Spatial Regression part 2

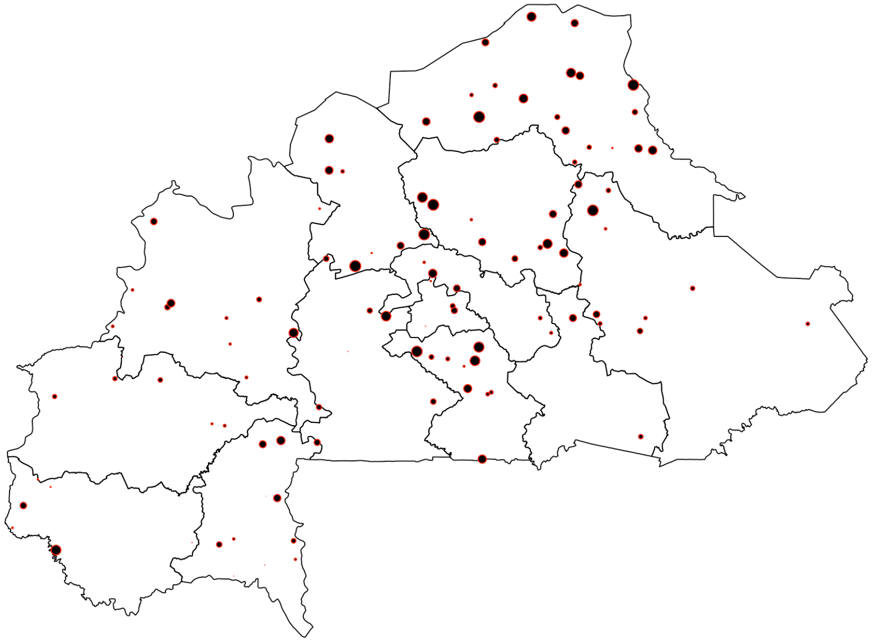
This week

- Review and follow up from last week
- Modeling areal data using CAR models
- CARBayes and INLA

- Point prevalence data
- Accounting for spatial autocorrelation
  - GAMs and INLA-SPDE
- Prediction

Point prevalence data

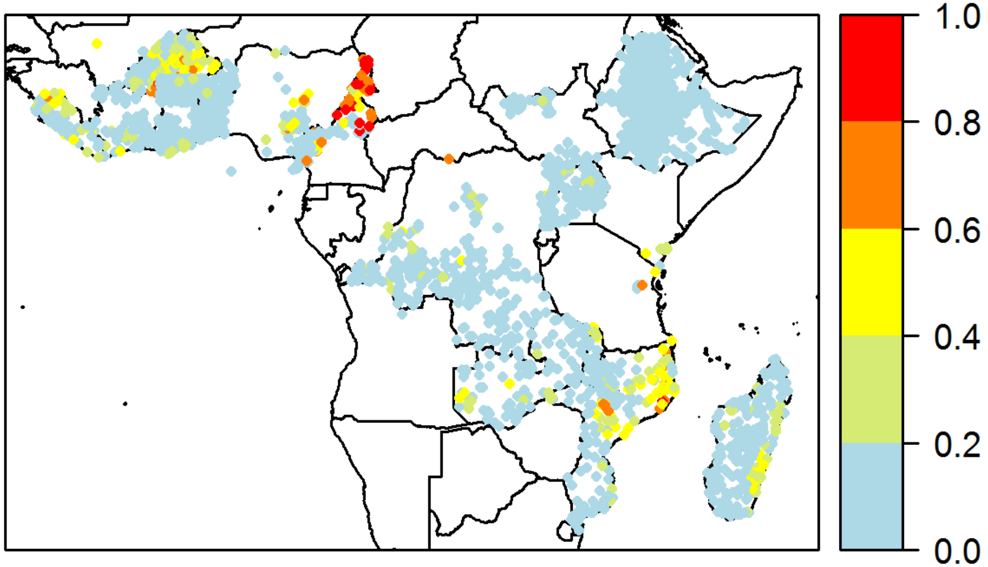
- Active measurement taken from a population
- For health studies, often as part of a survey
- Data are often geo-referenced at household or village level
- Take the format of 0/1 for individual level, or #pos/#tested at the aggregate level (household, village, etc)
- For spatial problems, we want to understand the underlying spatial process that determined the prevalence at the measured points



