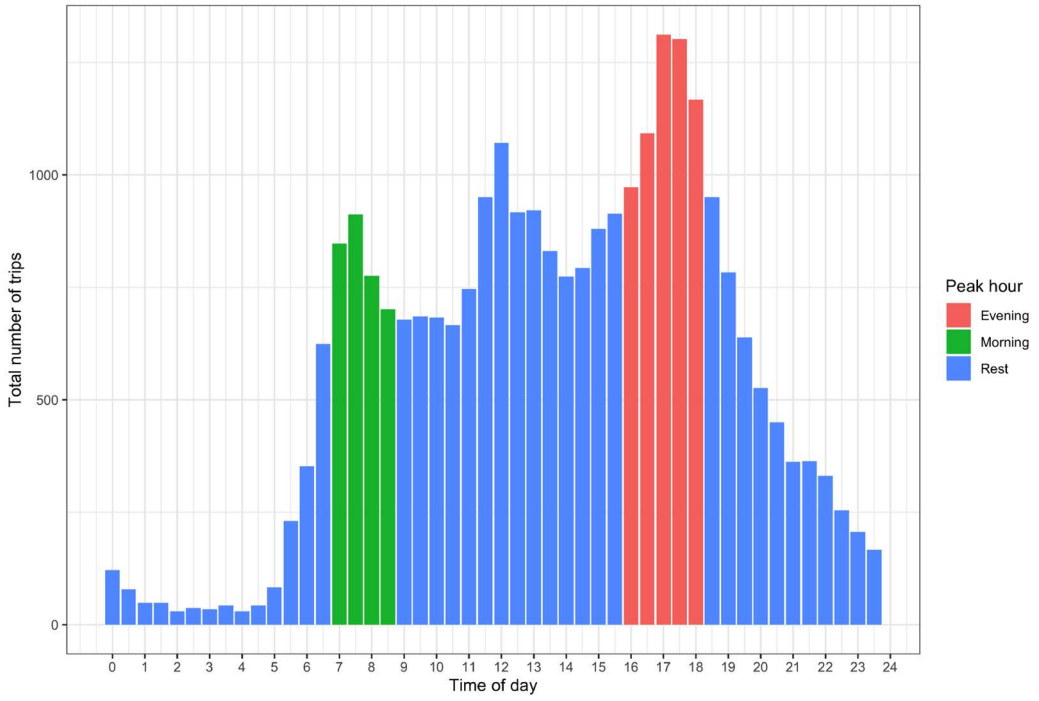

Assignment: Transport Planning Methods

Prof. KW Axhausen

Lucas Meyer de Freitas

Assignment **Fall 2023**

IVT Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

ETH zürich

1 Overview

This exercise has three parts and builds on the open-source software R and R-Studio, which you can download from https://stat.ethz.ch/CRAN and https://www.rstudio.com.

We highlight that basic knowledge of R is required for a successful completion of the exercise in this course. If you don't have it, there is no need to panic. In this case we would recommend you to complete the following R tutorials which will give you the basic knowledge necessary for starting with the assignments (total time investment: ca. 4h):

- Free R (programming language) Tutorial - R Basics - R Programming Language Introduction | Udemy

Also, a great deal of work (if not most of it) in working with data analytics and modelling is actually done before the actual data analysis, namely data processing. While there are many R-packages and functions available, we strongly recommend you to get familiarized with the packages of the tidyverse family because of its structure which makes it very easy to read to code (and therefore understand what it is doing).

Therefore, we strongly recommend all the course participants to go through chapters 1-8 of the tidyverse book (and do the exercises), this will potentially save you a lot of time later on in the course: 1 Introduction | R for Data Science (had.co.nz)

In the first part of the exercise you are asked to clean and describe data using tables and graphs in R. In the second part, you will have to develop and apply different sub-models of the traditional 4-step modelling approach under some assumptions. In the last part, you will carry out a cost-benefit analysis (CBA) concerning the appraisal of a policy scenario.

