3-GS Spatial statistics and machine learning

INFOMSSML, period 3B, MSc Applied Data Science

Course manual

Dr. Menno Straatsma
Department of Physical Geography
m.w.straatsma@uu.nl
course coordinator

Dr. Anae Sobhani
Department of Human Geography and Spatial Plannin
a.sobhani@uu.nl

Faculty of Geoscience,
Utrecht University

Contents

1 Course Description

1.1 Course Objectives

At the end of this cours, the student is able to:

1. Manipulate vector and raster data, including spatial interpolation
2. Extract information from remote sensing scenes using different classification methods
3. Find patterns in spatial data using decision trees and random forests.
4. Adapt machine learning techniques by analysing real world problems in the context of group assignments and term project.

1.2 Course Content

The course will consist of a series of lectures on specific techniques, followed by a computer practical where the techniques are practiced on a given data set. The topics addressed in this course include:

1. Spatial data wrangling of
   a. vector data: projections, spatial and geometric operations
   b. raster data: remote sensing (lidar and multispectral data
2. Spatial interpolation: inverse distance and kriging
3. Classification of remote sensing images, including nearest neighbor, support vector machines and random forests.
4. Basic terminology and concepts behind statistical learning
5. Regression modelling with spatial data:
   a. Multiple linear regression
   b. Non-linear regression
      i. Polynomial regression
      ii. Regression spline
      iii. Generalized additive model
6. Tree-Based Methods
   a. Regression tree
   b. Bagging, random forest, and boosting

Term project

From the third week of the course, students will start to work on a term project, in which they answer a research question based on a their own or a given data set. The term project team should be carried out by **three students** and should be approved by Dr. Sobhani and Dr. Straatsma. During the term project, you will study a real life problem, which includes the choice of the appropriate technique(s), the appropriate variables, implementing the method, interpretation of the results. The work is presented in a **short paper** (1500 words) and a **15 minutes presentation** and a **5 minutes Q&A** during the closing seminar. Feedback is given by peers and the supervisors on both the paper and the presentation.

1.3 Language of Instruction

All lectures and practical will be given in English. The exam and assignments have to be written in **English**.

1.4 *Course Materials*

- Campbell, J.E., Shin, M. (2012) Geographic Information System Basics, 248 pp, Open source at lardbucket.org.
- De Jong, S.M., Addink, E.A., Heuff, F. (2015) Remote sensing lecture notes. 113 pp, Fac. of Geosciences, Utrecht University.
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R, Volume 103, Springer.
- Various handout papers.

1.5 *Software*

For this course, you will use different software and software packages to carry out the data wrangling, and the statistical and Machine learning techniques. The variety of methods does not allow to use one package. The following software is used:

- QGIS, long term release v3.10.14, available at https://www.qgis.org/en/site/
- Lastools, https://rapidlasso.com/lastools/ , license provided within the course
- Python, available through anaconda.org.
- R is available free of charge via the following link: https://towardsdatascience.com/how-to-install-r-and-rstudio-584eeefb1a41. Install R and RStudio.

Students need to install the software on their laptop for the practical sessions. The practical with Lastools, which runs on Windows, may be done on myworkplace (www.myworkplace.uu.nl) as well.

1.6 *Participation Requirements*

Students should complete the assignment in time. Attendance is compulsory during the data presentations at the start of the course and the seminar at the end of the course. The student must notify the lecturer concerned beforehand of an absence, if that is possible (see ‘contact with lecturers’). Attendance at the formal lectures and practicals is not required, although the material covered in the lectures will be needed for the exam and final project presentation.

1.7 *Assessment*

Term Project (report (20%) and presentation (20%), Assignments (15% + 15%), Exam (30%). The final grade is a weighted average of the four components of the assessment.

A once-only rounding off takes place in the determination of the final grade. An average grade of 5.49 leads to an incomplete final grade (5); the average grade 5.50 leads to a satisfactory final grade (6).

What happens when the final grade is below the pass mark?