Geostatistical data

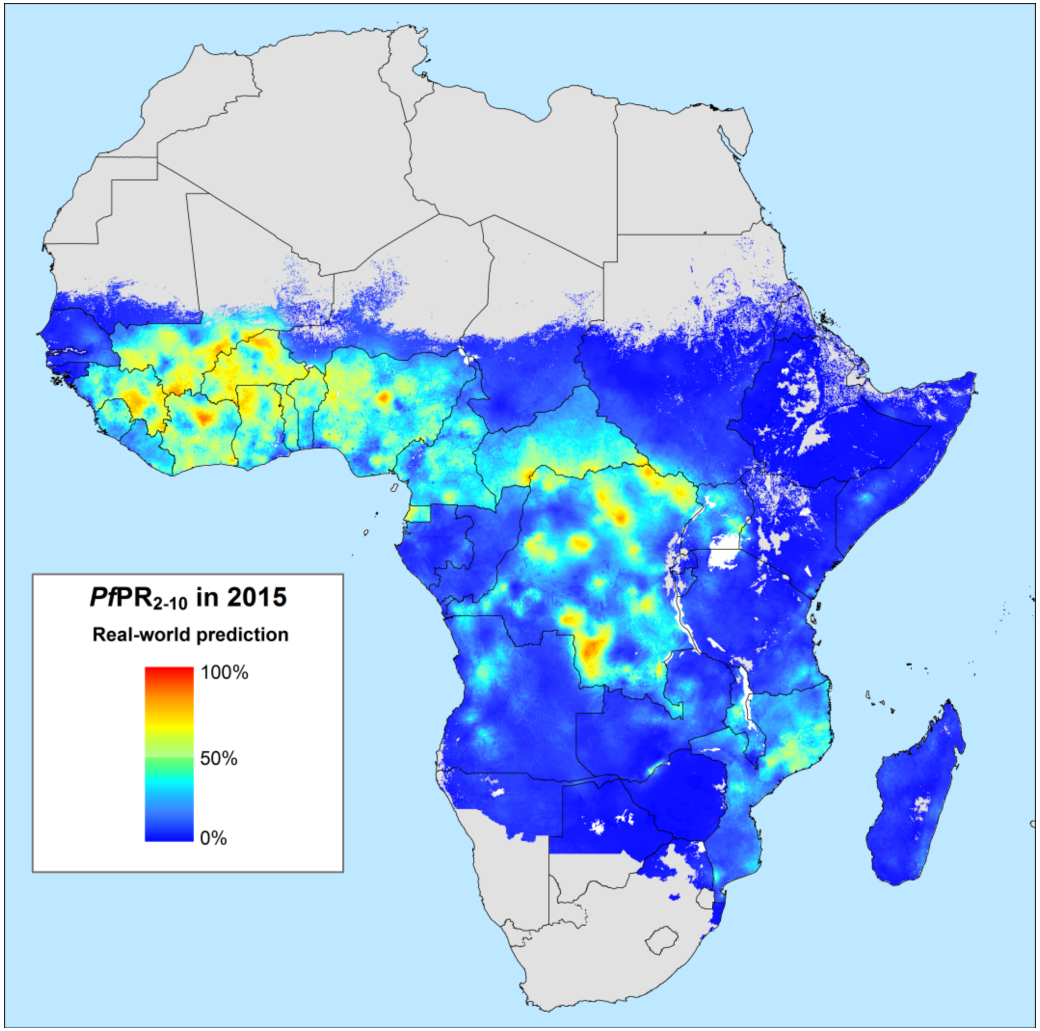
The disease risk is a spatially continuous phenomenon but it is measured only at particular sites

