Model Based Geostatistics

ISAIR 2025

Oliver Brady

Previously

Objectives

Process and model **areal** spatial data

The session covered:

1. Constructing spatial weights matrices
2. Testing the residuals of non spatial models for spatial correlation using Moran's *I*
3. Adjusting models to account for spatial structure using neighbourhood matrices

This session's learning objectives

Objectives

Process and model **point** spatial data

This practical will cover:

1. Identifying spatial autocorrelation in point data using variograms
2. Constructing models to account for autocorrelation using Gaussian processes
3. How to analyse point pattern data to estimate relative risk
4. Make predictive risk maps from both kinds of geostatistical data

What is spatial data?

Spatial data are:

- observed variables
- with spatial coordinates
- whose values may vary according to some unobserved spatial process

Types of spatial data

The main classes of spatial data:

1. **Areal data** – data within a polygon
2. **Point data**
   - Geostatistical data – discrete measurements of a phenomenon usually with a denominator (e.g. malaria prevalence)
   - Point pattern data – events observed at a given point (e.g. a malaria case)

Geostatistics

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Geostatistical data

Data that vary continuously over space, but are measured only at discrete locations

Consist of pairs of data (Y_i, s_i), where:

- Y_i is the value observed at a fixed location s_i.
- s_i = (x_i, y_i) where x and y define a coordinate system, e.g.
    - x longitude and y latitude
    - UTM coordinate reference system

Examples could be vaccine coverage, malaria prevalence, rainfall

Geostatistical data: an example

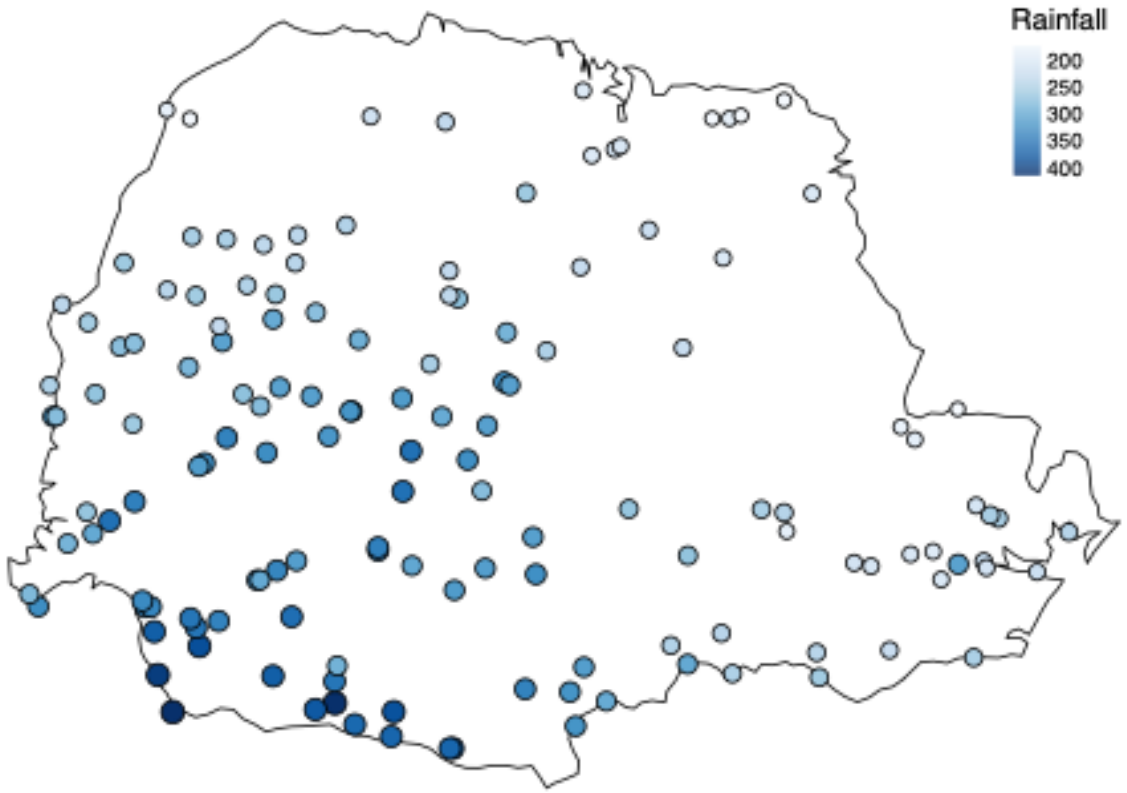


Figure 1: Average rainfall over different years for the period May-June collected at 143 recording stations throughout Parana state, Brasil

Working with geostatistical data

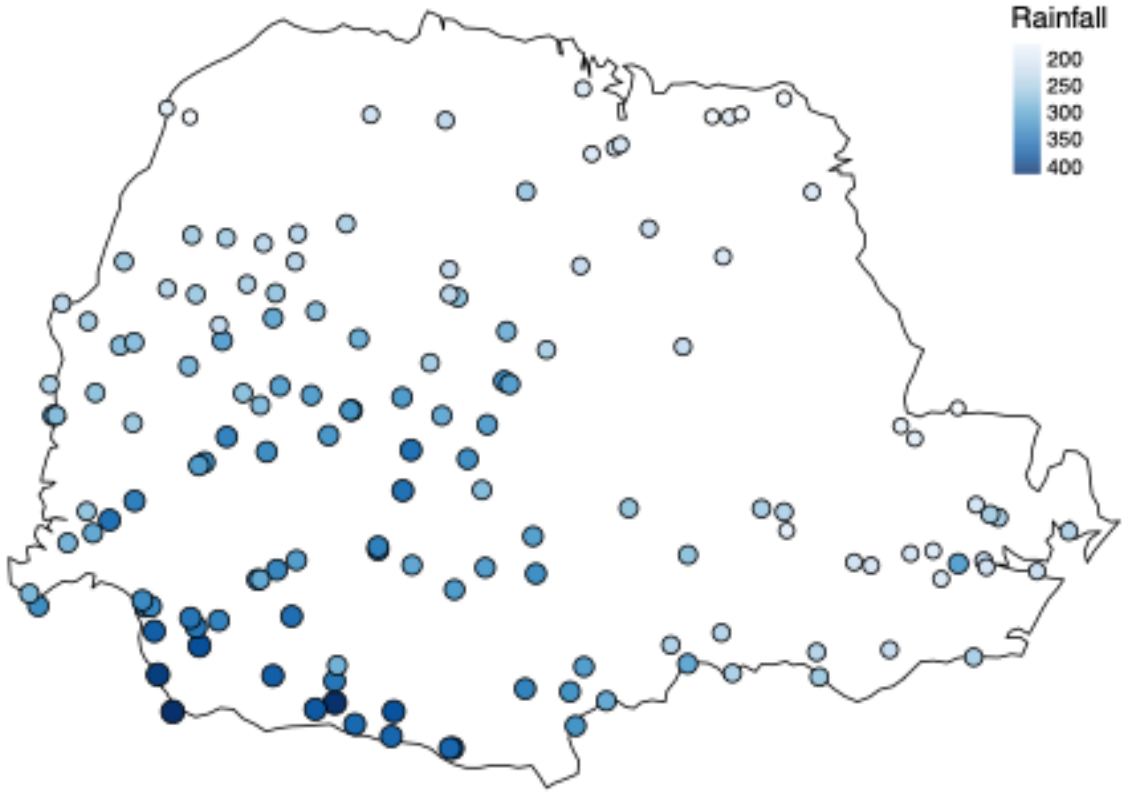
- Aim: describe spatial relationship between pairs of data points in terms of distance between observations

- ! Distance needs to be in metres/km/miles rather than degrees
    - degrees of longitude /≡ degrees of latitude
        - From Charing Cross, moving North 10 takes you 111 km
        - From Charing Cross, moving East 10 takes you 69 km

- Ensure you project to a planar coordinate system, e.g. UTM

Early geostatistics

- Nearer observations more correlated (Tobler 1970)

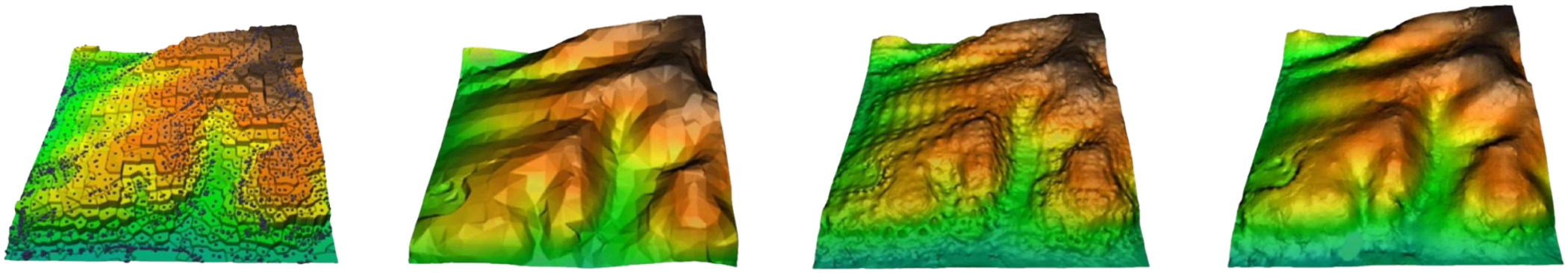


Average rainfall over different years for the period May-June collected at 143 recording stations throughout Parana state, Brasil

Early geostatistics

- Nearer observations more correlated (Tobler 1970)
  - Covariance between observations is a function of distance

- Focus was on different methods of kriging (Matheron 1963)
  - Origins in geology



Polygons | Linear interpolation | Inverse distance weighted | Kriging

Semivariance

- Describes covariance as a function of distance

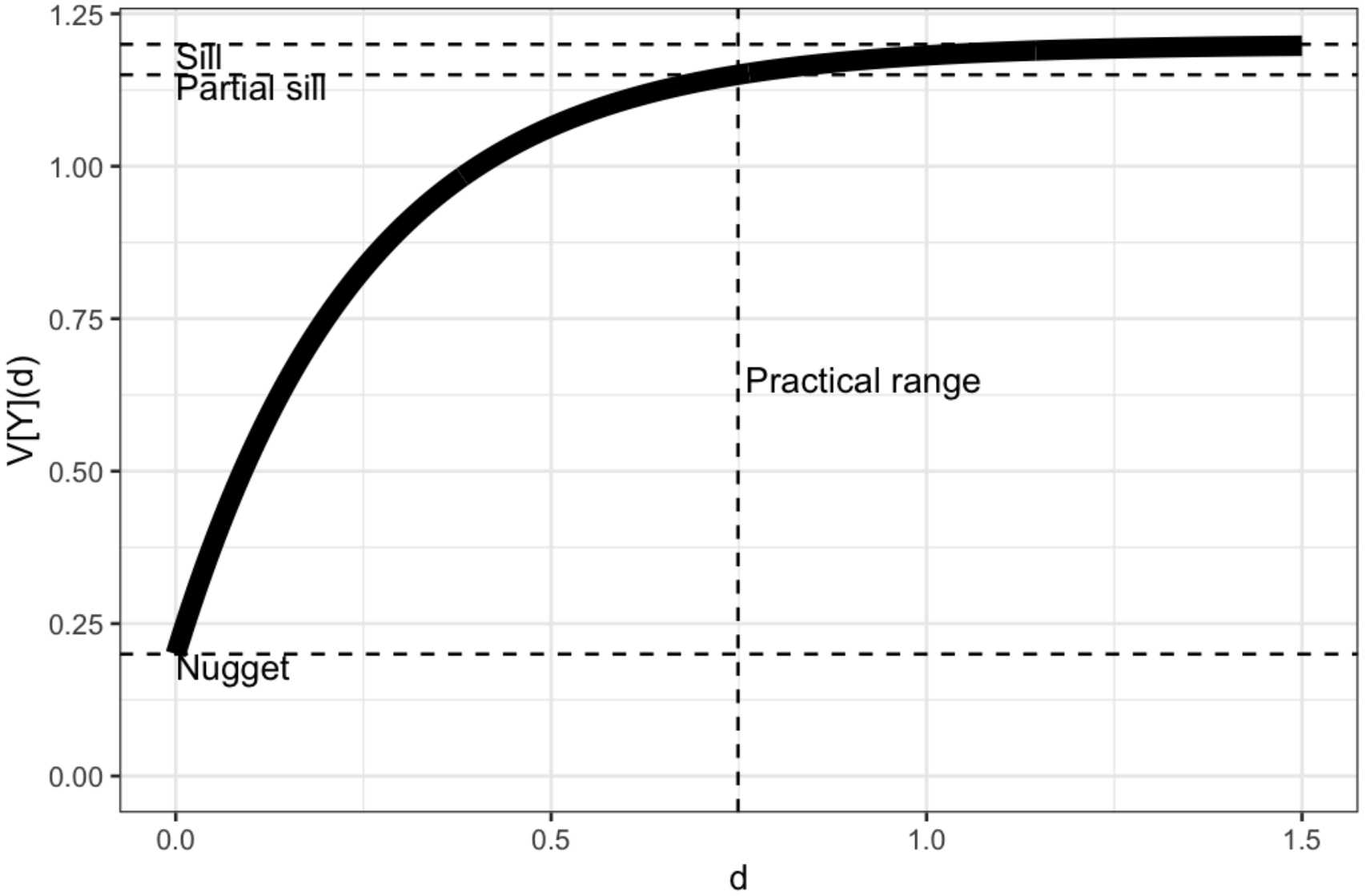
- Variance of spatial process, Y , given by

$\mathbb{V}[Y](d) = \tau^2 + \sigma^2 (1 - \rho(d))$

Described with parameters:

