Get your dataset ready!

Using R and GIS

Rosa Félix Gabriel Valença Rafael Pereira

2024-09-03

Table of contents

1 Introduction **4**
- 1.1 Mobility data . . . 4
  - Why R and GIS . . . 6
- 1.2 Course objectives . . . 6
  - Introduce R Programming Basics . . . 6
  - Teach Data Manipulation Techniques . . . 6
  - Spatial Data Visualization . . . 7
  - Perform Basic Spatial Analysis . . . 7
- 1.3 Target audience . . . 7

2 Course Structure **8**
- 2.1 Day 1 . . . 8
  - Morning . . . 8
  - Afternoon . . . 8
- 2.2 Day 2 . . . 8
  - Morning . . . 8
  - Afternoon . . . 9

3 Detailed schedule (TBC) **10**

4 Location **12**

5 Resources **14**

I Day 1 **15**

6 Software **16**
- 6.1 R and RStudio . . . 16
  - 6.1.1 R . . . 16
  - 6.1.2 RStudio . . . 16
  - 6.1.3 Rtools . . . 16
  - 6.1.4 R packages . . . 16
- 6.2 QGIS . . . 16
  - 6.2.1 Download . . . 17
  - 6.2.2 Plugins . . . 18

7 R basics **19**
7.1 Simple operations 19
7.2 Practical exercise 21

II Day 2 **23**

8 Introduction **24**

References **25**

1 Introduction

This course aims to provide tools to deal with exploring and treating transportation datasets using R programming, an open-source and widely used tool for data analytics in urban mobility.

Additionally, this course provides guidance towards the use of reproducible methods to deal with large datasets that require manipulation and/or spatial analysis.

The course has a **hands-on** approach, where participants will learn the basics of **coding**, **data manipulation**, and **spatial analysis** for urban mobility and transportation.

1.1 Mobility data

There is an emerging increase in mobility data, through new forms of technology, which result in very large and diverse datasets.

