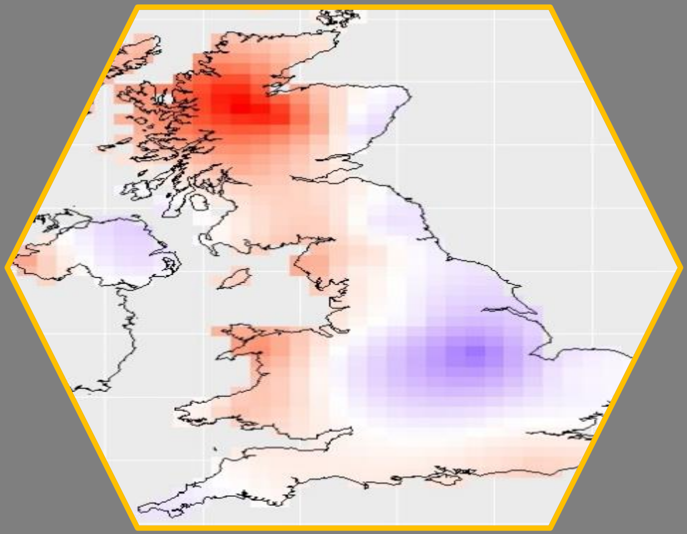


SPATIAL & SPATIO-TEMPORAL ANALYSIS – PART I

BAYESIAN COURSE, CEH SEPTEMBER 2019

CONTENT & LEARNING OUTCOMES

- Spatial data types
- Introduction to Geostatistics (focus on Gaussian data)
- Why ‘go Bayesian’ for spatial/spatio-temporal problems
- Examples
- In practice (software/packages)

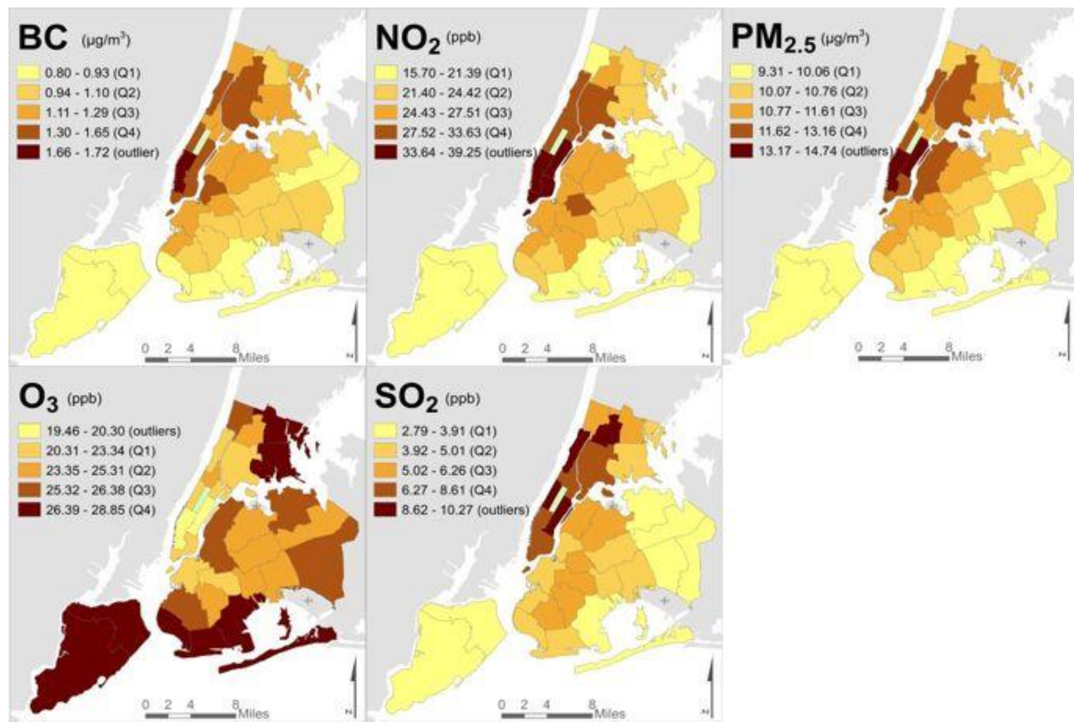
SPATIAL DATA '101'

- Data collected in space
- i.e. Sampling location is known and relevant
  - Space is of primary interest – we want to analyse and describe the spatial structure, how the data vary over space
  - Spatial structure is nuisance – we need to account for it in order to draw proper inference from our models (spatial autocorrelation)
  - Often data collected from more similar locations are more similar so are not independent
  - Upscaling

SPATIAL & SPATIO-TEMPORAL

SPATIAL DATA FORMS

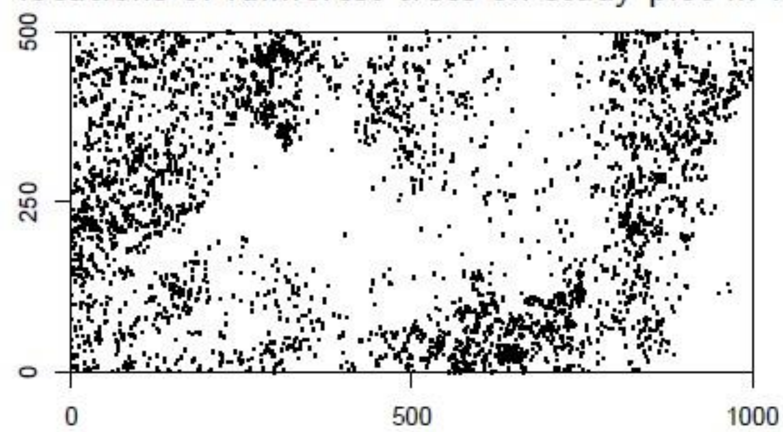
- Areal data – partitioned region with discrete spatial units, each with one value

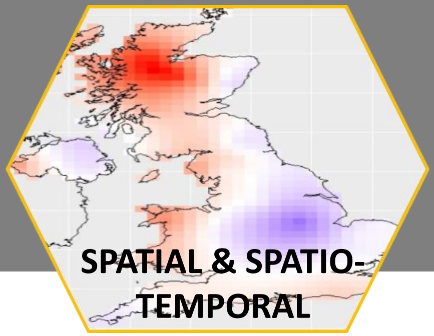


SPATIAL DATA FORMS

- Areal data
- Spatial point pattern data – Pattern formed by location of objects/events. Is the spatial pattern of points random or clustered/structured? Are the patterns determined by covariates?

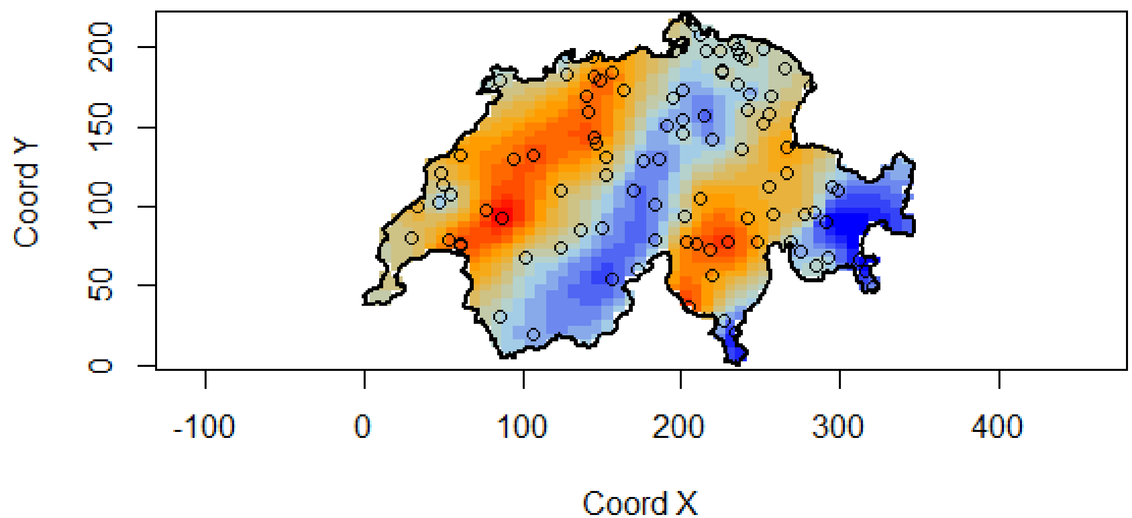
- locations of rainforest trees on study plot in Panama





