Lecture 4

Spatial point-referenced (geostatistical) data analysis

MSc in Epidemiology @ Imperial College London

March 11, 2019

Lecture outline

1. Introduction
2. Gaussian fields
3. Variogram and Semivariogram
4. Spatial regression with correlated random effects
   - Case study: Childhood malaria in the Gambia
5. Spatial Prediction - Case Studies

Point-referenced data

- The difference between models for *point referenced (or geostatistical) data* and the spatial models presented in the previous lectures is that here we treat space as continuous, not discretised (areas).
- We are concerned here with spatial data structures where the process of interest (response) is an underlying spatial field, $Z(\boldsymbol{s})$, i.e. a stochastic process taking real values at each point $\boldsymbol{s}$, (e.g $\boldsymbol{s} = (x, y)$ where x and y represent spatial coordinates):

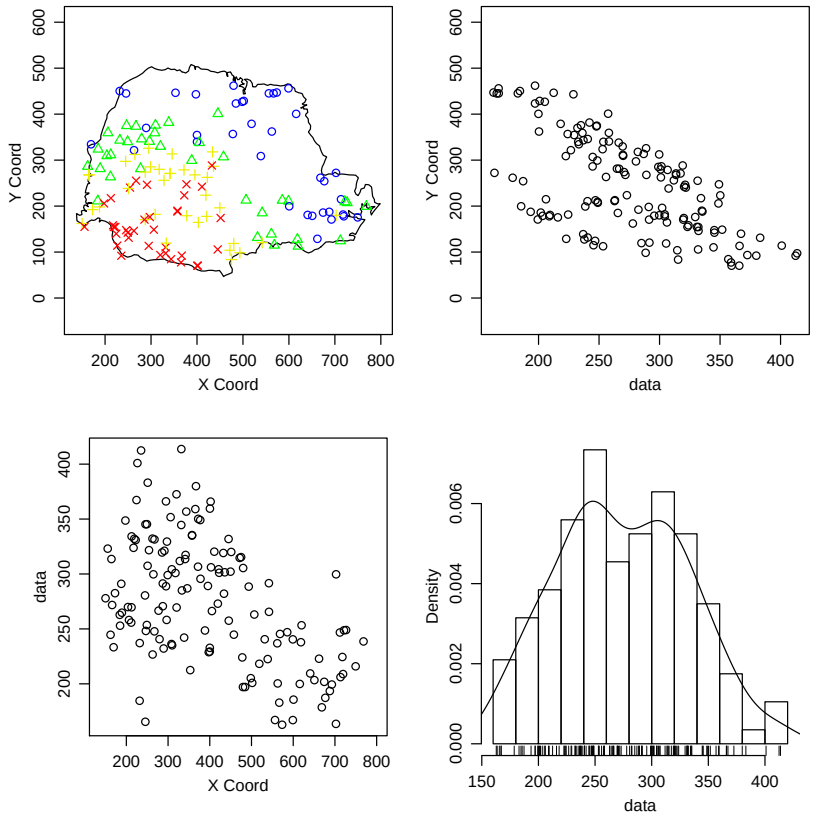
$$\{Z(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\} \text{ where} \mathcal{D} \text{ is a domain in } \mathbb{R}^2$$

- Examples:
  - in the field of environmental science: rainfall, air pollution concentrations, radioactive emission in soil, etc.
  - in epidemiology when considering the risk of disease at different locations
- Data are measured (possibly with error) at the spatial locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$

Rainfall in Parana State (Brazil), Diggle and Ribeiro, 2002

- Amount of rainfall has implications for viability of particular kinds of agricultural activity
- Average rainfall over different years for the dry season (May-June)
- 143 recording stations
- Top left plot shows measured locations coloured by wuartile of the observed rainfall distribution (blue=Q1, green=Q2, yellow=Q3, red=Q4)

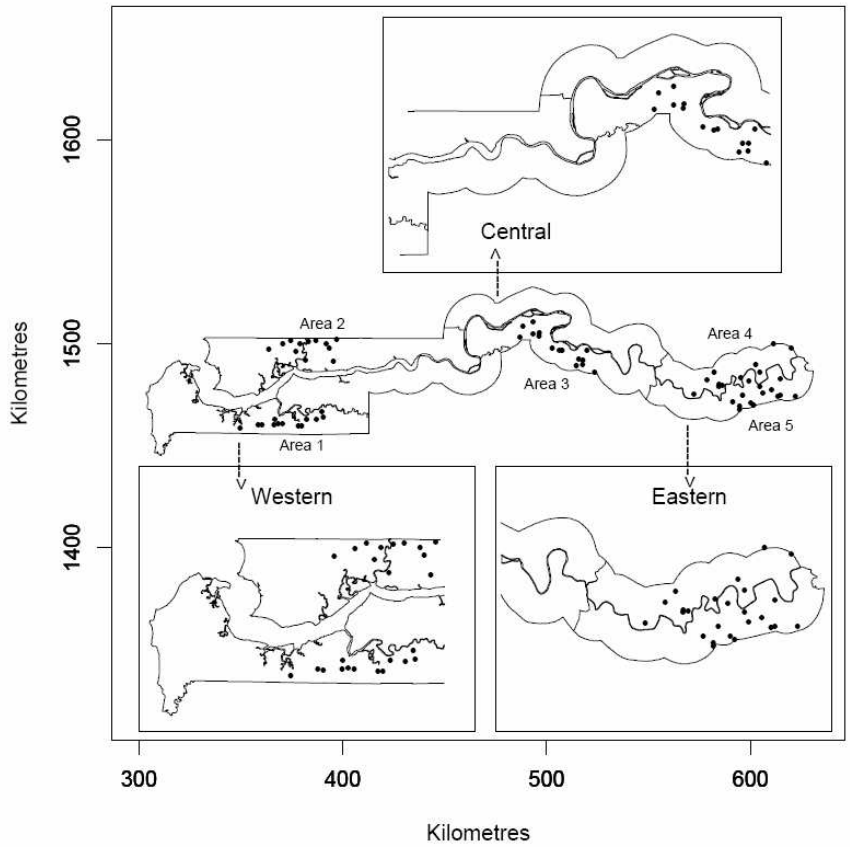
```
require(geoR)
data(s100)
plot(s100, lowess=TRUE)
```



Childhood Malaria in Gambia, Diggle et al, 2002

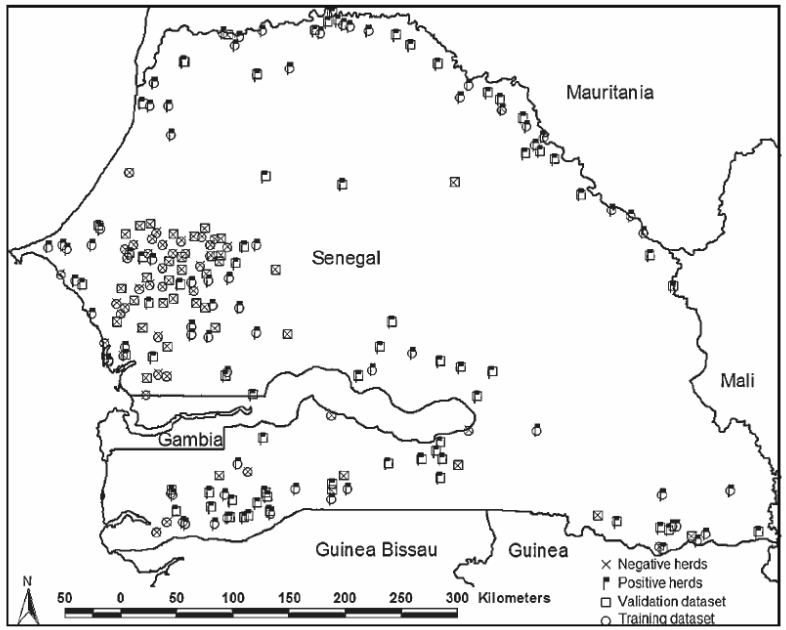
- Aim: measuring the effectiveness of the National Impregnated Bednet Programme in reducing child morbidity and mortality
- Describe the variation in the prevalence of malaria among a sample of village resident children in the Gambia (65 villages)
- Evaluate effectiveness of bednet use in reducing risk of malaria

```
require(geoR)
plot(gambia.borders, type="l", asp=1)
points(gambia[,1:2], pch=19)
gambia.map()
```



