Introduction to STK 4150/9150 Environmental and spatial-temporal statistics

16 .January 2017

Odd Kolbjørnsen

Today

- Who are we?
- Book
- Form
- Introduction
- Scope
- Begin @ the end
- Statistical preliminaries

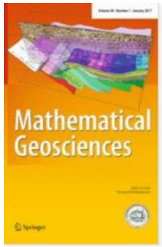
Odd Kolbjørnsen

NTNU

- Ph.D from NTNU in 2002
  - Nonlinear topics in the Bayesian approach to inverse problems
- 2002-2014: Norsk Regnesentral

NR Norsk Regnesentral NORWEGIAN COMPUTING CENTER

- 2010-2015: Associate Editor in Mathematical Geosciences



- 2014 - : Senior Exploration Analyst Lundin-Norway

Lundin Norway

Desired background

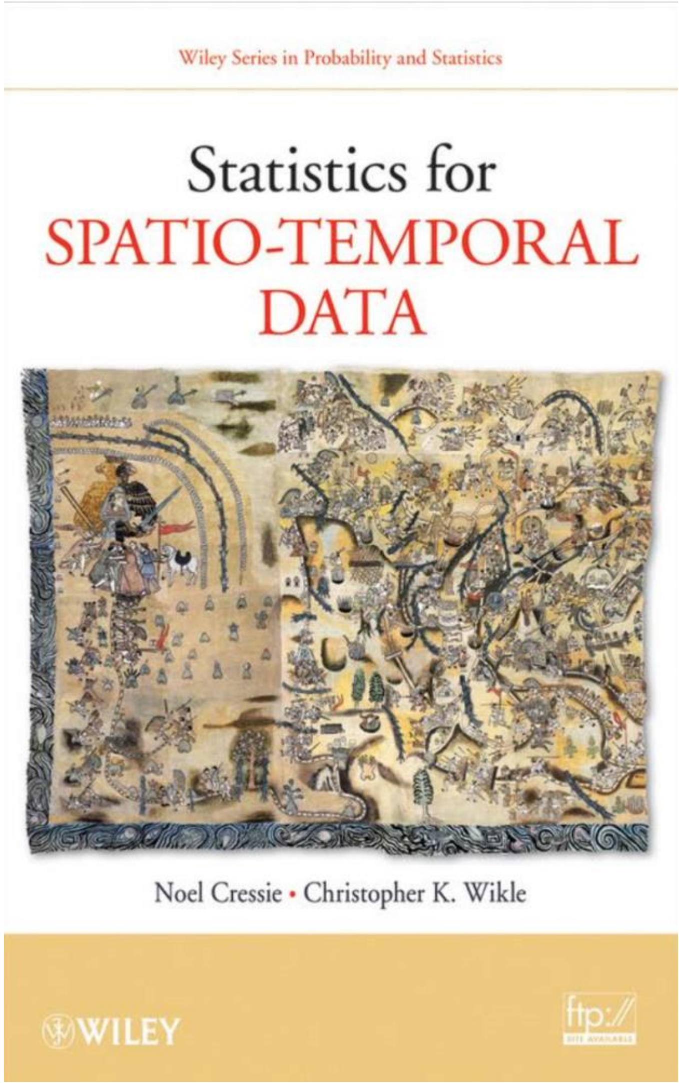
- General knowledge of quantitative methods/modelling
- General knowledge of mathematics (matrix/vector calculations)
- Probability theory
  - Expectations/Covariance matrices
  - Conditional densities/probabilities
  - Bayes theorem
- (Linear) Regression
- Maximum likelihood
- Knowledge of common distributions (normal, binomial, Poisson, Gamma, t)

STK 4150/9150

- Course made by:
  - Geir Storvik

- Lectured by
  - Odd Kolbjørnsen

- Book by:
  - Noel Cressie
  - Christopher K. Wikle


Wiley Series in Probability and Statistics
Statistics for
SPATIO-TEMPORAL
DATA
Noel Cressie • Christopher K. Wikle
WILEY
ftp://

Course form

- Lectures
- Weekly exercises
- Compulsory project (Given in May)
- Written exam (30.May 9:00-13:00)