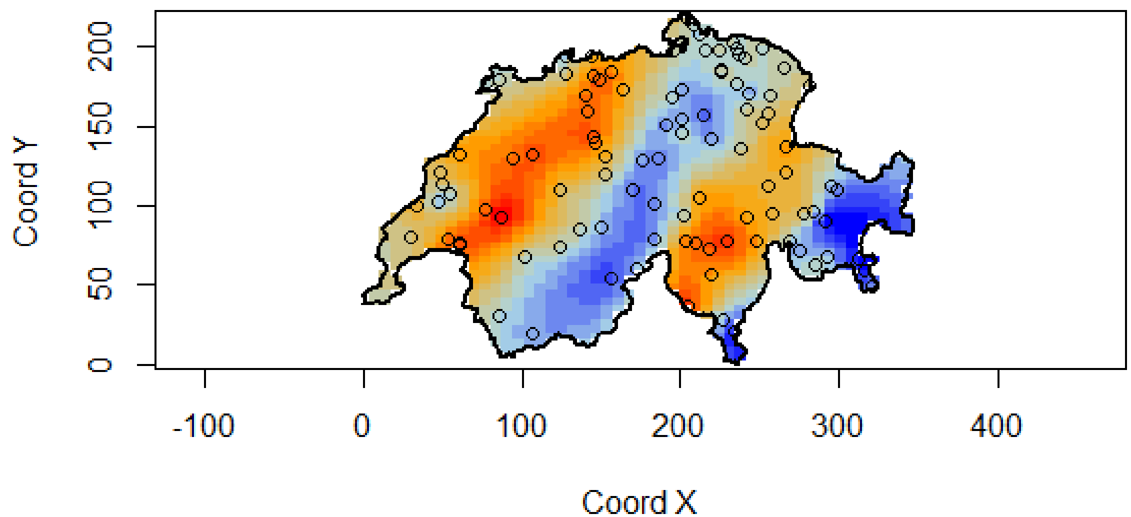
SPATIAL DATA FORMS

- Areal data
- Spatial point pattern data
- Geostatistical data – spatially continuous phenomena, based on observations at a finite number of locations

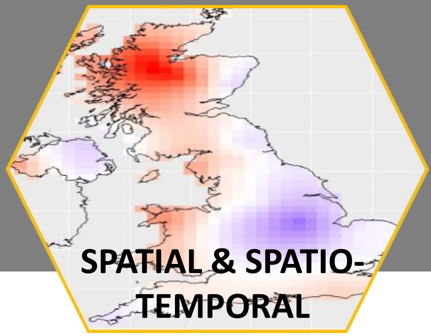


SPATIAL DATA FORMS

- Areal data
- Spatial point pattern data
- Geostatistical data

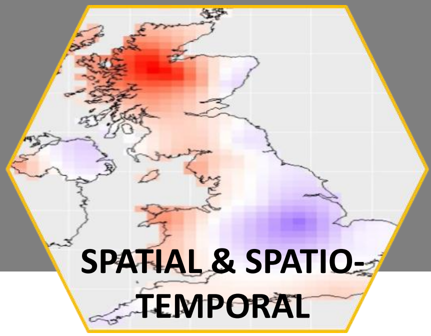


- Sometimes this is not our end-point -> covariates for other environmental/ecological processes



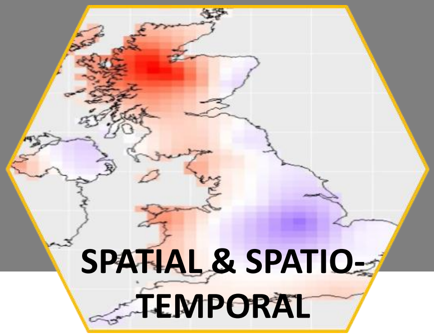
GEOSTATISTICS

- SAMPLES drawn from a CONTINUUM in space
- Often a goal is to PREDICT values in intervening spaces
- Without BIAS and MINIMISING UNCERTAINTY



SPATIAL DEPENDENCE or CORRELATION

- **Correlations** between sites are a function of **distance**
- Typically, geostatistical data will display **positive** correlation
- The closer two observations are the more similar their values are likely to be
- Arises because variables of interest being affected by other **unmeasured processes** which are themselves spatially correlated
- For example:
  - Air pollution
  - Soil nutrients
  - Your examples……



WHY 'GO BAYESIAN'…

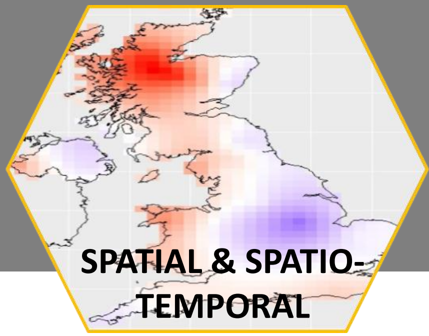
…FOR GEOSTATISTICAL ANALYSIS

- Correctly allows for variation in the parameters - Parameters of correlation function – ***random***, not fixed (have an associated probability distribution)
- Uncertainty intervals are easy to obtain for all parameters, not just regression parameters
- Appropriate propagation of uncertainty means prediction intervals will be wider

…WE WILL REVISIT!

A NOTE ON SOFTWARE

- **geoR**, and geoRglm for non-Gaussian data
- SpBayes
- R-INLA/inlabru
- OpenBUGS
- JAGS

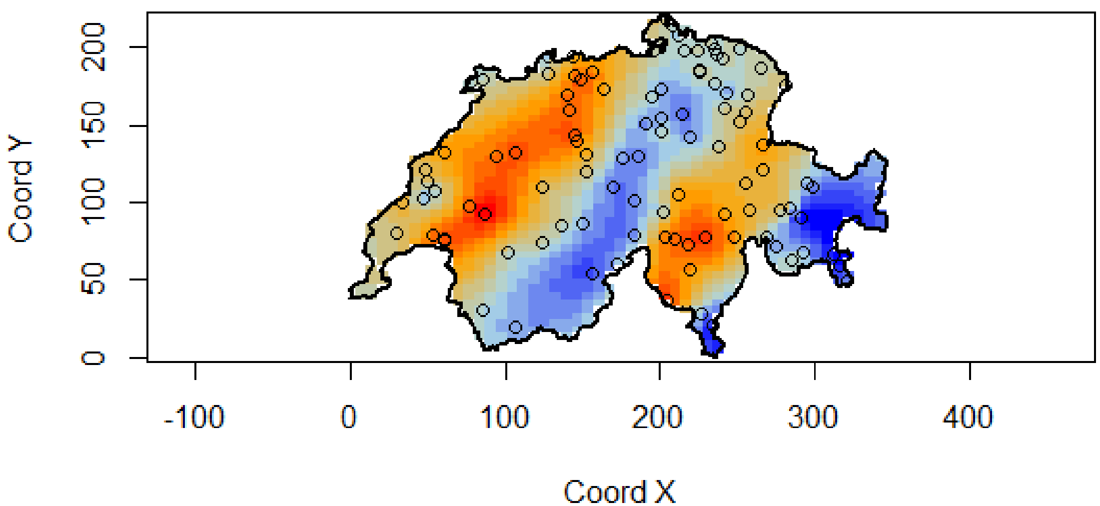


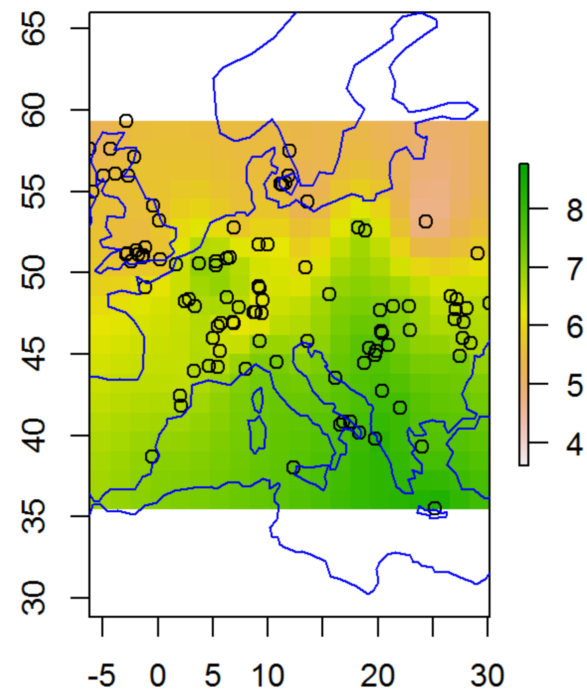
TODAY'S PRACTICALS

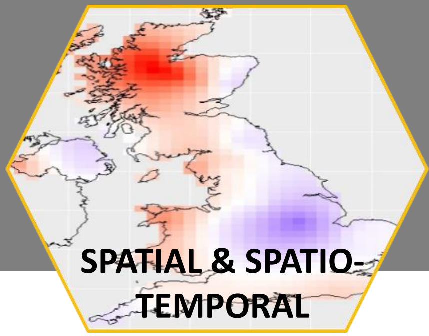
In geoR….