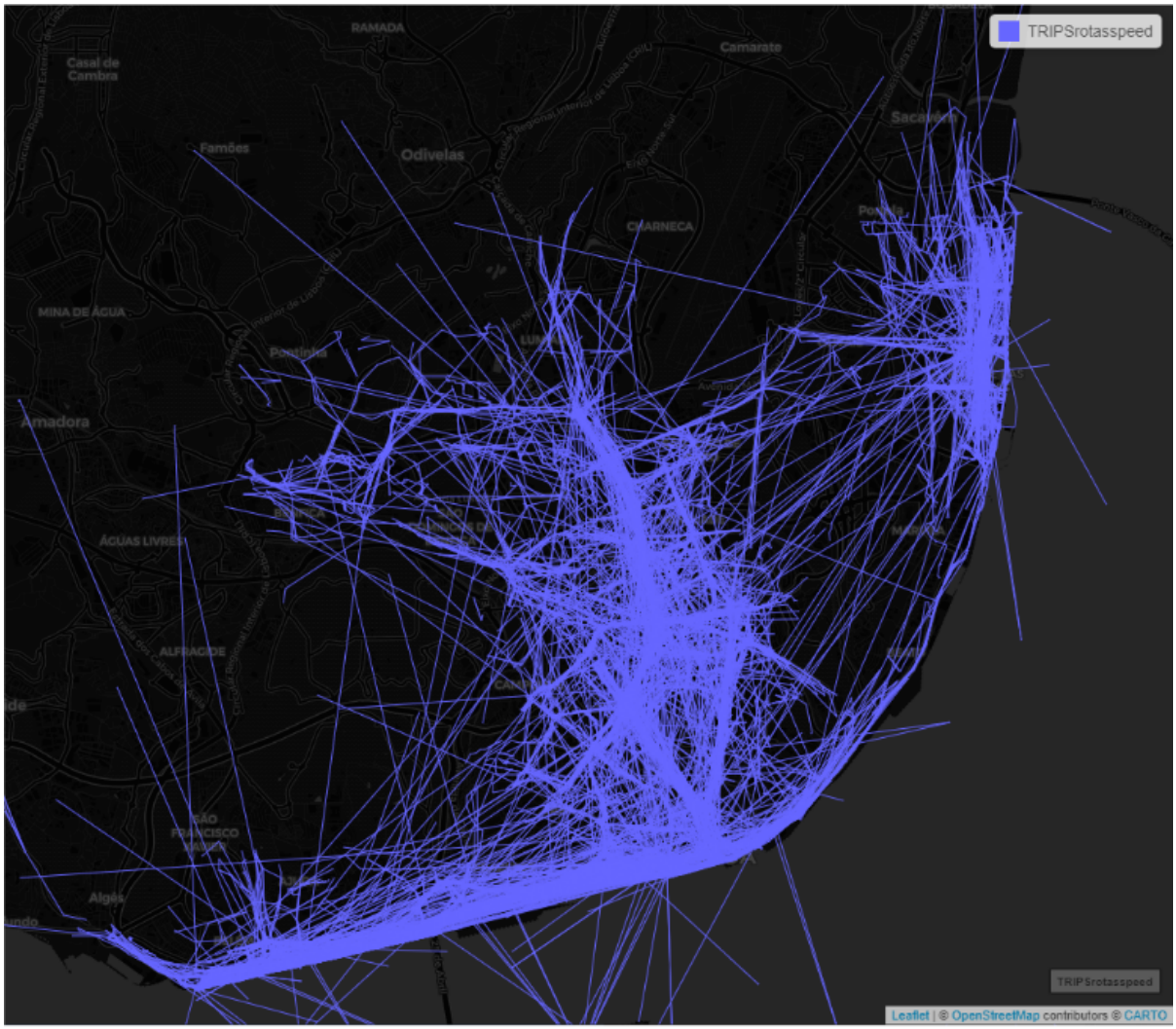


Figure 1.1: E-Scooter trip data in Lisbon. How to deal with it?

Knowing how to get, treat and analyze complex datasets with the up-to-date technologies is extremely relevant for academia, policy makers and start-ups, since it allows them to:

1. acquire critical view on urban mobility based on data;
2. spatially identify locations in the city that require policy priorities;
3. and improve the efficiency of data analysis processes.

Why R and GIS

Most academic programs focus on teaching modelling and deep analysis of data. However, there is a need to learn how to explore and prepare a dataset for modelling. The use of **programming and GIS** techniques have enormous advantages, including their flexibility; reproducibility; and transparency and understanding the step-by-step process.

The use of GIS techniques in transportation is, traditionally, not considered in transportation learning programs, despite being of enormous relevance when doing accessibility analysis or reeling with georreferenced transportation data, such as bike sharing route trips' datasets, origin-destination flows datasets, home/work locations, GTFS public transit data, and so on. There is a need to learn how to locate these open datasets, how to explore them and how to integrate them into transportation and urban analysis. Additionally, the use of open source software and datasets allows researchers to perform methods that are reproducible and transparent.

TLDR

- Open-source tools widely used in data analytics and spatial analysis
- Flexibility and reproducibility in data manipulation and visualization
- Critical for urban mobility and transportation research, with spatial relevance
- Large transportation datasets are becoming increasingly common

1.2 Course objectives

Introduce R Programming Basics

- Equip participants with foundational skills in R programming
- Emphasize reproducible research practices to ensure transparency and replicability in analyses

Teach Data Manipulation Techniques

- Use key R packages for data cleaning, manipulation, and summarization of datasets
- Enable participants to efficiently handle large and complex transportation datasets

Spatial Data Visualization

- Introduce methods for quick and effective spatial data visualization using R and GIS tools
- Provide hands-on experience with creating interactive maps and visualizations

Perform Basic Spatial Analysis

- Teach participants how to perform spatial analysis of transportation datasets using GIS techniques with R
- Cover practical applications such as georeferencing data, accessibility analysis, and routing ODs
- Utilize real-world transportation data for practical, hands-on learning

1.3 Target audience

- Ph.D. candidates from DTN and other researchers
- Policy makers and practitioners in urban mobility
- Beginners to intermediate R users, no prior experience needed

2 Course Structure

The course consists of an in-person 2-day course, taking place during the EIT DTN Annual Meeting on the **19th and 20th September 2024**.

The first day will focus on learning the basics of R programming and how to treat and explore datasets. The second day will focus on analyzing spatial datasets, and routing origins to destinations.