We want to predict the disease risk at unobserved locations and construct a spatially continuous risk surface

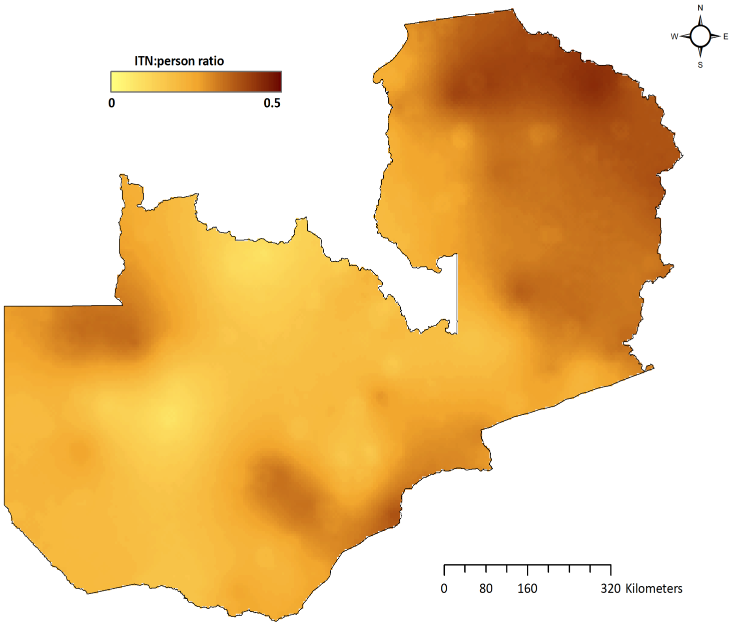


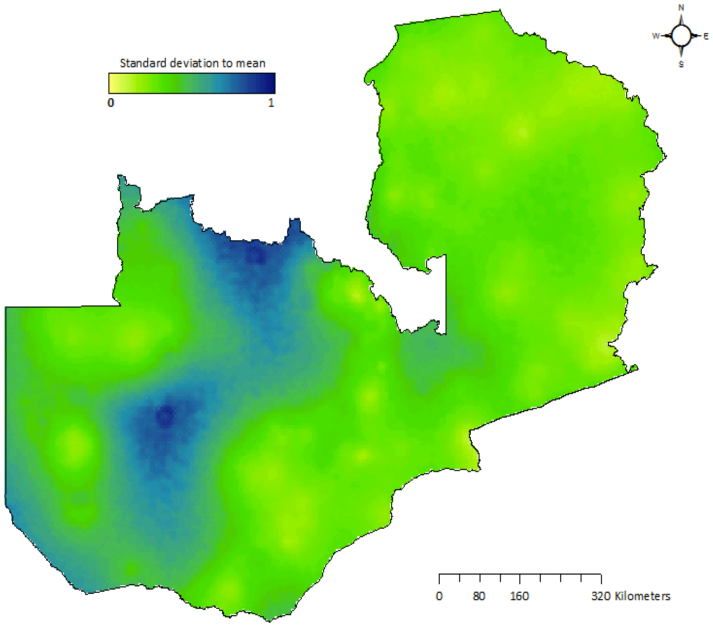
Risk mapping

- http://www.map.ox.ac.uk/



Mapping uncertainty





Distributions

Poisson distribution

$Y \sim Poisson(\mu)$

$y = 0, 1, 2, \ldots$

expected number of occurrences $\mu > 0$

$p(y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!}$

$E(Y) = \mu, \; Var(Y) = \mu$

Binomial distribution

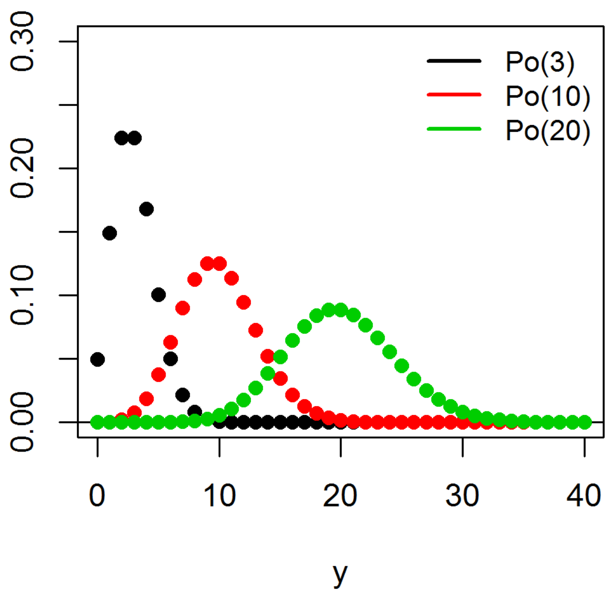
$Y \sim Binomial(n, \pi)$

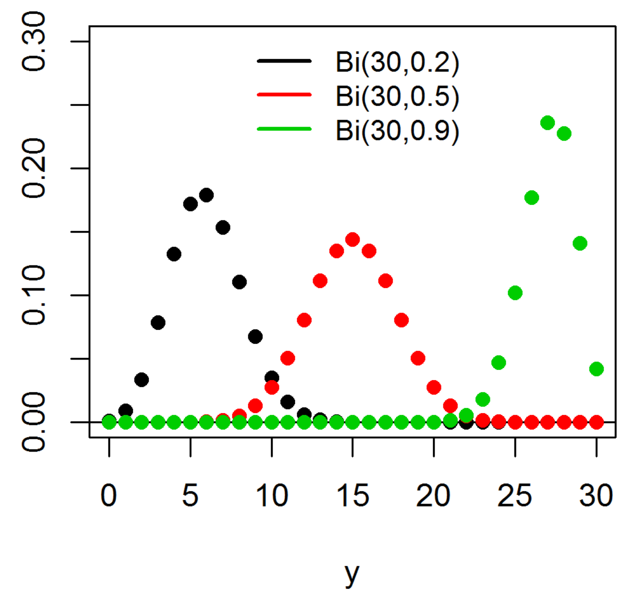
$y = 0, 1, \ldots, n$

number of trials $n \in \{1, 2, \ldots\}$,
probability of success $\pi \in [0, 1]$

$p(y|n, \pi) = \binom{n}{y} \pi^{y}(1-\pi)^{(n-y)}$

$E(Y) = n\pi, \; Var(Y) = n\pi(1-\pi)$





Regression models

Log-linear model

$Y_i \sim Po(\mu_i),\ i = 1, ..., n$

$\log(\mu_i) = \alpha + \beta x_i$

For a one unit increase in x, the mean increases by a factor of $exp(\beta)$ (holding all other covariates constant)

Logistic model

$Y_i \sim Binomial(n_i, \pi_i),\ i = 1, ..., n$

$\text{logit}(\pi_i) = \alpha + \beta x_i$

log-odds: $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

$\pi_i = \text{logit}^{-1}(\alpha + \beta x_i) = \frac{exp(\alpha + \beta x_i)}{1 + exp(\alpha + \beta x_i)}$

β is the change in the log-odds associated with one-unit increase in the x covariate (holding all other covariates constant)

