Applied Geodata Science

Benjamin Stocker (lead), Koen Hufkens (contributing), Pepa Aran (contributing), Pascal Schneid

2022-11-22

Contents

About this book **5**

Course plan **7**

1 Getting started **9**
1.1 Learning objectives 9
1.2 Tutorial 9
1.3 Exercises 9
1.4 Solutions 9

2 Programming primers **11**
2.1 Learning objectives 11
2.2 Tutorial 11
2.3 Exercises 11
2.4 Solutions 11

3 Data wrangling **13**
3.1 Learning objectives 13
3.2 Tutorial 13
3.3 Exercises 13
3.4 Solutions 13

4 Data visualisation **15**
4.1 Learning objectives 15
4.2 Tutorial 15
4.3 Exercises 15
4.4 Solutions 15

5 Data variety **17**
5.1 Learning objectives 17
5.2 Tutorial 17
5.3 Exercises 17
5.4 Solutions 17

6 Code management **19**
6.1 Learning objectives . . . 19
6.2 Tutorial . . . 19
6.3 Exercises . . . 19
6.4 Solutions . . . 19

7 Open science practices **21**
7.1 Learning objectives . . . 21
7.2 Tutorial . . . 21
7.3 Exercises . . . 21
7.4 Solutions . . . 21

8 Regression **23**
8.1 Learning objectives . . . 23
8.2 Tutorial . . . 23
8.3 Exercises . . . 23
8.4 Solutions . . . 23

9 Supervised machine learning **25**
9.1 Learning objectives . . . 25
9.2 Tutorial . . . 25
9.3 Exercises . . . 25
9.4 Solutions . . . 25

10 Random forest **27**
10.1 Learning objectives . . . 27
10.2 Tutorial . . . 27
10.3 Exercises . . . 27
10.4 Solutions . . . 27

11 Neural networks **29**
11.1 Learning objectives . . . 29
11.2 Tutorial . . . 29
11.3 Exercises . . . 29
11.4 Solutions . . . 29

12 Interpretable machine learning **31**
12.1 Learning objectives . . . 31
12.2 Tutorial . . . 31
12.3 Exercises . . . 31
12.4 Solutions . . . 31

About this book

This book accompanies the course(s) *Applied Geodata Science*, taught at the Institute of Geography, University of Bern.

The course introduces the typical data science workflow using various examples of geographical and environmental data. With a strong hands-on component and a series of input lectures, the course introduces the basic concepts of data science and teaches how to conduct each step of the data science workflow. This includes the handling of various data formats, the formulation and fitting of robust statistical models, including basic machine learning algorithms, the effective visualisation and communication of results, and the implementation of reproducible workflows, founded in Open Science principles. The overall course goal is to teach students to tell a story with data.

Course plan

1. Getting started
2. Programming primer
3. Data wrangling
4. Data visualisation
5. Data variety
6. Code management
7. Open Science practice

   MILESTONE 1: Communicating a reproducible workflow (→ LO1)

8. Regression
9. Supervised machine learning fundamentals
10. Random Forest
11. Neural Networks
12. Interpretable machine learning
13. Unsupervised machine learning

    MILESTONE 2: Identify patterns and demonstrate how explained (→ LO2)

Chapter 1

Getting started

Chapter lead author: Pepa Aran

TBC

Contents:

- Lecture (Beni): Data revolution, opportunities, challenges; explain relevance and why new methods are required
- installing environment
- workspace management
- R, RStudio
- R libraries, other libraries and applications

1.1 Learning objectives

1.2 Tutorial

1.3 Exercises

1.4 Solutions

Chapter 2

Programming primers

Chapter lead author: Pepa Aran

TBC

Contents:

- Lecture (Beni): Models and data
- Base R
- variables, classes
- data frames
- loops
- conditional statements
- functions
- input and output
- intro to visualisation
- Performance assessment: link to my exercise, link to Dietze exercise

