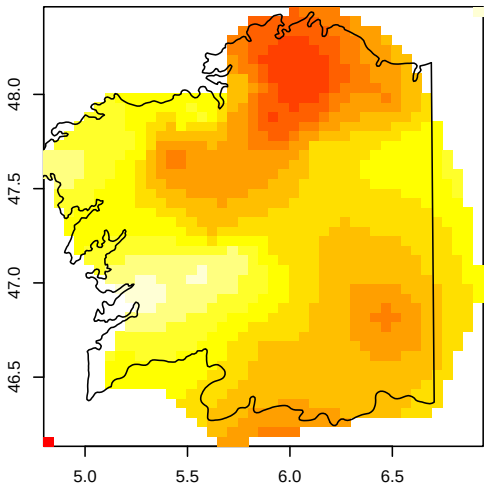
	untransformed		log-transformed	
	1997	2000	1997	2000
Number of locations	63	132	63	132
Mean	4.72	2.15	1.44	0.66
Standard deviation	2.21	1.18	0.48	0.43
Minimum	1.67	0.80	0.52	-0.22
Maximum	9.51	8.70	2.25	2.16

Marginal distributions of log-transformed lead concentrations

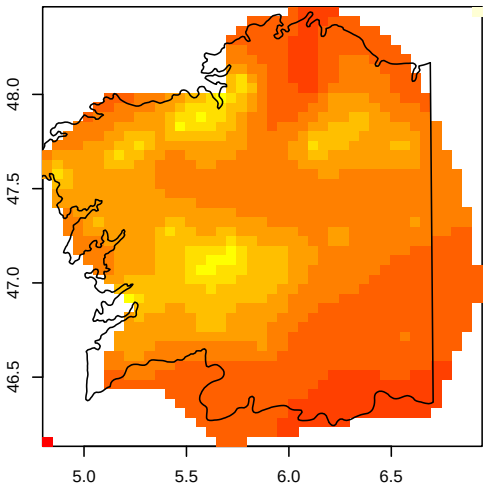


Spatial prediction of lead concentrations

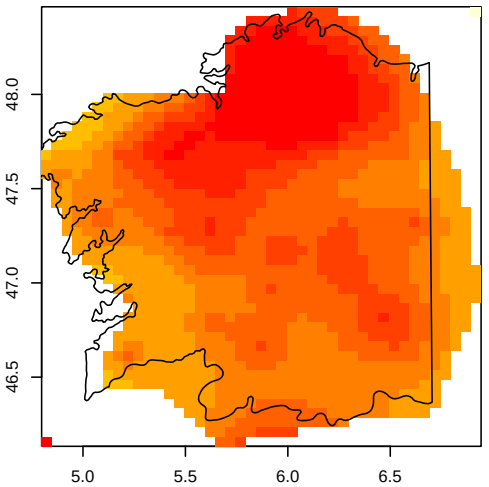
Common scale runs from −0.756 (red) to 8.358 (white)