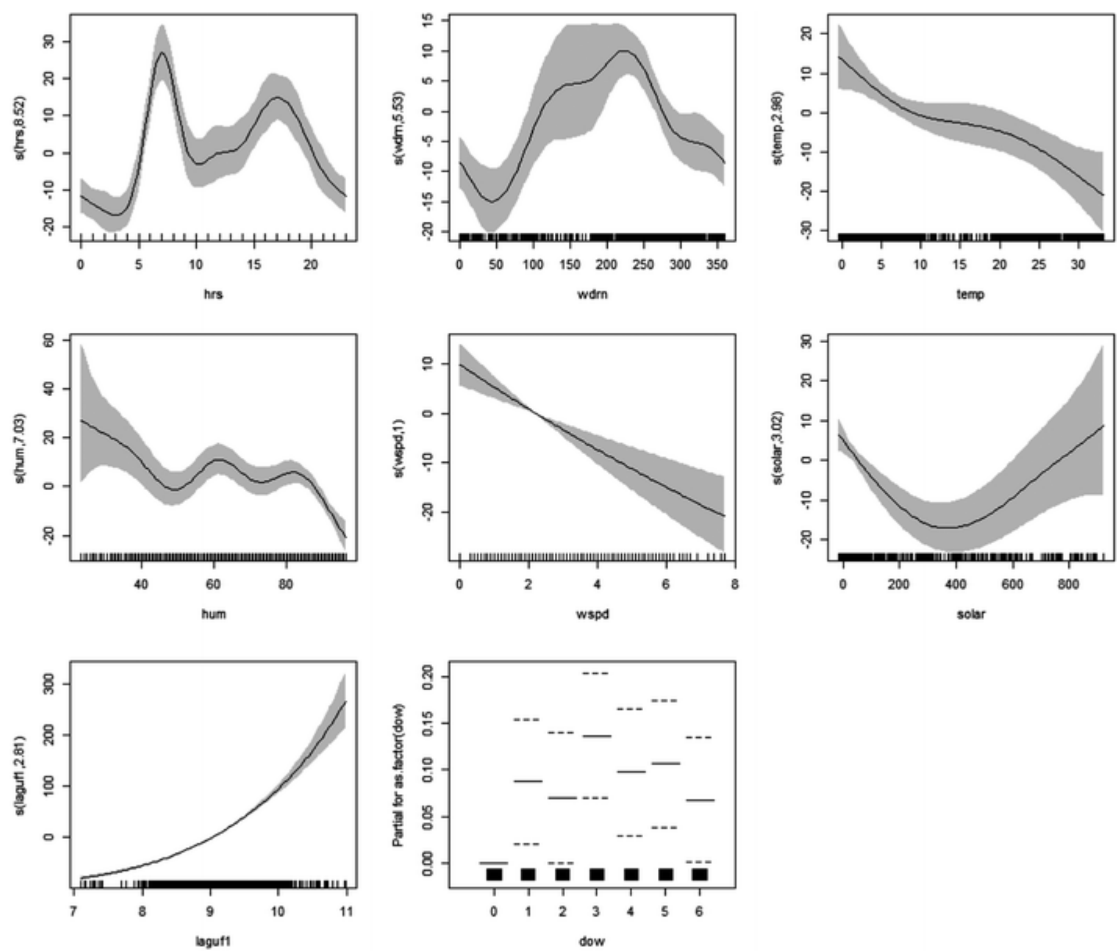
$exp(\beta)$ is the odds ratio

$$OR = \frac{\frac{\pi}{1-\pi}\big|_{x=1}}{\frac{\pi}{1-\pi}\big|_{x=0}}$$

For a one unit increase in x, the odds increase by a factor of $exp(\beta)$ (holding all other covariates constant)

Generalized additive models ('GAM')

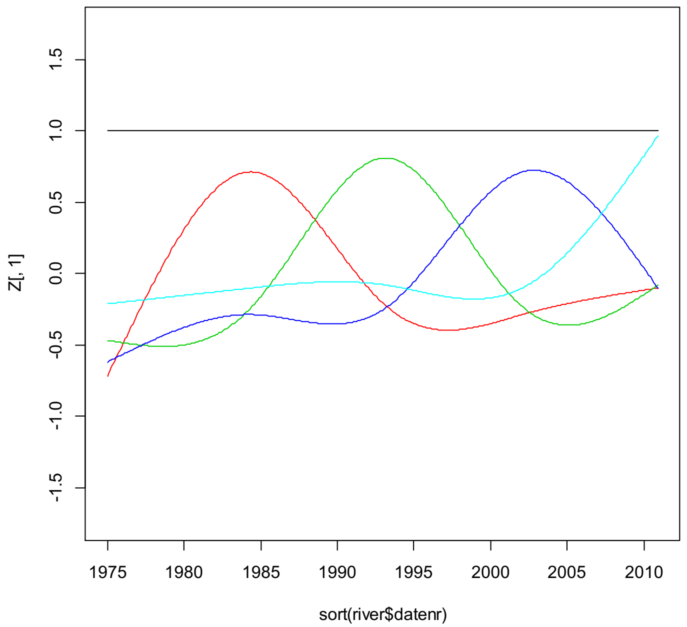
- Highly flexible modeling structure
- Estimate smoothed relationships
- Non-parametric
  - Models best relationship between predictor and outcome using splines
- Useful for spatial patterns



