

STK4150/9150 - Environmental and Spatial Statistics

Introduction

Thordis Thorarinsdottir

2024-01-17

Environmental data includes, among other things,

- meteorological data (temperature, wind, humidity)
- pollution data (ozone, sulfur, carbon monoxide)
- biological data (species, plants)
- human data (mortality, health)

Environmental data includes, among other things,

- meteorological data (temperature, wind, humidity)
- pollution data (ozone, sulfur, carbon monoxide)
- biological data (species, plants)
- human data (mortality, health)

Common features

- vary in space and time
- complicated correlation structures

Environmental data includes, among other things,

- meteorological data (temperature, wind, humidity)
- pollution data (ozone, sulfur, carbon monoxide)
- biological data (species, plants)
- human data (mortality, health)

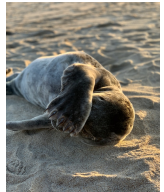
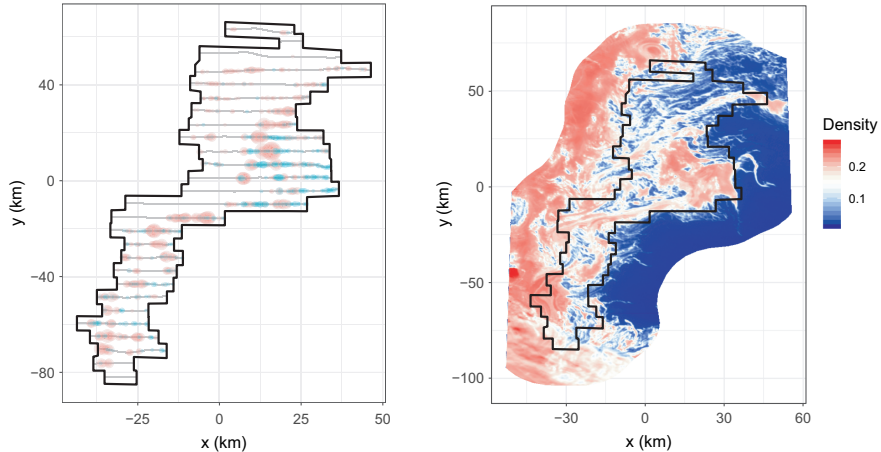
Common features

- vary in space and time
- complicated correlation structures

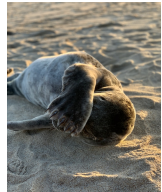
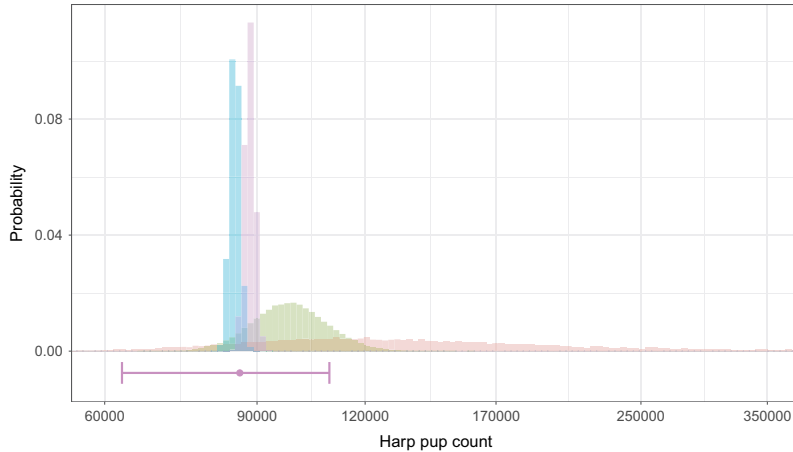
Typical aims

- understand relationships (inference)
- prediction

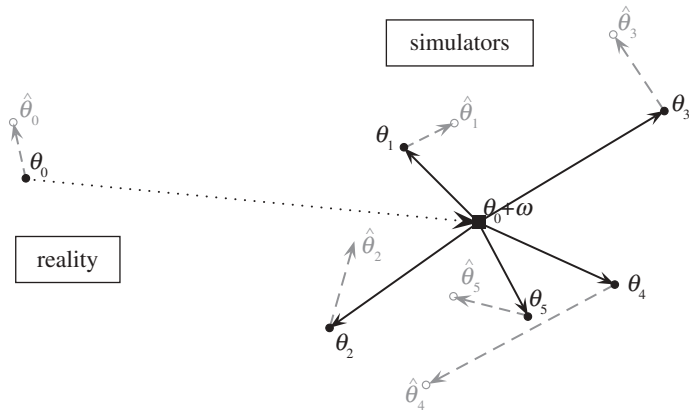
Example 1: How many seal pups? (Jullum *et al.*, 2020)