- A student who has fulfilled all attendance requirements and completed all assignments during the course (as indicated in the study guide) and whose average grade for all course components *before* rounding up is greater than 3.99 and lower or equal to 5.49 is given one opportunity to complete a supplementary assignment as a compensation assignment in order to complete the course with a satisfactory grade (6). If the supplementary assignment receives a satisfactory grade (5.50 or higher), the final grade for the course is registered as 6. The lecturer decides on the form and content of the supplementary assignment.

- When the average final grade is 3.99 or lower no supplementary assignment is possible. The administrative office for students' progress will process the assigned final grade and all part results attained in the course will then lapse

1.8 Instructors

-	Menno W. Straatsma, Ph.D.	: M.W.Straatsma@uu.nl, room 4.40, 253 2754
-	Anae Sobhani, Ph.D.	: a.sobhani@uu.nl, room 6.68, tel. 253 2041
-	Francisco Bahamonde Birke, Ph.D.	: bahamondebirke@gmail.com
-	Jessie Lensing, teaching assistant	: j.lensing@students.uu.nl

1.9 Assumed Knowledge

Participants should have knowledge of the following techniques, and be able to apply them using statistical and Machine Learning packages:

- descriptive statistics (frequency distributions, measures of central tendency and dispersion, cross tabulation)
- deductive statistics (t-test, chi-square test, ANOVA, linear regression, correlation analysis)

Our advice is to test your statistical skills prior to the first lecture!

1.10 Important dates:

- Term project:
    - February 23, 2021, kickoff and matchmaking, bring your data/topic
    - March 2, proposal submission on BB 5 pm
    - March 9: Feedback on proposal and Q&A
    - March 25: Feedback term project Q&A (on request)
    - April 13: Peer review on 1500 word report term project
    - April 15: Term project presentation
    - April 16: 5 PM submission in BB (term project report and presentation)
- March 23, 2021, by 5 PM: Submit assignment 1 on Blackboard
- April 6, 2021: by 5 PM: Submit assignment 2 on Blackboard

2 Detailed schedule 2021

Week	**Date**	**Time**	**Location**	**Topic**	**Content**	**Staff**
6	Thu 11 Feb	13:15-15:00	Teams	Introduction to the course and spatial data wrangling	Lecture	MS, AS
	Thu 11 Feb	15:15-17:00	Teams	Spatial data wrangling	Practical	MS, JL
7	Tue 16 Feb	9:00 – 9:45	Teams	Presentations dataset examples by students	Pitch	MS, JL
	Tue 16 Feb	10:00-13:00	Teams	Spatial data wrangling	Practical	MS, JL
	Thu 18 Feb	13:15-15:00	Teams	Remote sensing (RS)	Lecture	MS
	Thu 18 Feb	15:15-17:00	Teams	Remote sensing spectral satellite data	Practical	MS, JL
8	Tue 23 Feb	9:00-9:45	Teams	Kickoff term project and matchmaking	Practical	MS, AS
	Tue 23 Feb	10:00-13:00	Teams	Remote sensing lidar data	practical	MS, JL
	Thu 25 Feb	13:15-15:00	Teams	Spatial interpolation	Lecture	MS
	Thu 25 Feb	15:15-17:00	Teams	Spatial interpolation	Practical	MS, JL
9	Tue 2 March	9:00-9:45	Teams	Presentation of RS data examples by students	Pitch	MS, JL
	Tue 2 March	10:00-13:00	Teams	Spatial interpolation	Practical	MS, JL
	Thu 4 March	13:15-15:00	Teams	Machine learning with RS data	Lecture	MS
	Thu 4 March	15:15 – 17:00	Teams	Classification of remote sensing data 2: k-means, (un-)supervised, svm	Practical	MS, JL
10	Tue 9 March	9:00-10:45	Teams	Feedback on term project proposal and Q&A	Q&A	MS, AS
10	Tue 9 March	9:00-13:00	Teams	Classification of remote sensing data 2: deep learning, random forest	Practical	MS
10	Thu 11 March	13.15-15.00	Teams	Introducing the basic terminology and concepts behind statistical learning	Lecture	BB
	Thu 11 March	15.15-17.00	Teams	Regression and sampling	Lecture	BB
11	**Tue 16 March**	**10.00-12.00**	**Remindo**	**Exam online**	**Exam**	**MS**
	Thu 18 March	13.15-15.00	Teams	Multiple linear regression, Polynomial regression, Regression spline	Practical	BB
	Thu 18 March	15.15-17.00	Teams	Generalized additive model	Lecture	BB
12	Tue 23 March	9.00-10.45	Teams	Generalized additive model	Practical	BB
	Thu 25 March	13.15-17.00	Teams	Work on the term project, feedback on project on request	Practical	AS
13	Tue 30 March	9.00-10.45	Teams	The basics of decision trees	Lecture	AS
	Thu 1 April	13.15-15.00	Teams	Fitting regression decision tree	Practical	AS
	Thu 1 April	15.15-17.00	Teams	Bagging, random forest, and boosting	Lecture	AS
14	Tue 6 April	9.00-10.45	Teams	Bagging, random forest, and boosting	Practical	AS
	Thu 8 April	13.15-15.00	Teams	Work on the term project	Practical	
	Thu 8 April	15.15-17.00	Teams			
15	Tue 13 April	9.00-12.45	Teams	**Peer review of term project paper**	Practical	MS, AS
	Thu 15 April	13.15-15.00	Teams	**Term project presentation**	Practical	MS, AS
	Thu 15 April	15.15-17.00	Teams	**Term project presentation**	Practical	MS, AS
15	Fri 16 April	17:00	Blackboard	**Submit term project paper and presentation**	---	