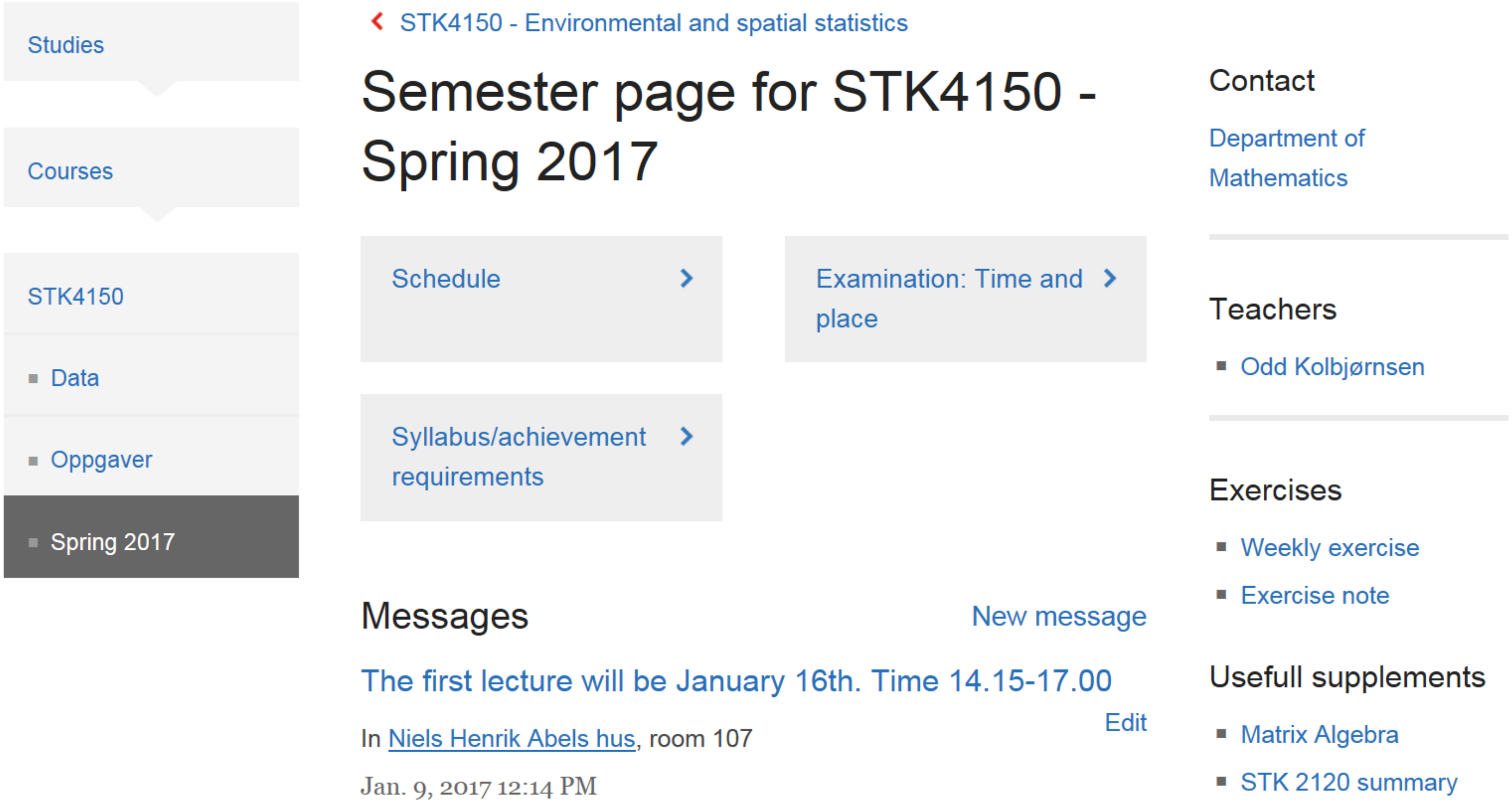
Grading

- Monday:

  14:15 – 16:00 Lectures

  16.15 – 17:00 Last week Exercise Q&A

http://www.uio.no/studier/emner/matnat/math/STK4150/v17/



STK4150 - Environmental and spatial statistics

Semester page for STK4150 - Spring 2017

Messages

The first lecture will be January 16th. Time 14.15-17.00

In Niels Henrik Abels hus, room 107

Jan. 9, 2017 12:14 PM

Contact

Department of Mathematics

Teachers

- Odd Kolbjørnsen

Exercises

- Weekly exercise
- Exercise note

Usefull supplements

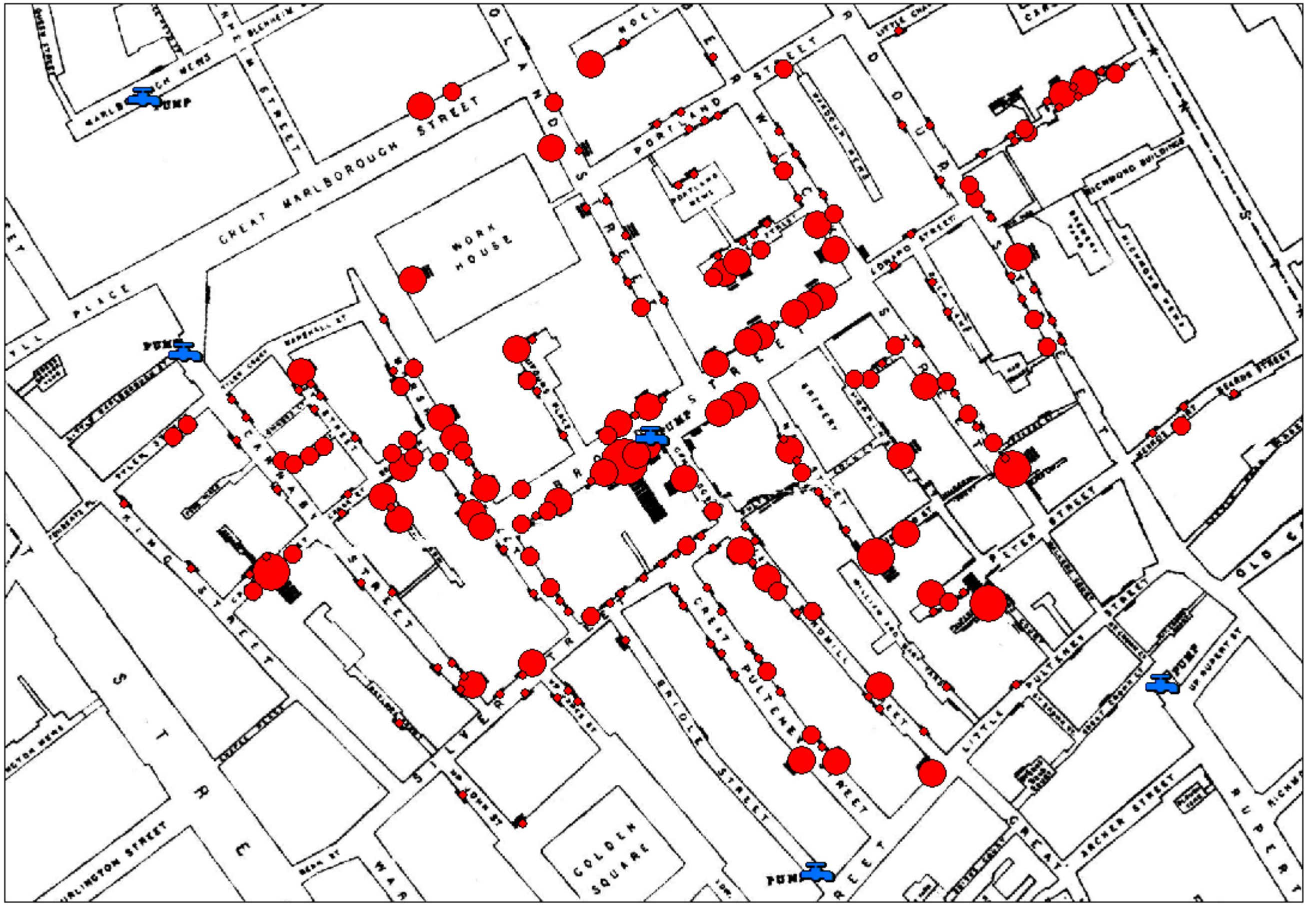
- Matrix Algebra
- STK 2120 summary

Environmental statistics (Wide definition)

- Meteorology (temperature, wind, humidity)
- Climate (CO2, Temperature)
- Pollution (ozone, sulfur,)
- Biological data (species, plants, Sustainable population management)
- Human data (Mortality, Disease surveillance)

John Snow on Cholera

1854 London outbreak of Cholera



First use of spatial statistics in medicine. **Disease surveillance**

Environmental statistics

- Variability in space and time
- Correlation in data
- Aim to infer cause- effect relationship
- Need **where** and **when!**
  - Recording of where and when is a powerful tool for obtaining covariates