Rift Valley Fever in Senegal, Clements et al, 2007

- Rift Valley Fever is a mosquito-transmitted disease affecting a wide range of animals
- Aim: define ecological areas that are potential endemic foci of RVF and from which future epidemics may arise
- 16,738 animals



Aims

- To reconstruct a latent spatial field from a finite set of noisy observations taken at a finite number of spatial locations.
- To use the spatial dependence to predict values of the spatial field (together with associated uncertainty) at locations where there are no observations.
- The common framework to geostatistics models is that of Gaussian fields (or processes) which are based on the Multivariate Normal (MVN) distribution).

Multivariate Normal distribution

$$\begin{aligned} \boldsymbol{Z} &\sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \text{where } \boldsymbol{\mu} &= N \times 1 \text{ vector with elements } \mu_i \\ \boldsymbol{\Sigma} &= N \times N \text{ covariance matrix with elements } \sigma_{ij} \end{aligned}$$

- The element σ_{ij} of $\boldsymbol{\Sigma}$ is the covariance between Z_i and Z_j
  $\sigma_{ij} = 0 \iff Z_i$ and Z_j are independent Gaussians.

- We can write $\boldsymbol{\Sigma} = v^2\rho$ where v^2 is the overall variance and ρ is the correlation matrix, with elements ρ_{ij} representing the correlation between Z_i and Z_j.

Gaussian fields

- A spatial process $\{Z(\boldsymbol{s})\}$ is a Gaussian field (GF) if for any n and for each set of locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, the vector $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))$ follows a multivariate Normal distribution with mean $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))$ and spatially structured covariance matrix $\boldsymbol{\Sigma}$.
- The generic element of $\boldsymbol{\Sigma}$ is defined by a covariance function $\mathcal{C}(\cdot, \cdot)$ such that $\Sigma_{ij} = \text{Cov}\left(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\right) = \mathcal{C}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j))$.
- The spatial process is called second-order stationary if
  - $\boldsymbol{\mu}$ is constant (i.e. $\mu(\boldsymbol{s}_i) = \mu$ for each i)
  - the spatial covariance function depends only on the distance vector $(\boldsymbol{s}_i - \boldsymbol{s}_j) \in \mathbb{R}^2$, i.e. $\text{Cov}\left(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\right) = \mathcal{C}(\boldsymbol{s}_i - \boldsymbol{s}_j)$.
- Moreover, a stationary process is isotropic if the covariance does not depend on the direction but just on the Euclidean distance $||\boldsymbol{s}_i - \boldsymbol{s}_j|| \in \mathbb{R}$.
  Several functions are available for the spatial covariance function (eg exponential, Matérn, spherical, etc.) parameterized by some parameters. See Banerjee et al. 2014 (Chapter 2).

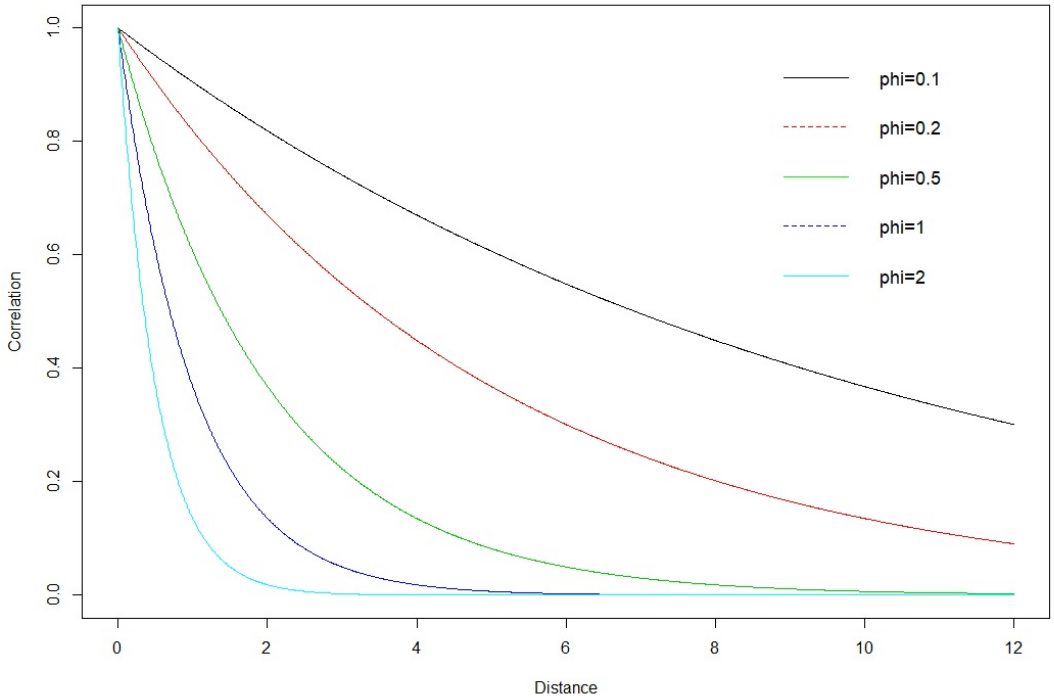
Modelling the covariance matrix

In geostatistical modelling, our aim is to construct a statistical model for the correlation or covariance matrix that reflects how spatial dependence between any pair of points varies as a function of distance (we will only consider here stationary, isotropic covariance models $\rightarrow$ correlation just depends on distance and not on direction or location)

- Assume parametric form $\sigma_{ij} = v^2 f(d_{ij}; \phi)$ where
  - d_{ij} = distance between areas i and j
  - $f(d_{ij}; \phi)$ models the correlation between sites i and j
- Choice of $f(.)$ is essentially restricted to functions that ensure $\mathbf{\Sigma}$ is positive definite (so that it is invertible) and that correlation decreases with increasing distance (see Ripley, 1981)
- A commonly used model (1 parameter):