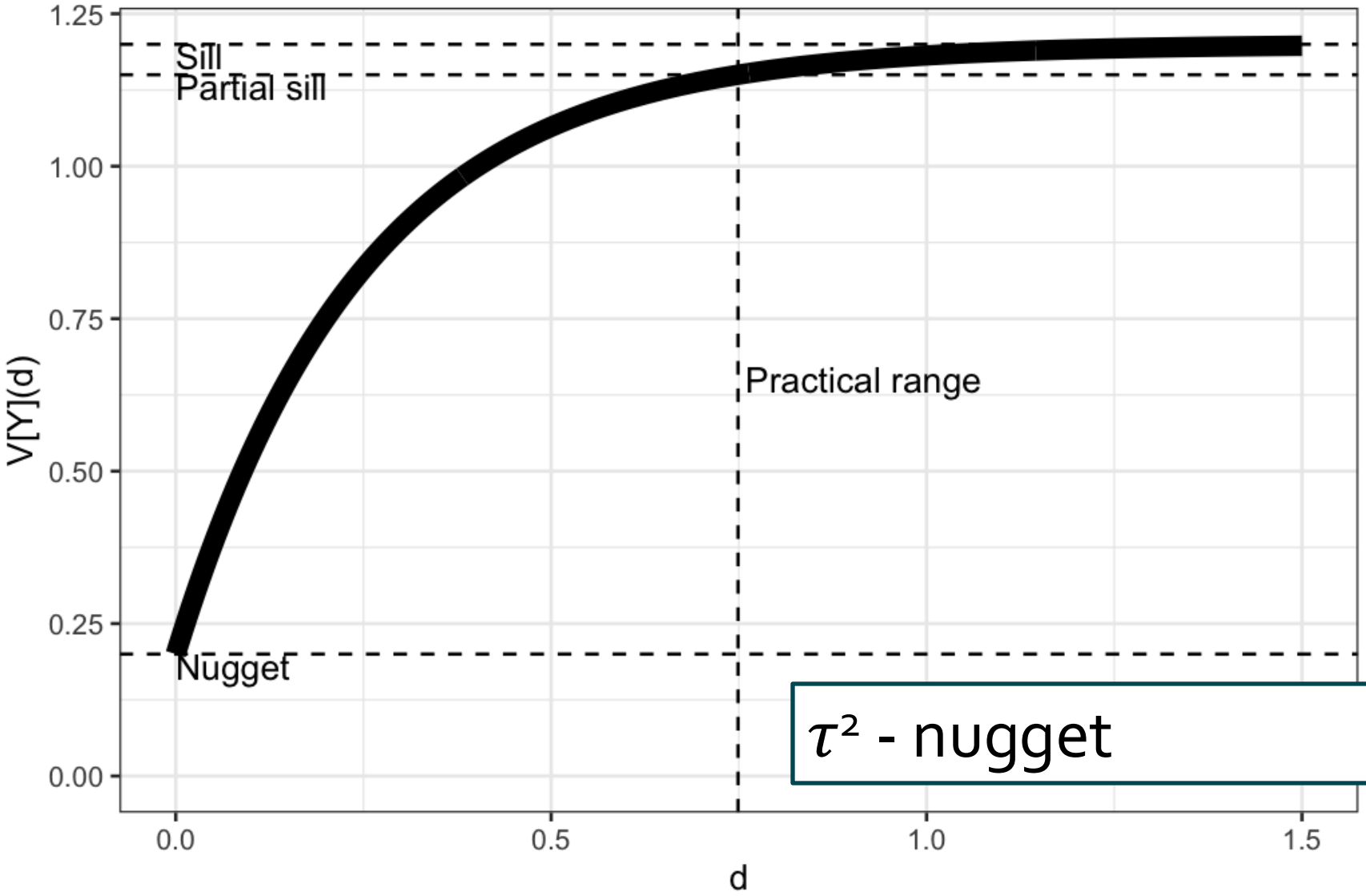
- σ^2 - variance of spatial process
- τ^2 - nugget, measurement error (when $d = 0$)
- $\rho(d)$ - correlation function

The theoretical variogram



The theoretical variogram

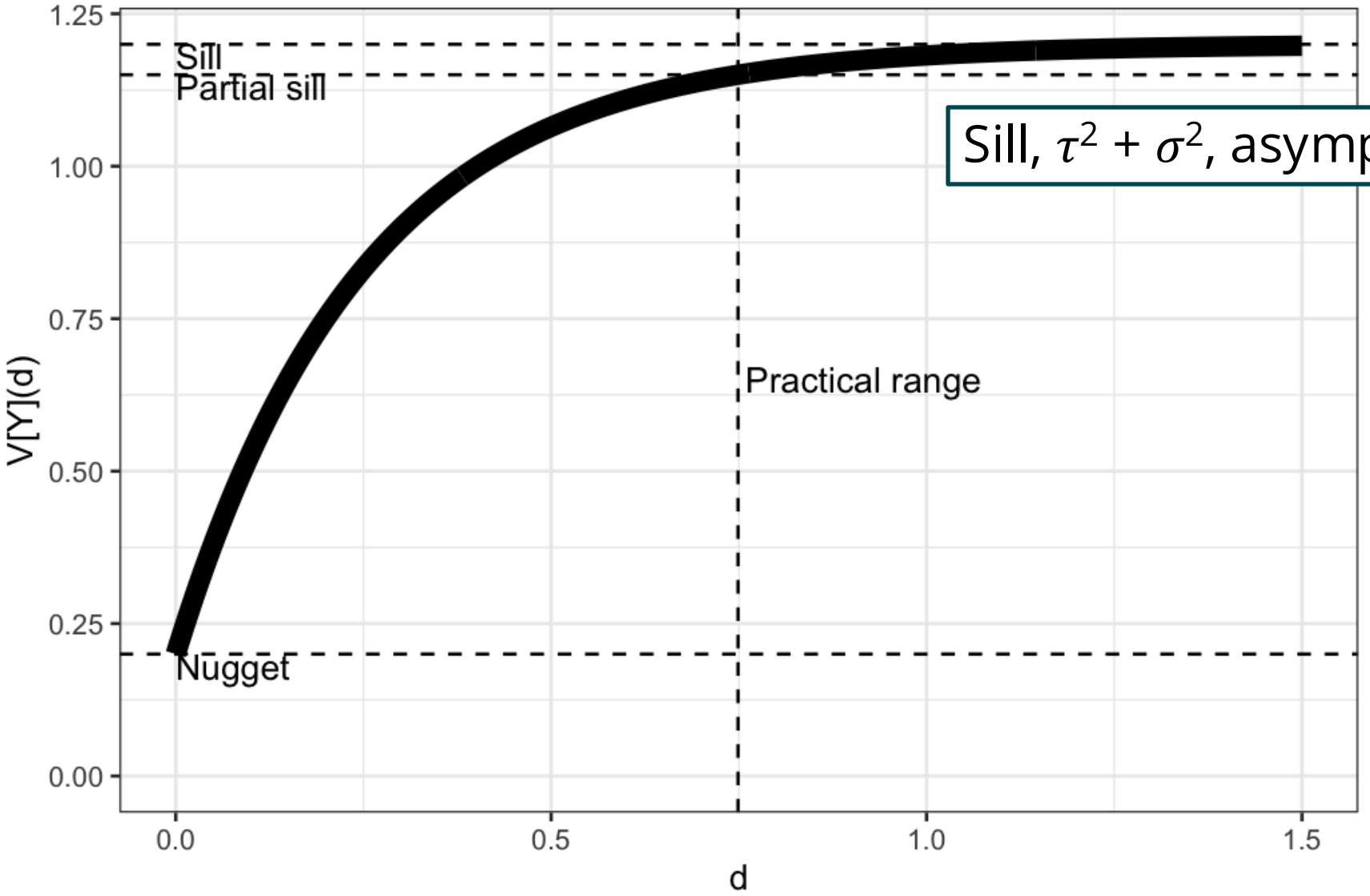
$$\mathbb{V}[Y](d) = \tau^2 + \sigma^2(1 - \rho(d))$$



τ^2 - nugget

The theoretical variogram

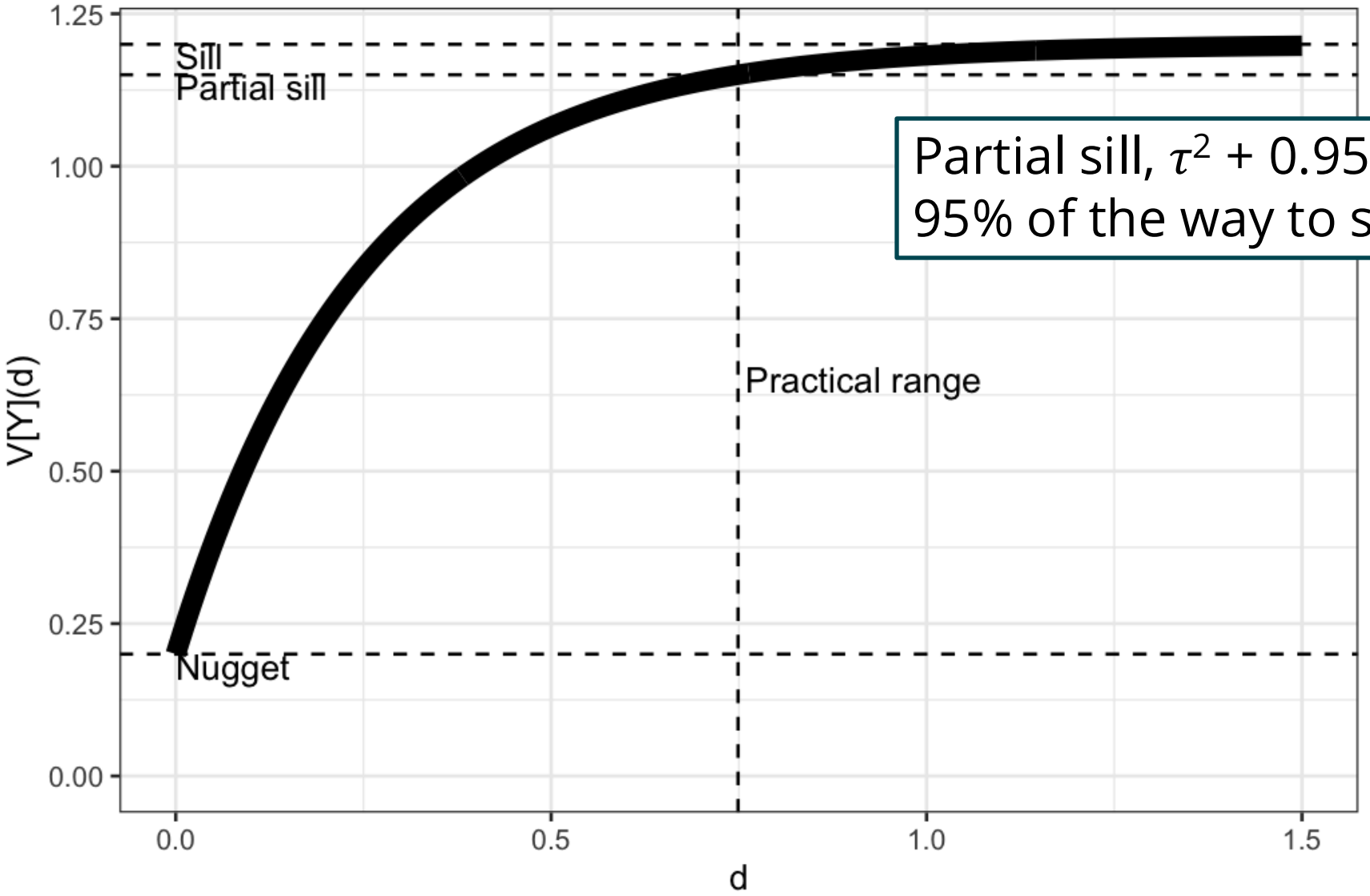
$$\mathbb{V}[Y](d) = \tau^2 + \sigma^2(1 - \rho(d))$$



Sill, $\tau^2 + \sigma^2$, asymptotic limit of $\mathbb{V}[Y](d)$

The theoretical variogram

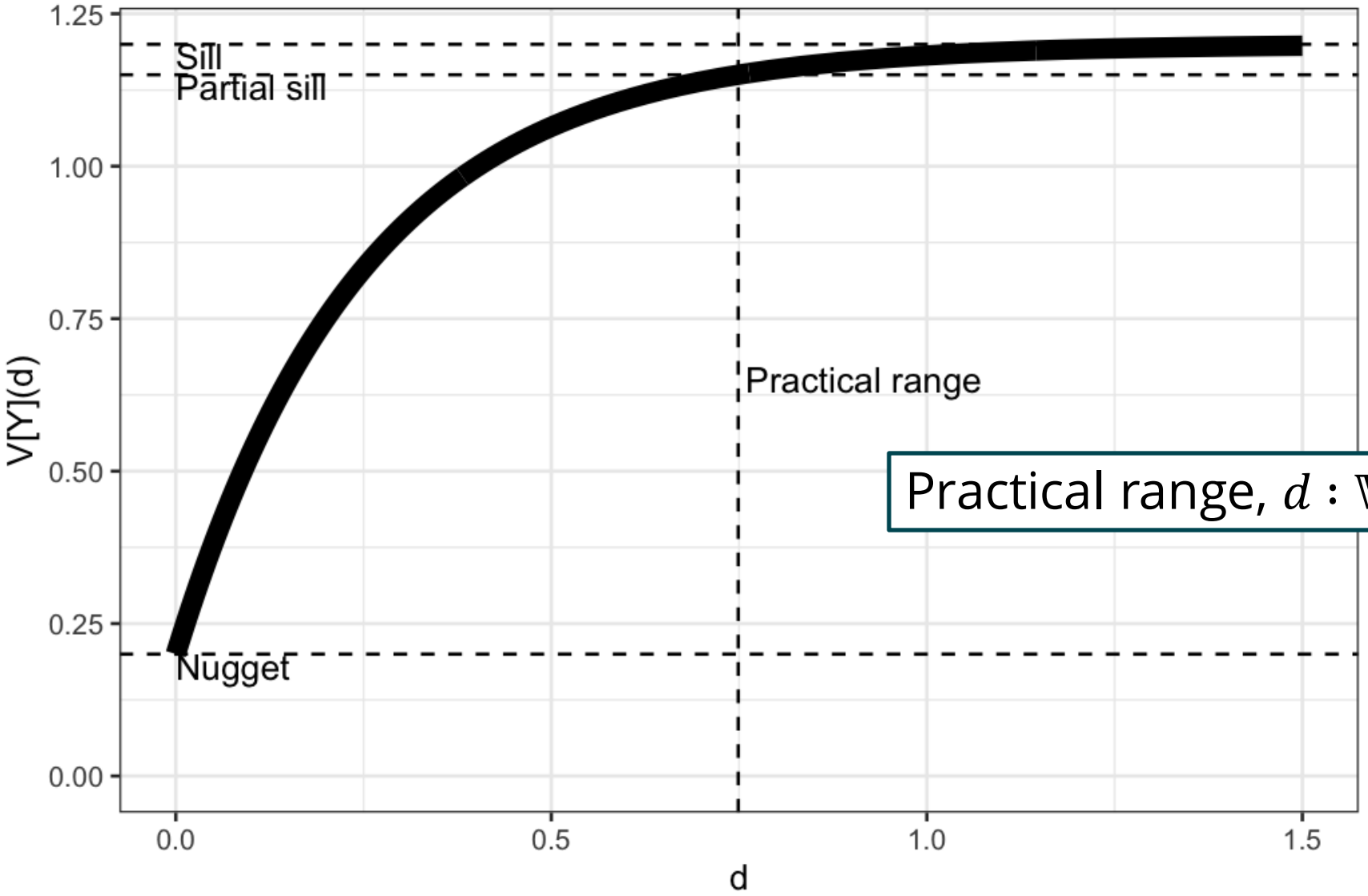
$$\mathbb{V}[Y](d) = \tau^2 + \sigma^2(1 - \rho(d))$$