Example 1: How many seal pups? (Jullum *et al.*, 2020)

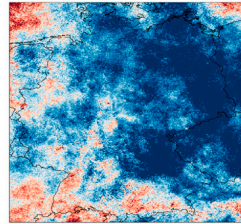
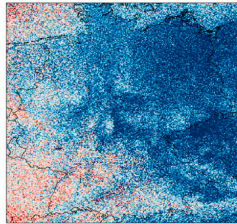
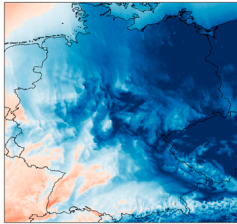


More generally: Combining information (Chandler, 2013)



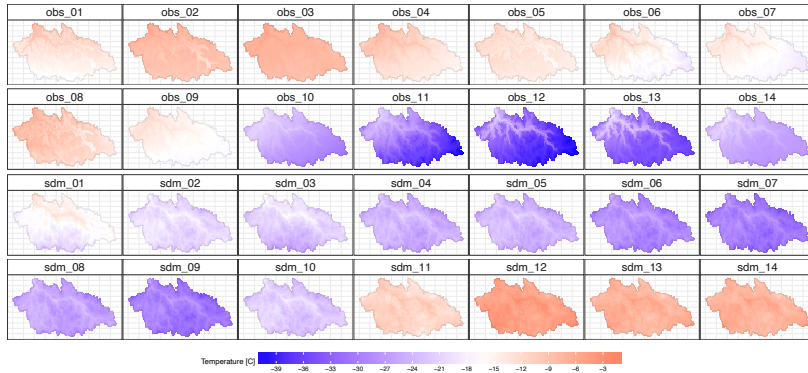
Uncertainty in data and uncertainty in models

Example 2: Temperature forecasts (Feldmann *et al.*, 2015)



NWP forecast → univariate bias-correction → multivariate bias-correction

Example 3: Temperature in space and time (Yuan *et al.*, 2019)



Course content

- Statistical preliminaries (Chapter 1, 2)
- Temporal processes (Chapter 3.1, 3.3, 3.4)
- Spatial geostatistical processes (Chapter 4.1)
- Spatial lattice processes (Chapter 4.2)
- Spatial point processes (Chapter 4.3)
- Exploring spatio-temporal data (Chapter 5.1, 5.3)
- Spatio-temporal statistical models (Chapter 6.1, 6.2, 6.4)
- Univariate extremes
- Multivariate & spatial extremes

Good to know

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

Course format

- Thursday morning: Lecture
- Wednesday afternoon: Exercises

Course format

- Thursday morning: Lecture
- Wednesday afternoon: Exercises

Course schedule is available at

www.uio.no/studier/emner/matnat/math/STK4150/v24/timeplan/index.html

Course format

- Thursday morning: Lecture
- Wednesday afternoon: Exercises

Course schedule is available at

www.uio.no/studier/emner/matnat/math/STK4150/v24/timeplan/index.html

Note: No lectures in week 8 (February 21 & 22) and week 17 (April 24 & 25)

Course evaluation

- Mandatory assignment (due on April 25?)
- Oral or written exam (June 5)

Statistical preliminaries

- Multivariate Gaussian distribution
- Conditional probabilities and hierarchical modelling
- Bayesian hierarchical modelling
- Model evaluation
- Computation

Multivariate Gaussian distribution

Central in modelling dependence

Multivariate Gaussian distribution

Central in modelling dependence

Definition: A random vector

$$\mathbf{Y} = (Y_1, \dots, Y_m)^T$$

is said to have the *multivariate normal distribution* if every *linear combination* of its components

$$V = a_1 Y_1 + \dots + a_m Y_m$$

is *normally distributed*. That is, for any constant vector $\mathbf{a} \in \mathbb{R}^m$, the random variable $V = \mathbf{a}^T \mathbf{Y}$ has a univariate normal distribution.

Multivariate Gaussian distribution

We write $\mathbf{Y} \sim \text{Gau}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{Y}$ has density

$$[\mathbf{Y}] = (2\pi)^{-m/2} |\boldsymbol{\Sigma}|^{-1/2} \exp[-(1/2)(\mathbf{Y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu})]$$

Multivariate Gaussian distribution

We write $\mathbf{Y} \sim \text{Gau}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{Y}$ has density

$$[\mathbf{Y}] = (2\pi)^{-m/2} |\boldsymbol{\Sigma}|^{-1/2} \exp[-(1/2)(\mathbf{Y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu})]$$

Common application that allows for flexible multivariate modelling:

- $\mathbf{Z}$ observations
- $\mathbf{Y}$ latent variables, assumed multivariate Gaussian

$$[\mathbf{Z}|\mathbf{Y}] = \prod_i [Z_i|Y_i]$$
$$[\mathbf{Z}] = \int_{\mathbf{Y}} [\mathbf{Z}|\mathbf{Y}] [\mathbf{Y}] d\mathbf{Y}$$