Running GAMs in R

$$Y = a + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) + \varepsilon$$

- f are smooth functions – different types of loess or splines can be incorporated
- We can implement these very simply using the mgcv package in R
- gam package also available
- Specification is similar to glm models
  - We can model binomial data
  - Mixed models can be fit with gamm4
  - We can fit 'big data' using bam



Splines available in mgcv

- Thin plate spline (‘tp’)
  - does not use knots; can be used for multiple covariates   (the default)
  - we can model lat/long to account for spatial trend
- Cubic regression spline (‘cr’)
  - uses knots; only for single covariates
- Tensor products (‘te’)
  - can be used for multiple covariates (interactions etc)

Model-based geostatistics

Geostatistics is that part of spatial statistics which deals with data obtained by spatially discrete sampling of a spatially continuous process $S(\cdot) = \{S(x) : x \in A \subset \mathbb{R}^2\}$

Data consist of measurements $Y_1, \ldots, Y_n$ taken at locations $x_1, \ldots, x_n$ sampled within A, and Y_i is a noisy version of $S(x_i)$

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

Prediction problems are based on explicitly declared statistical models for the data. Likelihood-based methods of inference are applied to fit the model and the fitted model is used to make predictions

Model

Distribution for the number of positive results Y_i out of N_i people sampled conditional on the true prevalence $P(x_i)$ at locations x_i,

$$Y_i | P(x_i) \sim \text{Binomial}(N_i, P(x_i)),$$

$$\text{logit}(P(x_i)) = z_i \beta + S(x_i)$$

$z_i = (1, z_{i1}, \ldots, z_{ip})$ vector of the intercept and covariates
$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^t$ coefficient vector

Fixed effects quantify the effects of the covariates on the disease

$S(\cdot)$ is a zero-mean Gaussian process with
$Var[S(x)] = \sigma^2$ and $\rho(u) = Corr[S(x_i), S(x_j)]$, $u = ||x_i - x_j||$

Correlation functions: Matérn, Powered exponential, Spherical families

Random effects represent residual spatial variation which is not explained by the available covariates

Correlation functions

Matérn family

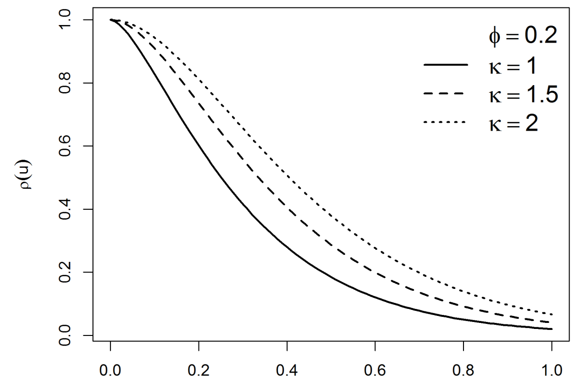
$$\rho(u) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)}(u/\varphi)^{\kappa}\ K_{\kappa}\ (u/\varphi)$$

$\varphi > 0$ scale, $\kappa > 0$ order, $K_{\kappa}(\cdot)$ modified Bessel function of order κ