Partial sill, $\tau^2 + 0.95\sigma^2$
95% of the way to sill from

The theoretical variogram

$$\mathbb{V}[Y](d) = \tau^2 + \sigma^2(1 - \rho(d))$$



Practical range, d : $\mathbb{V}[Y](d) = \tau^2 + 0.95\sigma^2$

More complex covariance functions

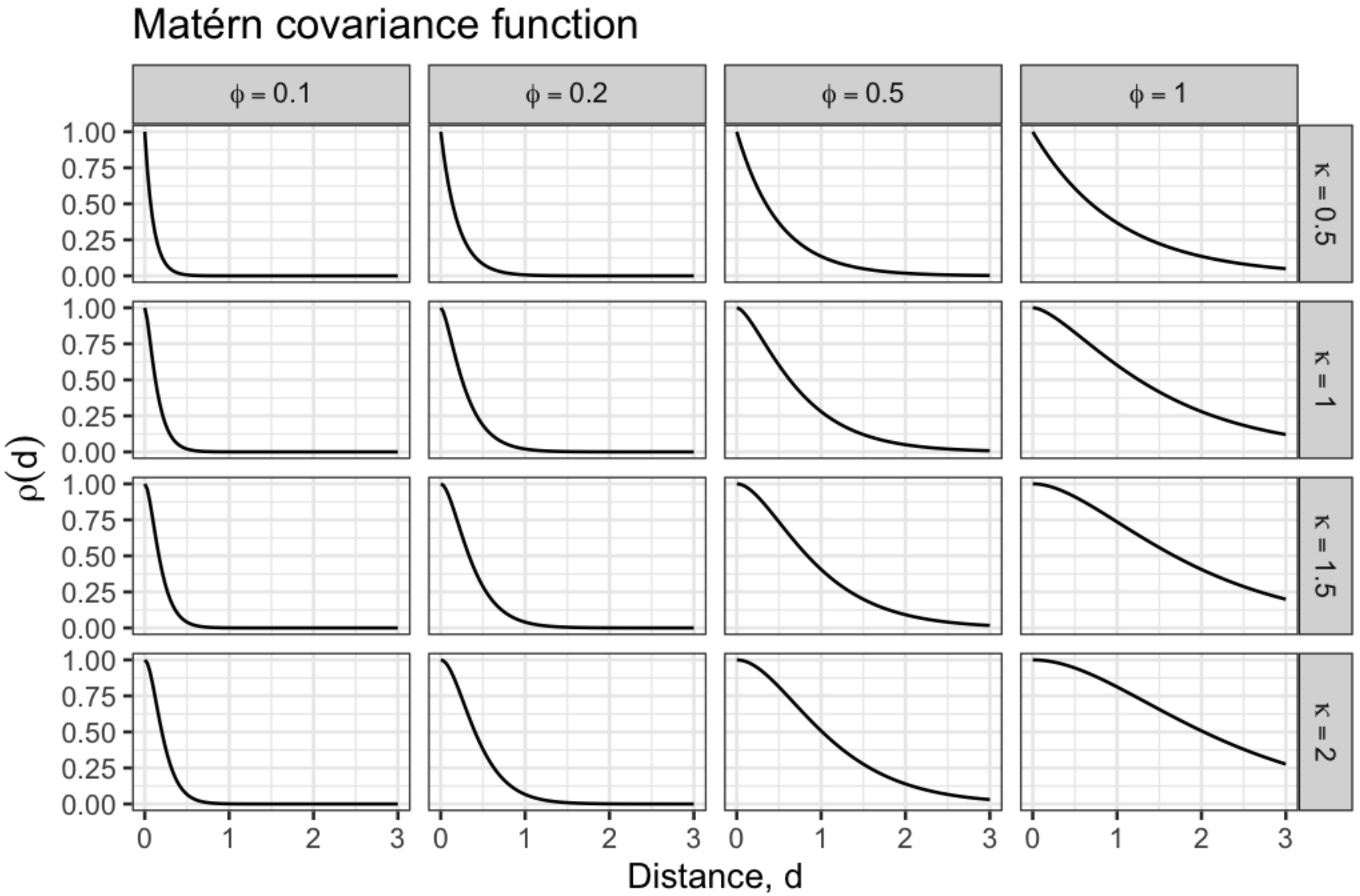
- The Matérn covariance function captures a wide variety of covariance behaviour and can be used to estimate more complex variograms

$$\rho(d) = \frac{2^{1-\kappa}}{\Gamma(\nu)} \left(\frac{d}{\phi} \right)^{\kappa} K_{\kappa} \left(\frac{d}{\phi} \right)$$

where: - ϕ - the range of the spatial correlation - κ - the differentiability (smoothness) - K_{κ} the modified Bessel function of the second kind of order κ

- Many classic covariance functions can be captured with careful choice of K and ϕ

More complex covariance functions



Strengths and limitations of the variogram approach

Advantages:

- Easily interpretable
- Empirical variograms a good first step for spatial data exploration

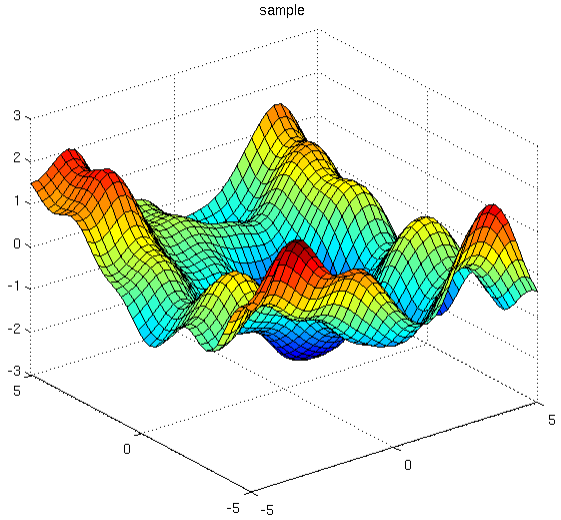
Disadvantages:

- Assumes stationarity
- Can’t include covariates
- Can be noisy for small datasets

Gaussian Processes

- The Gaussian process (O'Hagan 1978) aims to overcome these limitations
- GPs are a stochastic process which generalises the Markov Random Field to continuous space (or time)
    - each random variable in the continuous field has a Gaussian distribution
    - GP gives the distribution over these functions
    - Can build up a complex joint multivariate normal distribution

- Basis of much machine learning (Seeger 2004;
- Rasmussen and Williams 2005) and spatial inference



- Defined by a mean function and a covariance function

Gaussian Processes

$$g\left(\mathbb{E}\left(y\right)\right) = X\beta + u(s)$$

- $X\beta$ is a spatial trend from explanatory variables
- $u(s)$ is a spatial Gaussian process with a structure given by its variance-covariance matrix:

$$u \sim N\left(0, \sigma^2 \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,n} \\ \rho_{2,1} & 1 & \rho_{2,3} & \cdots & \rho_{2,n} \\ \rho_{3,1} & \rho_{3,2} & 1 & \cdots & \rho_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \rho_{n,3} & \cdots & 1 \end{bmatrix}\right)$$

- for $\rho_{i,j} = \rho_{j,i} = \rho\left(d\left(s_i, s_j\right)\right)$ with some parameters

Implementing GPs

In R's `mgcv` package, we can specify a Gaussian Processes in a similar format to splines

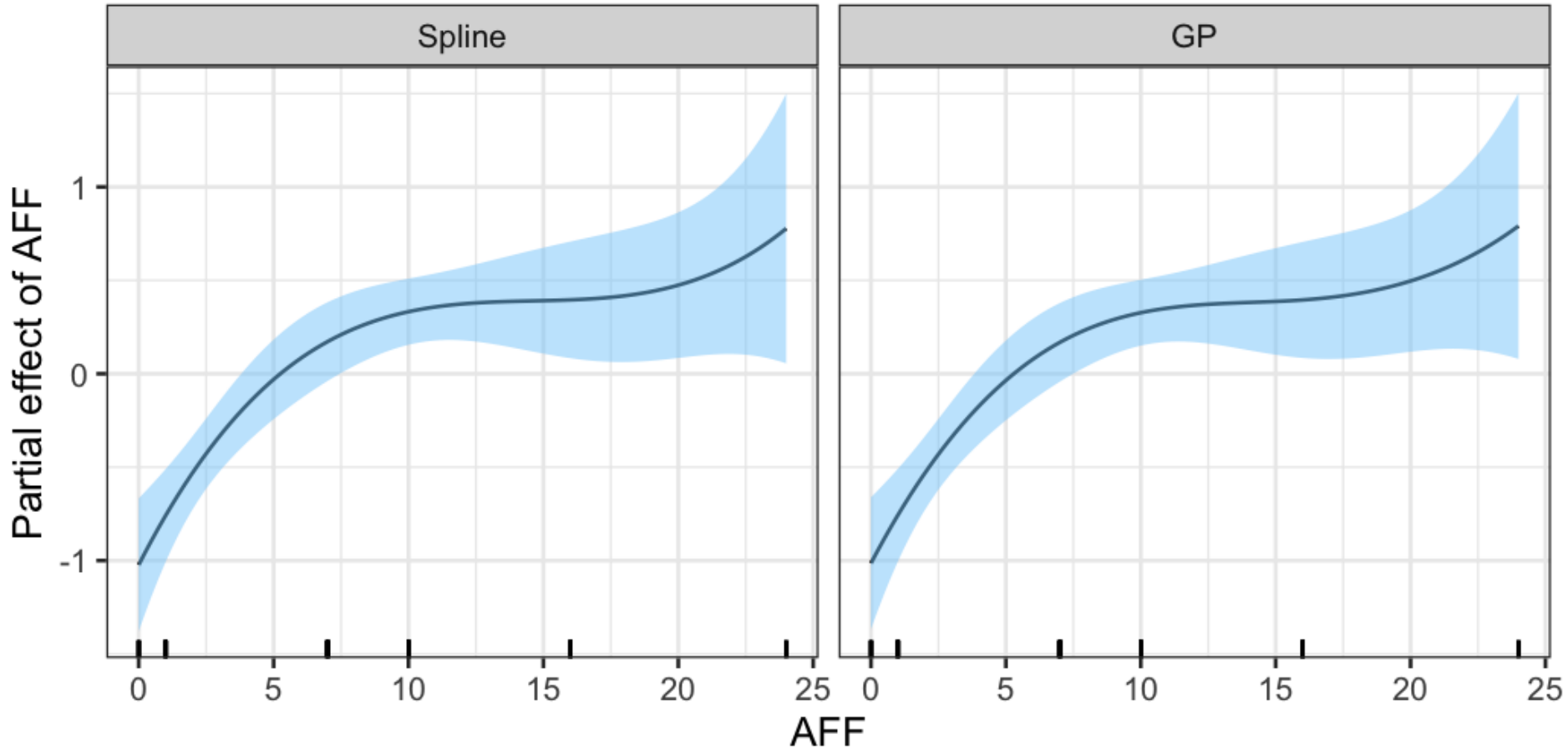
```
... + s(x, bs="gp") + ...
```

Non-spatial GPs

- The GP can be used similar to a smoothing spline
- For example, if we don't know the relationship between AFF and lip cancer and want to fit a non-spatial model, we can fit the models

```
library(mgcv)
ps_model <- gam(data = scotland_lip,
                CANCER ~ offset(log(CEXP)) +
                  s(AFF, k=4, bs='ps'),
                # penalised b-spline
                family = poisson())

gp_model <- gam(data = scotland_lip,
                CANCER ~ offset(log(CEXP)) +
                  s(AFF, k=4, bs='gp'),
                # gaussian process
                family = poisson())
```

Non-spatial GPs



Spline regression a special case of Gaussian processes

Spatial GPs

In R's `mgcv` package, we can specify a spatial Gaussian process smooth:

```
... + s(x, y, bs="gp") + ...
```

- The default behaviour is to use a Matérn covariance function with $\kappa = 3/2$, as recommended by Kammann and Wand (2003)
- We can specify other covariance functions/kernels with

```
?mgcv::smooth.construct.gp.smooth.spec()
```