  Exponential $\quad f(d_{ij}; \phi) = \exp(-\phi d_{ij}); \quad \phi > 0$

Exponential model: $f(d_{ij};\phi) = \exp(-\phi d_{ij})$

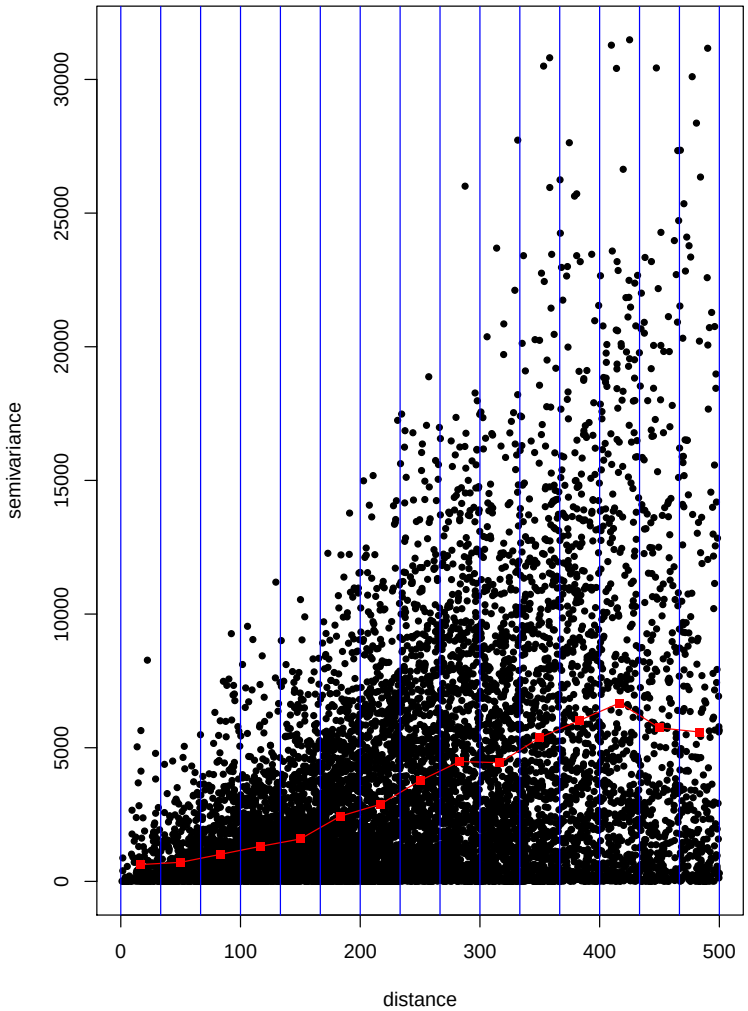


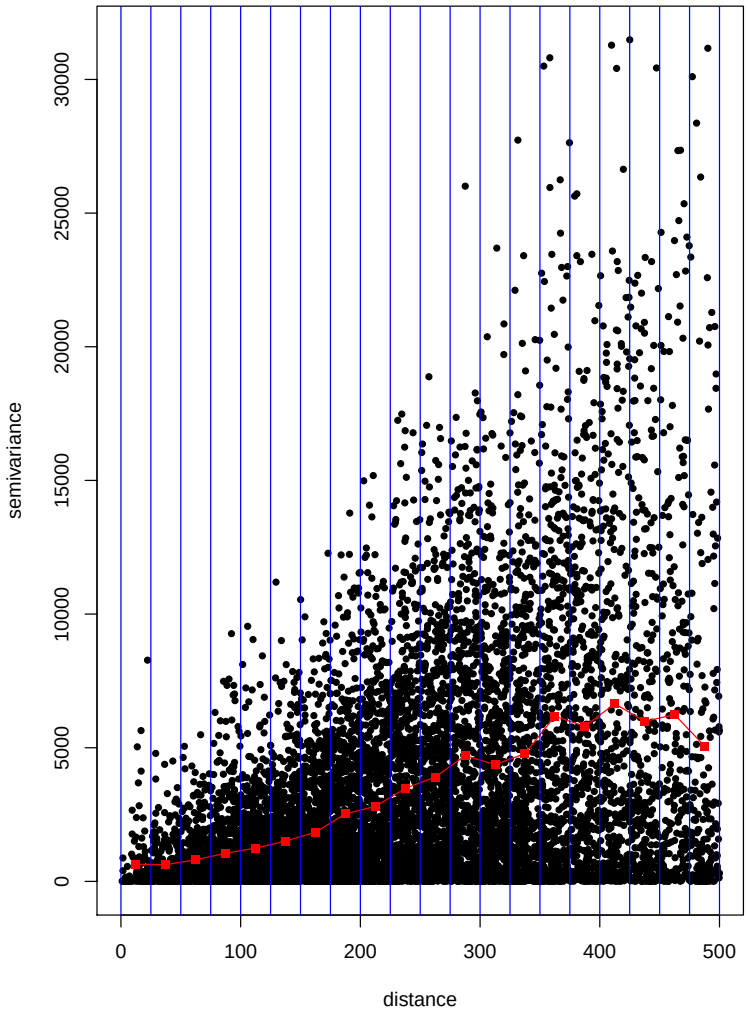

```
# R script for distance-decay curves for the exponential decay with distance
d = seq(0,12,0.01)
phi = c(0.1,0.2,0.5,1,2)
plot(exp(-max(phi)*d)~d,type="n",xlab="Distance",ylab="Correlation", cex=1.1)
for(i in 1:length(phi)) lines(exp(-phi[i]*d)~d,col=i)
legend(8.5, 1, legend=c("phi=0.1", "phi=0.2", "phi=0.5","phi=1","phi=2"),
       col=c("black","red","green","blue","cyan"), lty=1:2, cex=1.1, bty = "n")
```

Variogram and Semivariogram

- In order to explore the decay of the spatial covariance with distance, empirical plots are useful.
- The empirical variogram measures the similarity of values as a function of the distance between their locations.
  - $\text{Var}[Z(s+h) - Z(s)] = E[(Z(s+h) - Z(s))^2]$
  - This is the expected squared difference between values, which generally increases as a function of the distance between the locations.
  - $\text{Var}[Z(s+h) - Z(s)] = 2\gamma((s+h) - s) = 2\gamma(h)$
  - $2\gamma(h)$ is the variogram and $\gamma(h)$ is the semivariogram
- Values of $\gamma(h)$ are then plotted against the distances between s and $s+h$ for every pair of locations to produce a variogram cloud.
- To aid interpretation the empirical variogram is often computed by averaging $\gamma(h)$ within distance bands.
- It can be computed by using the function `variog` in the package `geoR` (Ribeiro and Diggle, 2006).

Variogram for the rain fall in Brazil example (15 (left) or 20 (right) distance classes)