preferential



non-preferential

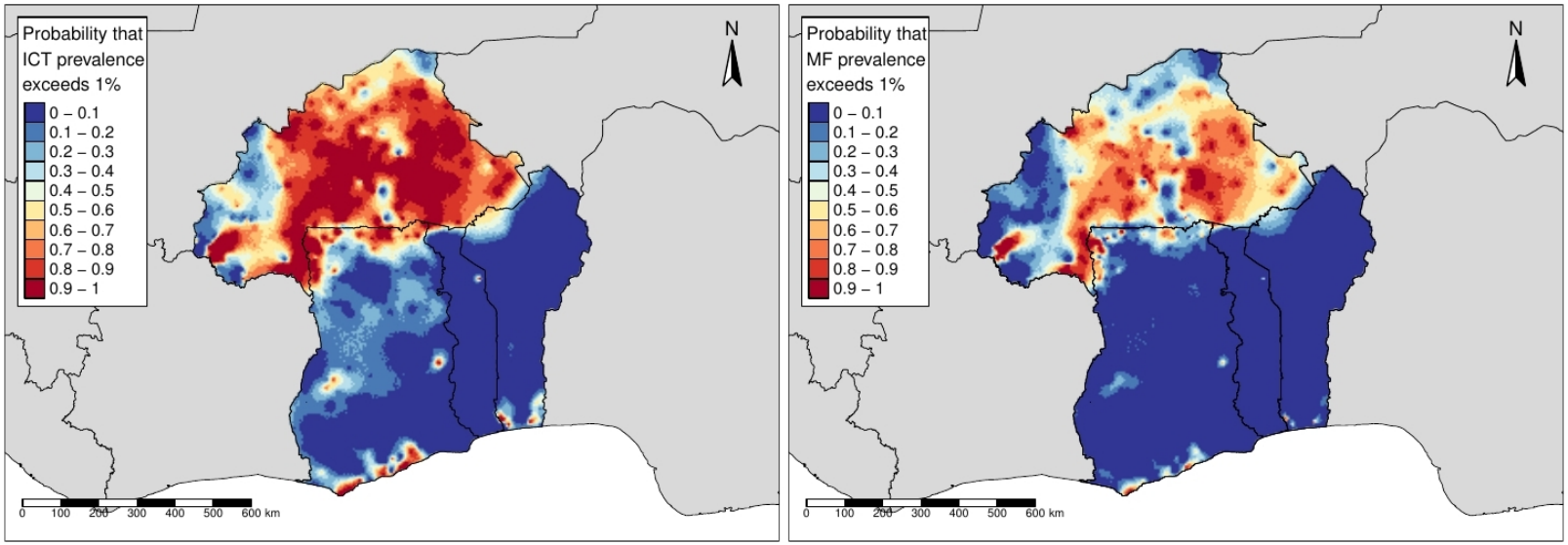


difference

Hybrid geospatial/mechanistic models

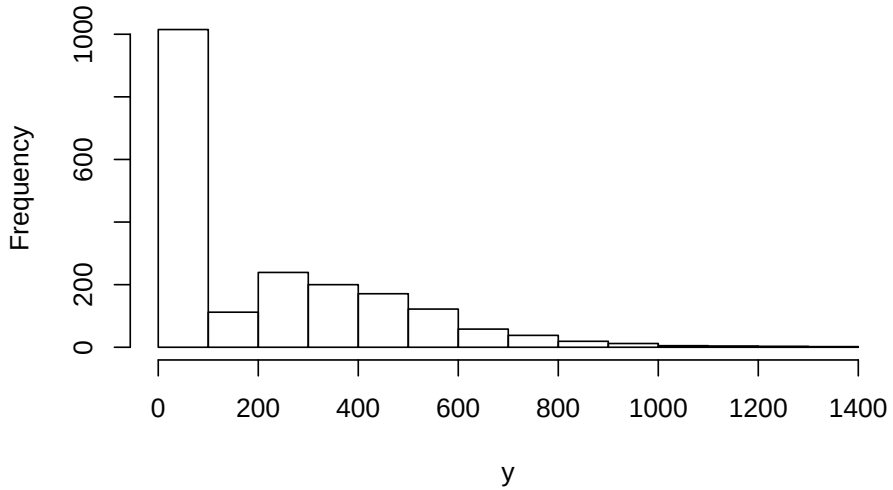
- $\lambda(x)$ = **Worm density at location** x
- α **reproductive rate**
- γ **sensitivity of the ICT**

$$\begin{cases} p_{MF}(x) = 1 - \exp\{\lambda(x)[1 - \exp\{-\alpha\}]\} \\ p_{ICT}(x) = \gamma(1 - \exp\{\lambda(x)\}) \end{cases}$$



Spatially structured zero-inflation

Entomological data often look something like this:



Non-spatial model

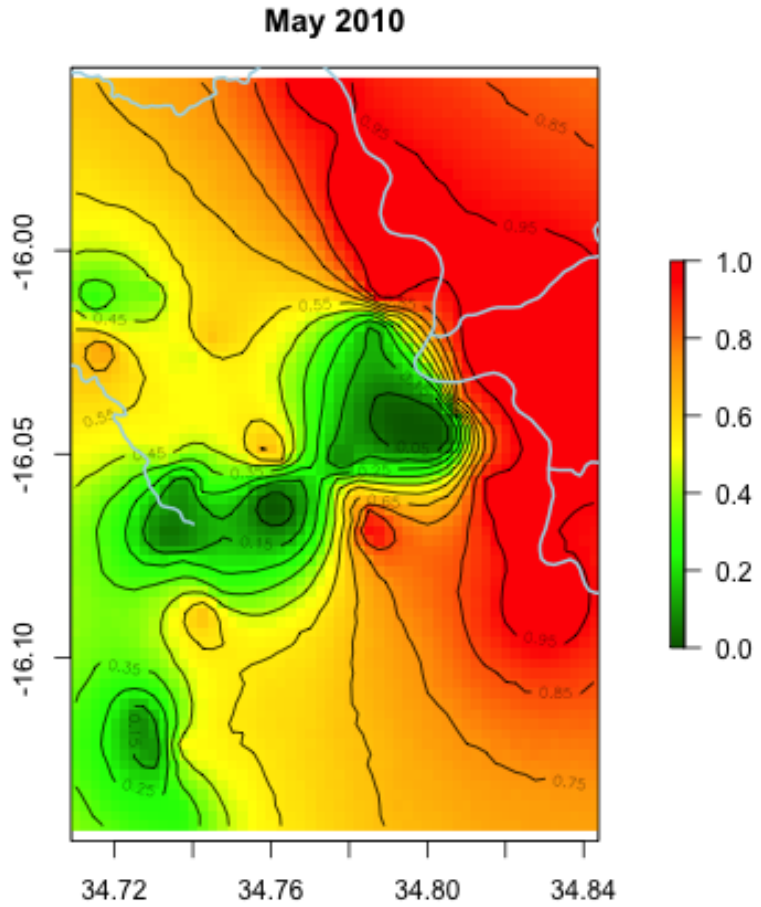
$$Y_i \sim \begin{cases} 0 & : \quad \text{wp } \phi \\ f(y;\theta) & : \quad \text{wp } 1-\phi \end{cases}$$

Spatial model

$\{\phi, \theta\} \rightarrow \{P(x), T(x)\} : x \in \mathbb{R}^2 \sim$ **bivariate stochastic process**

Spatio-temporal mapping: rolling malaria indicator surveys

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



Great words from great statisticians

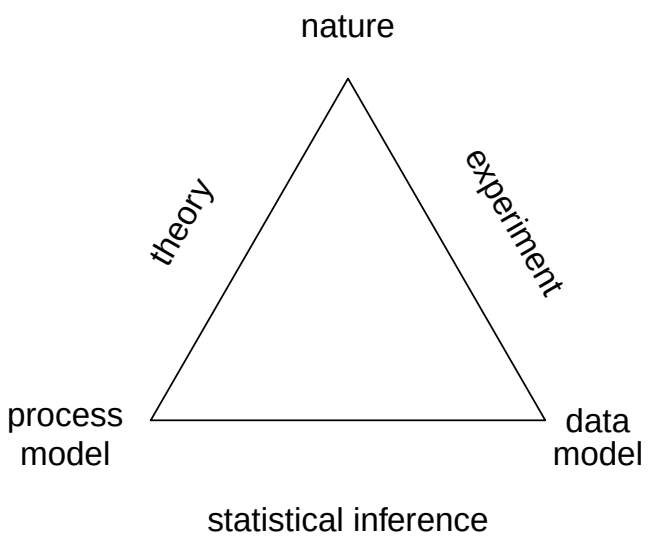
"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

John Tukey (1915–2000)

"...the importance of making contact with the best research workers in other subjects and aiming over a period to establish genuine involvement and collaboration in their activities."

Sir David Cox (b 1924)

Conclusion: the role of statistical method in science



A statistical model is:

- **a device to answer a question**
- **a bridge between theoretical and applied science**
- **a framework to enable principled inference in the presence of uncertainty**

- **Scientific purpose is more important than data-format**
- **Analyse problems, not data**

Diggle, P.J. (2018). Analyse problems, not data. *Spatial Statistics*, **28, 4–7**

Diggle, P.J. and Chetwynd, A.G. (2011). *Statistics and Scientific Method: an Introduction for Students and Researchers*. **Oxford: Oxford University Press.**