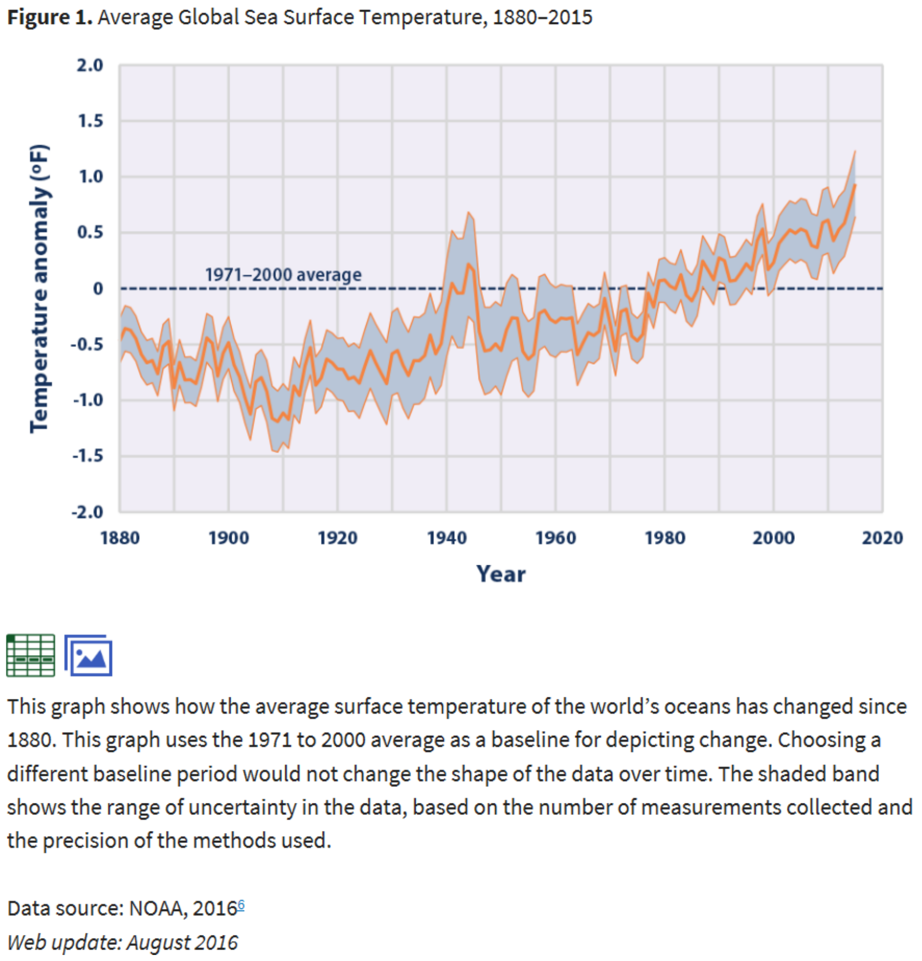


Figure 1. Average Global Sea Surface Temperature, 1880–2015

This graph shows how the average surface temperature of the world's oceans has changed since 1880. This graph uses the 1971 to 2000 average as a baseline for depicting change. Choosing a different baseline period would not change the shape of the data over time. The shaded band shows the range of uncertainty in the data, based on the number of measurements collected and the precision of the methods used.

Data source: NOAA, 2016[6]

Web update: August 2016

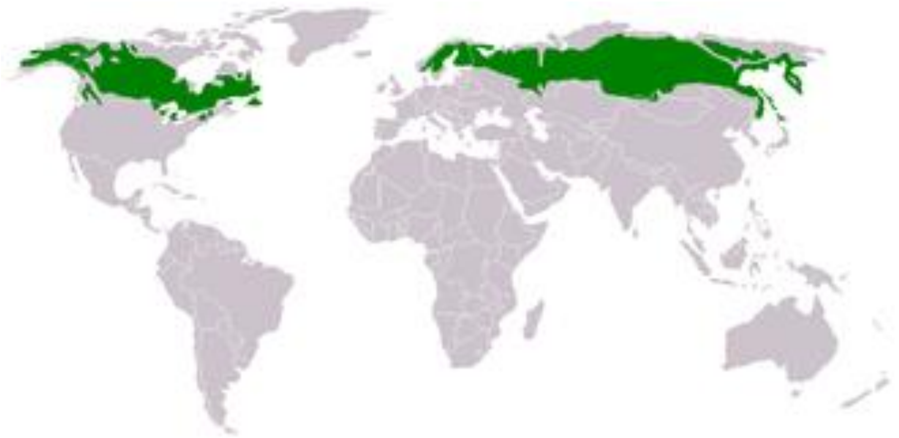
Change of support:

- Measure at locations (~ 0.01m^2)
- Average over globe~ (361 132 000 km^2)

https://www.youtube.com/watch?v=e0vj-0imOLw

Spatial statistics

- "Predict region of boreal forest (taiga) from satellite data.
- Explanatory variables:
  - Wetness index
  - Vegetation index
  - Temperature
  - Greenness index



- Boreality index
  - Number of boreal spices / total number of spices

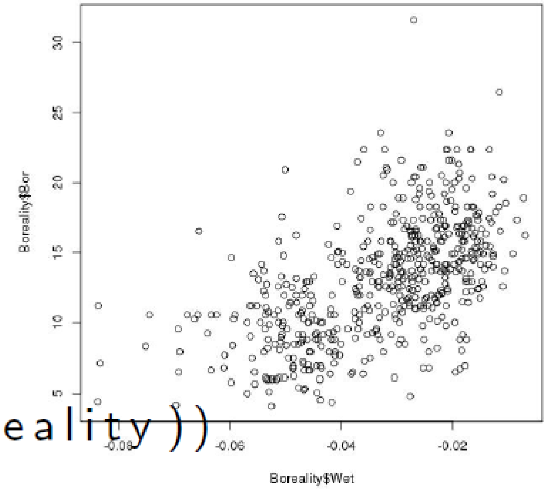
Spatial data - Example

Typical question

- Do an explanatory variable influence a response?
- Example:
  - Response: Number of species that belong to a set of boreal species divided by the total number of species *at a site*.
  - Explanatory variable: Index of wetness