MS=Menno Straatsma, BB=Francisco Bahamonde Birke, AS=Anae Sobhani, JL-Jessie Lensing

3 Topics in detail

February 11, 2021. Topic 1: Introduction to the course and spatial data wrangling

Working with spatial data brings its own set of challenges and solutions. In this introductory topic, we will cover the basics of working with vector data.

1. Introduction to the course
2. Example term project
3. Spatial data wrangling basics
4. Practical spatial data wrangling, to be continued at home.

Required reading:

- Campbell, J.E., Shin, M. (2012) Geographic Information System Basics, 248 pp, Open Source at lardbucket.org. Study chapter 4, 5, 7 and 8

Practicals: after the lecture, two online practicals are scheduled on spatial data wrangling with Python. The answers are distributed after both practicals have been covered.

Assignment group 1: Make 2 minute powerpoint pitch on one of the freely available datasets. List of datasets will be provided.

February 18, 2021. Topic 2: Remote sensing, wrangling raster data

Airborne and spaceborne remote sensing creates an ever increasing amount of data in raster format. We introduce the basics of multispetral and lidar remote sensing to derive meaningful descriptions of the spatial distribution of surface attributes.

Required reading:

- De Jong, S.M., Addink, E.A., Heuff, F. (2015) Remote sensing- a tool for environmental observations, lecture notes. 113 pp, Fac. of Geosciences, Utrecht University. Read chapters 1, 2, and 5, excluding 5.8
- Lohani, Bharat, and Suddhasheel Ghosh. "Airborne LiDAR technology: a review of data collection and processing systems." Proceedings of the National Academy of Sciences, India Section A: Physical Sciences 87.4 (2017): 567-579.

Practicals: after the lecture, two online practicals are scheduled on working with satellite data from Landsat with Python and airborne lidar data with Lastools and batch scripting.

Assignment group 2: make 2 minute powerpoint pitch on one of the available remote sensing datasets. List of sensors will be provided.

February 25, 2021. Topic 3: Spatial interpolation

Field measurements at the Earth's surface acts as reference data for remote sensing and proces-based modelling. However, they are expensive and even though they act as reference data, the data still contains an error. The conversion of point measurements to spatial fields of the same attribute (e.g. soil carban, or surface temperature) requires spatial interpolation. We cover inverse distance weighting and simple and ordinary kriging.

Required reading:

- Lecture notes spatial statistics by Sterk
- Davis, John C., and Robert J. Sampson. Statistics and data analysis in geology. Third edition New York: Wiley, 2002.
    - Book chapter on kriging.

Practicals: after the lecture, two online practicals are scheduled on working with the semivariogram and interpolation.

Assignment: none

February 25, 2021. Topic 4: Machine learning with remote sensing scenes.

Classification of remote sensing scenes into land cover or land use is the basis of mapping and monitoring the state of the landscape. This topic covers different classification methods.