2.1 Day 1

Morning

- Introduction to **programming** techniques and **data structures**
- Introduction to R, and RStudio: **software installation** and main packages
- **R base and basics**: examples and exercises

Afternoon

- **Data manipulation**: using the dplyr package to select, filter, left-join, group and summarize
- Introduction to **GIS** and **spatial data**: import and visualize vector data
- R markdown and **interactive maps**

2.2 Day 2

Morning

- **Desire lines** from OD and transport zones
- **Georeference** coordinates: examples from surveys
- **Accessibility analysis**: from buffers to road networks

Afternoon

- **Open Transportation data**: where to find it
- **Routing with R**: multimodal and intermodal (*r5r demo* - Rafael Pereira)
- Group **exercise**

3 Detailed schedule (TBC)

Day 1	
9.30	Introductions and Presentation of the course contents
10.00	Introduction to programming techniques and data structures
10.30	Introduction to R and RStudio: hands-on to install software and main packages
11.00	*Coffee break*
11.15	(cont.)
11.30	R basics: examples and exercises
12.30	*Lunch break*
13.30	Data manipulation: examples and exercises (select, filter, left-join, subset, group and summarize, using dplyr package)
15.00	Introduction to GIS and spatial data: import and visualize vector data
15.30	*Coffee break*
15.45	(cont.)
16.15	View and export interactive maps
17.00	*End of day 1*

Day 2	
9.30	Desire-lines from OD pairs and transport zones: examples and exercises
10.30	Georeferenced coordinates from survey responses: example and exercises
11.00	*Coffee break*
11.15	(cont.)
11.30	Euclidean distance and buffers: example and exercises
12.30	*Lunch break*
13.30	Open Transportation data: where to find it (OSM and GTFS)

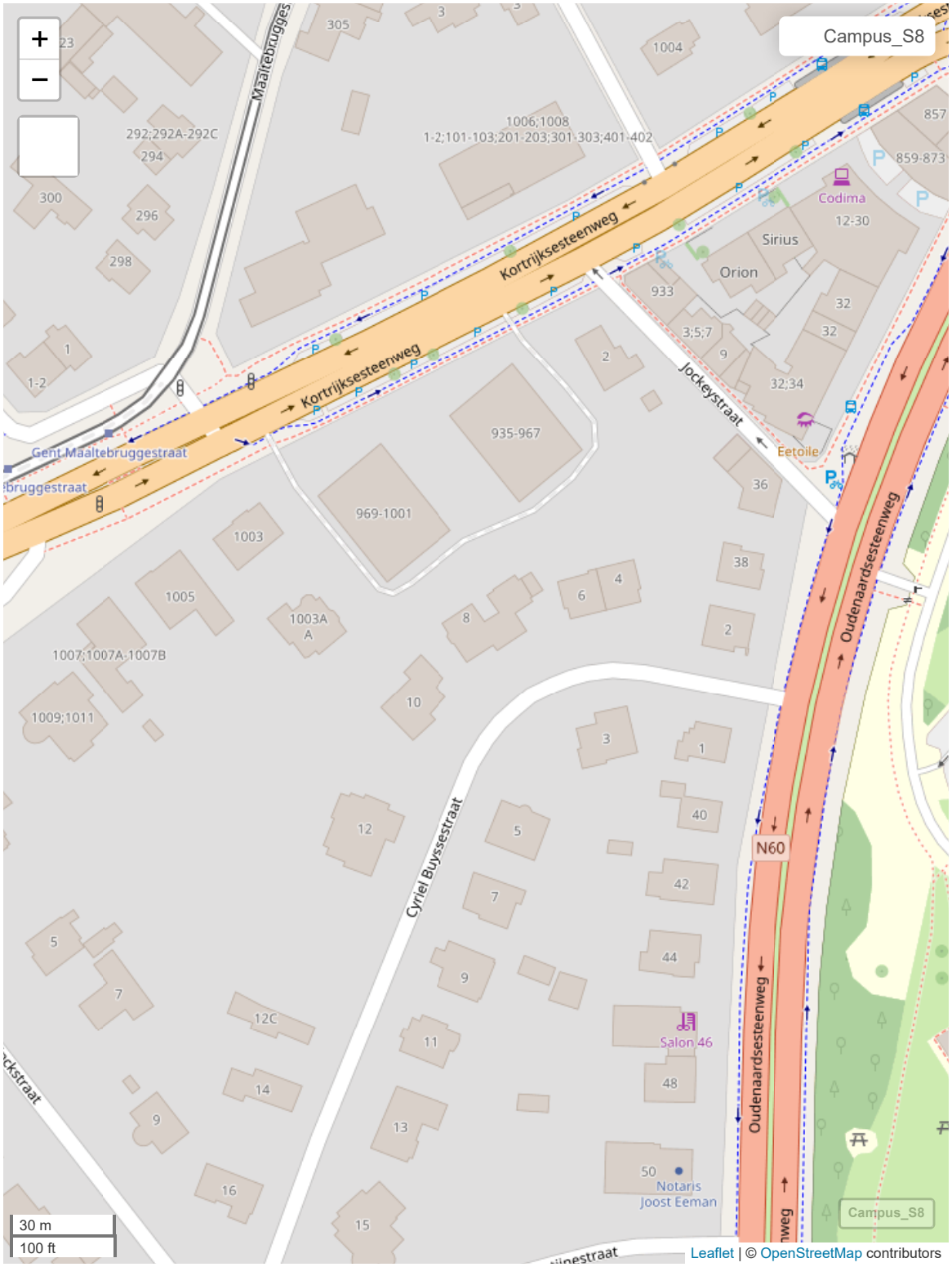
Day 2	
14.30	Uni-modal and Inter-modal Routing with r5r
15.30	Accessibility analysis with r5r
16.00	*Coffee break*
16.15	Using you data: manipulation and spatial analysis methods and further applications
16.45	Survey and feedback from participants
17.00	*End of day 2*