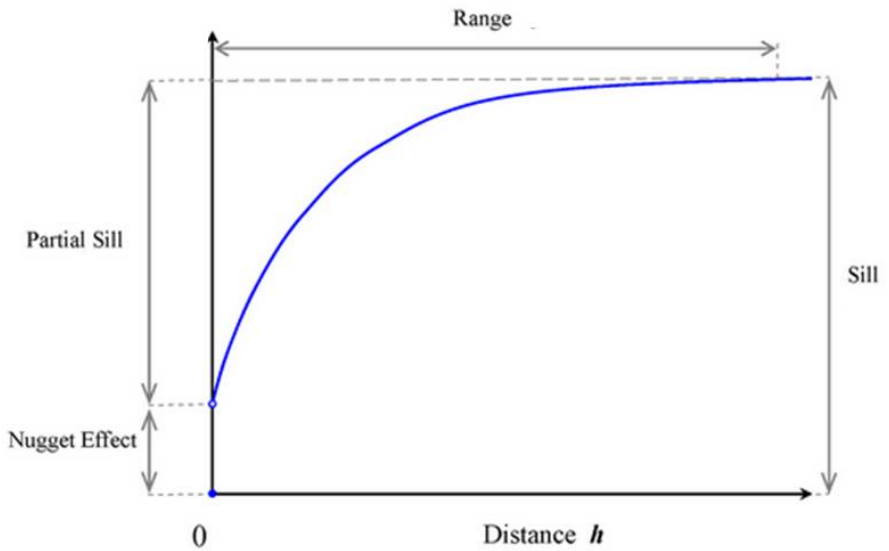


Interpreting the variograms

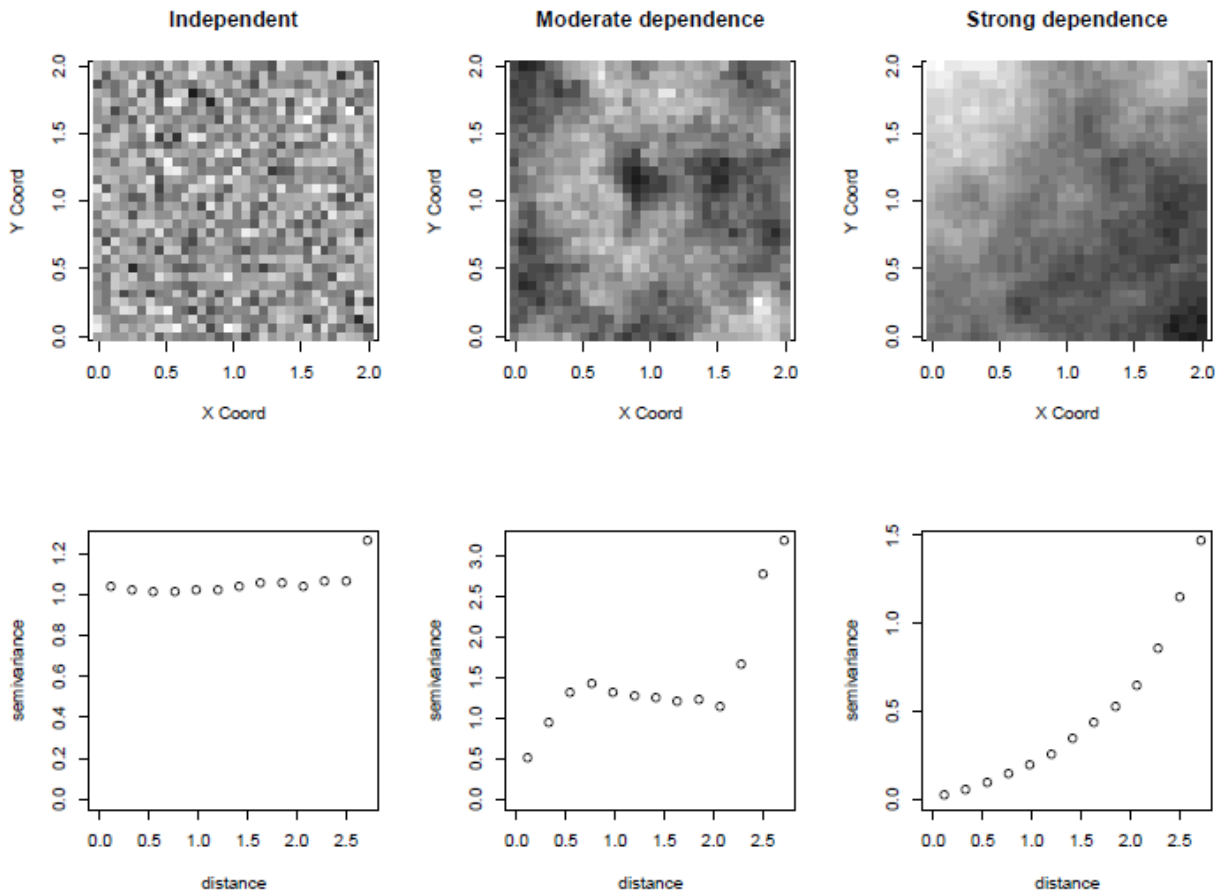
- The semivariogram ($\gamma(h)$ measures half the squared difference between each pair of locations separated at distance h
- If the data are spatially correlated, we would expect observations close together to be more similar than observations far apart (squared differences for observations close together will be smaller than for observations further apart so expect the variogram to gradually increase with increasing distance between observations)
- If data are independent, then on average we would expect the squared difference of a pair of observations that are close together to be similar to that of observations further apart (variogram should be approximately constant as distance increases)

Interpreting the variograms

- Sill:value at which the variogram levels off (corresponds to the overall variance of the data)
- Range: distance at which the semivariogram reaches the sill (corresponds to the distance at which observations become independent)
- Nugget effect: measurement error and microscale variations

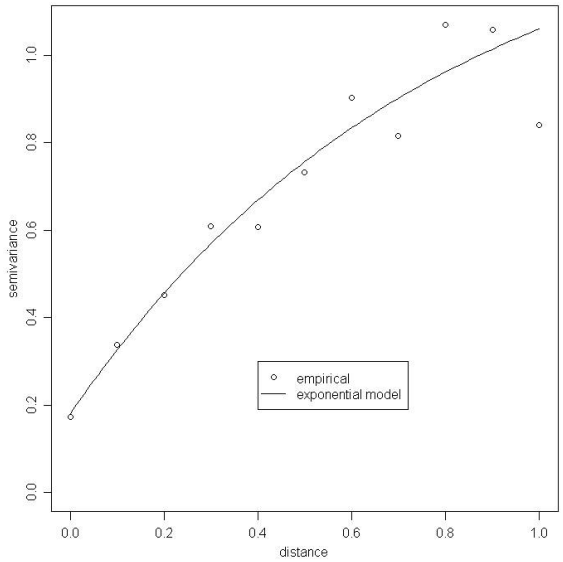


Variogram examples



Estimating the variogram parameters

- We can fit a parametric model to the variogram, to get a smooth fit



- This is done by specifying a suitable parametric model for the covariance matrix (e.g. exponential model) and estimating the parameters of the covariance model (using either max likelihood or Bayesian inference).
- For the exponential model, $\sigma_{ij} = v^2 \exp(-(\phi d_{ij}))$
  - variogram sill $= v^2$
  - variogram range $= 3/\phi$

Uses of geostatistical models in spatial epidemiology

- A geostatistical model is essentially a multivariate normal distribution for a set of observations measured at point locations, where the covariance matrix is modelled as a parametric function of distance
- We would use such a model directly as a sampling distribution for spatially correlated data

$$\boldsymbol{Z} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

  $\rightarrow$ widely used for modelling and predicting environmental exposures
  $\rightarrow$ only appropriate for continuous data
- For modelling spatially correlated health outcomes we use hierarchical GLM framework and model random effects using a spatial multivariate normal prior distribution

Correlated Normal random effects distributions

- In disease mapping and spatial regression over a continuous space, we can use directly a MVN specification for the set of area specific parameters w_i, with spatially structured covariance in terms of distance between the areas
- Alternative to the CAR structure that is based on a discretized space and a notion of neighbourhood.
- E.g. simple disease mapping model:

$$\begin{aligned} \mathrm{O}_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log \lambda_i &= \alpha + w_i \end{aligned}$$

- We now specify directly a MVN structure for $\boldsymbol{w} = \{w_1, ..., w_N\}$ using one of the isotropic specification described previously:

$$\boldsymbol{w} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ follows, for example, an exponential model with correlation decreasing with distance.