2.1 Learning objectives

2.2 Tutorial

2.3 Exercises

2.4 Solutions

Chapter 3

Data wrangling

Chapter lead author: Benjamin Stocker

Contents:

- Lecture (Beni): Tidy data, "bad" data
- Data frame manipulations with tidyverse
- Tidy data
- Dealing with missingness, bad data, outliers
- Imputation (note also imputation as part of the modelling workflow)
- Performance assessment: **CAT 1,** link, Make table tidy

3.1 Learning objectives

3.2 Tutorial

3.3 Exercises

3.4 Solutions

Chapter 4

Data visualisation

Chapter lead author: Benjamin Stocker

Contents:

- Lecture (Isabelle Bentz?): The art of visualising data, grammar of graphics
- Exercise: Develop decision tree for what type of visualisation to apply
- Performance assessment: Interactive work sequence

4.1 Learning objectives

4.2 Tutorial

4.3 Exercises

4.4 Solutions

Chapter 5

Data variety

Chapter lead author: Koen Hufkens

Contents:

- Lecture (Mirko): Mapping data
- Data formats, standards, metadata
- Geographic data
- Scraping, wget
- APIs

5.1 Learning objectives

5.2 Tutorial

5.3 Exercises

5.4 Solutions

Chapter 6

Code management

Chapter lead author: Koen Hufkens

Contents:

- git: repositories, stage, commit, push, fork, pull request, fetch upstream
- Performance assessment: **CAT 2**

6.1 Learning objectives

6.2 Tutorial

6.3 Exercises

6.4 Solutions

Chapter 7

Open science practices

Chapter lead author: Koen Hufkens

Contents:

- Lecture (Koen): Open science - history, motivation, reproducibility crisis, current initiatives, overview of practices
- Environmental data repositories
- Methods to create visualised reproducible workflow
- RMarkdown files
- Performance assessment: **CAT 3**, link to Dietze exercise on pair coding

7.1 Learning objectives

7.2 Tutorial

7.3 Exercises

7.4 Solutions

Chapter 8

Regression

Chapter lead author: Benjamin Stocker

Contents:

- Linear regression
- Regression metrics
- Logistic regression
- classification metrics
- Comparing models (AIC, …)
- Feature selection, stepwise regression, multi-colinearity (vif)
- Performance assessment: Exercise for stepwise regression link

8.1 Learning objectives

8.2 Tutorial

8.3 Exercises

8.4 Solutions

Chapter 9

Supervised machine learning

Chapter lead author: Benjamin Stocker

- Lecture (Beni): Overfitting, training, and cross-validation (link)
- K nearest neighbour models
- Data splitting
- Preprocessing, standardization, imputation, dimension reduction, as part of the model training workflow
- formula notation, recipes, generic train()
- Training and loss function
- Hyperparameters
- Resampling
- Performance assessment: Exercise comparing performance on test set of linear regression and KNN with different hyperparameter choices (like this), discuss link to overfitting example

9.1 Learning objectives

9.2 Tutorial

9.3 Exercises

9.4 Solutions

Chapter 10

Random forest

Chapter lead author: Benjamin Stocker

Contents:

- Lecture (Beni): Wisdom of the crowds, from decision trees to random forests
- Performance assessment: Competition for best-performing model, given training-testing split of data; others should be able to reproduce performance

10.1 Learning objectives

10.2 Tutorial

10.3 Exercises

10.4 Solutions

Chapter 11

Neural networks

Chapter lead author: Benjamin Stocker

Contents:

- Lecture (Beni): General introduction
- Performance assessment: Competition for best-performing model, given training-testing split of data; others should be able to reproduce performance

11.1 Learning objectives

11.2 Tutorial

11.3 Exercises

11.4 Solutions

Chapter 12

Interpretable machine learning

Chapter lead author: Benjamin Stocker

Contents:

- Variable importance
- Partial dependency
- Performance assessment: Compare partial dependency to a given predictor, detected with RF and with NN.

12.1 Learning objectives

12.2 Tutorial

12.3 Exercises

12.4 Solutions