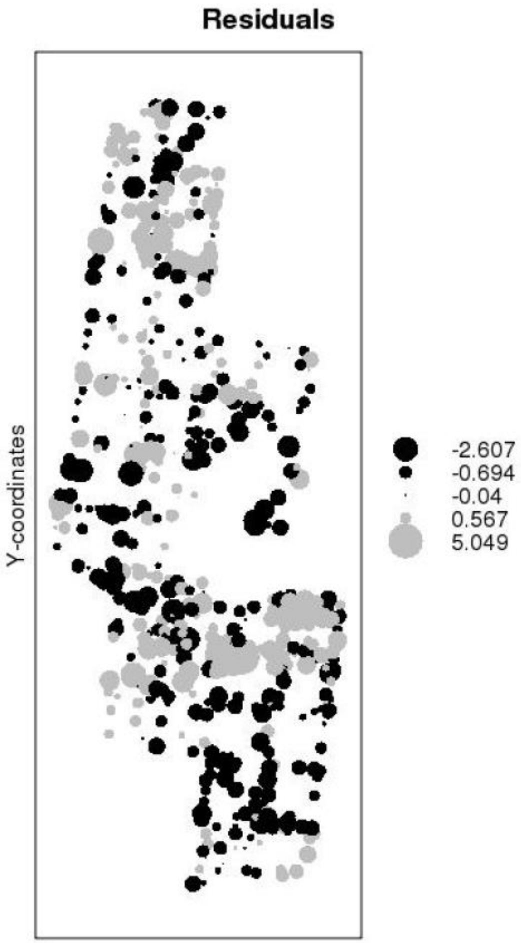
Possible approach

- Linear regression
- Assumes independence in residuals. Realistic here?



```
> summary(lm(Bor~Wet, data=Boreality))
 Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   18.4880      0.3787    48.82    <2e-16 ***
Wet          165.8036     10.5991    15.64    <2e-16 ***
```

Spatial correlation



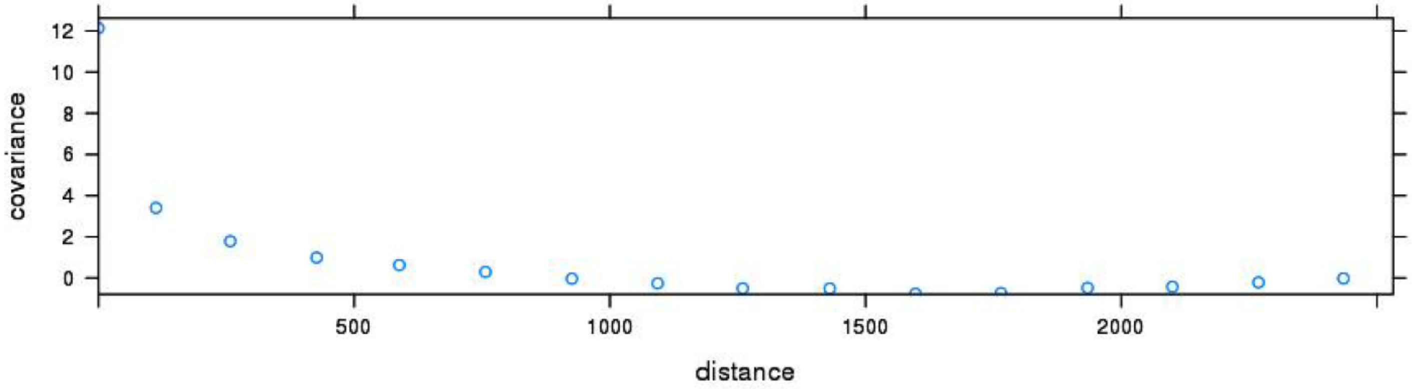
- Bubble plot of residuals
- Clustering of high and low values
  - Spatial correlation in the residual process
  - Missing explanatory variable with spatial characteristics

Spatial correlation - Covariance function

Covariance function for process $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{R}^2\}$:

$$\mathrm{Cov}[Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})] = C_Y(\mathbf{h})$$

Plot of estimate (Boreal data)



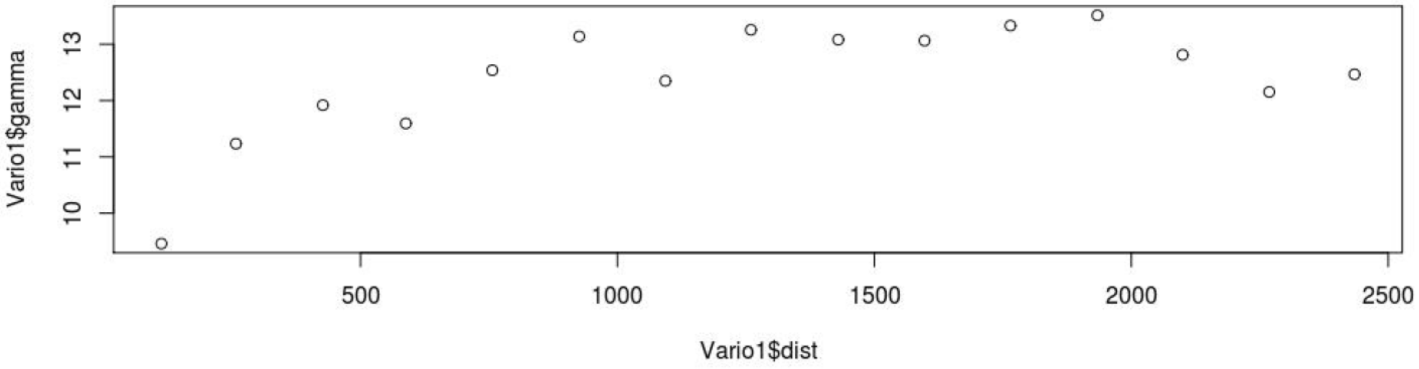
Spatial correlation - Variogram function

Alternative: Variogram function for process $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{R}^2\}$:

$$\begin{aligned}
\gamma(\mathbf{h}) =& 0.5\text{Var}[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})] \\
=& 0.5\{\text{Var}[Y(\mathbf{s}+\mathbf{h})] + \text{Var}[Y(\mathbf{s})] - 2\text{Cov}[Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})]\} \\
\overset{\mathbf{h}\text{ large}}{\approx}& 0.5\{\text{Var}[Y(\mathbf{s}+\mathbf{h})] + \text{Var}[Y(\mathbf{s})] \\
=& \text{Var}[Y(\mathbf{s})] \quad \text{if stationarity}
\end{aligned}$$

Note: $\text{Cov}[Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})] = C_Y(\mathbf{h}) = \gamma(\mathbf{0}) - \gamma(\mathbf{h})$

Plot of estimate (Boreal data)



What should we do with the boreal data?

- What goes wrong if we assumes independence?
  - Too certain on a large scale
    - Data redundancy (information content repeated)
  - Too uncertain on a short scale
    - Local refinement due to correlation (interpolation)

Correlation - does it matter?

Simulation example: X_i and Y_i are independent, $i = 1, ..., n$:
Model

$$Y_i | X_i = x_1 \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Assume $\beta_1 = 0$.
Wald test on $H_0 : \beta_1 = 0$ should reject hypothesis a fraction of $100\alpha\%$.
Repeated experiment 1000

- All x's and y's independent: Rejection 48 times
- Dependence:

$$\begin{aligned} X_i &= a x_{i-1} + \eta_i & a = 0.9 \\ Y_i &= b y_{i-1} + \varepsilon_i & b = 0.9 \end{aligned}$$

  Rejection 530 times!
- Dependence: Effective number of observations much lower

Modelling temporal dependence

Assume $Y_1, ..., Y_n$ time series
Simple model

$$Y_1 \sim N(\mu, \sigma^2/(1-a^2))$$
$$Y_t = \mu + a(y_{t-1} - \mu) + \delta_t, \quad \delta_t \sim N(0, \sigma^2), \quad t = 2, ...,$$

Autoregressive model of order 1, AR(1).