BUGS model specification using the power exponential model $\exp(-\phi d_{ij})$ called `spatial.exp`

```
model {
 for(i in 1:N) {Y[i] ~ dpois(mu[i])
  log(mu[i]) <- log(E[i]) + alpha + w[i]
  RR[i] <- exp(alpha + w[i])
}
w[1:N] ~ spatial.exp(mu[], x[], y[], v2.inv, phi, kappa)

 # Prior on overall inverse variance
 v ~ dunif(0, 10)    # variance of the log relative risks
 v2.inv <- 1/pow(v,2)

alpha ~ dnorm(0, 0.0001)  # vague prior (normal with small precision)
                          # on overall mean log risk

for(i in 1:N) { mu[i] <- 0}

# Prior for phi
 range ~ dunif(lower, upper)
 phi <- 3/range
 kappa <- 1
}
```

Arguments to `spatial.exp()` distribution

`spatial.exp[]` distribution is a multivariate normal distribution with covariance matrix having elements $\sigma_{ij} = v^2 \exp(-(\phi d_{ij}))$

- `x[]` and `y[]` are the vectors of x and y coordinates
- `mu[]`: vector of length N (number of spatial locations) giving value of mean in each area (note: we recommend this is set to zero for all areas)
- `v2.inv`: a scalar parameter representing inverse of variance parameter, $1/v^2$
- `phi`: a scalar parameter representing the rate of decline of correlation with distance between points
- `kappa`: a scalar parameter controlling the amount of spatial smoothing (note: we recommend this is set to 1)

Choosing a prior for the range parameter of the structured MVN distribution

- Data often contain little information about ϕ, so need careful choice of prior to ensure model is identifiable and chains mix well; exploratory analysis can help
- Remember that magnitude of ϕ depends on the units of measurement of distance (i.e. of x and y co-ordinates)
- One strategy for specifying a fairly non-informative prior for ϕ:

$$\log \phi \sim \text{Uniform}(\log \phi_{\min}, \log \phi_{\max})$$

where $\phi_{\min}$ and $\phi_{\max}$ are chosen to give a sensible range of correlations at both the minimum and maximum distances between any pair of areas in the study region

Choosing a prior for the range parameter of the structured MVN distribution

- Recall that variogram range $= 3/\phi$ where the range is the distance at which correlation drops to zero.
- The prior on ϕ can be transformed in the prior for the range (more interpretable)
- Another strategy consists in specifying prior on range and then work out the implied values of ϕ for a given range, e.g.

$$\text{range} \sim \text{Uniform}(\text{range}_{min}, \text{range}_{max})$$

where $\text{range}_{\text{min}}$ and $\text{range}_{\text{max}}$ are chosen to give a sensible range of correlations at both the minimum and maximum distances between any pair of areas in the study region. E.g. set minimum range as the minimum distance in the dataset; maximum range should be several times larger than the range of empirical variogram (to be non informative) but no more than 3-4 times the maximum distance in the dataset (otherwise potential confounding with the overall variance v^2)

Case study: Childhood malaria in the Gambia

- Random sample of 2035 children selected from 65 villages
- Response = Binary indicator of presence of malarial parasites in blood sample taken from each child
- Covariates include child's age and use of bed nets, inclusion/exclusion of village from primary health care system and greenness of surrounding vegetation (from satellite information)

Questions of interest:

- Does sleeping under bed nets reduce risk of malaria?
- Do other child or village level covariates help explain risk of malaria?
- Is there evidence of additional spatial structure in malaria prevalence between villages that is not explained by the measured covariates?

Original data set (https://rdrr.io/cran/geoR/man/gambia.html)

Data on 2035 children (M) sampled from 65 villages (N). Variables available:

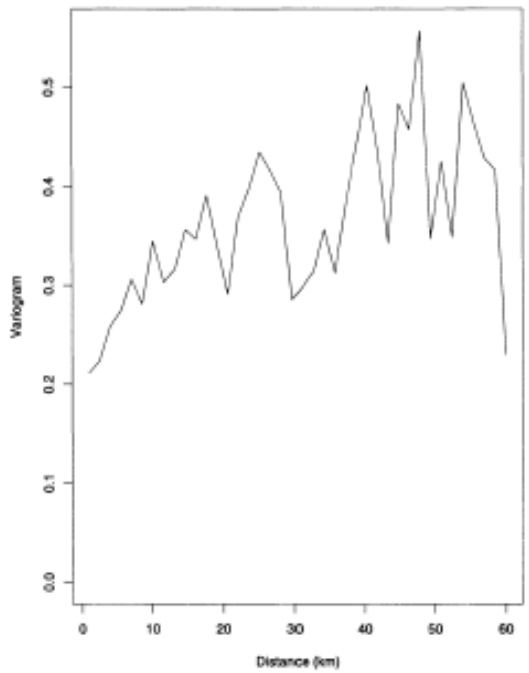
- x: x-coordinate of the village (UTM)
- y: y-coordinate of the village (UTM)
- pos: presence (1) or absence (0) of malaria in a blood sample taken from the child
- age: age of the child, in days
- netuse: indicator variable denoting whether (1) or not (0) the child regularly sleeps under a bed-net
- treated: indicator variable denoting whether (1) or not (0) the bed-net is treated (coded 0 if netuse=0)
- green: satellite-derived measure of the green-ness of vegetation in the immediate vicinity of the village (arbitrary units)
- phc: indicator variable denoting the presence (1) or absence (0) of a health center in the village

Non-spatial generalized linear mixed model for childhood malaria

$$
\begin{aligned}
O_{ji} &\sim \text{Bernoulli}(p_{ji}); \quad j = 1, ..., 65, i = 1, ..., 2035 \\
\text{logit}(p_{ji}) &= \alpha + \boldsymbol{\beta}' X_{ji} + \eta_j \\
\eta_j &\sim N(0, v^2) \\
\alpha, \boldsymbol{\beta}, v^2 &\sim \text{vague priors}
\end{aligned}
$$

Here, $\boldsymbol{\eta}$ are independent normal random effect distribution for the village-level random effects (on the log odds scale)

Empirical Variogram for the posterior means of the random effects



- Rising trend up to a distance of about 30km $\rightarrow$ the covariates do not adequately explain the short-range spatial structure, we need to include a spatial effect

Spatial generalized linear mixed model for childhood malaria

$$
\begin{aligned}
O_{ji} &\sim \text{Bernoulli}(p_{ji}); \quad j = 1, ..., 65, i = 1, ..., 2035 \\
\text{logit}(p_{ji}) &= \alpha + \boldsymbol{\beta}' X_{ji} + w(\boldsymbol{s}) \\
\boldsymbol{w} &\sim \text{MVN}(\boldsymbol{\mu}, v^2 \boldsymbol{\Sigma}) \\
\sigma_{jk} &= \exp^{-\phi d_{jk}} \quad (d_{jk} = \text{distance between villages } j \text{ and } k) \\
\alpha, \boldsymbol{\beta}, v^2 &\sim \text{vague priors}
\end{aligned}
$$

Results for the spatial generalized linear mixed model for childhood malaria

Parameter	*2.5% quantile*	*97.5% quantile*	*Mean*	*Median*
α	−2.966473	2.624348	−0.131214	−0.077961
β_1 (age)	0.000455	0.000933	0.000689	0.000685
β_2 (untreated)	−0.673143	−0.042011	−0.357825	−0.359426
β_3 (treated)	−0.753803	0.088418	−0.32954	−0.325853
β_4 (greenness)	−0.085675	0.047924	−0.020068	−0.020834
β_5 (PHC)	−0.787913	0.129883	−0.344846	−0.349915
β_6 (area 2)	−1.14419	0.51023	−0.324665	−0.331634
β_7 (area 3)	−1.40862	0.558616	−0.5321	−0.559229
β_8 (area 4)	−0.109472	2.425342	1.049441	1.016969
β_9 (area 5)	0.164828	2.606357	1.309553	1.325129
σ^2	0.311756	1.050227	0.585592	0.553477
ϕ	0.915789	10.20069	2.522294	1.422975
δ	0.079522	2.784646	1.084108	0.937436

