TMA4250 Spatial Statistics, spring 2017

Course page:
https://wiki.math.ntnu.no/tma4250/2017v/start

- Our data
- Thea (Precipitation)
- Torstein (Lithology)
- Haakon (Seals in Scotland)
- Avalanches in Sogn
- Lightning strikes over Norway
- Methylation
- Doctor-prescription in France
- Scots pine in Sweden

Hierachical statistical models (HM)

Data model: $[Z|Y,\theta]$

Process model: $[Y|\theta]$

Parameter model: $[\theta]$

For precipitation (Thea):

Data model: $[Z|Y,\theta]$ Distribution for observations given true precipitation Independent $Z(s_i) = N(Y(s_i), \sigma_d^2)$

Process model: $[Y|\theta]$ Distribution for precipitation.
$Y(s) \sim GRF(\theta_p)$

Parameter model: $[\theta]$ Prior for parameters. $[\theta] = [\sigma_d^2][\theta_p]$

- Bayesian HM (BHM): θ considered random variable, given prior
- Empirical HM (EHM): θ considered fixed, but unknown

Inference

BHM: We want posterior distributions

- $[Y|Z]$, process given data
- $[\theta|Z]$, parameters given data
- $[Y, \theta|Z]$

How:

- MCMC (TMA4300, we will do)
- For some models INLA (developed at NTNU)

EHM: We want posterior distributions

- $[Y|Z, \hat{\theta}]$, process given data

How? Estimate θ ($\hat{\theta}$) by

- Maximum-likelihood (we will do)
- Expectation-Maximalization (EM), pseudo-likelihood,...

Why do we want to do spatial statistics?

- Want to predict for locations not observed. I.e. $[Y(s_0)|Z(s_{obs})]$ (or $[Z(s_0)|Z(s_{obs})]$)
- Want to understand underlaying processes, i.e. $[Y(s_0)|Z(s_{obs})]$ or $[\theta|Z(s_{obs})]$
- Want to account for spatial dependency (not independent observations)
- Want to use as proxy for 'lurking variables' (Simpsons's paradox / Ecological fallacy)

Success treatment of Kidney stones (pg 12)

For all surgery
- Open surgery: Success rate 78 %
- Ultra sound: Success rate 83 %

For small stones:
- Open surgery: Success rate 93 %
- Ultra sound: Success rate 87 %

For large stones:
- Open surgery: Success rate 73 %
- Ultra sound: Success rate 69 %

Lurking variable: Patients kidney stone size

The Ecological Fallacy

Foreigne born and litteracy in US, 1930s (pg 197)

At individual level: Correlation: -0.11

At state level: Correlation 0.53

Same as *Simpsons's paradox*, due to *change-of-support*, also named *ecological bias*

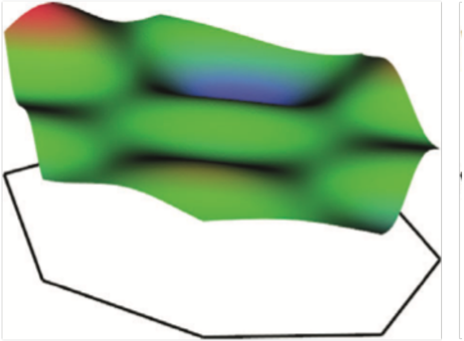
Three parts:

Part 1: Geostatistical Processes (chapter 4.1)

Part 2: Spatial Point processes (chapter 4.3 ++)

Part 3: Lattice processes (chapter 4.2, focus on discrete Markov random fields)

We need a stochastic models for random variables that are defined for a domain (in space) for the process model.

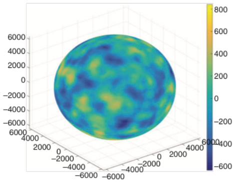


Gaussian Random Field (GRF)

The random field $Y(s)$, $s \in D$ (f.ex. in R^2) is a Gaussian Random field if, for any n, and any set of locations $s_1, s_2, \ldots, s_n$, all finite collections $(Y(s_1), Y(s_2), \ldots Y(s_n))$ are multivariate Normal distributed.