Spread of Agriculture

Swiss Rainfall Data

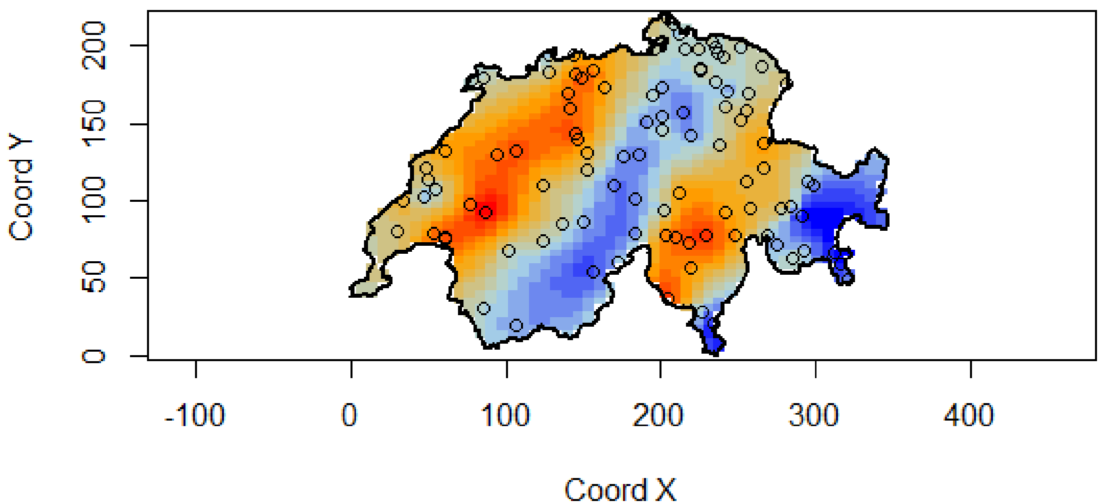






GEOSTATISTICS

Swiss Rainfall Data



Stochastic (random) process
- Variable of interest, Z, at locations set (s) within total space D

$$\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$$

Locations s at which data *could* occur vary continuously over D.
But data are observed at a finite number of locations, denoted by:

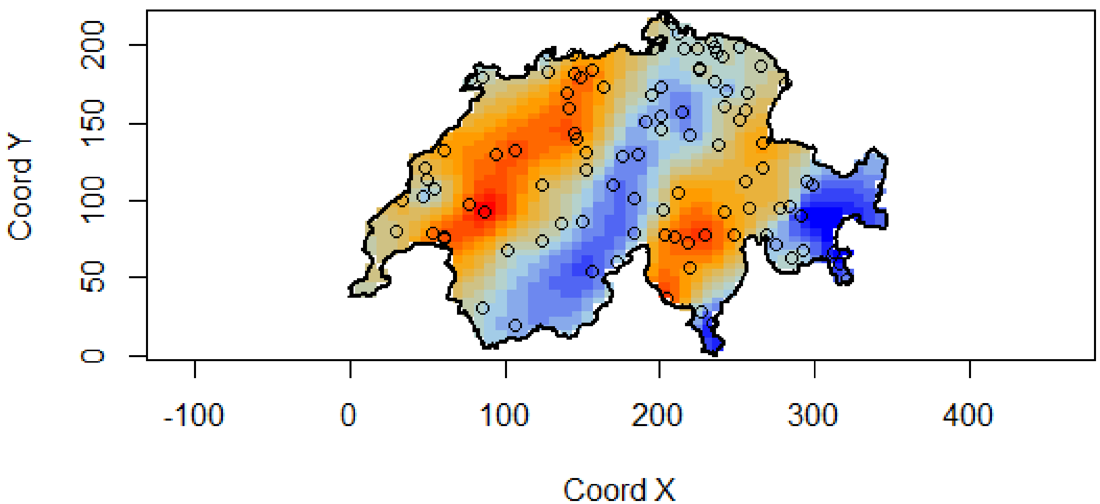
$$\mathbf{z} = \{z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_m)\}$$

i.e. a particular realisation of random variables Z(s)

SPATIAL & SPATIO-TEMPORAL

GEOSTATISTICS

Swiss Rainfall Data



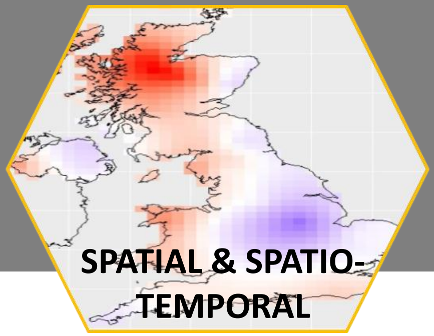
Data are observed at *m* locations:
What drives variation in *Z*?
Often want to predict the unknown stochastic process at locations where we have not sampled – i.e. produce a map across domain D.

Stochastic (random) process: Variable of interest, Z, at locations (s) within total space D

$$\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$$

Locations s at which data *could* occur vary continuously over D.
But data are observed at a finite number of locations, denoted by:

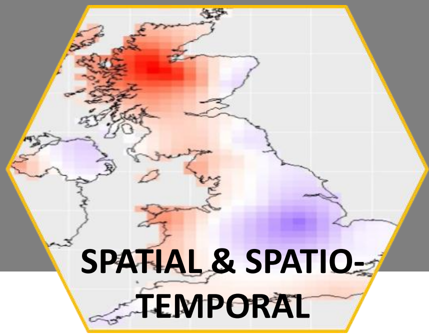
$$\mathbf{z} = \{z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_m)\}$$



STATIONARITY

- Data are observed at *m* locations but we often want to predict the unknown stochastic process at locations where we have not sampled – i.e. produce a map across domain *D*.
- BUT we only have one REALISATION – inference requires many realisations
- We must make some assumptions to simplify
- STATIONARITY – such that each *observation* can be treated as a random variable
  - Strict stationarity – all characteristics of the random function remain the same (not practicable)
  - Weak/Intrinsic stationarity – certain moments are invariant, whereas others are allowed to vary
- Put simply, the process Z(s) has the same degree of variation from place to place
- These assumptions underlie the theoretical variogram

e.g. Oliver & Webster 2014 *Catena*



RANDOM PROCESS MODEL

$$Z(\boldsymbol{s}) = \mu + \varepsilon(\boldsymbol{s})$$

Mean of the process μ, and random quantity $\varepsilon(\boldsymbol{s})$ with mean 0 and covariance C(h), where h is the separation between samples in distance and direction.

$$\mathrm{C}(\boldsymbol{h}) = \mathrm{E}[Z(\boldsymbol{s})Z(\boldsymbol{s}+\boldsymbol{h}) - \mu^2]$$

e.g. Oliver & Webster 2014 *Catena*

THE VARIOGRAM

Spatial association as a function of separation distance

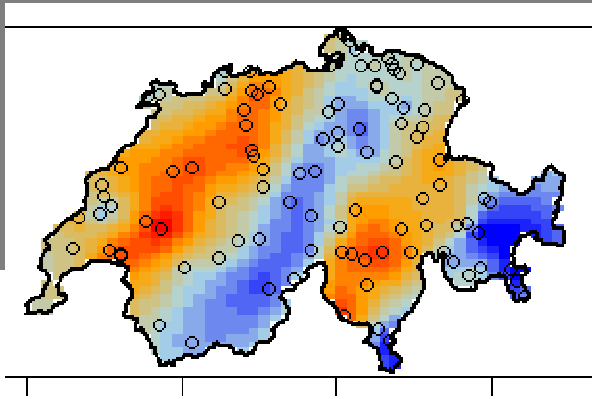
The semi-variance is a function denoted by:

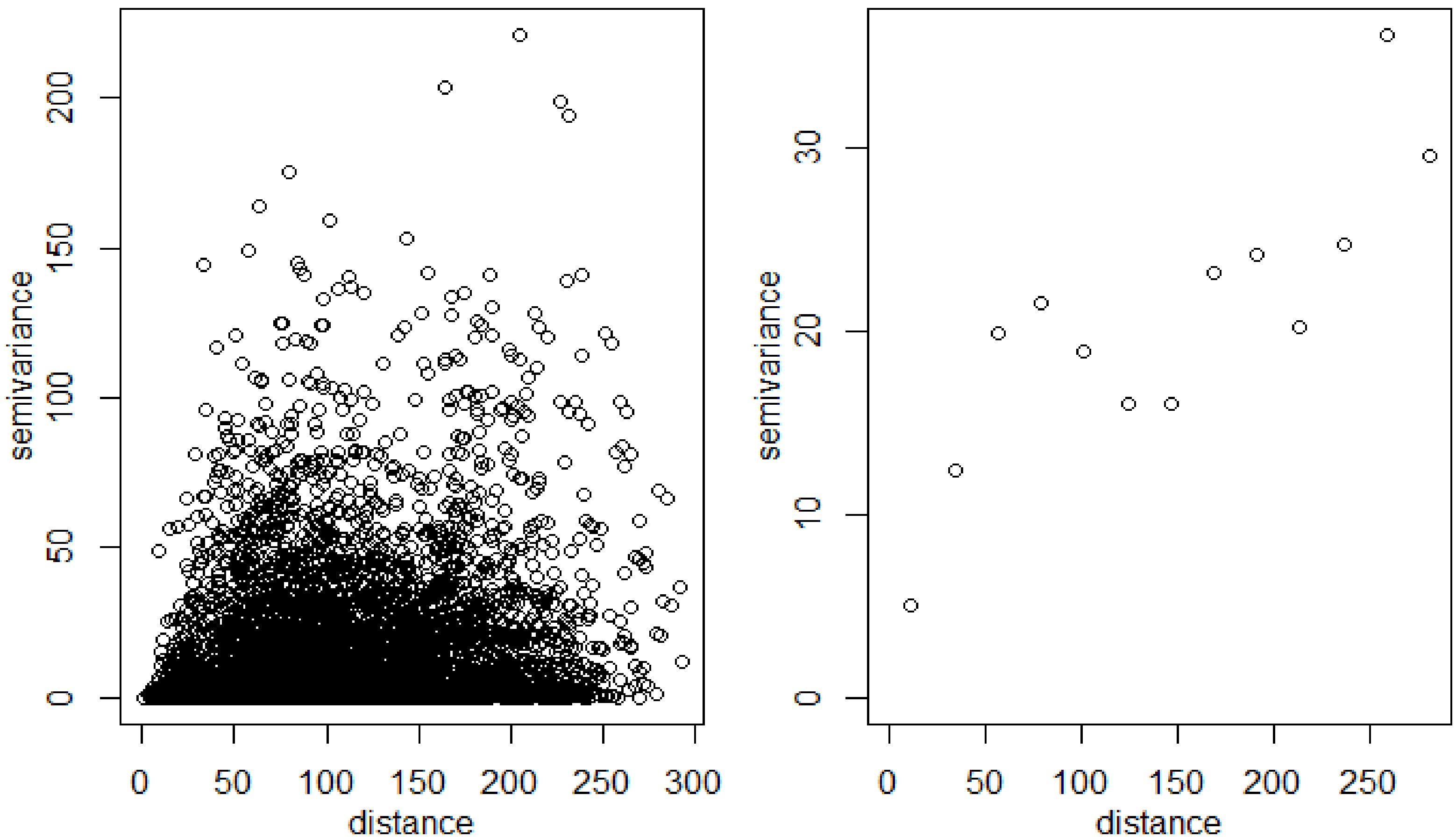
$$\gamma_z(\boldsymbol{h}) = \frac{1}{2}\,\mathrm{Var}[Z(\boldsymbol{s}) - Z(\boldsymbol{s}+\boldsymbol{h})]$$

Variance in difference in the process Z between two locations – when this is small, locations are spatially correlated

Semi-variance depends only on h (separation) – as a function of h, it is *the variogram*

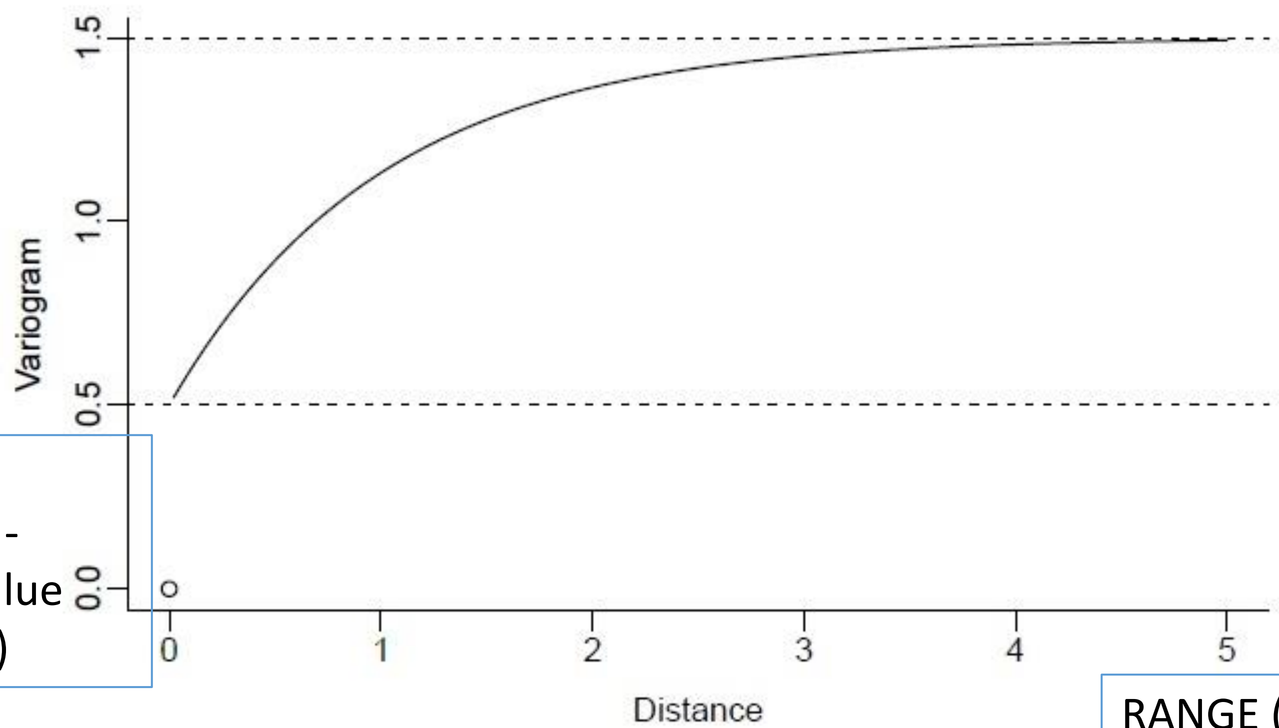
EXPERIMENTAL VARIOGRAM





THE VARIOGRAM

Theoretical shape of the variogram

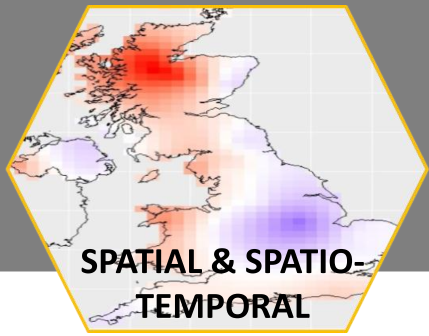


SILL (max variance between two zs)

PARTIAL SILL (sill minus nugget)

NUGGET (intercept - limiting value towards 0)

RANGE (distance at which sill is reached)



PARAMETRIC MODELS OF THE VARIOGRAM

- Features:
  - Nugget – $\boldsymbol{\tau^2}$ [tau^2] > 0
  - Partial sill – $\boldsymbol{\sigma^2}$ [sigma^2] > 0
  - Range parameter (rate at which covariance decays to zero) – $\boldsymbol{\varphi}$ [phi] > 0

SPATIAL & SPATIO-TEMPORAL

PARAMETRIC MODELS OF THE VARIOGRAM

- Some commonly used correlation functions

Floch 2018 *Handbook of Spatial Statistics*

Definition 5.3.2 — Exponential model.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s\left[1 - exp\left(-\frac{h}{a}\right)\right] & h > 0 \end{cases} \tag{5.16}$$

Definition 5.3.3 — Gaussian model.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s\left[1 - exp\left(-(\frac{h}{a})^2\right)\right] & h > 0 \end{cases} \tag{5.17}$$

Definition 5.3.4 — Power model.