Figure: Estimates and 95% intervals for the model parameters

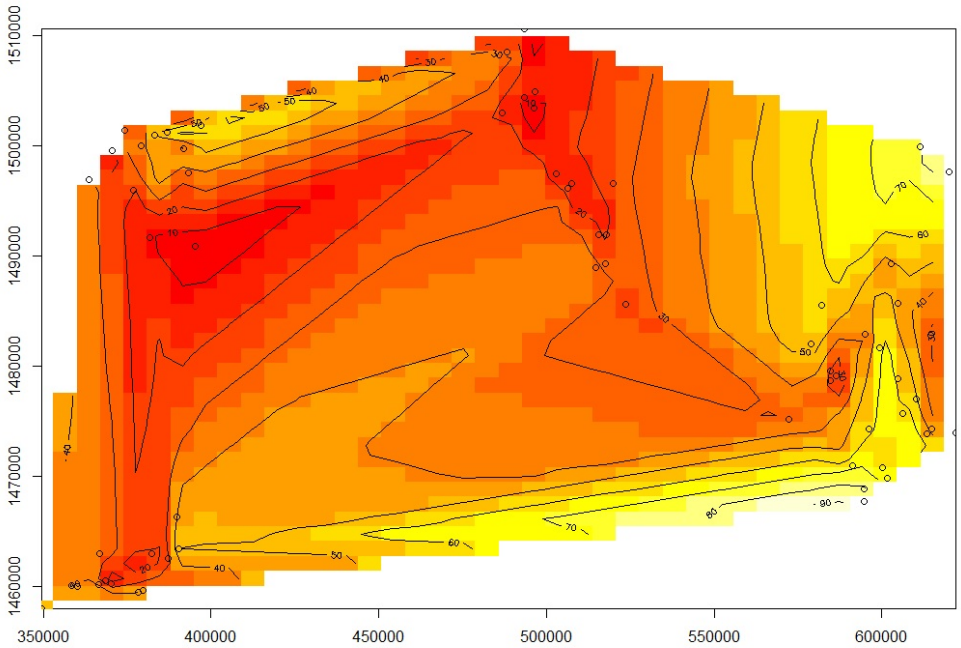
Introducing Practical 4, where we will analyse the data on childhood malaria in the Gambia

We manipulate these data and we write a model for BUGS. In particular, we use

- a categorical variable for bed net use, bednet: (1) doesn't sleep under bed net; (2) sleeps under untreated bed net; (3) sleeps under bed net impregnated with permethrin insecticide
- a categorical variable for age, age: (1)=0-2 yrs, (2) 2-3 yrs, (3) 3-4 yrs, (4) 4-5 yrs
- the variable village identifies the village of the child

```
> head(gambia)
          x       y pos  age green phc bednet village   age
1 349631.3 1458055   1 1783 40.85   1      1       1     4
2 349631.3 1458055   0  404 40.85   1      2       1     1
3 349631.3 1458055   0  452 40.85   1      2       1     1
4 349631.3 1458055   1  566 40.85   1      2       1     2
5 349631.3 1458055   0  598 40.85   1      2       1     2
6 349631.3 1458055   1  590 40.85   1      2       1     2
```

- During the explorative analysis, is often helpful to create image plots and place contour lines on the plot.
- For example, using `akima` invoking the library: `library(akima)` we can *interpolate* the prevalence of children with malaria (more in this lecture and in practical 4).



Model code in BUGS

```
model {
   for(i in 1:M) {
      O[i] ~ dbern(p[i])
      logit(p[i]) <- alpha + beta.age[age[i]] + beta.bednet[bednet[i]] +
                                    beta.green*(green[i] - mean(green[])) +
                                    beta.phc*phc[i] +  w[village[i]]
   }

   # spatial village-level random effect
   w[1:65] ~ spatial.exp(mu[], x[], y[], v2.inv, phi, kappa)

   for(j in 1:N) {  OR.village[j] <- exp(w[j])}   # odds ratio of malaria
                                                  in village j relative to average

   # priors
   for(i in 1:N) { mu[i] <- 0 }

   # alternative priors for random effects variance to check sensitivity
   v2.inv ~ dgamma(0.5, 0.0005)
   v <- 1/sqrt(v2.inv)                     # sd of spatial random effects
# priors on parameters of spatial exponential covariance function
kappa <- 1
range ~ dunif(0.9, 273)
phi <- 3/range
}
```

Model code in BUGS [2]

```
# Priors on regression coefficients
   alpha ~ dnorm(0, 0.0001)

   beta.age[1] <- 0          # set coefficient for baseline age group to zero (corner point constraint)
   beta.age[2] ~ dnorm(0, 0.0001)
   beta.age[3] ~ dnorm(0, 0.0001)
   beta.age[4] ~ dnorm(0, 0.0001)

# Other method to specify the regression coefficients
   #for (k in 2:4){
   #beta.age[k] ~ dnorm(0, 0.0001) }
   #beta.age[1] <- 0

   beta.bednet[1] <- 0     # set coefficient for baseline bednet group to zero (corner point constraint)
   beta.bednet[2] ~ dnorm(0, 0.0001)
   beta.bednet[3] ~ dnorm(0, 0.0001)

   beta.green ~ dnorm(0, 0.0001)

   beta.phc ~ dnorm(0, 0.0001)
```

Model code in BUGS [3]

```
# calculate functions of interest
# odds ratio of malaria for age group k vs age group 1
 for(k in 2:4) { OR.age[k] <- exp(beta.age[k]) }
 ...

# odds ratio of malaria for children using treated bednets
# vs children using untreated bednets
 OR.bednet[4] <- exp(beta.bednet[3] - beta.bednet[2])

# odds ratio of malaria per unit increase in
# greenness index of village
OR.green <- exp(beta.green)

# odds ratio of malaria for children living in
# villages belonging to the primary health care
# system versus those who don't
OR.phc <- exp(beta.phc)

# baseline.prev = prevalence of malaria in baseline group (i.e. child
# in age group 1 (<2yrs), sleeps without bednet, and lives in a village
# with average greenness index and not in the health care system)
logit(baseline.prev) <- alpha

# probability that using treated vs
# untreated bed net reduces risk of malaria
PP.treated <- step(1 - OR.bednet[4])
```

Posterior means and 95% CI for the odds ratios

spatial exponential model

Predictor	OR (posterior mean)	95%CI
baseline.prev	0.30	(0.19, 0.42)
OR.age[2]	2.20	(1.47, 3.20)
OR.age[3]	2.82	(1.88, 4.04)
OR.age[4]	3.17	(2.16, 4.51)
OR.bednet[2]	0.65	(0.47, 0.88)
OR.bednet[3]	0.44	(0.28, 0.66)
OR.bednet[4]	0.69	(0.45, 1.02)
OR.green	1.05	(1.01, 1.08)
OR.phc	0.78	(0.47, 1.22)

- $e^{\beta_{bednet[2]}}$ = OR of malaria for children sleeping under untreated bed nets vs no bed nets.
  Posterior mean and 95% interval = 0.65 (0.47, 0.88)
- e^{w_j} = OR of malaria in village j after adjusting for covariates. Posterior mean range from 0.23 to 4.70
- $e^{-1\phi}$ = correlation in residual log odds ratios for villages 1km apart. Posterior median and 95% interval = 0.95 (0.66, 0.99)

Bayesian Kriging and Spatial Prediction

- Kriging: method for spatial interpolation or prediction based on a set of observations at point locations.
- Aim: to predict values of a target variable over the whole area of interest, which typically results in image or maps.
- The problem: Given observations at some locations, $Z(\boldsymbol{s}_i), i = 1, \ldots, n$ we want to make statements about the value at unobserved location $(\boldsymbol{s})$, $Z(\boldsymbol{s}_0)$ for $\boldsymbol{s}_0 \notin \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$
- MVN models with parameterised covariance correspond to the variogram models often used for kriging in geostatistics.