A Gaussian Process model for *loa loa*

- As an example, consider the data for *loa loa* prevalence
- As prevalence is between 0 and 1, we will use the binomial likelihood,

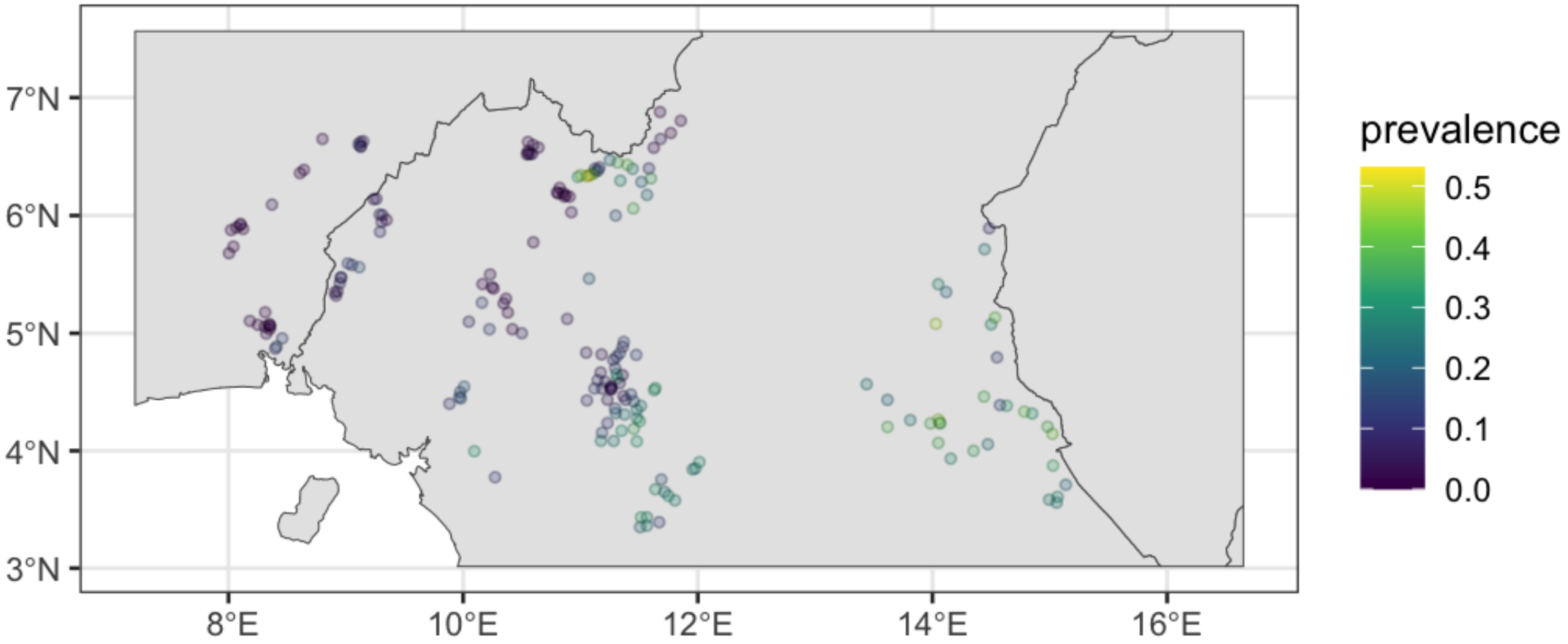
$$f(y_i; p_i, n_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

$$\log\left(\frac{p_i}{1 - p_i}\right) = X_{i,*}\beta + u_i$$

$$u \sim GP(0, \Sigma)$$

where Σ is the variance-covariance matrix described previously and X is the design matrix of predictor variables

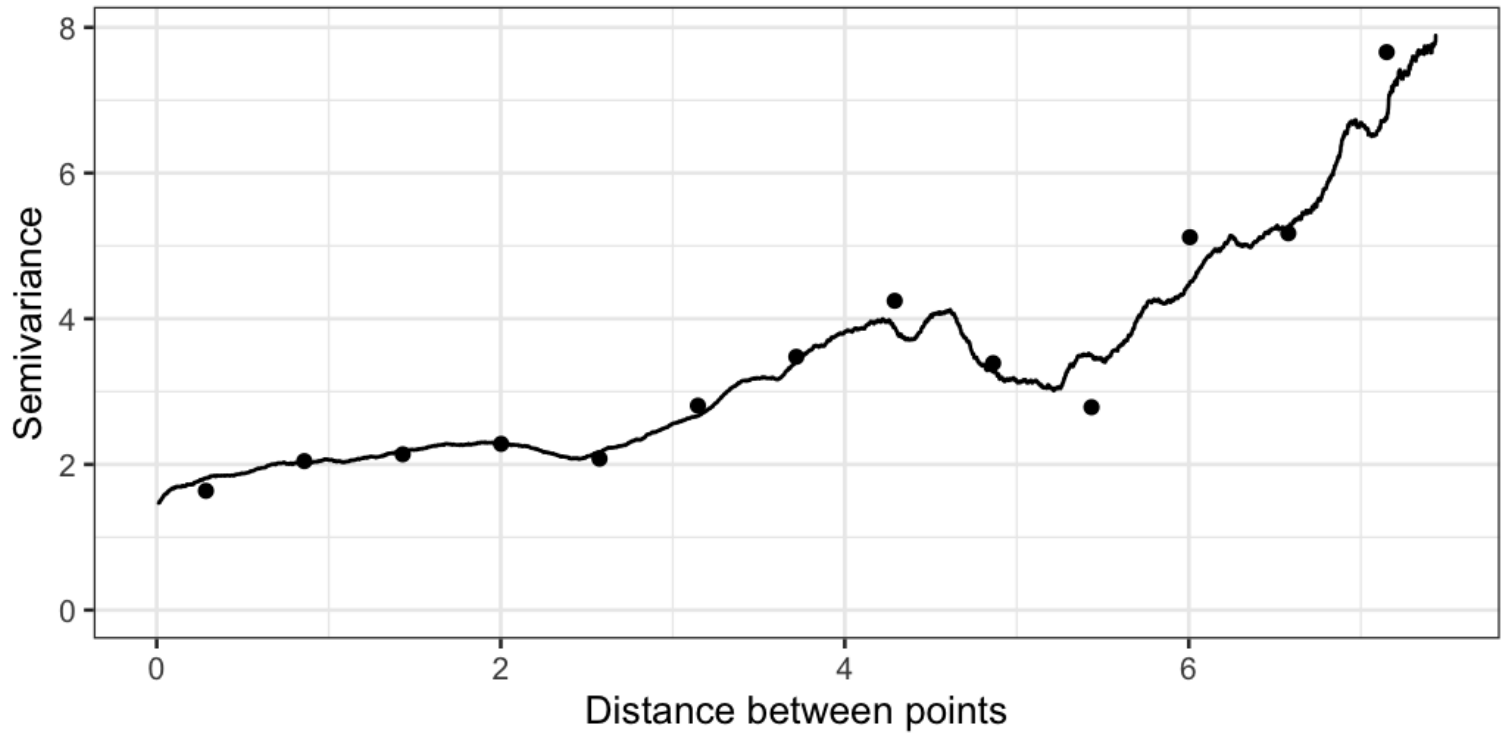
A Gaussian Process model for *loa loa*



Loa loa prevalence at survey locations in Cameroon

A Gaussian Process model for *loa loa*

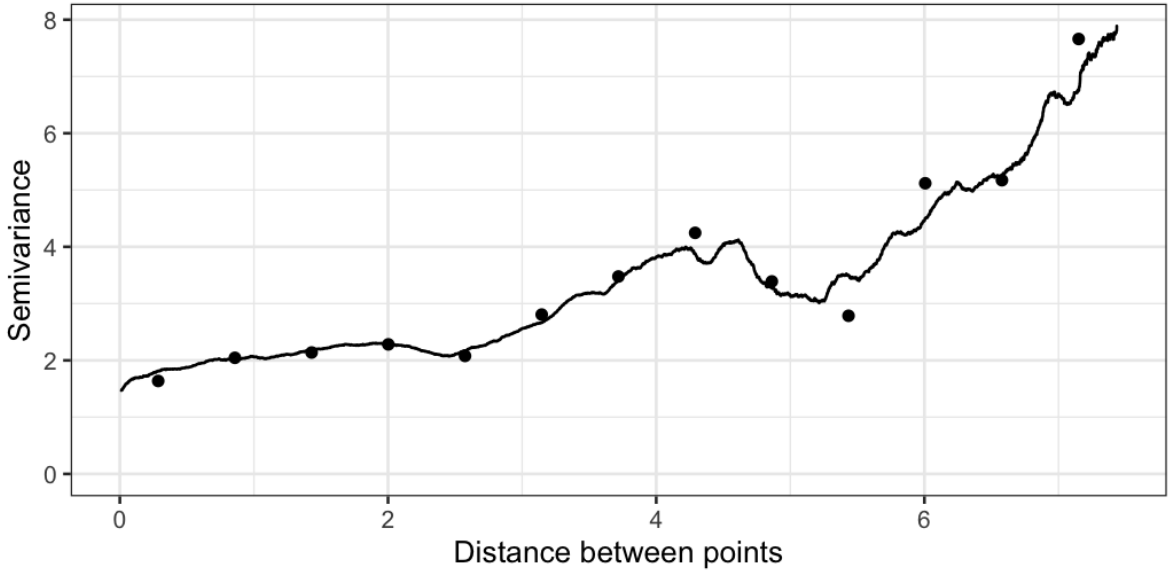
As a first step we want to plot the empirical semi-variogram of the empirical logit to indicate if there may be spatial dependence



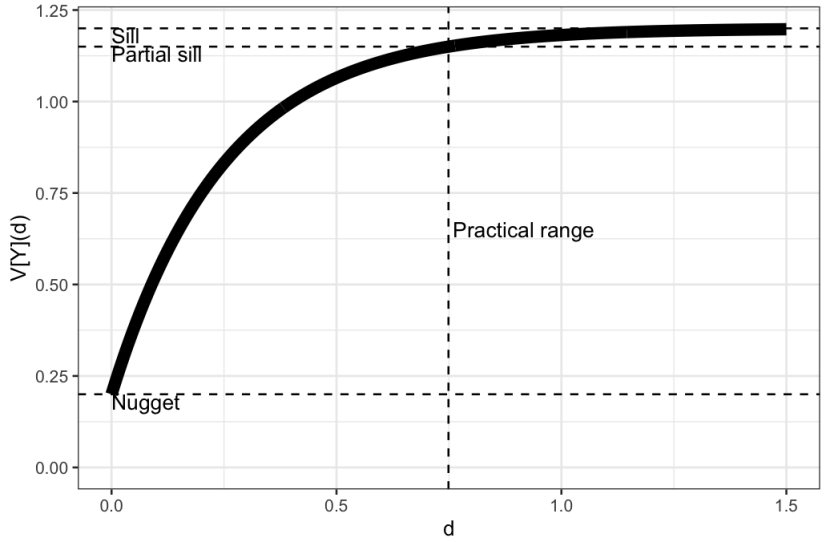
Empirical semivariogram of *loa loa* prevalence

A Gaussian Process model for *loa loa*

As a first step we want to plot the empirical semi-variogram of the empirical logit to indicate if there may be spatial dependence



Empirical semivariogram of *loa loa* prevalence



Theoretical variogram

A Gaussian Process model for *loa loa*

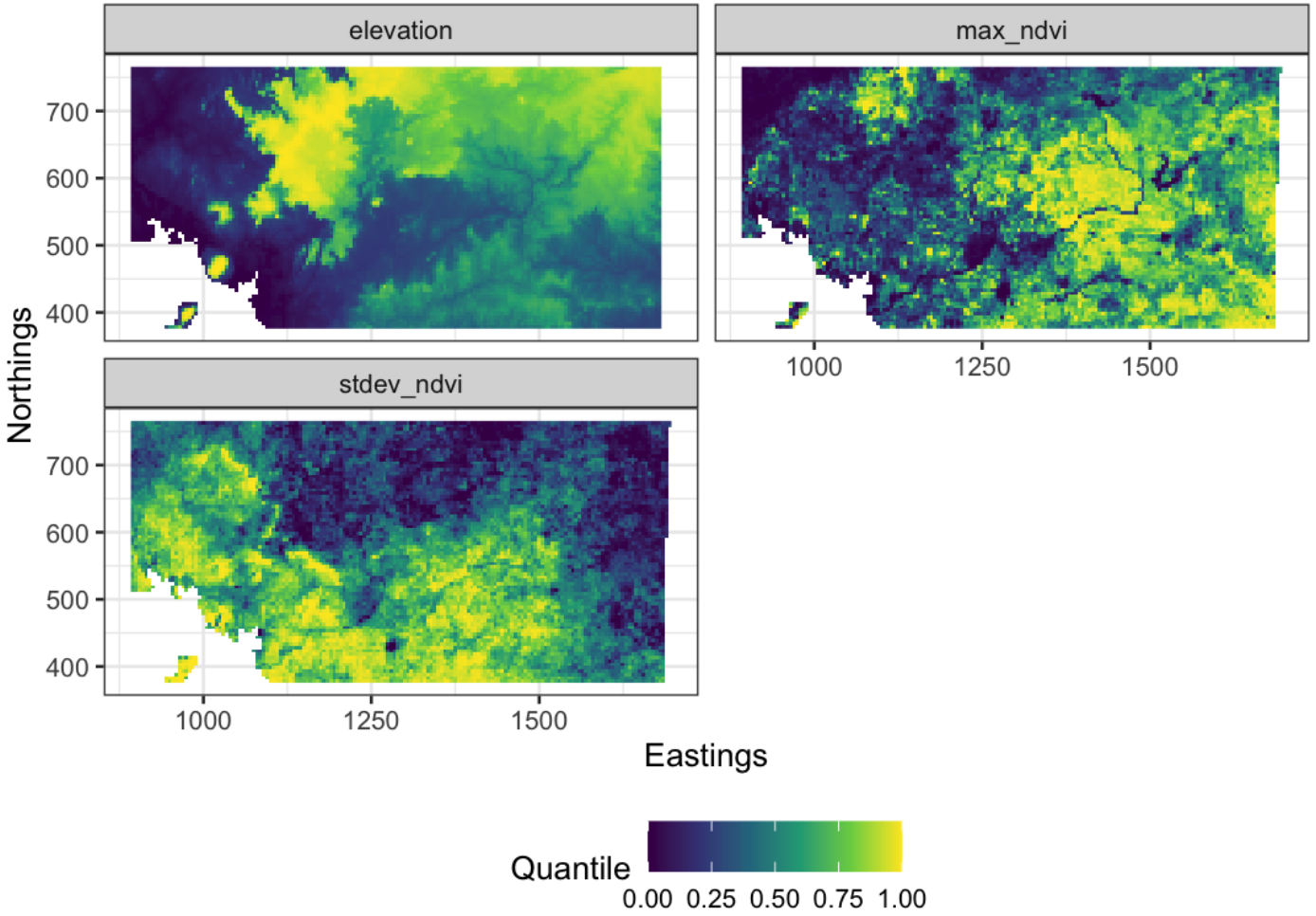
Predictor variables at the pixel level are