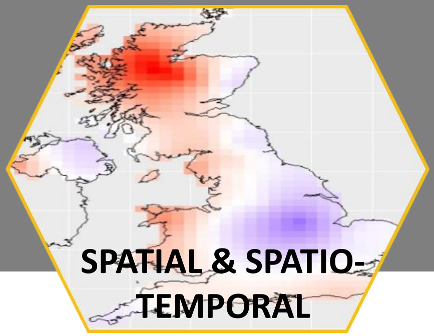
$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + bh^p & h > 0 \end{cases} \tag{5.18}$$

Definition 5.3.5 — Matern model.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_s\left[1 - \frac{\frac{h}{a}}{2^{\alpha-1}\Gamma(\alpha)} K_\alpha\left(\frac{h}{a}\right)\right] & h > 0 \end{cases} \tag{5.19}$$

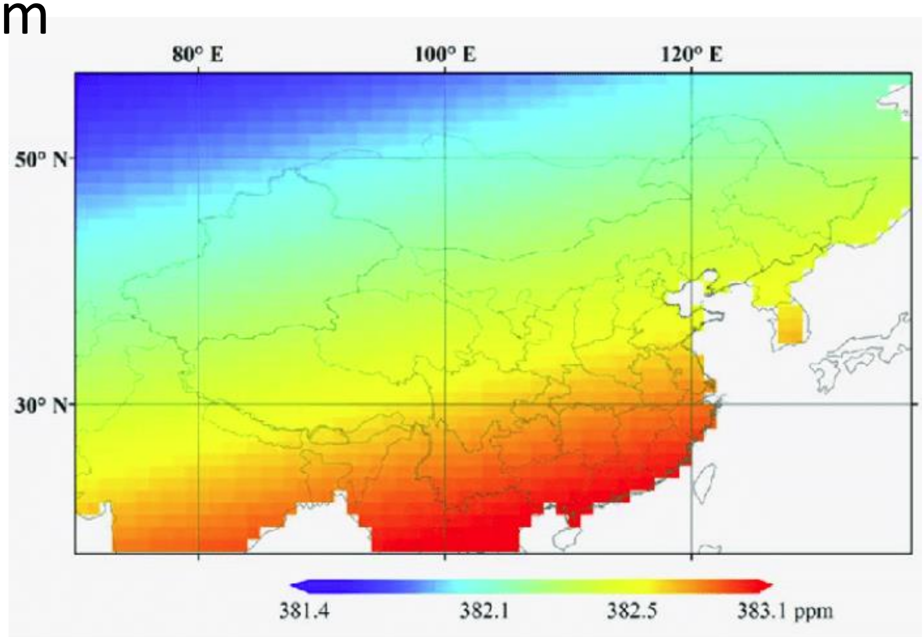
where Γ refers to the gamma function and K_α, the modified Bessel of the second kind of parameter α.

- Caution with Gaussian – can lead to bizarre predictions (Wackernagel 2003)
- **Choice of function -> influence results**



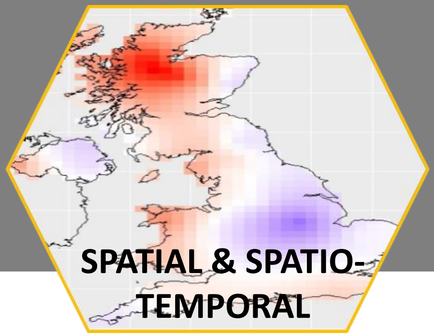
TRENDS IN SPATIAL DATA

- Recall stationarity assumption….
- Gradual variation/large-scale variation
- Modelled using covariates or coordinates
- Then, in essence, just modelling the *residual* spatial variability with the variogram

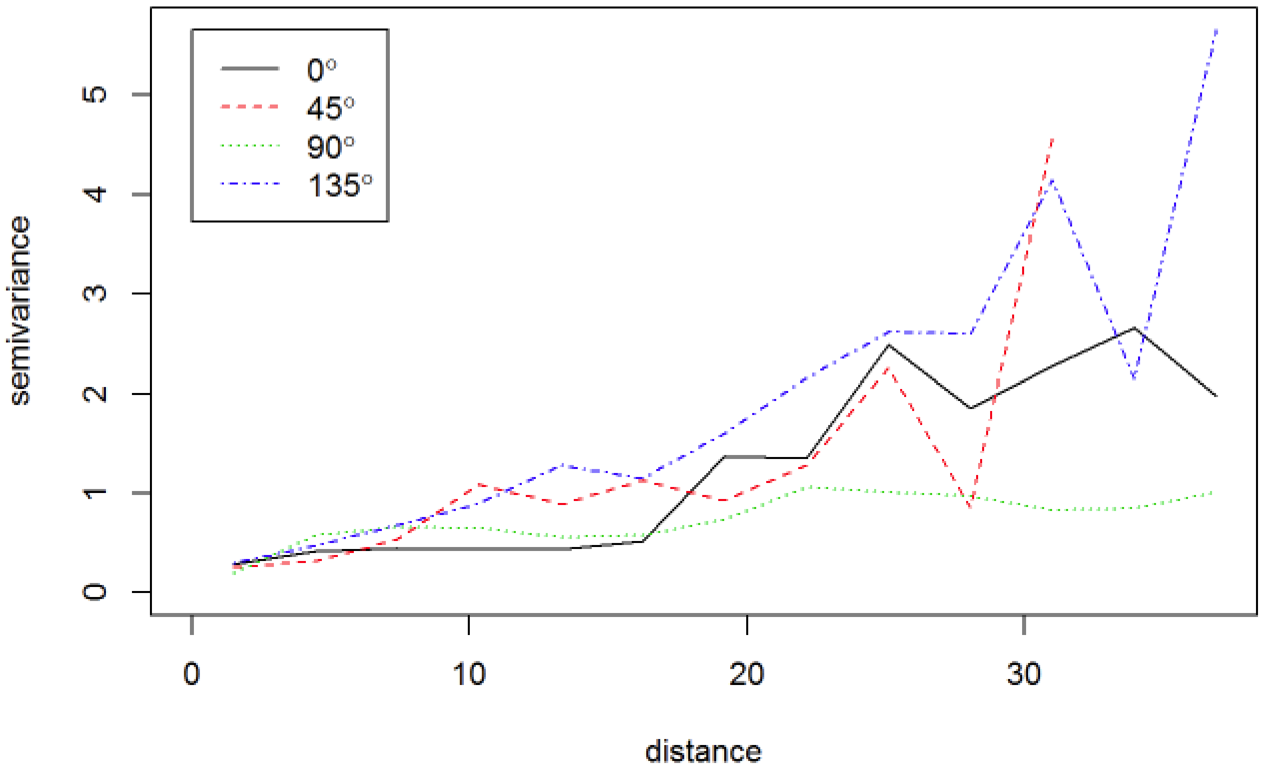


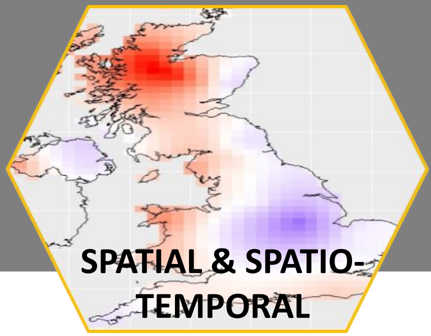
Linear spatial trend surface of Xco 2 in the study area derived from ACOS-GOSAT data in one month of September 2009.