Exponential: $\rho(u) = exp(-u/\varphi)$ $(\kappa = 0.5)$

Gaussian: $\rho(u) = exp(-(u/\varphi)^2)$ $(\kappa \to \infty)$

$S(\cdot)$ is $I\kappa - 1$ times mean-square differentiable



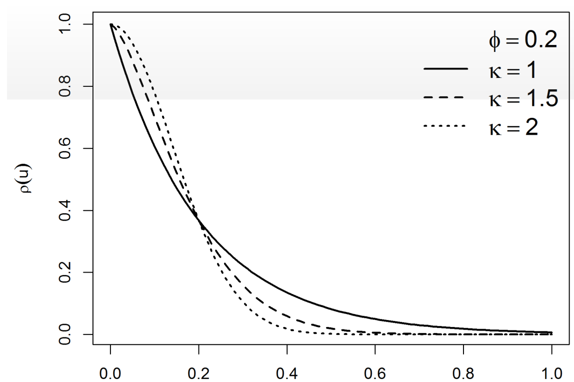
Powered exponential family

$$\rho(u) = exp\left(-(u/\phi)^{\kappa}\right)$$

$\varphi > 0$ scale, $0 < \kappa \leq 2$ shape

if $\kappa < 2$, $S(\cdot)$ mean-square continuous but not mean-square differentiable

if $\kappa = 2$, $S(\cdot)$ infinitely mean-square differentiable



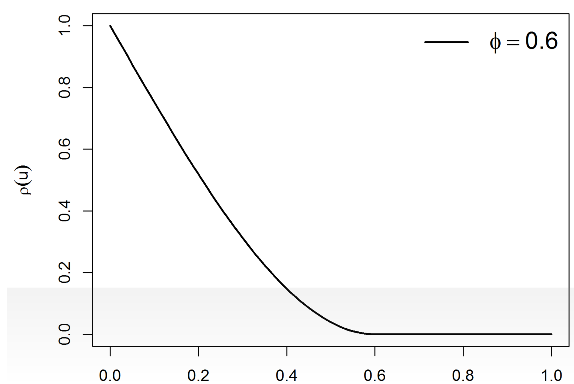
Spherical family

$$\rho(u) = \begin{cases} 1 - \frac{3}{2}(u/\phi) + \frac{1}{2}(u/\phi)^3 & : 0 \leq u \leq \phi \\ 0 & : u > \phi \end{cases}$$

$\varphi > 0$ single parameter

It has finite range, i.e. $\rho(u) = 0$ for sufficiently large u, namely $u > \varphi$

$\rho(u)$ is only once differentiable at $u = \varphi$



Example: Model

Conditional on the true prevalence $P(x_i)$ at location x_i, $i = 1, \ldots, n$, the number of positive results Y_i out of N_i people sampled at x_i follows a binomial distribution,

$$Y_i | P(x_i) \sim \text{Binomial}(N_i, P(x_i))$$

$$\text{logit}(P(x_i)) = z_i \beta + S(x_i)$$

$z_i = (1, z_{i1}, \ldots, z_{ip})$ vector of the intercept and the p covariates

$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^t$ coefficient vector

$S(x_i)$ spatially structured random effect

$S(x_i)$ zero-mean Gaussian process with covariance function

Fitting model-based geostatistical models

- WinBUGs – slow!!
- INLA – faster!
- For geostatistical models in INLA, we will use the 'SPDE' approach
- Spatial Partial Differential Equations
- There is a great tutorial here:
  - http://www.math.ntnu.no/inla/r-inla.org/tutorials/spde/html/
- Can also use geoR and geoRglm (not covered)

WinBUGS modeling of point prevalence data

Can be modeled as binomial or bernoulli
positives[i]~dbin(p[i], examined[i])
positive[i]~dbern(p[i])
logit(p[i])<-alpha + beta + …..