Required reading:

- De Jong, S.M., Addink, E.A., Heuff, F. (2015) Remote sensing- a tool for environmental observations, lecture notes. 113 pp, Fac. of Geosciences, Utrecht University. Read chapters 5.8 and 6.

Practicals: after the lecture, two online practicals are scheduled on working with k-means clustering, support vector machines, deep learning and random forests.

Assignment: none

March 11, topic 5: Basic terminology and concepts behind statistical learning

Develop accurate models that can be used for output predictions and inference. We will see a number of examples that fall into the prediction setting, the inference setting, or a combination of the two.

Topics that will be covered in this session:

a. What is statistical learning

There are two main reasons that we may wish to estimate effect of variables the outputs: prediction and inference. In this part we will lok into not only prediction but also dive into answering which predictors are associated with the response, and what is the relationship between the response and each predictor.

b. The trade-off between prediction accuracy and model interpretability

Of the methods that we examine in this lecture, some are less flexible, or more restrictive, in the sense that they can produce just a relatively small range of shapes to estimate outputs. One might reasonably ask the following question: why would we ever choose to use a more restrictive method instead of a very flexible approach?

c. Supervised versus unsupervised learning

Most statistical learning problems fall into one of two categories: supervised or unsupervised. In this lecture we will focus the supervised learning domain.

d. Assessing model accuracy

It is an important task to decide for any given set of data which method produces the best results. Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice. Hence, we will discuss some of the most important concepts that arise in selecting a statistical learning procedure for a specific data set.

Required reading:

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R, Volume 103, Springer. pp. 15-57

Assignment: none

March 11, topic 6: Regression modelling

As we know linear models are relatively simple to describe and implement, and have advantages over other approaches in terms of interpretation and inference. However, standard linear regression can have significant limitations in terms of predictive power. In this session we relax the linearity assumption while still attempting to maintain as much interpretability as possible. We do this by studying non-linear regression models. Topics that will be covered in this session:

a. Multiple linear regression

Multiple linear regression is a very simple approach for supervised learning. In particular, it is a useful tool for predicting a quantitative response. It has around for a long time. Though it may seem somewhat dull compared to some of the more modern statistical learning approaches described, Multiple linear regression is still a useful and widely used statistical learning method.

b. Non-linear regression

i. Polynomial regression

Historically, the standard way to extend linear regression to settings in which the relationship between the predictors and the response is non-linear has been to replace the standard linear model. Polynomial regression extends the linear model by adding extra predictors, obtained by raising each of the original predictors to a power.

c. Regression spline

It is more flexible than polynomial, and in fact is an extension of the it. It involves dividing the range of X into K distinct regions. Within each region, a polynomial function is fit to the data. However, these polynomials are constrained so that they join smoothly at the region boundaries, or knots. Provided that the interval is divided into enough regions, this can produce an extremely flexible fit.

d. Generalized additive model

It allows us to extend the methods above to deal with multiple predictors. I other words, it provides a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity. Just like linear models, it can be applied with both quantitative and qualitative responses.

Required reading:

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R, Volume 103, Springer. pp. 203-301

Assignment: TBA

March 30, topic 7: Tree-Based Methods: Regression tree

In this session, we describe tree-based methods for regression. In order to make a prediction for a given observation, we typically use the mean or the mode of the training observations in the region to which it belongs. Since the set of splitting rules used to segment the predictor space can be summarized in a tree, these types of approaches are known as decision tree methods. Topics that will be covered in this session:

a. The basics of decision trees

Decision trees can be applied to both regression and classification problems. In this Class, we only look at regression problems

i. Regression tree

Decision tree for regression has a number of advantages over the more classical approaches seen in the previous lecture. However, by aggregating many decision trees, using methods like bagging, random forests, and boosting, the predictive performance of trees can be substantially improved.

Required reading:

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R, Volume 103, Springer. pp. 303-335

Assignment: TBA

Tree-Based Methods: Bagging, random forest, and boosting

Bagging, random forest, and boosting use trees as building blocks to construct more powerful prediction models.

Required reading:

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R, Volume 103, Springer. pp. 303-335

Assignment: none