Zeng et al. 2014 *Transactions on Geoscience and Remote Sensing*


SPATIAL & SPATIO-
TEMPORAL

ISOTROPY


0°
45°
90°
135°
semivariance
0
1
2
3
4
5
0
10
20
30
distance



BAYESIAN GEOSTATISTICAL MODEL

Data:

$$\mathbf{z} = \{z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_m)\}$$

Gaussian model with likelihood:

$$\mathbf{z} \sim \mathrm{N}(X\boldsymbol{\beta}, \Sigma(\boldsymbol{\theta}))$$

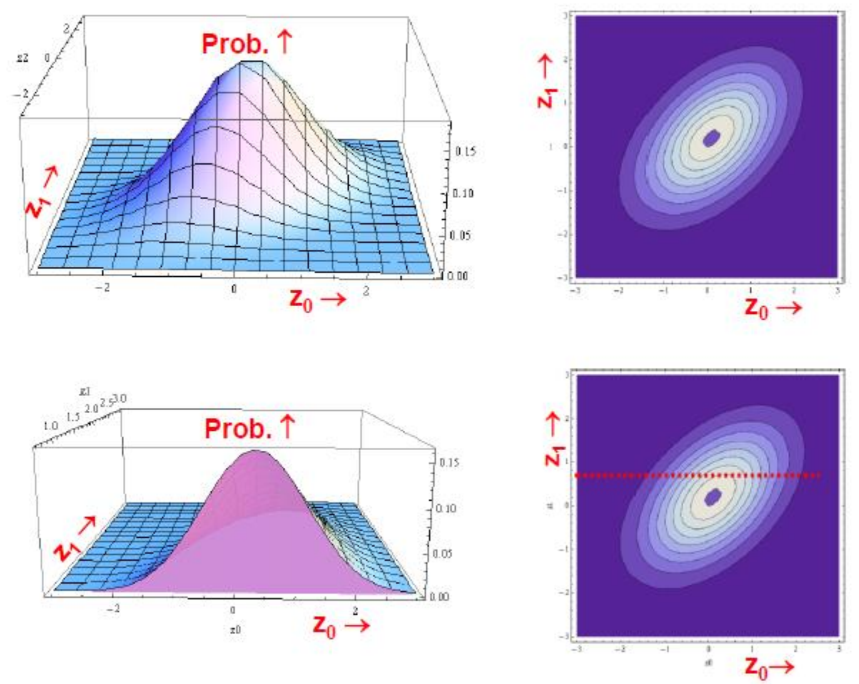
Where the mean function $X\boldsymbol{\beta}$ is a linear combination of known covariates (X) and $\Sigma(\boldsymbol{\theta})$ is m x m covariance matrix for the m observations, determined by a stationary and isotropic covariance function.
The unknown parameters are $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \sigma^2, \varphi, \tau^2, \kappa^2)$
Joint distribution:

$$f(\boldsymbol{\Theta}) = f(\boldsymbol{\beta}, \sigma^2, \varphi, \tau^2, \kappa^2)$$

SPATIAL & SPATIO-TEMPORAL

BAYESIAN GEOSTATISTICAL MODEL

Multivariate Gaussian distribution



Data:

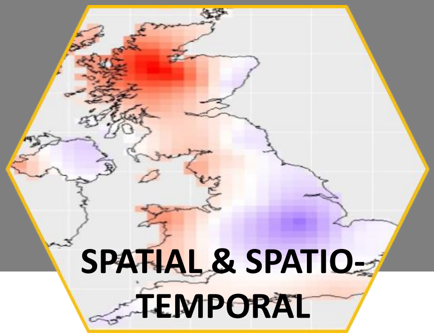
$$\mathbf{z} = \{z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_m)\}$$

Gaussian model with likelihood:

$$\mathbf{z} \sim \mathrm{N}(X\boldsymbol{\beta}, \Sigma(\boldsymbol{\theta}))$$

Where the mean function $X\boldsymbol{\beta}$ is a linear
(X) and $\Sigma(\boldsymbol{\theta})$ is m x m covariance matrix
determined by a stationary and isotropi
Unknown parameters $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \sigma^2, \varphi, \tau^2,$
Joint distribution:

$$f(\boldsymbol{\Theta}) = f(\boldsymbol{\beta}, \sigma^2, \varphi, \tau^2, \kappa^2)$$



BAYESIAN GEOSTATISTICAL MODEL

Data:

$$\mathbf{z} = \{z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_m)\}$$

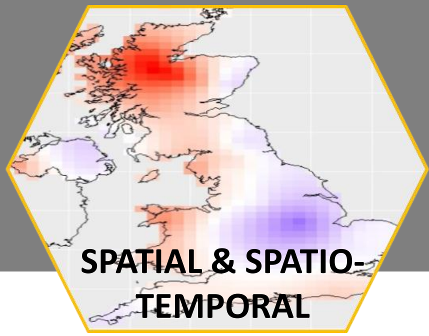
Gaussian model with likelihood:

$$\mathbf{z} \sim \mathrm{N}(X\boldsymbol{\beta}, \Sigma(\boldsymbol{\theta}))$$

Where the mean function $X\boldsymbol{\beta}$ is a linear combination of known covariates (X) and $\Sigma(\boldsymbol{\theta})$ is m x m covariance matrix for the m observations, determined by a stationary and isotropic covariance function.
The unknown parameters are $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \sigma^2, \varphi, \tau^2, \kappa^2)$
Joint distribution:

$$f(\boldsymbol{\Theta}) = f(\boldsymbol{\beta}, \sigma^2, \varphi, \tau^2, \kappa^2)$$

DATA + PRIORS -> POSTERIOR



PLAUSIBLE PRIORS

$$\boldsymbol{\beta} \sim \mathrm{N}(\mu_\beta, V_\beta)$$

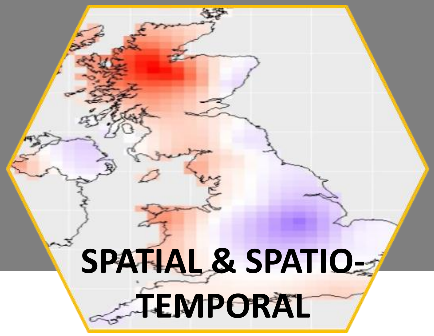
Take on any value

$$\sigma^2, \kappa^2 \sim \text{Uniform}(0, \text{*large}) \text{ or}$$
$$\sigma^2, \kappa^2 \sim \text{Inverse-Gamma}(a, b)$$

Must be positive
Inverse-Gamma conjugate for variance. Choice of (a,b) under research/debate.

$$\varphi \sim \text{Uniform}(c, d)$$

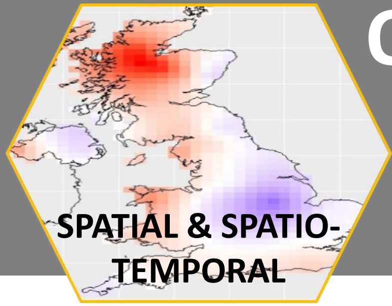
Chosen over likely range of correlations



WHY 'GO BAYESIAN'...

