DEGREE CURRICULUM

ST IN DATA MANAGEMENT AND VISUALIZATION WITH R

Coordination: AMÉZTEGUI GONZÁLEZ, AITOR

Academic year 2021-22

Subject's general information

<table>
<tr><td>Subject name</td><td>ST IN DATA MANAGEMENT AND VISUALIZATION WITH R</td></tr>
<tr><td>Code</td><td>111022</td></tr>
<tr><td>Semester</td><td>ANUAL CONTINUED EVALUATION</td></tr>
<tr><td>Typology</td><td>
<table>
<tr><th>Degree</th><th>Course</th><th>Character</th><th>Modality</th></tr>
<tr><td>Master's Degree Erasmus Mundus in Spatial and Ecological Modelling in European Forestry</td><td>2</td><td>OPTIONAL</td><td>Attendance-based</td></tr>
<tr><td>Master's Degree Erasmus Mundus in Spatial and Ecological Modelling in European Forestry</td><td></td><td>OPTIONAL</td><td>Attendance-based</td></tr>
</table>
</td></tr>
<tr><td>Course number of credits (ECTS)</td><td>5</td></tr>
<tr><td>Type of activity, credits, and groups</td><td>
<table>
<tr><th>Activity type</th><td>PRAULA</td><td>TEORIA</td></tr>
<tr><th>Number of credits</th><td>4</td><td>1</td></tr>
<tr><th>Number of groups</th><td>1</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>Coordination</td><td>AMÉZTEGUI GONZÁLEZ, AITOR</td></tr>
<tr><td>Department</td><td>AGRICULTURAL AND FOREST ENGINEERING</td></tr>
<tr><td>Teaching load distribution between lectures and independent student work</td><td>Lectures: 25%
Independent work: 75%</td></tr>
<tr><td>Important information on data processing</td><td>Consult this link for more information.</td></tr>
<tr><td>Language</td><td>English</td></tr>
</table>

Teaching staff	E-mail addresses	Credits taught by teacher	Office and hour of attention
AMÉZTEGUI GONZÁLEZ, AITOR	aitor.ameztegui@udl.cat	5	

Learning objectives

Introductory course on data science, focused on the collection, management, visualization and analysis of data. We cover the entire data science pipeline from data acquisition to publication. The course will focus on the R statistical computing language and is addressed to second-year students of the Master's Degree Erasmus Mundus in Spatial and Ecological Modelling in European Forestry at the University of Lleida. The aim is to teach the students how to use R to effectively manage, clean, analyze and visualize data. The course is based on a "hands-on" approach, so that the students can easily transfer the acquired knowledge to real case studies, and even use it to process and analyze their own set of data, helping them in the preparation of their Master Thesis

Subject contents

Part 1: Data Management

- Unit 1: Introduction: The tidyverse and the concept of tidy data. Good coding practices.
- Unit 2: Explore your data. Data transformation: create variables, filter, select & organize data
- Unit 3: Wrangle your data. Data import and export: reading and writing different file types. Modify your data frame and work with several tables.
- Unit 4: Program in R within the tidyverse. Functions and functional sequences. Iteration.

Part 2: Data visualization

- Unit 5: Fundamentals of data visualization. Types of data and aesthetics. Principles of figure design.
- Unit 6: Directory of visualizations. Visualizing amounts, proportions and distributions. Revealing trends and seeing relationships.
- Unit 7: Visualization of spatial information: maps. Accessing spatial data in the tidyverse. Projections and coordinate systems.
- Unit 8: Graphics for communication. Controlling axes and manipulating colours in ggplot. Adding text and annotations. Saving plots.

Methodology

"**Data management and visualization with R**" is designed as an asynchronous online course. That means that learning happens on your schedule. And it is based on a "hands-on" approach.This means that instructors will provide materials to introduce the main concepts, but you will learn by doing, that is, most of the learning will occur while you try to solve the exercises, assignments or reports. We know that coding can be hard at the beginning, and that you will make lots of mistakes. That's good! A key part of your success in using R lies in your ability to be self-relient and be able to get help and apply it to your own problem. This is why it it is important to become familiar from the beginning with the various alternatives for getting help (including, of course, asking the instructors!)

The course is organized around the following components:

- **Lectures:** we will try to keep them short, but some "theory" will be needed here and there. This will be provided as short videos introducing the main ideas.
- **Labs:** this is where most of the concepts will be presented. They're self paced tutorials with examples, and a final self-assessment questionnaire (does not count for grades) where students can evaluate the degree of

learning and understanding of the knowledge imparted.Each lab will contain some questions/exercises you will need to answer by creating a "lab report" (we'll see how).
- **Readings:** In some units we will recommend some external readings to expand or consolidate concepts covered in labs and lecture
- **Homework:** You will be assigned some larger data analysis tasks throughout the semester. These assignments will be completed individually (but can ask doubts of course!), and we will set the deadlines as the course progresses.
- **Exams:** Individual, two midterms. We will ask you to complete a number of small programming and or analysis tasks related to the material presented in the class.You will have access to all the materials provided in the course, and well... to the internet, so you will be able to check any source you wish. However, you MUST solve the exams individually, and all communication with classmates is explicitly forbidden.

A note on sharing/reusing code

- A huge volume of code is available on the web to solve any number of problems.
- Unless we explicitly tell you not to use something, the course's policy is that *you may make use of any online resources* (e.g. StackOverflow) but if you copy pieces of code from somewhere **you must cite** where you obtained it.
- Except for the exams, you are welcome to discuss the problems together and ask for advice, but you may not directly send or make **use of code from another student** in this class
- On the exams all communication with classmates is explicitly forbidden.

Development plan

All the info, links and materials of the course will be posted at datamanagement.netlify.com

We will use R as the software to solve all the data management and visualization problems. R is one of the most popular languages for data management, and is very versatile, multiplatform, and 100% free (in both senses). Moreover, analyses conducted in R are transparent, easily shareable, and reproducible.

To avoid the need to install any software in your computer, we will use RStudio Cloud, which is a cloud managed instance of the RStudio IDE. We have had very good experiences using RStudio Cloud in previous years, since it reduces friction at first exposure to R, and avoids the need for software local installation, or to install packages, circulate files with data or download files.

You will learn all what you need to know about R and Rstudio Cloud at the beginning of the course, but you can also check this introductory document: https://datamanagement.netlify.app/labs/lab01_intro-r.html

Evaluation

Component	Weight
Lab exercises	25%
Homeworks	25%
Midterm Exam 1	20%
Midterm Exam 2	20%
Participation and peer evaluation	10%

Bibliography

2021-22

These are the main texts used to prepare this course. All of them are either freely available, or available for loan in the library of the University of Lleida.

R for Data Science	Grolemund, Wickham	O'Reilly, 2016
Data visualization: a practical introduction	Healy	O'Reilly, 2019
Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures	Wilke	O'Reilly, 2019
The Functional Art: An introduction to information graphics and visualization	Cairo	New Riders Publishing, 2012
The truthful Art: Data, Charts, and Maps for Communication	Cairo	New Riders Publishing, 2016
Making data visual: a practical guide to using visualization for insight	Fisher , Meyer	O'Reilly, 2019
Good Enough Practices in Scientific Computing	Wilson et al.	PLoS Computational Biology 2016