Bayesian Kriging and Spatial Prediction

Classical kriging is a 2-step process:

- Estimate parameters of (parametric) variogram model
- Plug the estimated variogram parameters into the kriging equations to predict or interpolate data at arbitrary new locations
- Assumes data (or a transformation) are Gaussian and predictions are linear in the data

Bayesian or model-based kriging (see Diggle et al, 1998):

- Allows simultaneous estimation of the covariance function and prediction at new locations $\Rightarrow$ kriged predictions allow for parameter uncertainty
- Can be embedded within GLM framework to allow kriging with non-Gaussian data.

Study 1 (Linear model) : Prediction of Rainfall

Data:

- $Y_i, i = 1, ..., 143$ are average winter (dry season) rainfall at 143 recording stations throughout Parana State, Brazil

Questions of interest:

- Amount of rainfall has implications for viability of particular kinds of agricultural activity
  - Aim is to predict rainfall, T_j, at grid of 70×50 points (x_j, y_j) covering Parana State
  - Also aim to estimate subregion where rainfall exceeds 300mm with high probability

Essentials of BUGS code

```
# Model
for(i in 1:143) {
  Y[i] ~ dnorm(S[i], tau)  # measurement error
  mu[i] <- beta
  }
 S[1:143] ~ spatial.exp(mu[], x[], y[], v2.inv, phi, 1)
# Priors .......
# Prediction (single site)

for(j in 1:3500) {
T[j] ~ spatial.unipred(mu.T[j],x.T[j],y.T[j],S[])
mu.T[j] <- beta}

# Probability that rainfall will exceed 300mm
for(j in 1:3500) {
  P300[j] <- step(T[j] - 300)
}
```

`spatial.unipred` is a function for predicting values of the fitted surface at unsampled locations (it carries out single site prediction).

Exceedance probabilities that $S(s) > 300$. From Diggle and Ribeiro, 2002

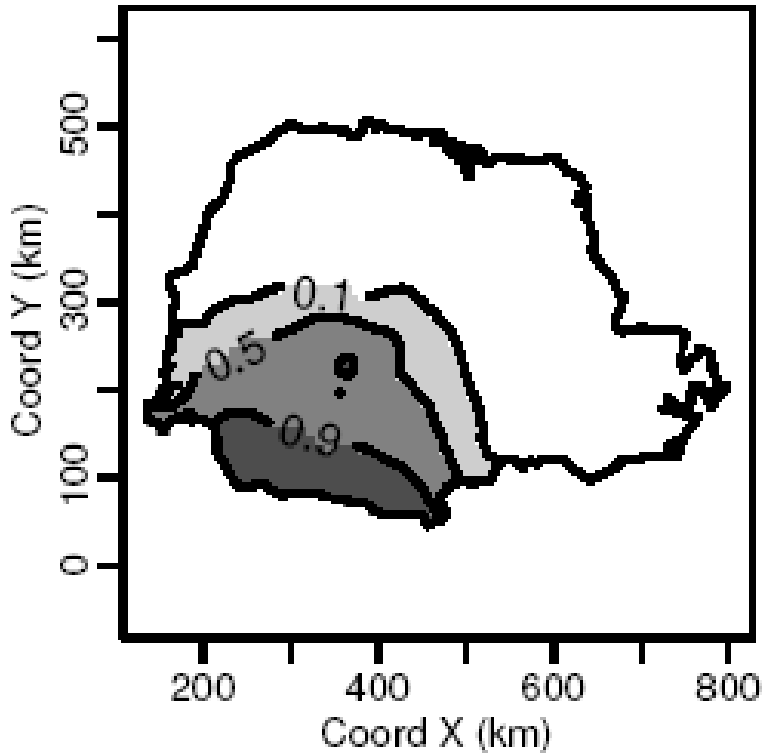


Figure 10. Posterior probability contours for levels 0.10, 0.50 and 0.90 for the random set $T = \{x : S(x) > 300\}$.

Study 2 : Spatial prediction of Rift Valley Fever in Senegal (Clements et al, 2007)

- Rift Valley fever (RVF) is broadening its geographic range and is increasingly becoming a disease of global importance with potentially severe consequences for human and animal health.
- Model of spatial risk assessment of RVF in Senegal using serologic data from 16,738 animals in 211 locations. Serologic status was defined on the basis of the virus neutralisation test
- Many spatial predictor variables are investigated (livestock density, land cover, weather variables,...) as well as non spatial variables (species present at each sampled location - cattle or mixed cattle/small ruminant, time of year during which the samples were taken,...)
- Training dataset (106 locations) and validation dataset (105 locations)
- Spatial prediction on additional areas

Essentials of BUGS code

```
model {
 for(i in 1:211) {
   O[i] ~ dbin(p[i], N[i])
   logit(p[i]) <- alpha + beta_1 X_{i1} + beta_2 X_{i2} + ... +
                  w[i]
 }
 # spatial locations random effect
 w[1:211] ~ spatial.exp(mu[], x[], y[], v2.inv, phi, 1)

 for(i in 1:211) {
 # OR of fever in area i relative to average
   OR.village[i] <- exp(w[i])
 # centre spatial random effects about zero
   mu[i] <- 0
 }
 # prior on sd of spatial random effects
 v2.inv ~ dgamma(1, 0.0005)
 v2 <- 1/inv.v2;  v <- sqrt(v2)
 # priors on parameters of spatial exponential covariance function
 logphi ~ dunif(-2.3,3.91)
 alpha ~dnorm(0,0.0001)
```