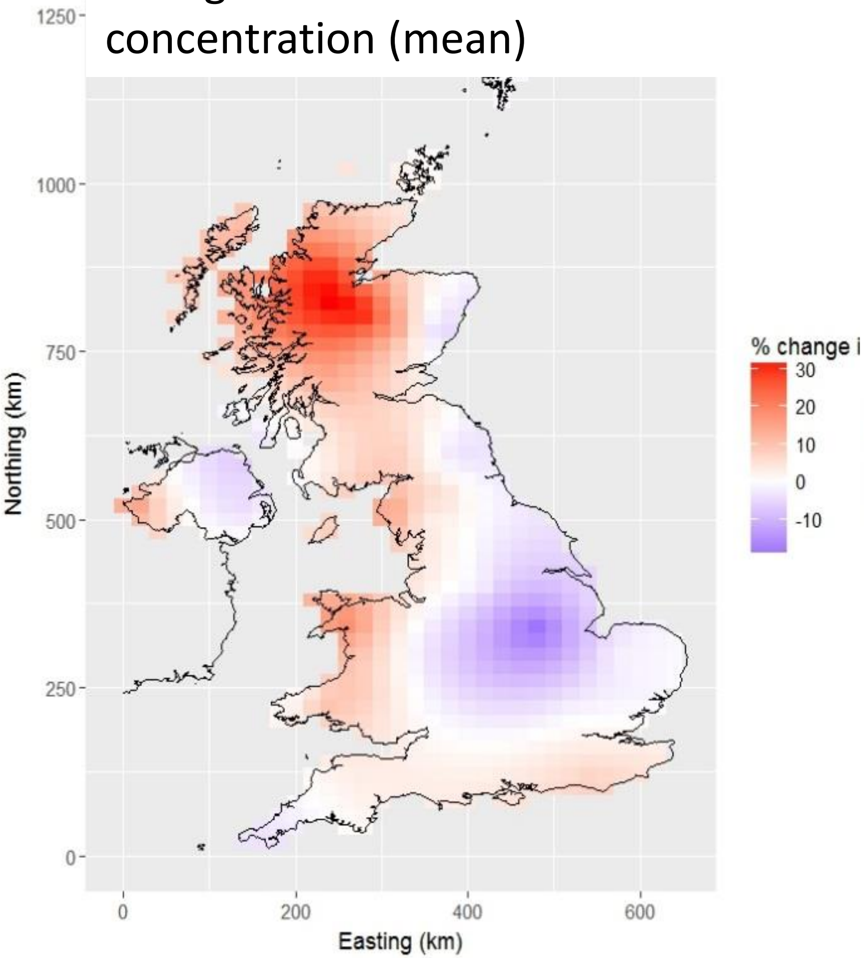
...FOR GEOSTATISTICAL ANALYSIS

- Correctly allows for variation in the parameters - Parameters of correlation function – ***random***, not fixed (have an associated probability distribution)
- Uncertainty intervals are easy to obtain for all parameters, not just regression parameters
- Appropriate propagation of uncertainty means prediction intervals will be wider

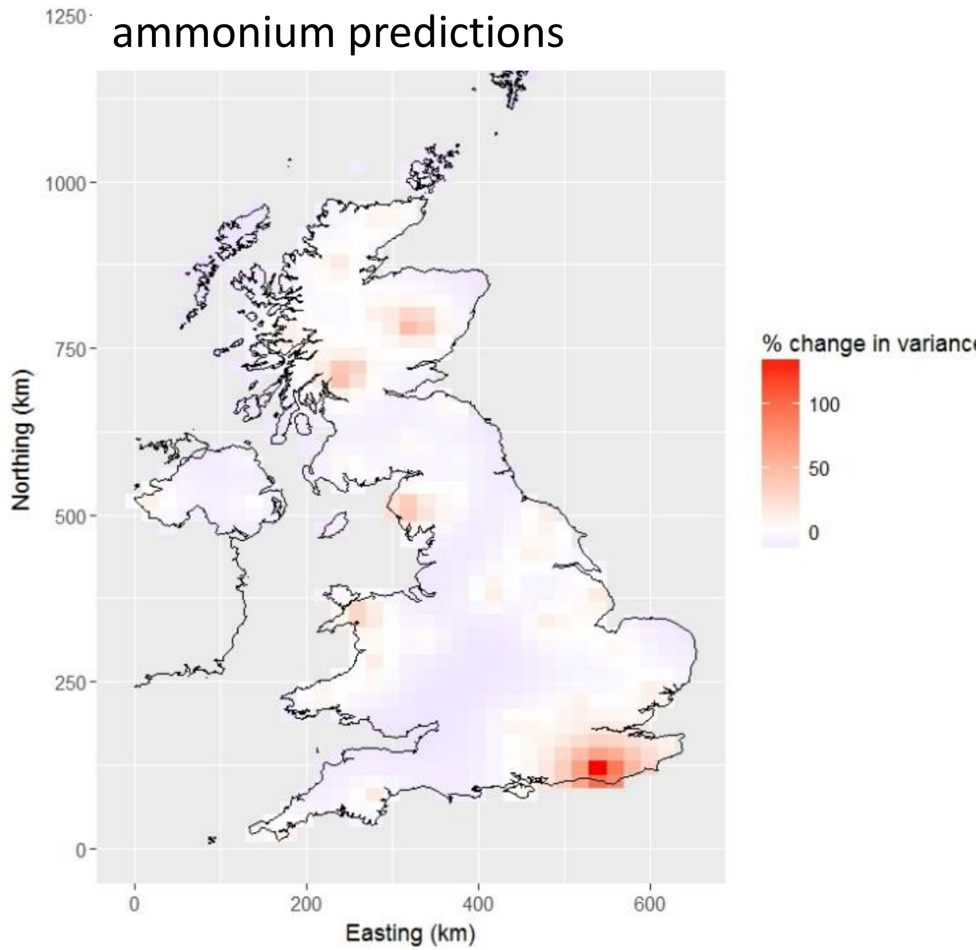


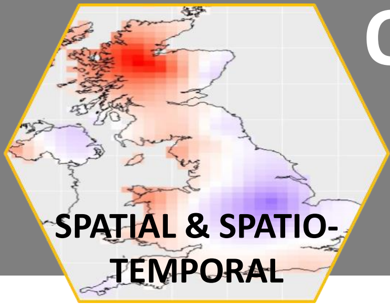
Q: What is the impact of losing 6 sites from an air pollution monitoring network?

Change in estimated ammonium concentration (mean)