Please note: Usually, there are many ways to come to a result. What is important is that you show how you obtain the results. Furthermore, the tables and graphs have to be properly formatted, meaning that the legends and values are **readable, well presented and consistent in style (color scheme, text formatting)** in the report. If you don't find solutions to your problems online (you are most likely not the first person to encounter a specific type of problem), feel free to get assistance from us.

The goal of this exercise is that you get a first idea on creating, coding and computing statistics as well as models in R and present them in tables and/or graphs. You have to write a report describing your work in detail. Note that your R-script has to be uploaded as well. A very good practice for yourself when using R studio is to thoroughly comment your code. This will make it much easier for you to debug it and to use it in the future.

1.1 Data

You will work with data from the 2021 Swiss mobility microcensus. You will be asked to sign a data use agreement with the Federal Statistical Office (provided on moodle). While all the specifics are described in the document we highlight some main points below:

- **Under no circumstances is the data to be shared with anyone.**
- **Under no circumstances is the data to be used for any purpose than this course (even for other uses within your studies at ETH).**
- **The dataset should be DELETED from your computed upon completion of the course.**

1.2 Group work

You must work in groups of 2 students (or one single group of 3 if there is uneven number of students in the course). The persons involved have to be named in the report and **you must inform the teaching assistant of who you will be working with by 27.09.2022.**

1.3 Report requirements

You have to hand in a report on this assignment. Each of the three parts of the assignment described in the following has a page limit specified (R-script does not count towards the total and should be submitted as a separate file). The report should describe your work and results in detail. Make sure that you add graphs or tables, if it is needed (and only if it is needed). Make sure that the submission fits the requirements of a scientific report (including the assumption that the reader does not know this assignment). You should use the template provided on https://www.overleaf.com/latex/templates/eth-ivt-strc-2020-template/bpmxdjypsjch to prepare your work if using Latex or the template provided in ivt-termpaper-word.dot (live.com) if working with MS Word. As an ETH student, you have free access to Overleaf, an online Latex editor. Please follow the style guidelines given in Axhausen (2016).

All your results and conclusions of this assignment have to be presented at the end of the semester.

1.4 Date of submission

The full report must be handed in by 05.01.2023, 11.59pm. Since this submission date is after the presentations, we expect each group to incorporate the feedback they receive during the presentations into their report.

However, note that each subtask has its own submission date (See each subtask heading in this document as well as the Schedule on Moodle). Only one member of the group should upload the report and the R script(s) to Moodle.

Only the final submission will be graded. The interim submissions have the purpose of giving you the chance to receive feedback on your work and improve it by incorporating the necessary changes.

2 Part 1: Descriptive analysis (due on: 11.10.2022)

Suggested section length: 10. Page limit for the entire report: 35.

For this exercise you will be needing the following datasets:

- Wegeinland.csv (trips): Contains information on the trips made within Switzerland
- Haushalte.csv (households)
- Zielpersonen.csv (target individual who was interviewed)

In the first part of this assignment you have to clean and describe the data set provided by answering the questions of the following subsections. You will have to decide on your own whether to use just text, tables or graphs to present the data. Note that tables and graphs need to be commented on with respect to their key insights. It is important that you follow the advice for table and graph formatting given in the respective lecture. The task of this assignment is that you practice using graphs and tables for scientific work. Graphs and tables are an important part of scientific work as they can visualize and support the findings of your research.

2.1 Data import and cleaning

Import the data to R-Studio and explore the data sets as well as the codebooks. For the report, only describe in a couple of sentences what cleaning steps you are supposed to do.

R-TIP for importing the data:

```
## Read in and process MZMV data
path_mc <- "local/path/to/data"
df_trips <- fread(paste0(path, path_mc, "wegeinland.csv"))

#R good practice: Always start the naming of data frames with df_
```

Get a feeling of the data-structure by looking into the variable names and types and have a look into a (numeric) variable of choice, by looking at the summary of the trip distances from the trips file.

```
glimpse(df_hh)
summary(df_trips$w_rdist)
```

2.2 Travel behaviour

This part of the assignment deals with the trips reported by the participants. Therefore, you will have to look at only the main mode legs of all trips. Task: Analyze the trips with respect to the following questions.