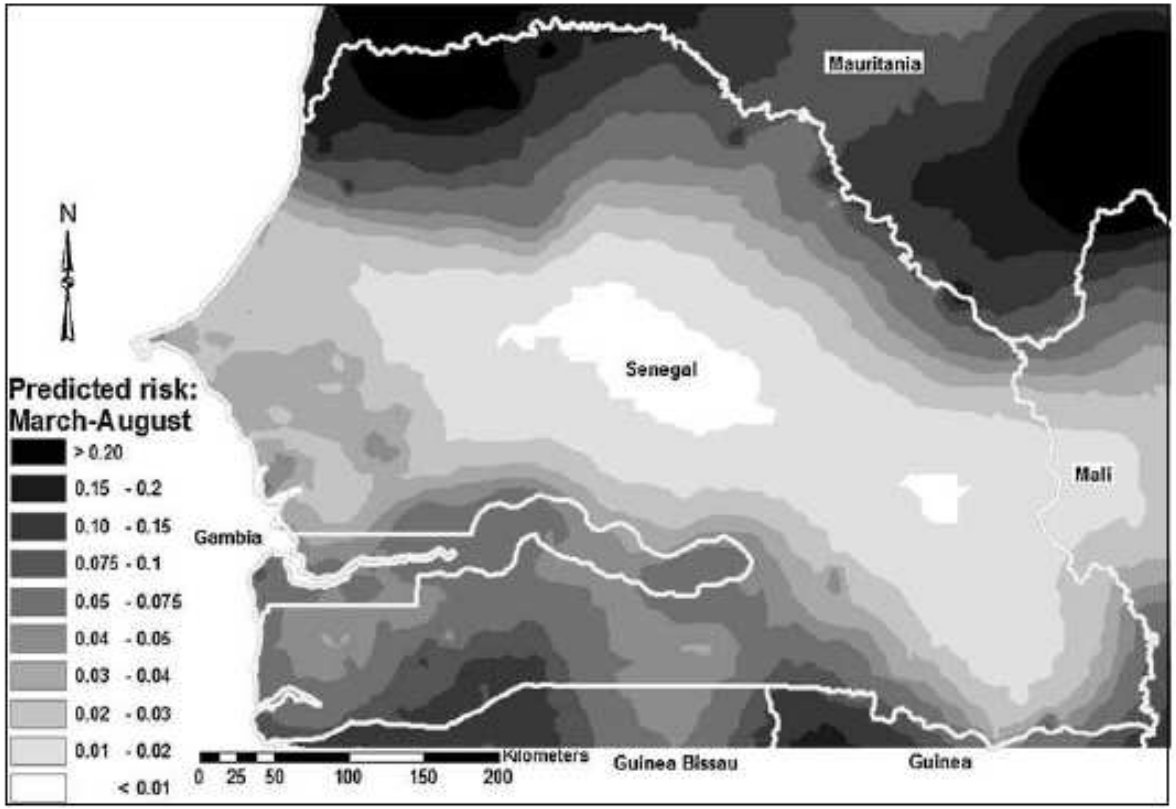
OpenBUGS code for prediction

```
for(i in 1:29){
 O.pred[i] ~ dbin(p.pred[i], N.pred[i])
 logit(p.pred[i]) <- alpha + beta_1 X.pred_{i1}
                           + beta_2 X.pred_{i2} + ... +
                           + w.pred[i]
#Spatial structure
w.pred[j] ~ spatial.unipred(mu.pred[j],x.pred[j],y.pred[j],w[])
mu.pred[j] <-0
}
```

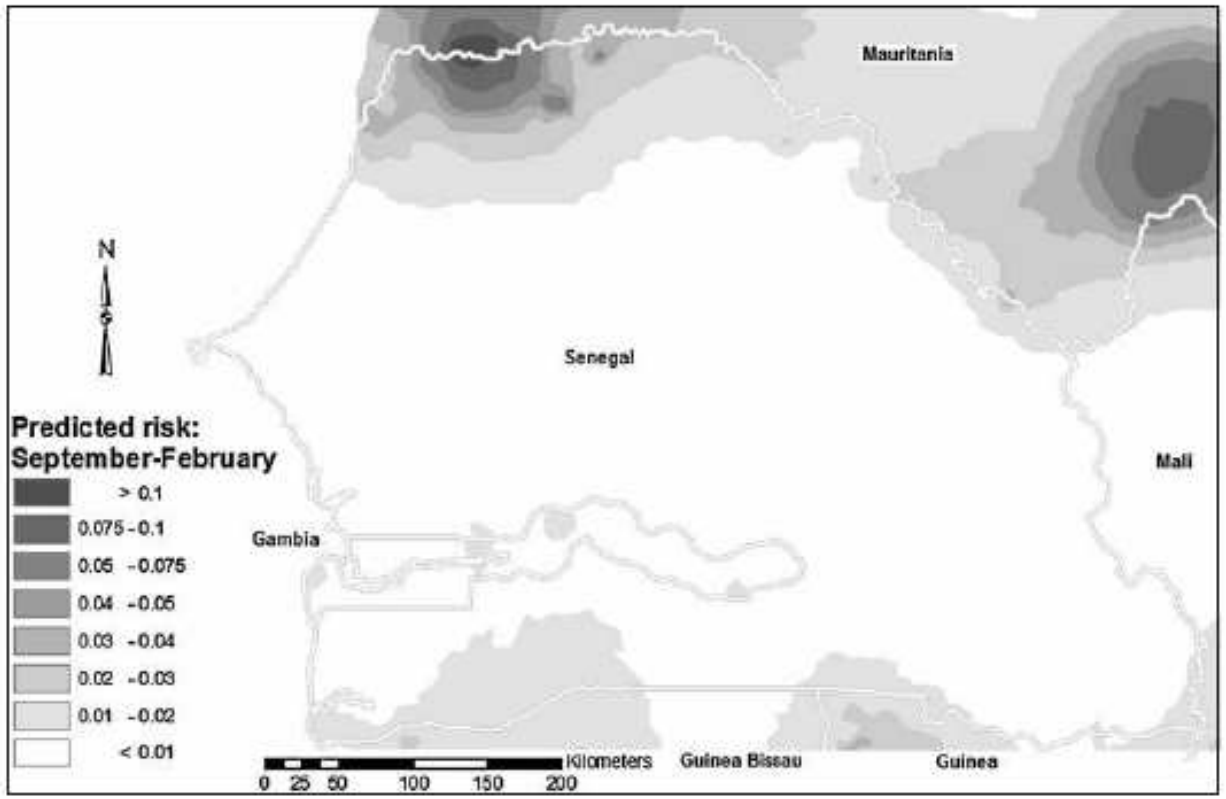
where

- `mu.pred` is the mean for the sites to be predicted
- `x.pred` and `y.pred` are the coordinates
- `w[]` is the vector of observations to which the spatial structure has been fitted

Prediction of the risk of disease for March-August



Prediction of the risk of disease for September-February



Some conclusions

- Foci of high risk in the lower Senegal River basin, Southern Mauritania, and the border regions between Senegal and Guinea and Guinea Bissau.
- The areas of lowest predicted risk were in central and eastern Senegal.
- Constant, overall higher risk in March-August compared to September-February, although the models used (which are not explicitly spatio-temporal) did not allow the spatial pattern to vary for the different time periods.

Summary

- Geostatistical models are used to model spatial dependency at point locations (instead of areas)
- Typical model consists of distance decay in the spatial correlation
- Prior on the hyperparameters ϕ (or range) needs to be chosen carefully - exploratory data using variogram can help
- Spatial prediction easy to carry out (on areas where data are not available) - but very computationally demanding in OpenBUGS

References and Further reading

- Clements et al (2007) Spatial risk assessment of Rift Valley Fever in Senegal. Vector-borne and zoonotic diseases 7(2), 203–216 (Application of Bayesian geostatistics models in WinBUGS to infectious disease mapping).
- Diggle and Ribeiro (2002) Bayesian inference in Gaussian Model-based Geostatistics, Geographical and Environmental Modelling, 6(2), 129–146.
- Diggle et al (2002) Childhood malaria in the Gambia: a case study in model-based geostatistics, Applied Statistics, 51, 493–506.
- Diggle and Ribeiro (2007) Model-based geostatistics. Springer (Springer Series in Statistics).
- Ribeiro and Diggle (2006) GeoR : Package for Geostatistical Data Analysis.
- Banerjee, Carlin and Gelfand (2014) Hierarchical Modeling and Analysis for Spatial Data. (2nd ed. CRC press).