- $Y_t \sim N(\mu, \sigma^2/(1-a^2)$ for all t!
- Independence if $a = 0$

Simulation example, 1000 simulations, $\alpha = 0.05$:

- Ignoring dependence: Rejected 530 times
- Correction for dependence using AR(1) model: Rejected 56 times.

AR(1) - alternative formulation

$$\begin{aligned} Y_1 \sim & N(\mu, \sigma^2/(1-a^2)) \\ Y_t = & \mu + a(y_{t-1} - \mu) + \delta_t, \quad \delta_t \sim N(0, \sigma^2) \end{aligned}$$

imply $\mathbf{Y} = (Y_1, ..., Y_n)$ is multivariate Gaussian where

$$\begin{aligned} E[Y_t] &= \mu \\ \text{var}[Y_t] &= \sigma^2/(1-a^2) \\ \text{cor}[Y_t, Y_{t+\tau}] &= a^\tau = C_Y(\tau) \end{aligned}$$

Can write

$$\begin{aligned} \mathbf{Y} \sim & N(\mu\mathbf{1}, \mathbf{\Sigma}) \\ \Sigma_{tt} = & \text{Var}[Y_t] \\ \Sigma_{t,t+\tau} = & \text{Cov}[Y_t, Y_{t+\tau}] = \Sigma_{tt}\text{cor}[Y_t, Y_{t+\tau}] \end{aligned}$$

AR(1) - extensions

The alternative formulation

$$\begin{aligned}\mathbf{Y} \sim & N(\mu\mathbf{1}, \boldsymbol{\Sigma}) \\ \Sigma_{tt} = & \sigma^2/(1-a^2) \\ \Sigma_{t,t+\tau} = & \Sigma_{tt} a^\tau = \Sigma_{tt} C_Y(\tau)\end{aligned}$$

allow for extensions:

- Including covariates

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \boldsymbol{\Sigma})$$

- Other correlation structures structures

$$\Sigma_{t,t+\tau} = \Sigma_{tt} C_Y(\tau)$$

- Higher dimensional processes

Spatial processes

Assume now $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{R}^2\}$ is a *spatial process*
Let $\mathbf{Y} = (Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))$
Assume $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{\Sigma})$ where

$$\text{cor}[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = C_Y(|\mathbf{s}_i - \mathbf{s}_j|)$$

Example boreal species:

- Using $C_Y(|\mathbf{s}_i - \mathbf{s}_j|) = \exp(-|\mathbf{s}_i - \mathbf{s}_j|/\theta)$
- Estimation of β_1

	Estimate	Std.Error	t-value	P-value
Independence	165.804	10.60	15.64	0
Spatial dependence	75.432	13.54	5.57	4.05e-08

Theoretical considerations for spatial processes

Assume now $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{R}^2\}$ is a spatial process
Let $\mathbf{Y} = (Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))$
Assume $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{\Sigma})$ where

$$\text{cor}[Y(\mathbf{s}_i), Y(\mathbf{s}_j)] = C_Y(|\mathbf{s}_i - \mathbf{s}_j|)$$

Problems:

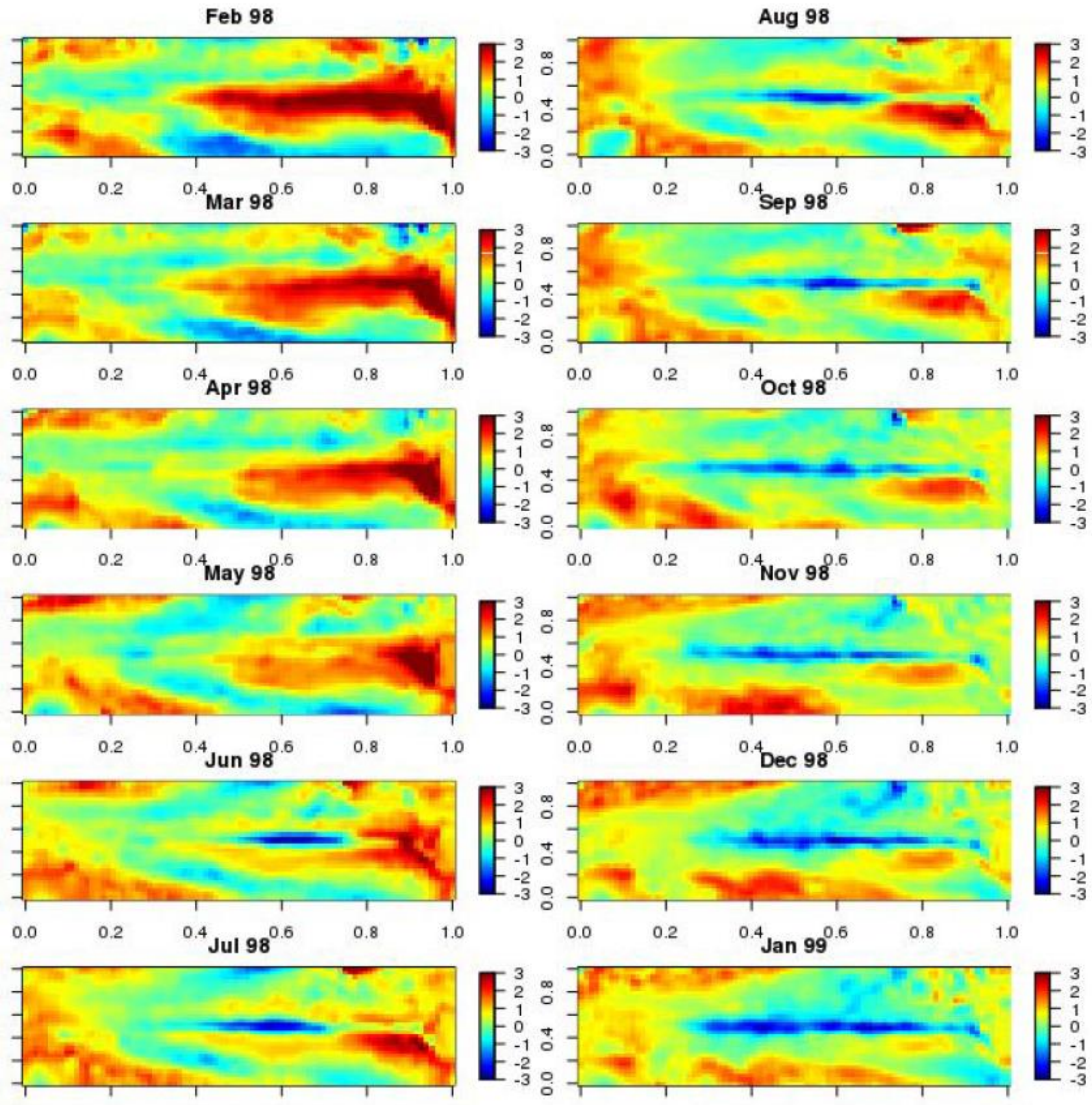
- Multivariate Gaussian process for a finite number of positions
- Want a simultaneous distribution for a continuous set of variables
- Possible to define distribution for $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{R}^2\}$ through all possible $\mathbf{Y}$?
- What restrictions are needed on $C_Y(\cdot)$?