R-TIP 1: *You will have to join datasets to perform these analyses. For this use the left_join() function which is available in tidyverse. At the "left" side of the join you usually want to have the most expanded dataframe (with most entries).*

R-TIP 2: *Do the analysis by using tidyverse pipes to produce new, aggregated tables. In the style of:*

Tip: *Standard ggplot color palettes make your plots look "cheap". Use different ones, here is a good reference: Top R Color Palettes to Know for Great Data Visualization - Datanovia. But to keep things simple and to be friendly to those with visual impairments, as long as you are not differentiating among too many classes, just differentiate by contrast by using grey color palettes. Don't make things too flashy!*

```
agg_df <- df %>%
 filter(!is.na(X) & AUSNR>0)%>%      #!!
 group_by(GROUPING VARIABLE) %>%
 summarise(PERFORM SUMMARIZING OPERATION)
```

Tip: *For this exercise and as a general professional recommendation:* ***Never use pie-charts!*** *People are very bad at estimating quantities from angles. Always prefer bar-charts. See examples why here: Don't use a Pie Chart | Hugh E. Williams (hughewilliams.com).*

Also, all graphs should be completely labelled, appealing to read and with actually readable legends and numbers. When you prepare your delivery always check if the legends and numeric values on charts and graphs are close to or preferably the same size of the text font in your document.

While the sampling technique used for recruiting individuals for this microcensus attempts to achieve a representative population sample, the sampled population is unbalanced. Therefore, a weighting of the participants in the sample has to be done to account for the weight of each individual with regard to the true population. To account for the weighting, multiply each calculated value with the weight of each individual "WP". For reference: Weights are calculated using a widely used method in sampling studies called "iterative proportional fitting".

1. Mobility by the household location type: What is the average number of trips by urbanization degree of the household location (W_DEGURBA variable). Present the result

in a table.

2. What is the average distance travelled by individual (tip: group in two steps) by urbanization degree of the household location? Present the result in a table.
3. Day-of-Week: What is the average number of trips reported per weekday? What do you observe? Do you have any explanation? (the weekday information is available in the "Tag" variable of the "zielpersonen" dataset). Plot this using a bar-chart.
   Tip: *Graphics should always be completely understandable on its own, without needing text references. Therefore, in the x-axis list the day of the week, not the numeric value of the "tag" variable. You want to have the bars ordered by the weekday. You want to use "mutate" twice here: once to create a new character variable with the name of the day and then to convert this new variable to a factor and then order it so that the plot is neat.*
4. Is this the correct way of representing such information? Compare the plot above to a new plot in which you calculate the average number of trips per individual per weekday. What is a possible reason for the different trend observed?
5. Trip distance 2: How does the travelled distances by individual vary with respect to household income (variable "f20601" of the "haushalte" dataframe)?
6. Mode share: How does the mode share vary by the public transport-accessibility of the household location? (use the aggregated mode definition from the variable "wmittel2a" and the PT Access definition (Güteklassen) from the Variable "W_OEV_KLASSE"). Produce a stacked bar-plot to show the results. To make the graph as informative as possible, show absolute pkm distances on the y-Axis but also include a text (geom_text) in each bin showing the relative percentage within it. Produce the graph once using individual averages and once using total pkm's. What conclusion can you draw from the differences between both graphs?
7. Trip purposes: Do the same analysis as above (for total pkms only) but instead of differentiating by PT Access, differentiate mode-share by trip purpose (Variable "wzweck3"). Not to cramp up the graph too much, only show percentages for the "individual motorized" mode. You also want to tilt the x-axis legend.
8. Make an analysis of your choice: Choose a relation you want to analyse from the dataset and choose an adequate graph to display it.

Present the results in a report where you show the graphics (make sure they are readable!) shortly discuss the results of the graphics.