- elevation above sea level
- Normalised Difference Vegetation Index (NDVI), a proxy for vegetation cover

Because vegetation cover changes over time we will need to create summary variables of the NDVI timeseries for each pixel:

- Max NDVI – how much vegetation is there at the peak
- Standard deviation of NDVI- how variable in vegetation cover over the timeseries

A Gaussian Process model for *loa loa*



Quantiles of each explanatory variable

A Gaussian Process model for *loa loa*

Fitting in `mgcv`, we transform our simple feature to include the point coordinates as X and Y and use the binomial model

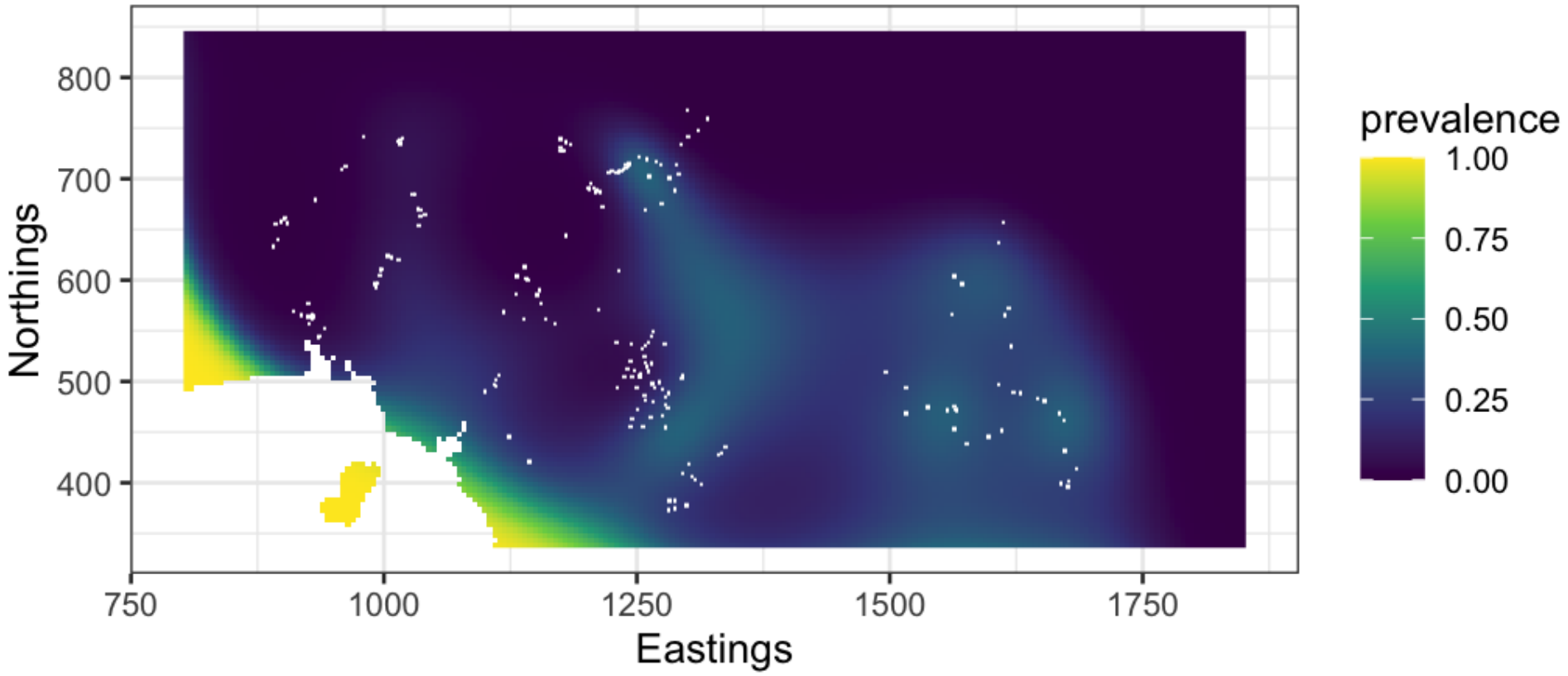
```
loaloa_sp_sf %<>% bind_cols(data.frame(st_coordinates(.)))

loaloa_sp_sf_df <- as.data.frame(loaloa_sp_sf) %>%
  dplyr::select(X, Y, geometry, prevalence, examined)

utils::head(loaloa_sp_sf_df)
```

```
##          X        Y                   geometry  prevalence examined
## 1 895.2158 639.6818 POINT (895.2158 639.6818) 0.000000000      162
## 2 891.0379 633.3642 POINT (891.0379 633.3642) 0.005988024      167
```

- Similar specification to MRF model but no need to pass neighbourhood matrix
- Can specify which covariance function used if we wish

```
loaloa_gp <- gam(data = loaloa_sp_sf_df,
                 prevalence ~ s(X, Y, bs="gp"),
                 family = binomial(),
                 weights = examined)
```

A Gaussian Process model for *loa loa*



Predicted *loa loa* prevalence from a Gaussian Process regression with no covariates

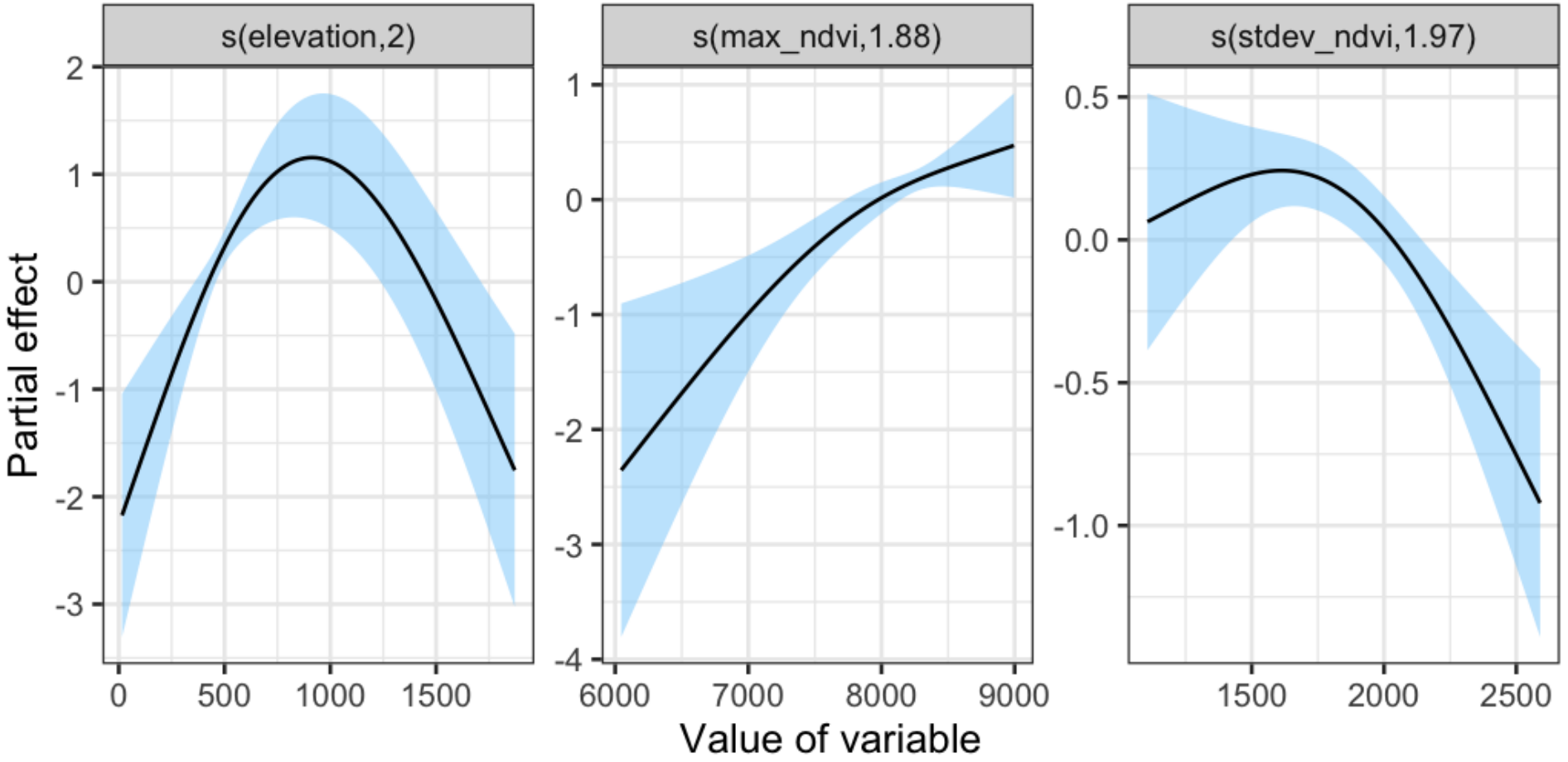
A Gaussian Process model for *loa loa*

Specify the model with smooth functions of predictors

```
loaloa_gp_x <- gam(
  data = loaloa_sp_sf,
  prevalence ~ s(X, Y,
                 bs="gp") +
    s(elevation,  k=3) +
    s(stdev_ndvi, k=3) +
    s(max_ndvi,   k=3) ,
  family = binomial(), weights = examined)
```

A Gaussian Process model for *loa loa*

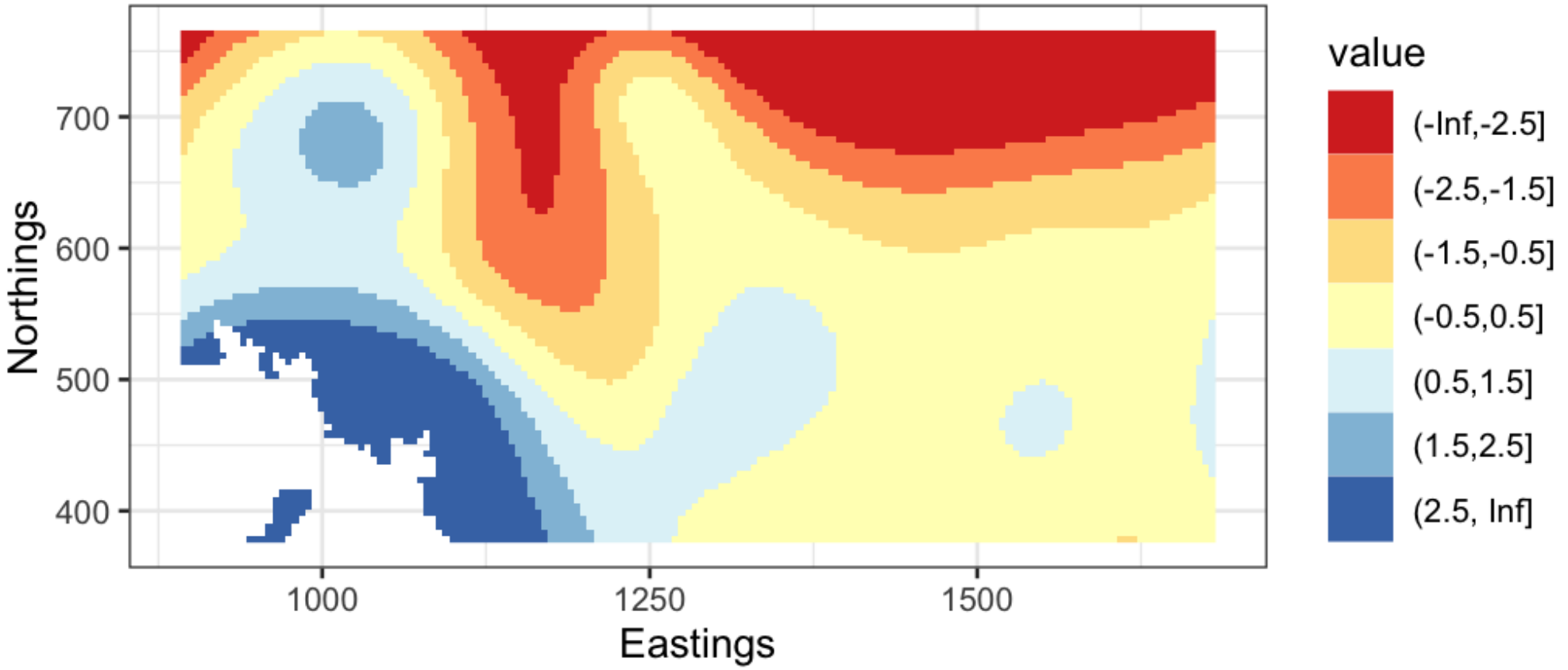
Smooth effects of predictors for model of *loa loa*



Do these relationships seem plausible?