Regional data

- Often, $Y(\mathbf{s})$ not available, but rather a summary for a region is given
- Regional data require special models
- Main idea: Neighbor regions have similar features
- Main approach: Markov models (extensions of models from STK2130)

Spatio-temporal Data

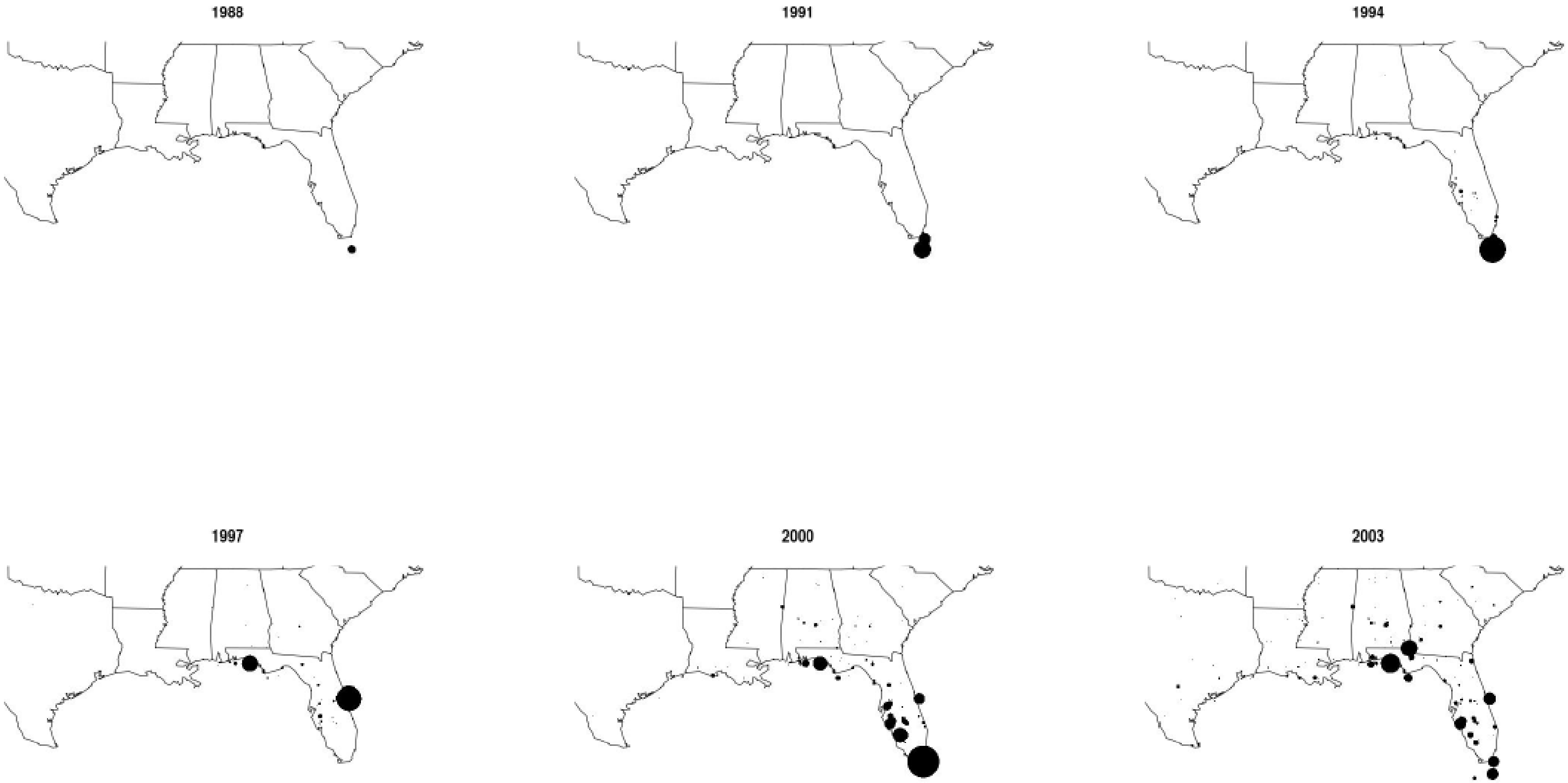
El Nino



Surface temperature in Pacific region from February 1998 to January 1999

La Nina

Invasive spread of Eurasian Collared Dove



North American Breeding Bird Survey data on the Eurasian Collared Dove (Streptopelia decaocto) for the years 1986 - 2008.

Spatio-temporal processes

Can in principle do the same as for spatial processes:
Let $\mathbf{Y} = (Y(\mathbf{s}_1, t_1), ..., Y(\mathbf{s}_n, t_n))$
Assume $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{\Sigma})$ where

$$\text{cor}[Y(\mathbf{s}_i, t_i), Y(\mathbf{s}_j, t_j)] = C_Y(|\mathbf{s}_i - \mathbf{s}_j|, |t_i - t_j|)$$

As for time-series, there is an alternative in writing the model as a dynamical process:

$$Y_t(\mathbf{s}) = \mathcal{M}_t(\mathbf{s}, \mathbf{Y}_{t-1}(\cdot)) + \delta_t(\mathbf{s})$$

$\mathcal{M}_t$ can be based on

- statistical properties in data and/or
- physical knowledge of the process (e.g differential equations)

Observation errors

Typically $\{\mathbf{Y}(\mathbf{s}, t)\}$, the physical/biological process of interest, is not directly observed but

$$\mathbf{Z}_t = \mathbf{Y}_t + \varepsilon_t, \quad \varepsilon_t \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$$

Further, there are unknown parameters involved

Hierarchical structure

$$\begin{tikzcd} & & \boldsymbol{\theta}_D \arrow[d] \\ \boldsymbol{\theta}_M \arrow[r] & \mathbf{Y} \arrow[r] & \mathbf{Z} \end{tikzcd}$$