Uses *spatial.exp* argument to model spatial random effects
#spatial effects
w[1:N]~spatial.exp(mu1[],longitude[],latitude[],tau.sp,phi,1)
Need latitude and longitude at each survey location
tau.sp is spatial ‘precision’ (inverse of variance)
phi is modeled estimate of spatial decay
“range” of spatial decay is roughly 3/phi

Model fitting using R-INLA

Triangulation to do the random field discretization – we build a 'mesh' on which to project the random field

```
bound2 <- inla.nonconvex.hull(as.matrix(cross.data.dfloc[,1:2]), convex=0.5, concave=-0.15)
```

```
mesh=inla.mesh.create.helper(as.matrix(cross.data.dfloc[,1:2]),boundary=bound2,offset=c(.1,.2),min.angle=c(15,15),max.edge=c(1,1))
```

Projector matrix to project the process at the mesh nodes to locations

```
A.est =inla.spde.make.A(mesh,loc=coords)
```

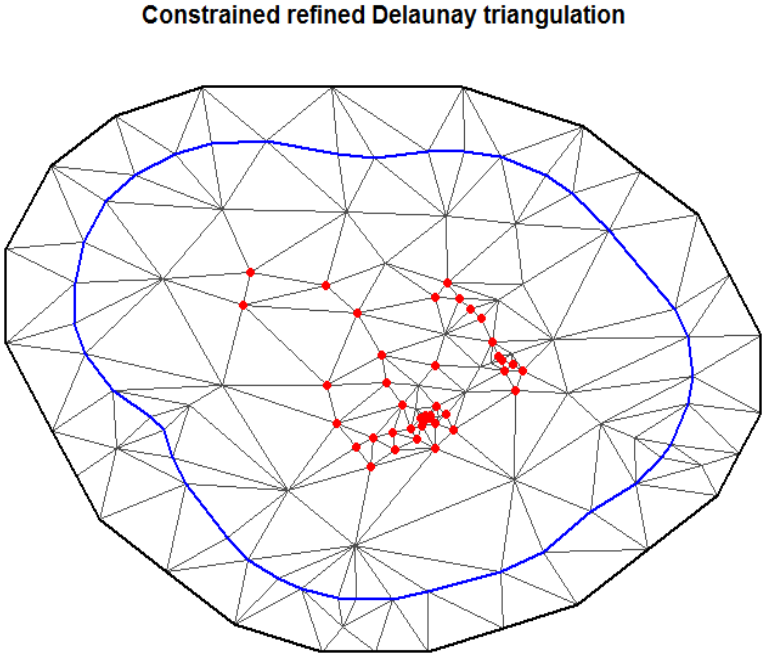
```
A.pred =inla.spde.make.A(mesh,loc=pred_points)
```

Build the SPDE model on the mesh

```
spde<-inla.spde2.matern(mesh,alpha=2)
```

Defining the ‘mesh’

- A ‘good’ mesh:
- has few very sharp angles
- has similar sized triangles
- extends outside the observation area to reduce edge effects



Model fitting using R-INLA

Stack the data

```
st.b<-
inla.stack(data=list(y=cross.data.dfull$positives,n=cross.data.dfull$examined),A=lis
t(A.est,1),
          effects=list(i=1:spde$n.spde,data.frame(b0=1,cross.data.dfull)),tag='est')
```

Call model

```
res.b<-inla(y ~ 0 + (b0) + distriver + evimay +
f(i,model=spde),family="binomial",
control.predictor=list(A=inla.stack.A(st.b)),Ntrials=examined,control.c
ompute=list(dic=T),control.fixed=list(expand.factor.strategy="inla"),da
ta=inla.stack.data(st.b))
```

Data

Malaria prevalence data from Burkina Faso
109 villages sampled
Altitude, rainfall, temperature, normalized difference vegetation index (NDVI) potential predictors

Summary

```
exp(res.b$summary.fix)
res.b$dic$dic
```

Predictions

```
igr<-inla.stack.index(stk.full,'pred')$data
mean<-res.p$summary.fitted.values[igr,"mean"]
low<-res.p$summary.fitted.values[igr,"0.025quant"]
high<-res.p$summary.fitted.values[igr,"0.975quant"]
```

Validation statistics

Mean absolute error (MAE)

Mean of (predicted-actual)

Mean percentage error (MPE)

Mean of (predicted-actual/actual)

Mean absolute percentage error (MAPE)

Mean of (abs(predicted-actual)/actual))

Root mean square error (RMSE)

Square root of mean of squared (predicted-actual)