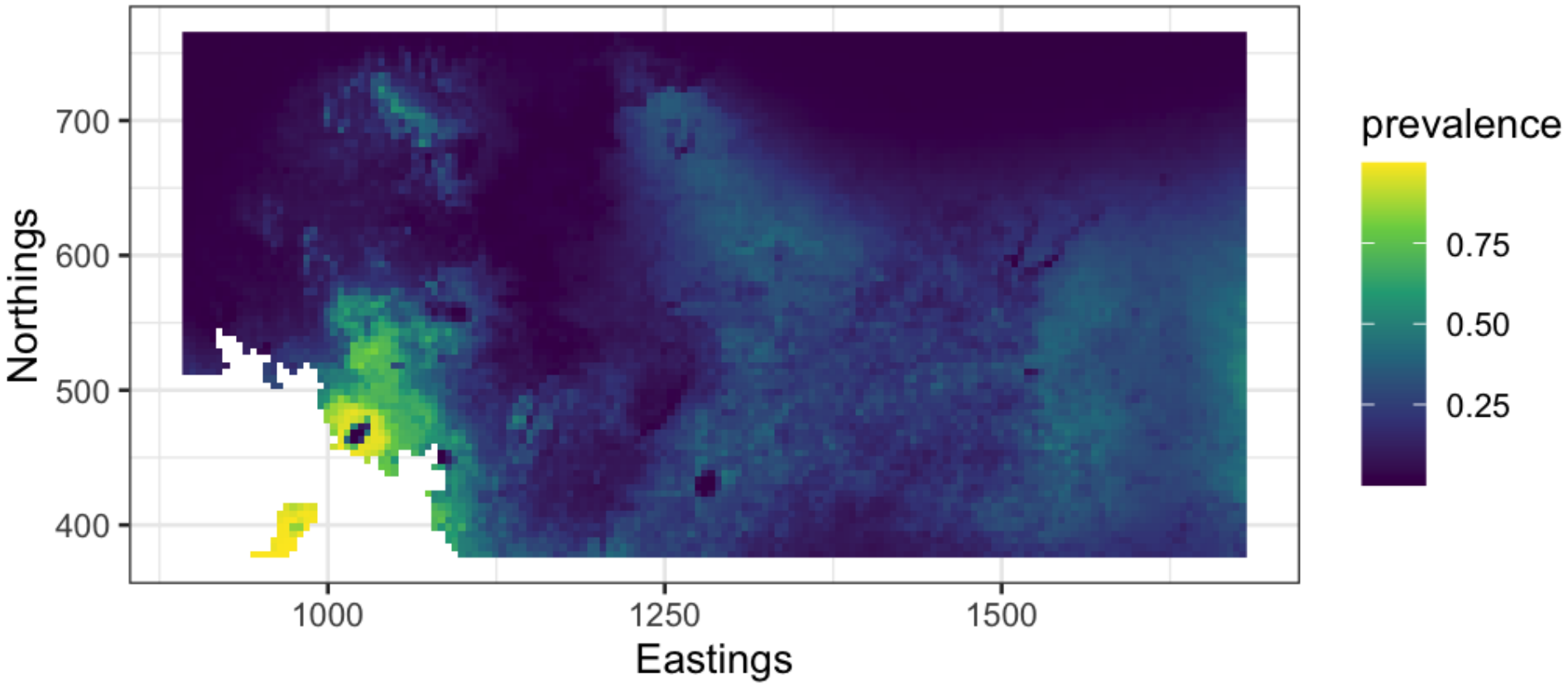
A Gaussian Process model for *loa loa*

Spatial random effect from *loa loa* model with predictors



A Gaussian Process model for *loa loa*

Spatial random effect from *loa loa* model with predictors



A Gaussian Process model for *loa loa*

We have seen that higher *loa loa* risk is associated with:

- A roughly quadratic effect of elevation with risk peaking around 800m
- Less variable NDVI
- Higher max NDVI

That the addition of a spatial random effect to the model accounts for smooth variability not due to predictors

A Gaussian Process model for *loa loa*

We can assess goodness of fit with the AIC as before

```
library(broom)
glance(loaloa_gp)
```

```
## # A tibble: 1 x 7
##      df logLik   AIC   BIC deviance df.residual  nobs
##   <dbl>  <dbl> <dbl> <dbl>    <dbl>       <dbl> <int>
## 1  32.0  -896. 1856. 1961.    1044.        165.   197
```

```
glance(loaloa_gp_x)
```

```
## # A tibble: 1 x 7
##      df logLik   AIC   BIC deviance df.residual  nobs
##   <dbl>  <dbl> <dbl> <dbl>    <dbl>       <dbl> <int>
## 1  36.7  -765. 1604. 1723.     799.        155.   192
```

Including these explanatory variables has reduced AIC

Loa loa prevalence mapping practical

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Point patterns

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

This session's learning objectives

Objectives

Process and model **point** spatial data

This practical will cover:

1. Identifying spatial autocorrelation in point data using variograms
2. Constructing models to account for autocorrelation using Gaussian processes
3. How to analyse point pattern data to estimate relative risk
4. Make predictive risk maps from both kinds of geostatistical data

Types of spatial data

The main classes of spatial data:

1. **Areal data** – data within a polygon
2. **Point data**
   - Geostatistical data – discrete measurements of a phenomenon usually with a denominator (e.g. malaria prevalence)
   - Point pattern data – events observed at a given point (e.g. a malaria case)

Point patterns

- Realisations of some spatial process that generates observations of a phenomena
    - Wish to understanding an intensity function, $\lambda(s_i)$, describing how many events per unit area are observed at s_i

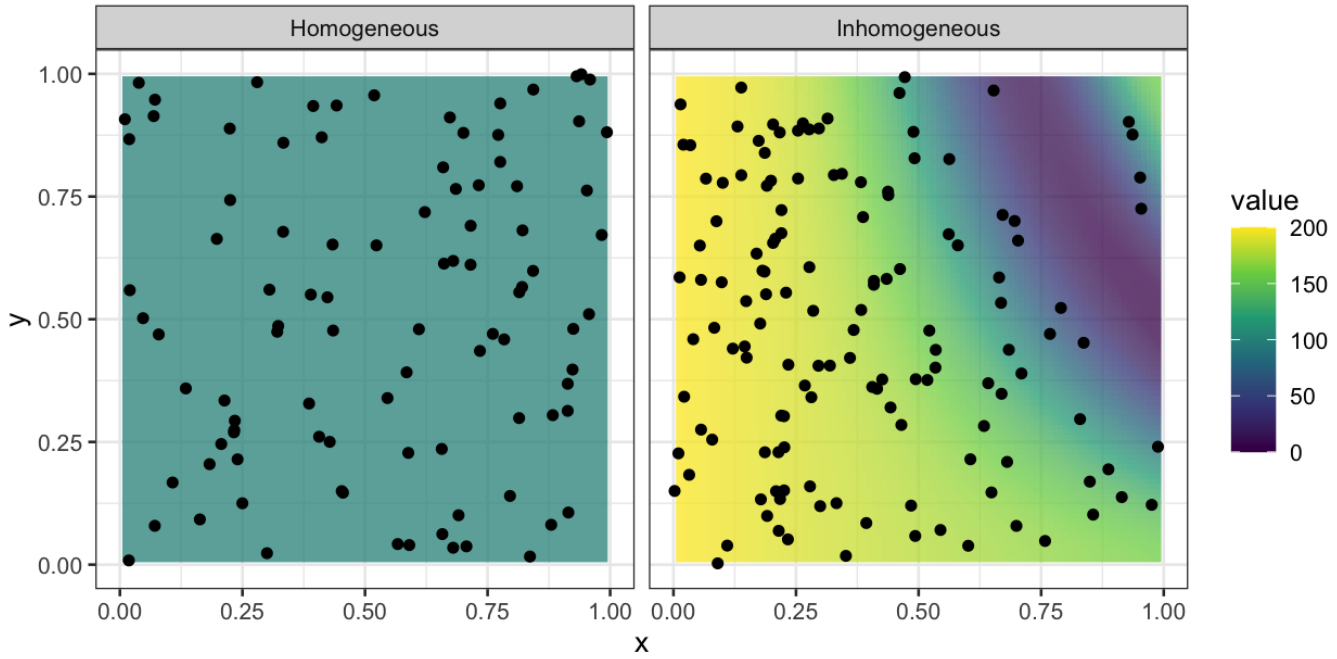
- Need to relate observed spatial locations to predictor variables and latent spatial field

- Spatstat package in R (Adrian Baddeley and Turner 2005; Adrian Baddeley, Rubak, and Turner 2015)

Point patterns

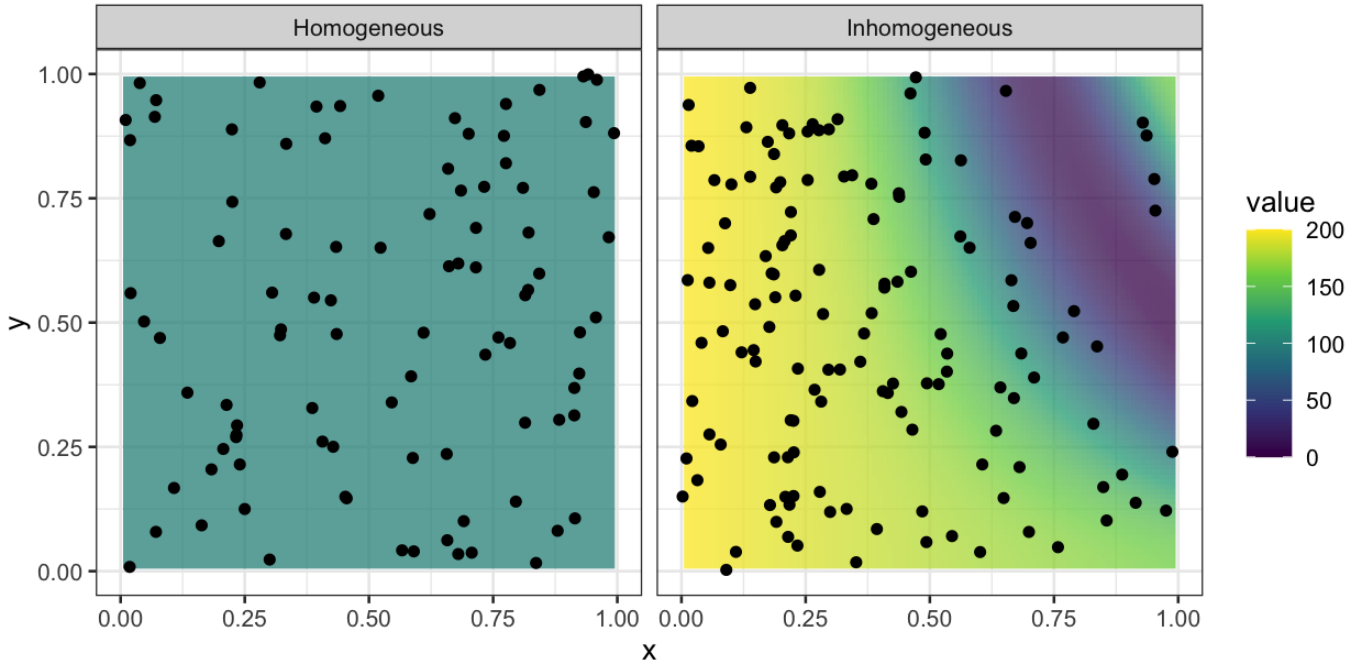
Some simulated point data for the functions:

- Homogeneous: $\lambda = 100$
- Inhomogeneous: $\lambda = 40\,(3 - x^2 + 2\cos(2\pi x^2 y))$