Estimation

If $\mathbf{Y} = \mathbf{y}$ is directly observed, likelihood

$$L(\boldsymbol{\theta}_M) = f(\mathbf{y}|\boldsymbol{\theta}_M) = \prod_{t=1}^{T} f(\mathbf{y}_t|\mathbf{y}_{<t};\boldsymbol{\theta}_M)$$

Typically multivariate Gaussian densities
(but more complicated if non-linearities are included)
$\mathbf{Z}$ observed data, $\mathbf{Y}$ latent structure, likelihood

$$L(\boldsymbol{\theta}_M,\boldsymbol{\theta}_D) = f(\mathbf{z}|\boldsymbol{\theta}_M,\boldsymbol{\theta}_D) = \int_{\mathbf{y}} f(\mathbf{z}|\mathbf{y};\boldsymbol{\theta}_D) f(\mathbf{y}|\boldsymbol{\theta}_M) d\mathbf{y}$$

High-dimensional integral
Dynamic modelling: Allow for sequential calculation of likelihood.

(Bayesian) Hierarchical modelling

Hierarchical model

	Variable	Densities	Notation in book
Data model:	$\mathbf{Z}$	$p(\mathbf{Z}\vert\mathbf{Y},\boldsymbol{\theta})$	$[\mathbf{Z}\vert\mathbf{Y},\boldsymbol{\theta}]$
Process model:	$\mathbf{Y}$	$p(\mathbf{Y}\vert\boldsymbol{\theta})$	$[\mathbf{Y}\vert\boldsymbol{\theta}]$
Parameter:	$\boldsymbol{\theta}$		

Simultaneous model: $p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})$
Marginal model: $p(\mathbf{z}|\boldsymbol{\theta}) = \int_{\mathbf{y}} p(\mathbf{z},\mathbf{y}|\boldsymbol{\theta})d\mathbf{y}$

Bayesian approach: Include model on $\boldsymbol{\theta}$

	Variable	Densities	Notation in book
Data model:	$\mathbf{Z}$	$p(\mathbf{Z}\vert\mathbf{Y},\boldsymbol{\theta})$	$[\mathbf{Z}\vert\mathbf{Y},\boldsymbol{\theta}]$
Process model:	$\mathbf{Y}$	$p(\mathbf{Y}\vert\boldsymbol{\theta})$	$[\mathbf{Y}\vert\boldsymbol{\theta}]$
Parameter model:	$\boldsymbol{\theta}$	$p(\boldsymbol{\theta})$	$[\boldsymbol{\theta}]$

Simultaneous model: $p(\mathbf{y},\mathbf{z},\boldsymbol{\theta})$
Marginal model: $p(\mathbf{z}) = \int_{\boldsymbol{\theta}} \int_{\mathbf{y}} p(\mathbf{z},\mathbf{y}|\boldsymbol{\theta})d\mathbf{y}d\boldsymbol{\theta}$

Spatio-temporal data and uncertainties

- Space-Time: The next Frontier
- Time-data give possibilities for causation
- Spatio-temporal data contain many sources of uncertainty
- Statistics: Science of Uncertainty!
    - Uncertainty in data
    - Uncertainty in models

Course @ a slide

Prelim/Jan

- Statistics preliminaries (Chapter 2)
- Temporal processes (Chapter 3)
  - Deterministic models
  - Stochastic models
  - Spectral representation

January-Mach

- Spatial processes
  - Geostatistical processes, sec 4.1
  - Lattice processes, sec 4.2
  - Point processes, sec 4.3

Mach-May

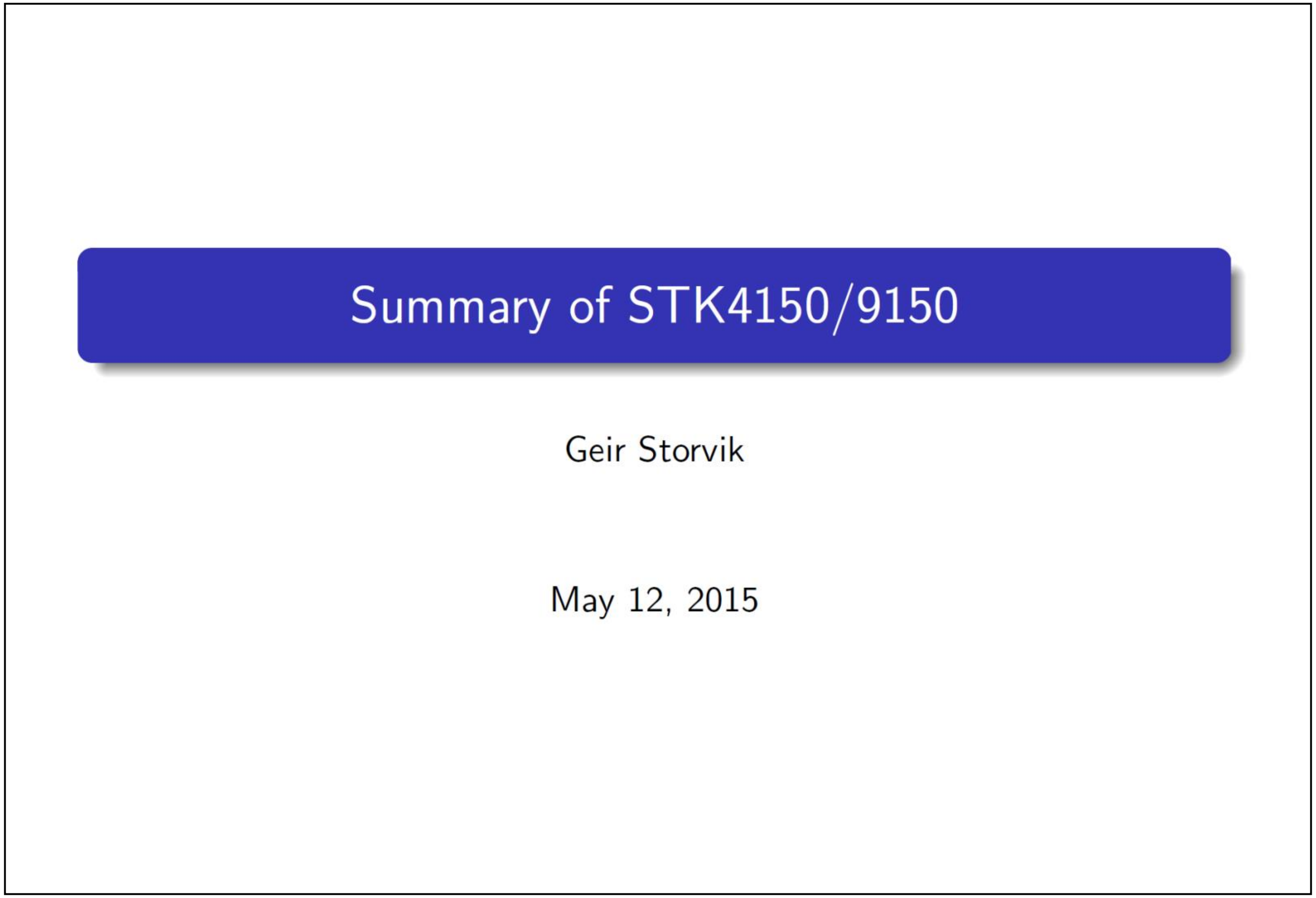
- Spatio-temporal processes
  - Exploratory methods (Chapter 5)
  - Models (Chapter 6)
  - Hierarchical models (Chapters 7 and 8)