Conditional probabilities and hierarchical modelling

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

$$[\boldsymbol{\theta}, \mathbf{Y}, \mathbf{Z}] = [\boldsymbol{\theta}] \times [\mathbf{Y}|\boldsymbol{\theta}] \times [\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}]$$

Conditional probabilities and hierarchical modelling

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

$$[\boldsymbol{\theta}, \mathbf{Y}, \mathbf{Z}] = [\boldsymbol{\theta}] \times [\mathbf{Y}|\boldsymbol{\theta}] \times [\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}]$$

- Model built up by several conditional probabilities/densities
- Easier than specifying a multivariate distribution on $(\boldsymbol{\theta}, \mathbf{Y}, \mathbf{Z})$ directly
- Model for $\mathbf{Y}$ often build up by physical/biological knowledge

Bayesian approach

Assume θ is *stochastic* with prior distribution $[\theta]$ which specifies our prior knowledge about θ .

Knowledge is *updated* when data $\mathbf{Z}$ is obtained (Bayes theorem)

$$[\mathbf{Y}, \theta | \mathbf{Z}] = \frac{[\mathbf{Z} | \mathbf{Y}, \theta] [\mathbf{Y}, \theta]}{\int_{\theta, \mathbf{Y}} [\theta, \mathbf{Y}, \mathbf{Z}] d\mathbf{Y} d\theta} = \frac{[\mathbf{Z} | \mathbf{Y}, \theta] [\mathbf{Y} | \theta] [\theta]}{[\mathbf{Z}]} \propto [\mathbf{Z} | \mathbf{Y}, \theta] [\mathbf{Y} | \theta] [\theta]$$

Bayesian approach

Can handle uncertainty in parameters coherently

Bayes		Frequentist
$\mathbb{E}[\mathbf{Z}^* \mathbf{Z}]$	$= \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{E}_{\mathbf{Z}^*}[\mathbf{Z}^* \boldsymbol{\theta}, \mathbf{Z}]]$	$\approx \mathbb{E}[\mathbf{Z}^* \hat{\boldsymbol{\theta}}, \mathbf{Z}]$
$\text{Var}[\mathbf{Z}^* \mathbf{Z}]$	$= \mathbb{E}_{\boldsymbol{\theta}}[\text{Var}_{\mathbf{Z}^*}[\mathbf{Z}^* \boldsymbol{\theta}, \mathbf{Z}]] + \text{Var}_{\boldsymbol{\theta}}[\mathbb{E}_{\mathbf{Z}^*}[\mathbf{Z}^* \boldsymbol{\theta}, \mathbf{Z}]]$	$\approx \text{Var}[\mathbf{Z}^* \hat{\boldsymbol{\theta}}, \mathbf{Z}]$

Multiple data sources

Assume several data sets $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$, all giving information about $\mathbf{Y}$, the process of interest.

$\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ may have a complex dependence structure due to their common relation to $\mathbf{Y}$.

Hierarchical modelling:

$$[\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3 | \mathbf{Y}, \boldsymbol{\theta}_D] = [\mathbf{Z}_1 | \mathbf{Y}, \boldsymbol{\theta}_{D,1}] \times [\mathbf{Z}_2 | \mathbf{Y}, \boldsymbol{\theta}_{D,2}] \times [\mathbf{Z}_3 | \mathbf{Y}, \boldsymbol{\theta}_{D,3}]$$

i.e. *conditional independence*

What is a good model?

- Goodness-of-fit: Does the model fit the data we have?
- Predictive performance: Can the model predict new data?
- Scientific interpretation: Does the model make sense?

Computation

Inference for space time-processes can be challenging due to

- Complex models
- Latent processes (likelihood not directly specified)
- Large data sets

Computation

Inference for space time-processes can be challenging due to

- Complex models
- Latent processes (likelihood not directly specified)
- Large data sets

We will use R and discuss several different approaches and packages, depending on the type of data and model

References

Chandler, R. E. (2013). Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1991), 20120388.

Jullum, M., Thorarinsdottir, T., & Bachl, F. E. (2020). Estimating seal pup production in the Greenland Sea by using Bayesian hierarchical modelling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 69(2), 327-352.

Feldmann, K., Scheuerer, M., & Thorarinsdottir, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review*, 143(3), 955-971.

Yuan, Q., Thorarinsdottir, T. L., Beldring, S., Wong, W. K., Huang, S., & Xu, C. Y. (2019). New approach for bias correction and stochastic downscaling of future projections for daily mean temperatures to a high-resolution grid. *Journal of Applied Meteorology and Climatology*, 58(12), 2617-2632.