From Wikipedia: *A random field is a generalization of a stochastic process such that the underlying parameter need no longer be a simple real or integer valued "time", but can instead take values that are multidimensional vectors, or points on some manifold.*

(a)



Multivariate Normal distribution

Multivariate Normal(MVN) density

$Y = (Y_1, Y_2, \ldots, Y_n)$ is MVN with expected value μ and covarance Σ, $Y \sim MVN(\mu, \Sigma)$ if

$$f(y) = \frac{1}{\sqrt{((2\pi)^n}} \exp(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(t-\mu))$$

Conditional MVN

Let $Y = (Y_1, Y_2)^T$, $\mu = (\mu_1, \mu_2)^T$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then the $[Y_1 | Y_2 = a] \sim MVN(\mu_{1|2}, \Sigma_{1|2})$ with

- $\mu_{1|2} = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$ and
- $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Covariance function

- Need a way to specify the covariance matrix Σ for any set of locations $(s_1, s_2, \ldots, s_n)$
- Σ has to be positive definite.
- Want locations close to have higher correlation that locations further away.

One valid correlation function:

Exponential correlation function

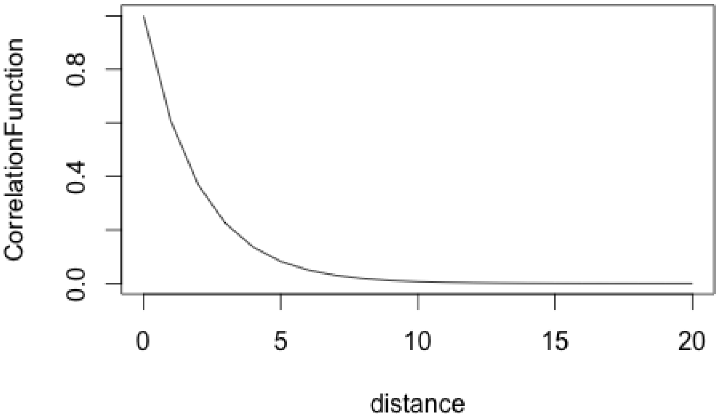
$$Corr(Y(s_1), Y(s_2)) = \exp(-d(s_1, s_2)/\theta_1)$$

where $d(s_1, s_2)$ is the (Euclidian) distance between s_1 and s_2.

Exponential covariance function :

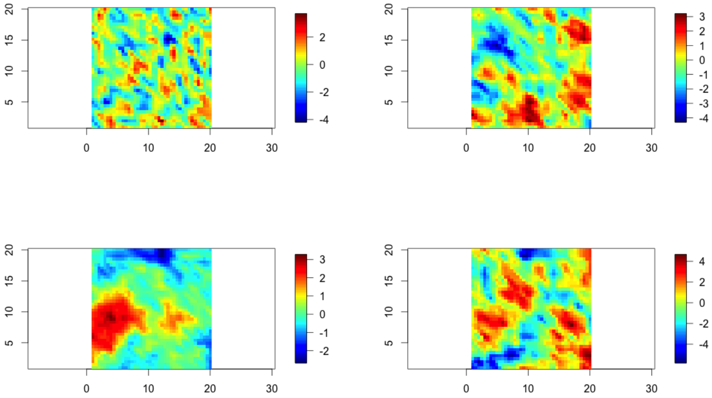
$$C(Y(s_1), Y(s_2)) = \sigma_1^2 Corr(Y(s_1), Y(s_2))$$

Exponential correlation function, $\theta = 2$



Examples samples from GRFs

How are these different?



Predictions

Now we assumes

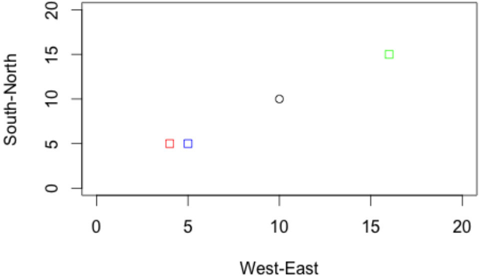
- Known parameters, i.e. mean μ and covariance function / Σ
- Perfect observations (observe $Y(s)$)

Example: Temperature outsid my home

- Want to predict, with uncertainty, the temperature at location s_0, i.e. $Y(s_0)$.
- Know the temperature at locations $s_1, s_2, \ldots s_p$.
- How to predict? Hint: Conditional MVN

Predictions, what design is best?

- Want to predict at $(10, 10)$ (black)
- Can observe $Y(s)$ at $(5, 5)$, $(4, 5)$ and $(16, 15)$.
- Which will you chose if you can chose 1 site?
- Which will you chose if you can chose 2 site?



Before Friday

- Read *stationary* and *isotropic* (page 34)
- Play with the code.