4 Location

The course will take place at Campus Sterre, Building S8, room 2.4.

```
Campus_S8_coord = c(3.7105372, 51.0241258)
Campus_S8 = sf::st_sfc(sf::st_point(Campus_S8_coord)) # create point
Campus_S8 = sf::st_as_sf(Campus_S8, crs = 4326) # assign crs

mapview::mapview(Campus_S8, map.types = "OpenStreetMap") # quick map view
```


Campus_S8
Kortrijksesteenweg
Kortrijksesteenweg
Gent Maaltebruggestraat
Jockeystraat
Oudenaardsesteenweg
Oudenaardsesteenweg
N60
Cyriel Buyssestraat
Codima
12-30
Sirius
Orion
Eetoile
Salon 46
Notaris
Joost Eeman
1006;1008
1-2;101-103;201-203;301-303;401-402
292;292A-292C
935-967
969-1001
1007;1007A-1007B
1009;1011
1003A
A
30 m
100 ft
Leaflet | © OpenStreetMap contributors

5 Resources

- You laptop, with any OS
- Github repository with all the materials (data, code and guidelines)
- Survey datasets, school locations and public transport operator datasets

Part I

Day 1

6 Software

In this chapter we will guide you through the installation of R and QGIS.

6.1 R and RStudio

You will need **R** installed on your computer (version 4.4.1 or higher) and also **RStudio**[1].

6.1.1 R

6.1.2 RStudio

6.1.3 Rtools

6.1.4 R packages

6.2 QGIS

QGIS is a geographic information system software that is free and open-source. QGIS supports Windows, macOS, and Linux. It supports viewing, editing, printing, and analysis of geospatial data in a range of data formats[2].

In this course QGIS will be used to geocode coordinates, make accessibility analysis using the street network, and visualize OD flows.

QGIS allows the visualization of data with a graphical user interface (GUI), which can be preferable for basic usage. For advanced computation with geo data, programming software such as Python or R can be more adequate.

[1] We will use RStudio, although if you already use other studio such as VScode, that's also fine.
[2] https://en.wikipedia.org/wiki/QGIS

6.2.1 Download

You should go for the Long-Term Version provided by QGIS. If you have installed the most up-to-date version, that's also fine.

https://qgis.org/download/

This download will be about 1.2-1.6 GB.

You should also have Python installed. Otherwise QGIS will install it for you [?].

6.2.1.1 Windows

Download and open the executable file.

6.2.1.2 Mac

Download and open the executable file.

6.2.1.3 Ubuntu

You can look for QGIS in the Ubuntu **Software Center** or install it via the terminal.

```
sudo apt install gnupg software-properties-common
sudo mkdir -m755 -p /etc/apt/keyrings  # not needed since apt version 2.4.0 like Debian 12 a
sudo wget -O /etc/apt/keyrings/qgis-archive-keyring.gpg https://download.qgis.org/downloads/
```

Add the QGIS repo for the latest stable QGIS (3.38.x Grenoble) to `/etc/apt/sources.list.d/qgis.sources`:

```
Types: deb deb-src
URIs: https://qgis.org/ubuntugis-ltr
Suites: jammy
Architectures: amd64
Components: main
Signed-By: /etc/apt/keyrings/qgis-archive-keyring.gpg
```

```
sudo apt update
sudo apt install qgis qgis-plugin-grass saga
sudo apt install python3-qgis
```

Consider change the language to English in the global options (easier to follow tutorials).