- Focus on
  - Modeling
  - Analysis in practice (using R)
  - Theoretical aspects

Computations

- Inference for space time-processes difficult due to
  - Complex models
  - Latent processes (likelihood not directly specified)
  - Huge amounts of data
- Possibilities
  - Use available software
    - Typically for specific models
  - Monte Carlo methods
    - Preferred method in the book, described in sec 2.3
  - Using integrated nested Laplace approximation (INLA)
    - Flexible software for latent Gaussian processes
    - Will be used throughout the course
  - Computation within R

Begin @ the end



Summary of STK4150/9150

Geir Storvik

May 12, 2015

Spatio-temporal processes

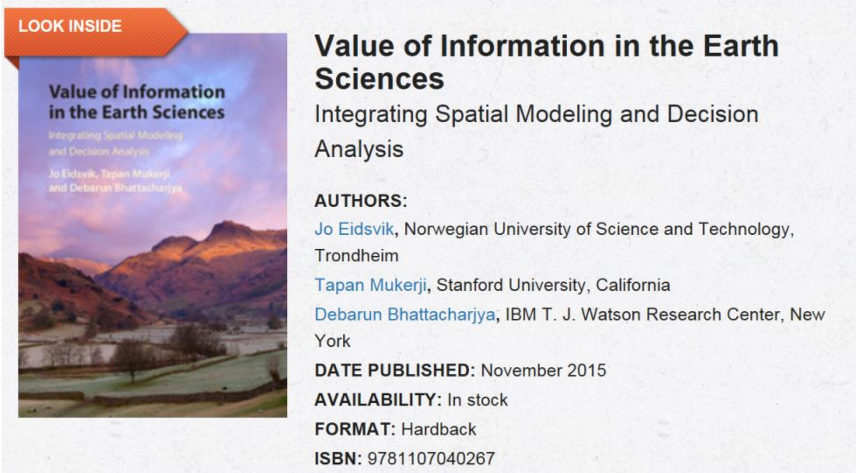
Aims

- Prediction/forecasting
- Learning about the processes

Challenges

- Complex dependence structures
- Huge amounts of data

Not focus on decisions theory.
If interested:


LOOK INSIDE
Value of Information in the Earth Sciences
Integrating Spatial Modeling and Decision Analysis
Jo Eidsvik, Tapan Mukerji and Debarun Bhattacharjya

Value of Information in the Earth Sciences

Integrating Spatial Modeling and Decision Analysis

AUTHORS:

Jo Eidsvik, Norwegian University of Science and Technology, Trondheim

Tapan Mukerji, Stanford University, California

Debarun Bhattacharjya, IBM T. J. Watson Research Center, New York

DATE PUBLISHED: November 2015

AVAILABILITY: In stock

FORMAT: Hardback

ISBN: 9781107040267

The importance of taking dependence into account

- Spatio-temporal data often contain dependence
- Ignoring dependence can give wrong results (too small uncertainty measures)
- Inclusion of dependence complicates
  - Modeling
  - Inference/computation
- Inclusion of covariates can reduce dependence
- Dependence can improve prediction/forecasting

Exploratory data analysis (EDA)

- Always do EDA!
- Various tools:
  - Direct plots of data/animations
  - Autocorrelation function (time series)
  - Variograms (spatial data)
  - Moran's I, STI (tests for dependence, spatial/spatio-temporal)
  - Empirical orthogonal functions (spatio-temporal)

Time series

- Differential equations (deterministic/stochastic) $Y_t = \mathcal{M}(Y_{t-1}) + W_t$
- ARMA models: $Y_t = \sum_{k=1}^{p} \alpha_k Y_{t-k} + \sum_{l=0}^{q} \beta_l W_{t-l}$
- Most useful: AR(1). NB: Sparse precision matrix
- Non-linear models - related to differential equations (?)

Spatial processes

- Geostatistical modeling - covariance functions
  - Need non-negative definite covariance matrix for any finite collection of spatial points
  - Challenge to construct legal covariance functions
    - Stationarity/isotrophy simplifies modeling *and* inference
  - Prediction through kriging
    - $\mathbf{x}_1|\mathbf{x}_2 \sim N(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$
- CAR/MRF-models - sparse precision matrices
  - $p(Y(\mathbf{s}_i)|\mathbf{Y}_{-i}) = p(Y(\mathbf{s}_i)|Y(\mathbf{s}_j), j \in \mathcal{N}_i)$
  - Simplifies modeling/inference
- Hierarchical modeling allow for non-Gaussian observations

Spatio-temporal processes

- Possible to define time as extra “spatial” dimension
  - More difficult to construct valid covariance functions
- Advantages in modeling time dynamically
  - Partial differential equations - stochastic versions
  - Time series approaches $\mathbf{Y}_t = \mathcal{M}(\mathbf{Y}_{t-1}) + \mathbf{W}_t$

Spatio-temporal processes - simplifications

- Spatio-temporal data involve a huge number of observations/variables
- Difficult to model, difficult to process
- Need simplifications
  - Additive models $Y_t(\mathbf{s}) = \mathbf{X}_t(\mathbf{s})^T \boldsymbol{\beta} + \alpha_t + \delta(\mathbf{s}) + \varepsilon_t(\mathbf{s})$
  - Separable covariance functions
  - Dimension reductions: $\mathbf{Y}_t = \boldsymbol{\Phi}\boldsymbol{\alpha}_t$ with dynamics in lower dimensional $\boldsymbol{\alpha}_t$.

Hierarchical modeling

- Distinguish between
  - process model
  - observation model
- Advantages
  - Simpler modeling
  - Allow for non-Gaussian observations
- Can include model for parameters in a Bayesian framework

Computational tools

- Kalman filtering (linear Gaussian models)
- MCMC
- INLA
- (Particle filters)
- (EM algorithm)

Challenges

- Efficient generic tools for models outside the INLA framework
- Tools for model selection
- More integration of physical and statistical models