Change in variance of ammonium predictions

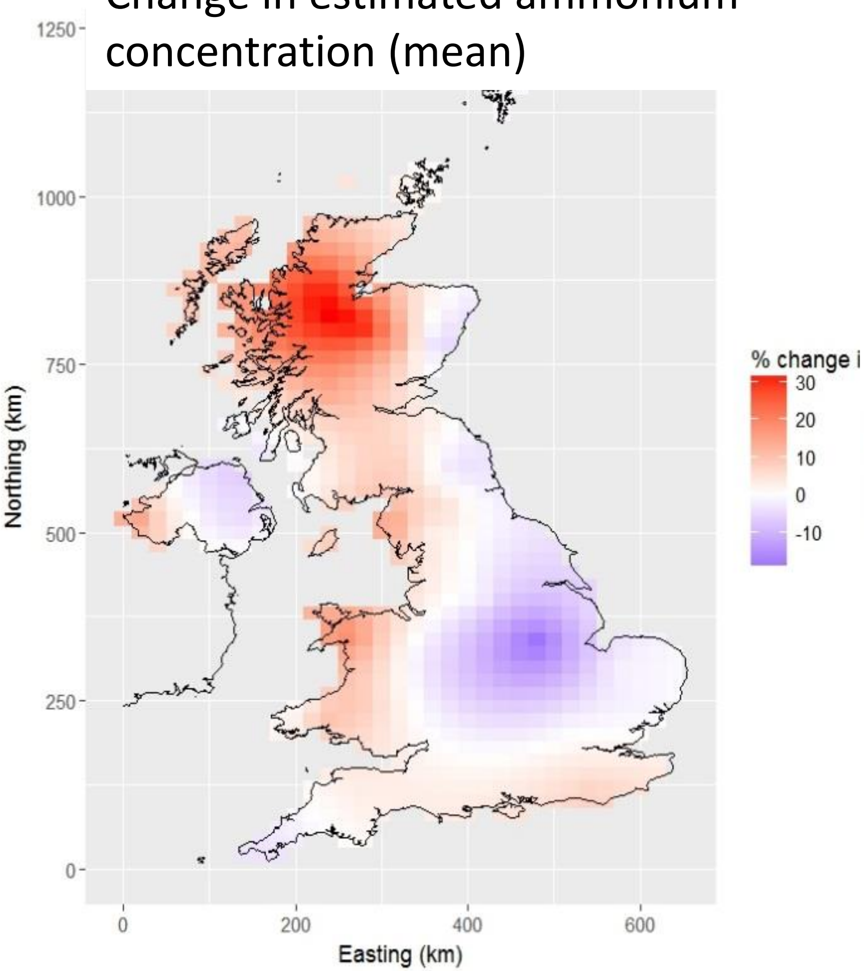




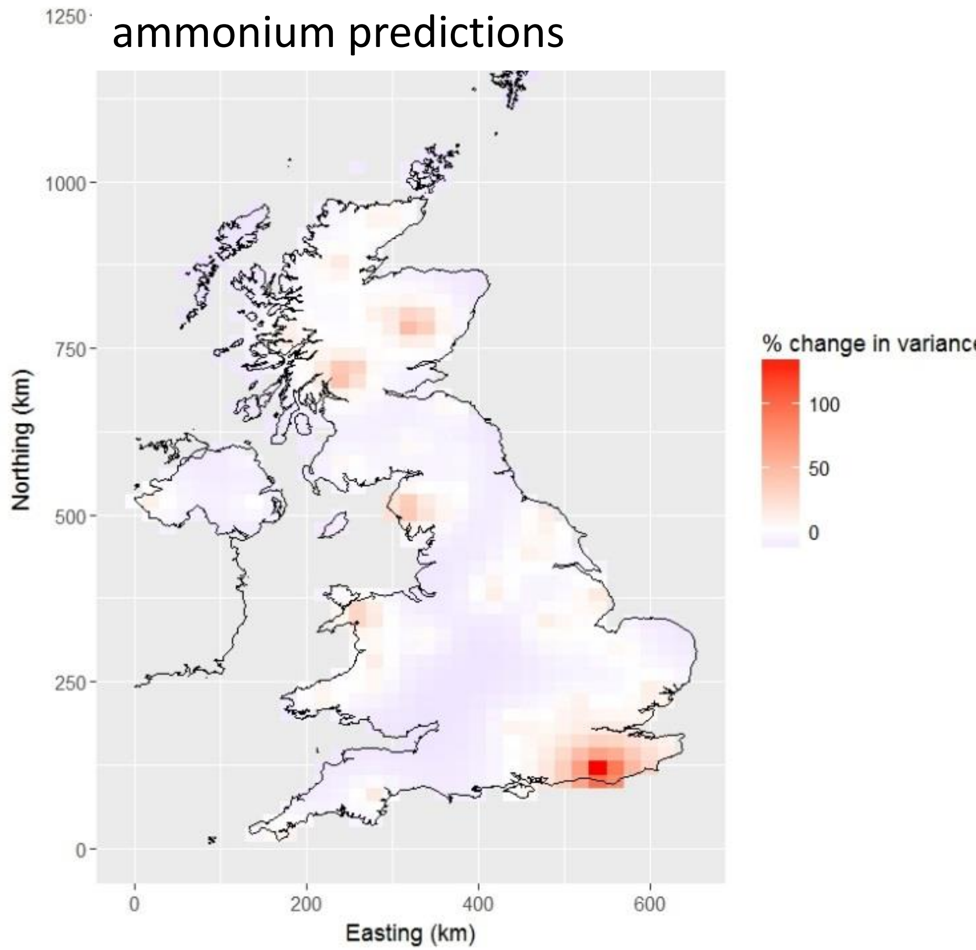
Q: What is the impact of losing 6 sites from an air pollution monitoring network?

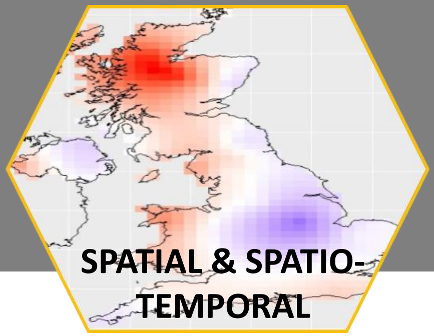
Generalise: Impact of sample size on spatial interpolation

Change in estimated ammonium concentration (mean)



Change in variance of ammonium predictions





ESTIMATING THE MODEL

- Data - how are they distributed; do we have covariates?
- Mean - constant; trends
- Covariance model
- Priors – which are fixed, which are random, which distribution?

Krige.bayes function
model.control

prior.control

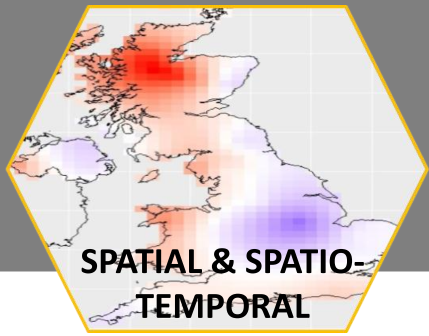
ESTIMATING THE MODEL

- Data - how are they distributed; do we have covariates?
- Mean - constant; trends
- Covariance model
- Priors – which are fixed, which are random, which distribution?
- Posterior distribution for each parameter
- Predictive distribution for each location
  - From which we can map the mean, variance etc…

Krige.bayes function
model.control

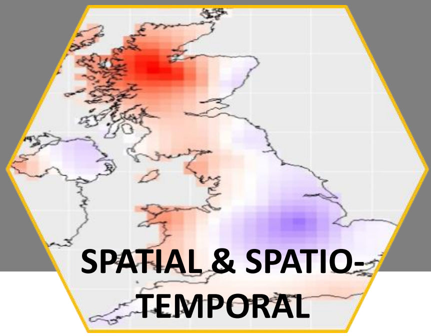
prior.control

output.control



SUMMARY

- In environmental data, things most often positively correlate in space
- Reflecting underlying processes and (un)measured data
- This may be of interest or a dependence for which we need to account
- Computational power is allowing us to do this in more sophisticated ways
- In Bayesian geostatistics the variogram parameters are random (defined by a probability distribution), not fixed (single value)
- As such we can propagate and more thoroughly estimate uncertainty, which may inform model selection as well as inference

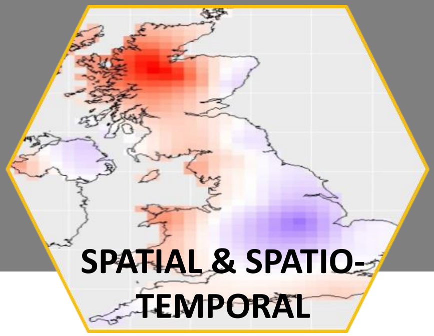


RESOURCES

- Banerjee & Fuentes (2012) Bayesian Modelling for Large Spatial Datasets. *Wiley Interdiscip Rev Comput Stat.* 4(1): 59–66.
- Oliver & Webster (2014) A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena* 113: 56-69
- Ribeiro et al. (2003) Geostatistical software – geoR and geoRglm. DSC 2003 Working Papers
- Ribeiro & Diggle Technical Report ST-99-08: Bayesian inference in Gaussian model-based geostatistics

See also references in Spatio-temporal slides

Supplementary slides



The mean function – the expected value at location **s** from the distribution of all possible values generated from stochastic process Z(**s**)

$$\mu_z(\boldsymbol{s}) = \mathbb{E}[Z(\boldsymbol{s})] \qquad \boldsymbol{s} \in D$$

When Z(s) is a continuous random variable:

$$\mu_z(\boldsymbol{s}) = \mathbb{E}[Z(\boldsymbol{s})] = \int_{-\infty}^{\infty} z f_{Z(S)}(z)\, dz$$

Where $f_{Z(S)}$ is the probability density function (pdf) for $Z(\boldsymbol{s})$

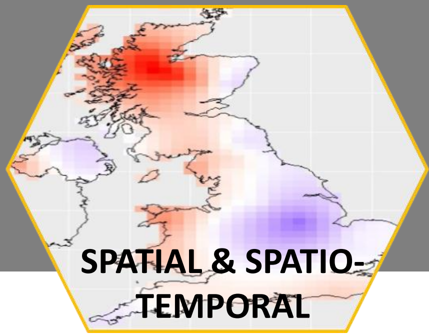
The covariance function – covariance measures strength of linear dependence between two random variables (Z(s) and Z(t))

$$C_Z(\boldsymbol{s},\boldsymbol{t}) = \mathrm{Cov}[Z(\boldsymbol{s}), Z(\boldsymbol{t})]$$

$$C_Z(\boldsymbol{s},\boldsymbol{t}) = \mathbb{E}[(Z(\boldsymbol{s}) - \mu_Z(\boldsymbol{s}))(Z(\boldsymbol{t}) - \mu_Z(\boldsymbol{t}))]$$

The variance function of Z(s) is the special case of the covariance s=t, giving:

$$\begin{aligned}\mathrm{Var}[Z(\boldsymbol{s})] &= \mathrm{Cov}[Z(\boldsymbol{s}), Z(\boldsymbol{s})] \\ &= \mathbb{E}[(Z(\boldsymbol{s}) - \mu_Z(\boldsymbol{s}))^2] \\ &= \varepsilon^2{}_{\mathrm{Z}}(\boldsymbol{s})\end{aligned}$$



GEOSTATISTICS

The correlation function – the strength of association between two random variables (Z(s) and Z(t)) is simply a scaled version of the covariance function:

$$\rho_z(\boldsymbol{s},\boldsymbol{t}) = \mathrm{Corr}[Z(\boldsymbol{s}), Z(\boldsymbol{t})]$$

$$= \frac{C_z(\boldsymbol{s},\boldsymbol{t})}{\sqrt{C_z(\boldsymbol{s},s)Cz(\boldsymbol{t},\boldsymbol{t})}}]$$