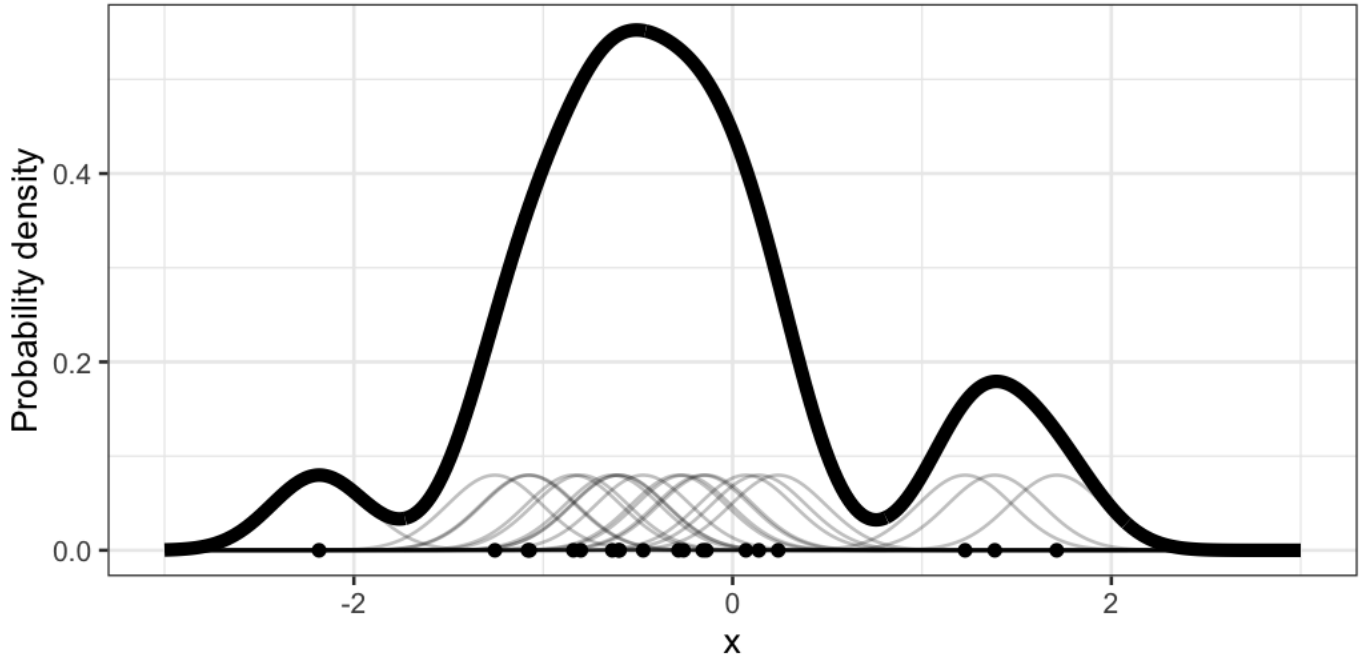
Point patterns - Smoothing

- Aim: to predict where we expect new points to be generated in future



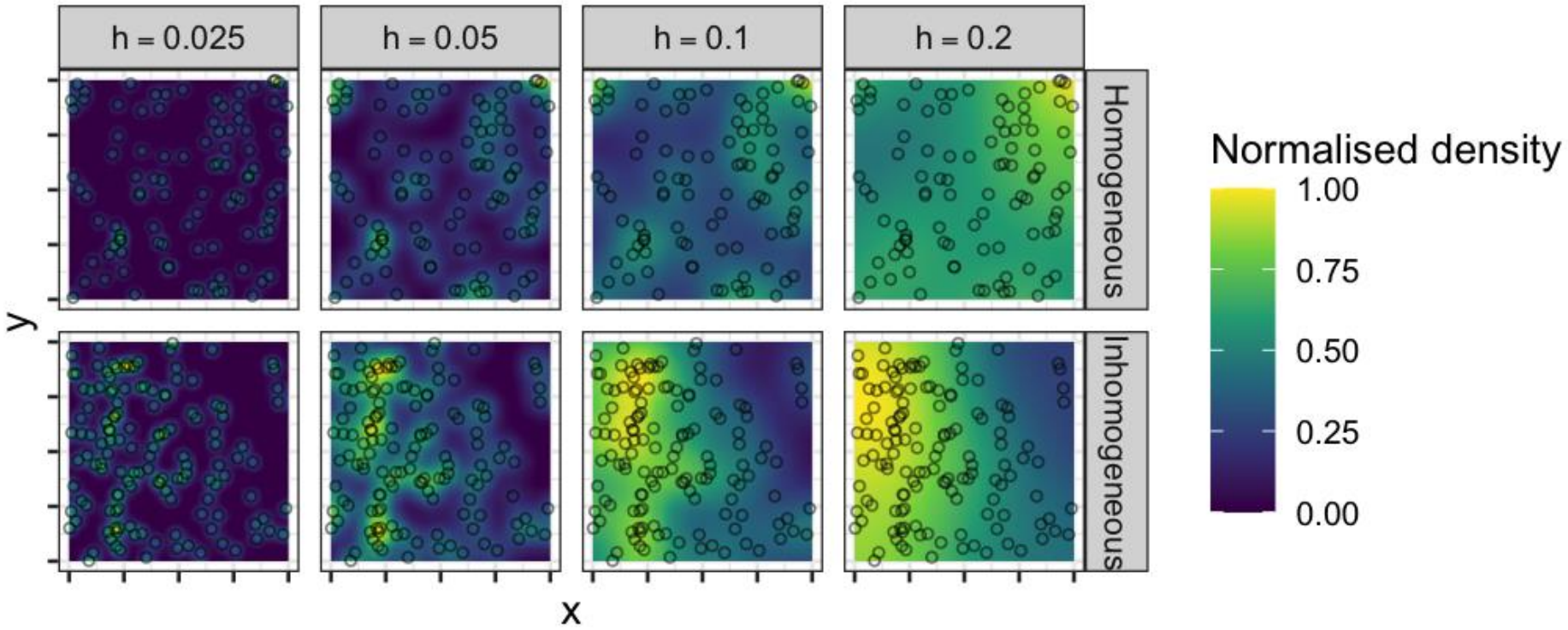
Point patterns - Smoothing

- Approach: Kernel-based smoothing replaces points with symmetric functions centred on the point, most commonly the Gaussian kernel



- Cross-validation techniques to choose optimal kernel

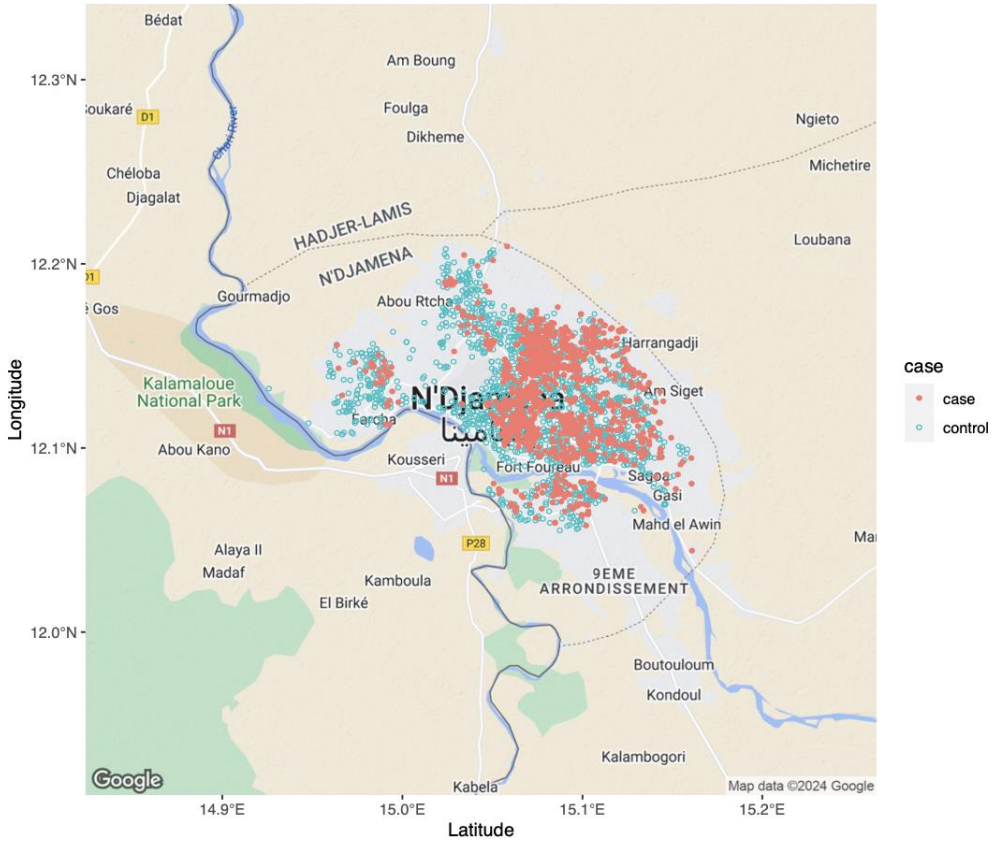
Point patterns - Smoothing



Smoothed densities of point patterns generated by homogeneous and inhomogeneous Poisson processes normalised by maximum density in each plot

Case study: Cholera in Chad

- Finger et al. (2018) consider a Cholera outbreak in Chad
- Cases cluster in certain areas, but so does the population and ”controls” (individuals seeking care at the same health centres as “cases” but with other diseases
- We want to know about the relative risk across N’Djamena



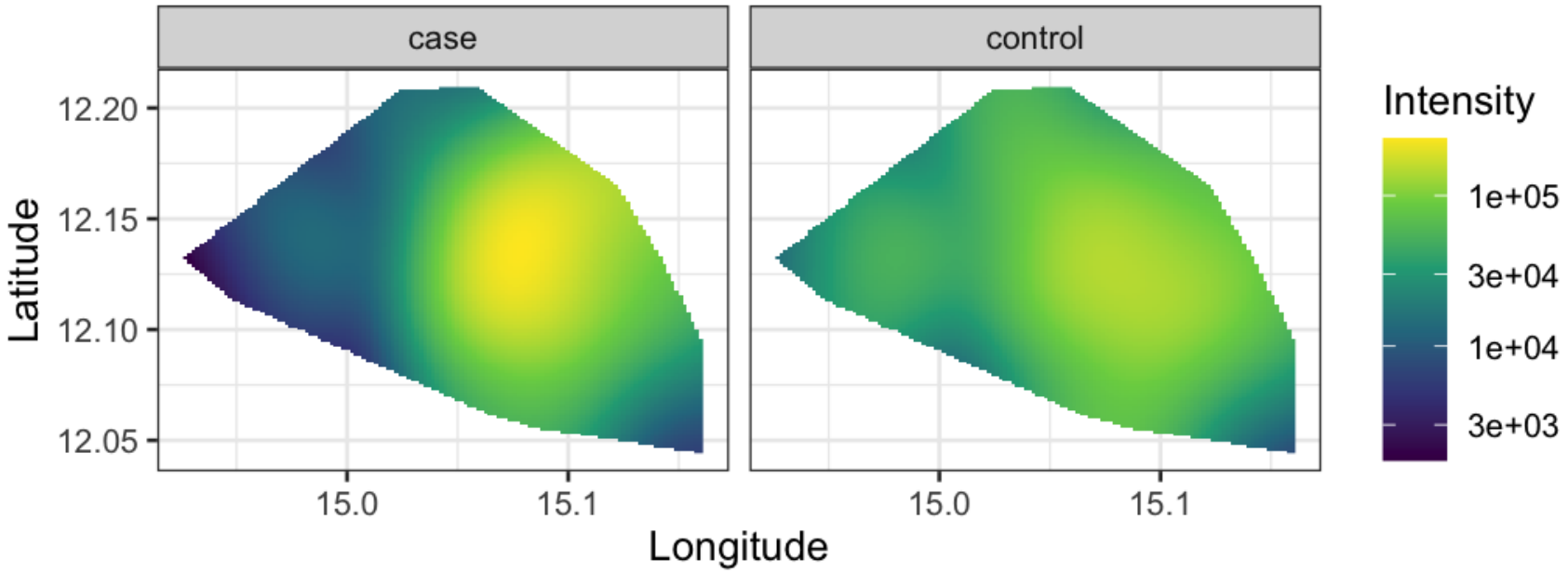
Cholera in Chad

- To calculate the relative risk of infection, we need a spatial estimate of the case density and control density, then relative risk is given as

$$RR(s_i) = \frac{\lambda_{\text{case}}(s_i)}{\lambda_{\text{control}}(s_i)}$$

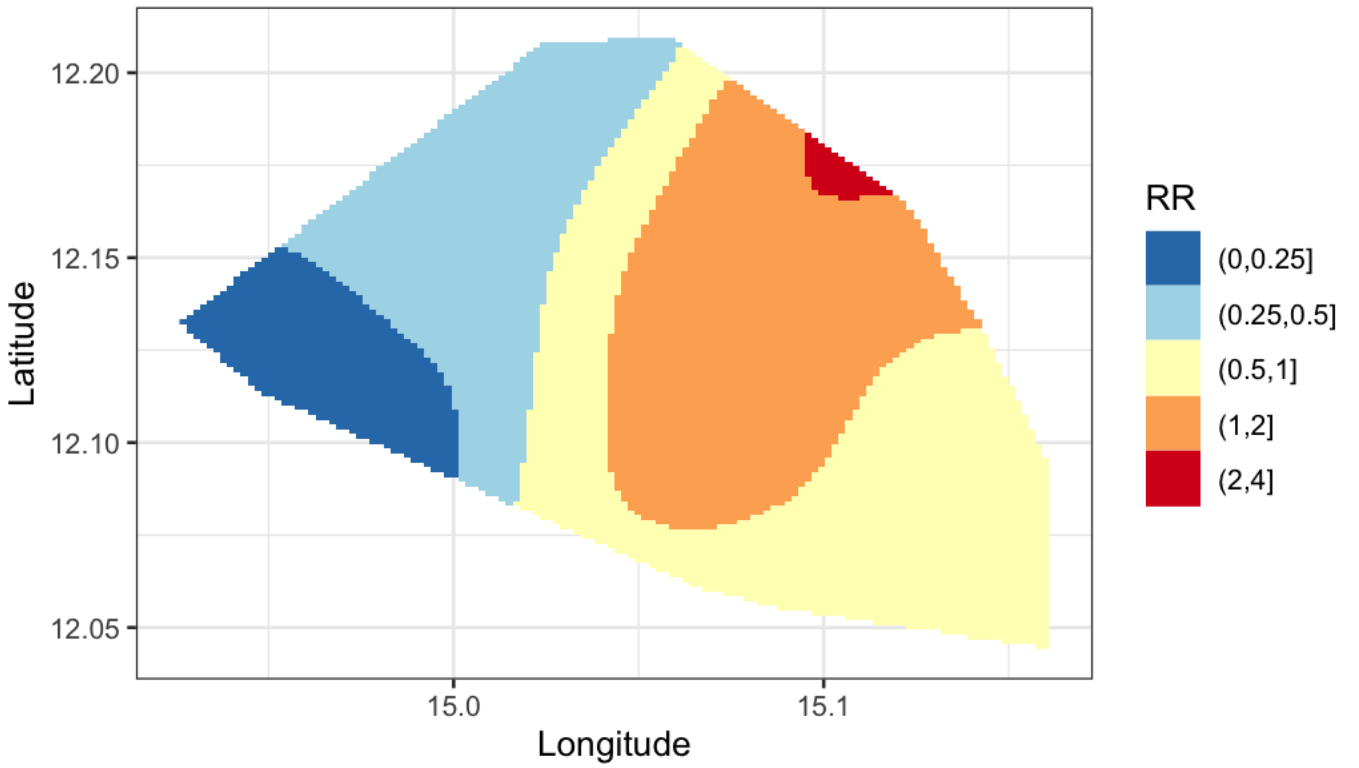
- Kernel smoothing the cases and controls gives us estimates of each $\lambda(s_i)$