6.2.2 Plugins

For this course we will use the Open Route Service plugin.

6.2.2.1 Basemaps

Install the useful free basemaps.

> Copy-past into the python console: https://raw.githubusercontent.com/klakar/QGIS_resources/master/collections/Geosupportsystem/python/qgis_basemaps.py

6.2.2.2 Open Route Service

Sign up for an account and create a token. Copy your API.

> In QGIS Plugins → Manage and install plugins → Search for **ORS Tools** → Install Open ORS Tools, provider, settings, and paste API key.

6.2.2.3 qgis2web

This is an useful pluggin to create interactive HTML maps.

> In QGIS Plugins → Manage and install plugins → Search for **qgis2web** → Install qgis2web.

Tutorial here: https://www.qgistutorials.com/en/docs/leaflet_maps_with_qgis2leaf.html

7 R basics

In this chapter we will introduce to the R basics and some exercises to get familiar to how R works.

7.1 Simple operations

- Math operations

```
# Sum
1+1
```

```
[1] 2
```

```
# Substraction
5-2
```

```
[1] 3
```

```
# Multiplication
2*2
```

```
[1] 4
```

```
# Division
8/2
```

```
[1] 4
```

```
# Round the number

round(3.14)
```

```
[1] 3
```

```
round(3.14, 1) # The "1" indicates to round it up to 1 decimal digit.
```

```
[1] 3.1
```

```
# You can use help ?round in the console to see the description of the function.
```

- Basic shortpaths

```
# Perform combinations

c(1, 2, 3)
```

```
[1] 1 2 3
```

```
c(1:3) # The ":" indicates a range between the first and second numbers.
```

```
[1] 1 2 3
```

```
# Create a comment with ctrl + shift + r
# Comments help you organize your code. The software will not run the comment.
```

- Create a table:

  A simple table with the number of trips by car, PT, walking, and cycling in a hypothetical street segment at a certain period.

```
# Define the variables
modes <- c("car", "PT", "walking", "cycling") # you can use "=" or "<-"
Trips = c(200, 50, 300, 150) # uppercase letters modify

# Join in table
table_example = data.frame(modes, Trips)
```

```
# Visualize the table by clicking on the "Data" in the "Environment" page.
# Or use the following function:
View(table_example)
```

7.2 Practical exercise

Import dataset with the number of trips between all municipalities in the Metropolitan Area of Lisbon, Portugal.

```
data = readRDS("data/TRIPSmode_mun.Rds")
```

Take a first look at the data

```
# Summary statistics
summary(data)
```

```
  Origin_mun        Destination_mun        Total             Walk       
 Length:315         Length:315         Min.   :     7   Min.   :     0  
 Class :character   Class :character   1st Qu.:   330   1st Qu.:     0  
 Mode  :character   Mode  :character   Median :  1090   Median :     0  
                                       Mean   : 16825   Mean   :  4033  
                                       3rd Qu.:  5374   3rd Qu.:     0  
                                       Max.   :875144   Max.   :306289  
      Bike              Car             PTransit            Other        
 Min.   :   0.00   Min.   :     0   Min.   :     0.0   Min.   :    0.0  
 1st Qu.:   0.00   1st Qu.:   263   1st Qu.:     5.0   1st Qu.:    0.0  
 Median :   0.00   Median :   913   Median :   134.0   Median :    0.0  
 Mean   :  80.19   Mean   :  9956   Mean   :  2602.6   Mean   :  152.4  
 3rd Qu.:   0.00   3rd Qu.:  4408   3rd Qu.:   975.5   3rd Qu.:   62.5  
 Max.   :5362.00   Max.   :349815   Max.   :202428.0   Max.   :11647.0  
```

```
# First 10 values of each variable
head(data)
```

```
  Origin_mun Destination_mun Total Walk Bike   Car PTransit Other
1  Alcochete       Alcochete 20478 6833  320 12484      833     7
2  Alcochete          Almada   567    0    0   353        0   214
3  Alcochete         Amadora   188    0    0   107       81     0
4  Alcochete        Barreiro   867    0    0   861        5     0
5  Alcochete         Cascais   114    0    0   114        0     0
6  Alcochete          Lisboa  2840   69    0  1994      775     0
```

```r
# Take a look at the data
View(data)

# Check the number of rows (observations) and columns (variables)
nrow(data)
```

```
[1] 315
```

```r
ncol(data)
```

```
[1] 8
```

Part II

Day 2

8 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

8.1

References

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.