Cholera in Chad



Kernel-smoothed spatial intensity

Cholera in Chad



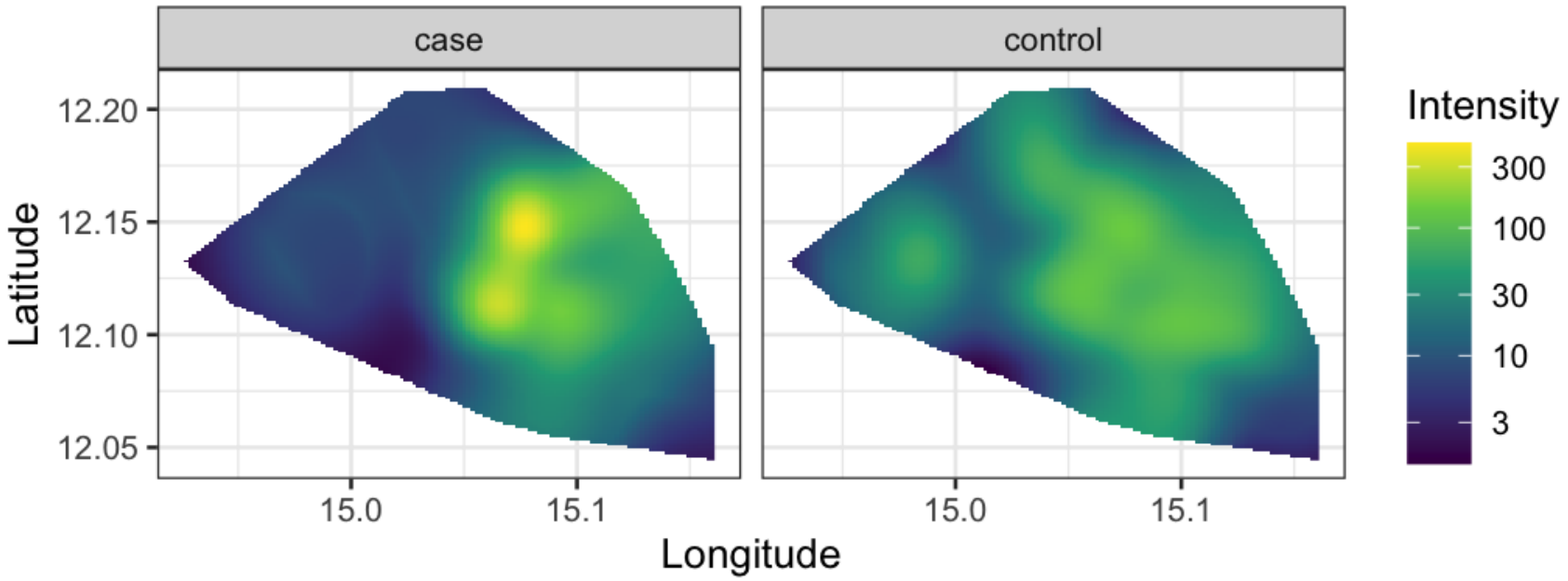
Relative risk of cholera

Point patterns - Smoothing

- Using the adaptive smoothing approach of Tilman M. Davies, Jones, and Hazelton (2016), T. M. Davies, Marshall, and Hazelton (2018)

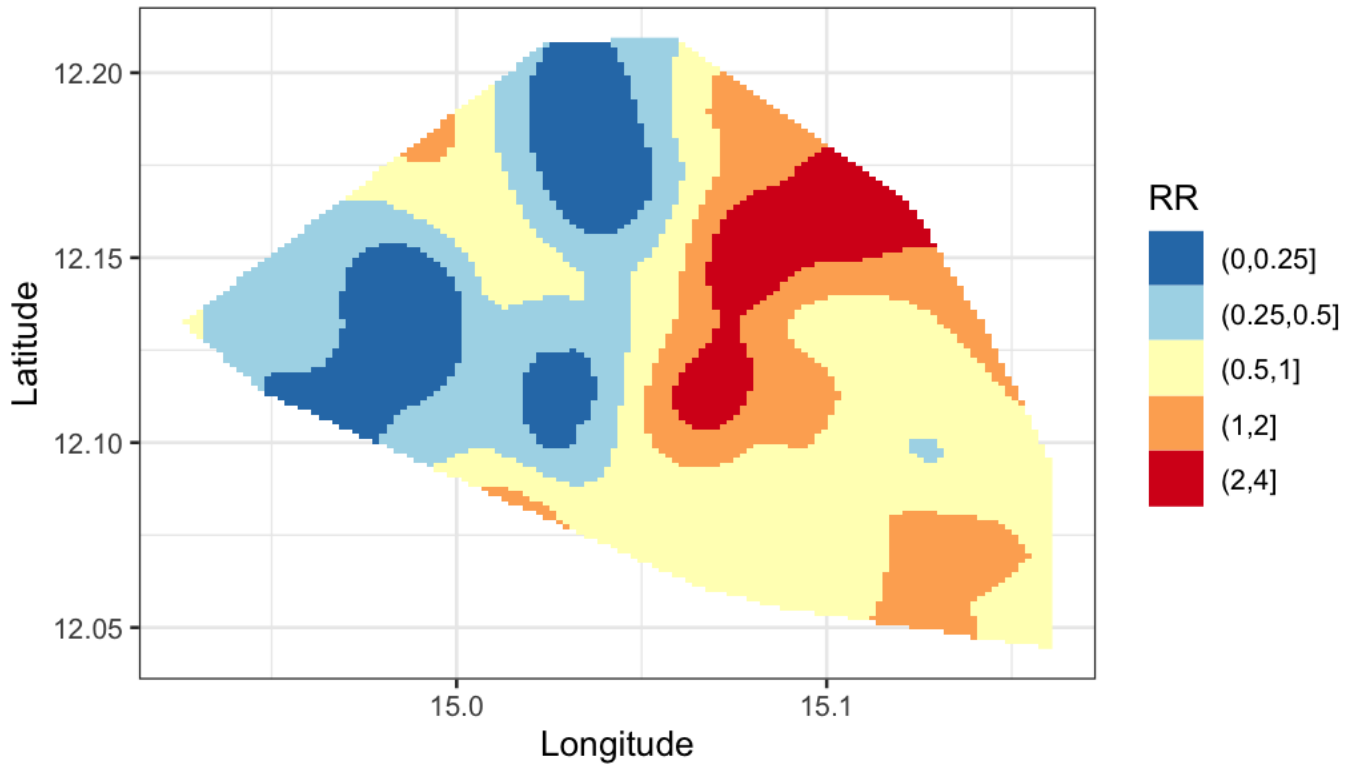
```
library(sparr)

cholera_risk <-
  # convert data frame to ppp
  ppp(x      = cholera$X,
      y      = cholera$Y,
      marks  = factor(cholera$case),
      window = cholera_window) %>% # domain extent
  risk(f = .,
       adapt = T,        # adaptive smoothing
       log = FALSE,      # don't return log of RR
       verbose = FALSE) # don't display messages
```

Cholera in Chad



Kernel-smoothed spatial intensity from adaptive smoothing

Cholera in Chad



Relative risk of cholera from adaptive smoothing

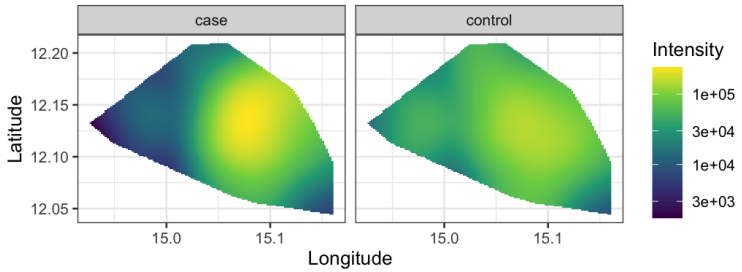
Cholera in Chad

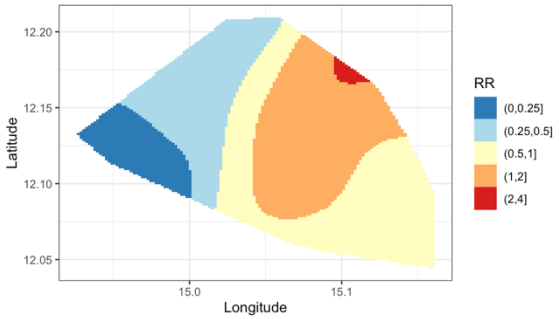
- Comparison:

Spatial patterns of cases and controls

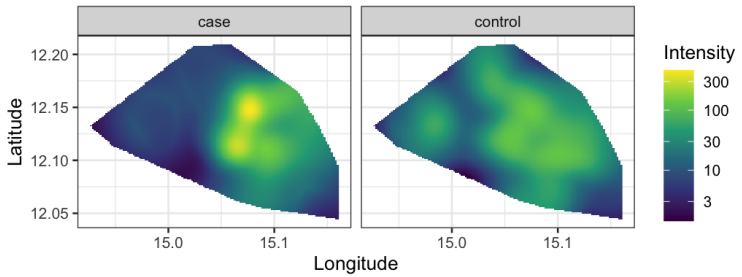
Spatial patterns of relative risk

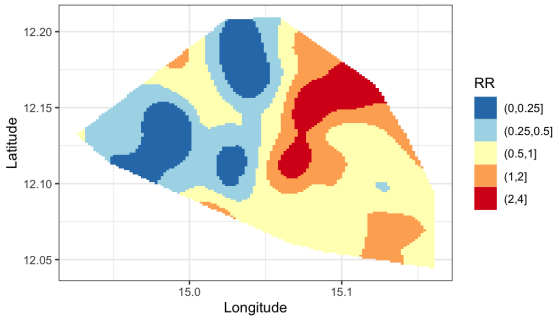
Defined kernel





Adapted kernel





Other types of point patterns

- Marked point process
  - Point data with counts (marks

- Inhomogeneous Point Process models that consider:
  - Spatial trend (e.g. using splines)
  - Interaction between points (A. Baddeley et al. 2014)
    - Clustering (infectious diseases)
    - Inhibition

- Species distribution models
  - Popular for presence only data
  - Models a (complex) relationship between risk and environmental covariates using machine learning methods
  - See https://rspatial.org for a dedicated species distribution modelling examples

References I

Baddeley, A., J.-F. Coeurjolly, E. Rubak, and R. Waagepetersen. 2014. “Logistic Regression for Spatial Gibbs Point Processes.” Biometrika 101 (2): 377–92. https://doi.org/10.1093/biomet/ast060.

Baddeley, Adrian, Ege Rubak, and Rolf Turner. 2015. Spatial Point Patterns: Methodology and Applications with R. London: Chapman; Hall/CRC Press. http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/.

Baddeley, Adrian, and Rolf Turner. 2005. “spatstat: An R Package for Analyzing Spatial Point Patterns.” Journal of Statistical Software 12 (6): 1–42. http://www.jstatsoft.org/v12/i06/.

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand. 2014. Hierarchical Modeling and Analysis for SpatialData. 2nd ed. Chapman; Hall/CRC.

Davies, T. M., J. C. Marshall, and M. L. Hazelton. 2018. “Tutorial on Kernel Estimation of Continuous Spatial and Spatiotemporal Relative Risk.” Statistics in Medicine 37 (7): 1191–221.

Davies, Tilman M., Khair Jones, and Martin L. Hazelton. 2016. “Symmetric Adaptive Smoothing Regimens for Estimation of the Spatial Relative Risk Function.” Computational Statistics & Data Analysis 101 (September): 12–28. https://doi.org/10.1016/j.csda.2016.02.008.

Elith, Jane, Steven J Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J Yates. 2011. “A Statistical Explanation of MaxEnt for Ecologists.” Diversity and Distributions 17 (1): 43–57.

Finger, Flavio, Enrico Bertuzzo, Francisco J. Luquero, Nathan Naibei, Brahima Touré, Maya Allan, Klaudia Porten, Justin Lessler, Andrea Rinaldo, and Andrew S. Azman. 2018. “The Potential Impact of Case-Area Targeted Interventions in Response to Cholera Outbreaks: A Modeling Study.” PLOS Medicine 15 (2): e1002509. https://doi.org/10.1371/journal.pmed.1002509.

Kammann, E. E., and M. P. Wand. 2003. “Geoadditive Models.” Journal of the Royal Statistical Society: Series C (Applied Statistics) 52 (1): 1–18. https://doi.org/10.1111/1467-9876.00385.

Matheron, Georges. 1963. “Principles of Geostatistics.” Economic Geology 58 (8): 1246–66. https://doi.org/10.2113/gsecongeo.58.8.1246.

References II

Mwase, Anna-Sofie AND Nsakashalo-Senkwe, Enala T. AND Stensgaard. 2014. “Mapping the Geographical Distribution of Lymphatic Filariasis in Zambia.” PLOS Neglected Tropical Diseases 8 (2): 1–13. https://doi.org/10.1371/journal.pntd.0002714.

O’Hagan, Anthony. 1978. “Curve Fitting and Optimal Design for Prediction.” Journal of the Royal Statistical
Society. Series B (Methodological), 1–42. Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. 2006. “Maximum Entropy Modeling of Species Geographic Distributions.” Ecological Modelling 190 (3-4): 231–59. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. 2006. “Maximum Entropy Modeling of Species Geographic Distributions.” Ecological Modelling 190 (3-4): 231–59.
https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2005. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.
http://www.gaussianprocess.org/gpml/chapters/.

Seeger, Matthias. 2004. “Gaussian Processes for Machine Learning.” International Journal of Neural Systems
14 (02): 69–106. https://doi.org/10.1142/s0129065704001899.

Simpson, Daniel, Finn Lindgren, and Håvard Rue. 2012. “In Order to Make Spatial Statistics Computationally Feasible, We Need to Forget about the Covariance Function.” Environmetrics 23 (1): 65–74. https://doi.org/https://doi.org/10.1002/env.1137.

Tobler, W. R. 1970. “A Computer Movie Simulating Urban Growth in the Detroit Region.” Economic Geography 46 (sup1): 234–40. https://doi.org/10.2307/143141.