SABD Project: Urban Mobility Analysis

Part 1 - Project Statement & Organisation

I. Joly, C Bobineau, B. Agard

2023

Table of Content

Statement

The MOBILITY project will be introduced to you through 2 parts statement:

1. First part focuses on the DS project organisation (team work, informatic work) and on the workflow for reproducible document writing
2. Second part focuses on data handling: variables creation, modification, and GIS

Objectives of the Part 1

This first part of the project is dedicated to :

- the data science work organisation: First step is to make this actual statement file reproducible on your own machine
- the team work organisation: Second step is to define team and choose a topic to work on during the semester
- Start looking at the literature relevant to help choosing the topic / discover the topic

Keywords of the literature for the following proposed topics are:

Urban mobility, public transport, motorized mode of transport, transport behavior, mode choice, traveled distance, travel time, residential location

Suggested scientific reviews on the topics are: Transportation; Transportation Research (part A to E); Transportation research record; Transport review; International journal of Transport Economics; International Journal of Choice Modeling

---

Project Statement

- This project is a **3 or 4** students group work.
- All groups work on the same database, with same expectations, but different angles and problematic.
- Precisely, each group has a specific **scientific question** (from the topics list at the end of this paper - topic assignment to group is randomly determined)
- Each topic points:
  1. A variable of interest Y
  2. A Set explicit variables to be **systematically included** in the analysis X and Z
  3. A topic's specific focus to be integrated in the analysis - variables describing this focus will have to be *identified / chosen* and **justified** by the students (new variable W)
- Each topic will be analyzed based on the following approach:
  1. **Present and describe** the dataset you are using (without giving priority to one variable over another): remarkable characteristics, properties of the different variables, underlying structures, creation of variables, factorization, etc.
  2. **Summary statistics**: Produce and draw conclusions from **Univariate and bivariate** study of each of the variables:
     - *Variable to be explained*
     - *Imposed variables* and their contribution to the explanation of the variable of interest
     - *Relevant variable to be chosen for topic's specific focus*)
  3. **Assumptions and Strategy Proposition**: Based on your *knowledge* of the topic, the **literature** (at least one relevant and well chosen paper is expected), and the former **summary statistics**:
     - *Define* and **justify** your strategy of analysis of the topic: what are the expected factors of the variable of interest?; what are your methodological choices (types of analysis, method, model,...)? what is your strategy (subsetting, algorithm, data handling,...)?
  4. The **specific focus** influence on the variable of interest will be examined. For that you will construct a new variable W that will synthesize a subset of other variables. You are asked to :
     - Select all pertinent variables related to the specific focus
     - Prepare all considered variables in order to create groups of users with similar specific focus situation
     - Cluster your dataset using those variables, clusters are stored in new variable W
     - Analyze your clusters relatively to the specific focus variables
     - Eventually iteratively improve your results reconsidering previous steps
  5. **Prediction** of variable of interest Y considering variables X, Z and W.
     - Select (with justifications) different prediction models
     - Make predictions
     - Evaluate the performance of each model.
     - Compare the different approaches
  6. **Explaination** the relationships between the variable of interest and the explanatory variables you have identified
     - *Discuss* the statistical significance (or equivalent notion) of the link between variables)
     - *Interpret* the link between variables in terms of mobility behaviors and analysis
     - *Conclude* on the assumptions you have drawn in the **point 3.**

You will produce a **10 pages maximum report** (eventually with few appendices) dealing with the following points (you will discuss them relatively to your topic):

For each point, specify if applicable if you worked on a subset of observations and / or if you deleted or created variables.

In addition, do not hesitate to collaborate with your comrades working on other variables, the issues being linked and the results of some can also be exploited by others.

The body of the report should be a synthesis of your analyzes, where you present the elements that you think are the most important. Concrete elements (values, tables, figures) should support your assertions, but do not give all the details of your analyzes in the body of the report (the appendices are made for this). For example, if you want to study the correlations of variables, indicate only the most significant in the report, and refer to the appendices for the complete matrix of correlations.

The report must be produced in Rmarkdown (or KnitR) format with all the elements allowing to **reproduce** your analysis. Your files (potentially zip archive) will be uploaded on Caseine and returned by email (iragael.joly@grenoble-inp.fr) on **20/01/2024** at 11:59 p.m. last deadline.

It is **imperative to ensure the reproducibility** of your work, both in the handling of the data and in the production of the report. So be sure to tell the user which packages to install, which intermediate files to run, etc.

Database description - mobility survey

Files list:

- Databases `allgre.PB_V2`
- Variables Dictionary **DICO_VAR_EMD_GRE2010_v22.xls**
- File **SABD_Project_MOBILITY_Part1_2023.Rmd** of this statement. This `Rmd` file creates files to be used in the project. It is advised to execute this file to initiate data handling and operations.
- GIS files: They are grouped in the folder: `FOND_MAP`
- GIS code are described in the variables dictionary

Topics and Data

Data are an extraction from mobility survey in Grenoble in 2010.

Those data are available in R format: `allgre.PB_V2.Rdata`.

Access to suitable MongoDB collection format which will be open during the first lesson on MongoDB. MongoDB collection will give you efficient access to the data and their organisation.

The R `allgre.PB_V2.Rdata` dataframe is needed today to start the project and reproduce this statement file. You will replace this file by the equivalent MongoDB collection.

The data describe the trips made by members of Grenoble households in 2010 (Cerema (2013)). They provide data relating to individual and household characteristics (household structure, motorization, sex, age, professional status, etc.) and mobility practices (numbers of trips, travel times, locations of trips, modes used, reasons for trips, etc.).

The structure of the database is described in the following sections.

List of proposed topics

1. Daily travel times

Analyze the **daily travel times** (variable BTT [1]). Tips: The literature has used the DTT to define *extreme commuters*: people traveling more than 2 hours per day.

- Variable of interest Y: **btt** - Travel Time Budget
- Variable X: **jourdepl** - day of the trip
- Variable Z: **permis** - driving licence holder
- Specific focus W: influence of **housing** on your decision variable Y

2. Mode of transport choice

Analyze the 'car' choice of the mode of travel of individuals and identify the determinants of this choice among the individual or household characteristics and modes of transport.

- Variable of interest Y: **Voiture** (variable to be created according to `mode_depl_ag = 'VP'` or not)
- Variable X: **permis** - driving licence holder
- Variable Z: **age** - age
- Specific focus W: influence of **family structure** on your decision variable Y

3. Mobility behavior

Propose a characterization of the population of mobile and immobile people and identify the determinants of mobility

- Variable of interest Y: **immobil** (variable to be created according to `nbd = 0` or `nbd > 0` - `nbd` is the number of trips)
- Variable X: **age** - age
- Variable Z: **VP_DISPO** - number of available cars in the household
- Specific focus W: influence of **car ownership** on your decision variable Y

4. Household motorization

What are the determinants of household motorization?

- Variable of interest Y: **motorisation** (variable to be created according to `VP_DISPO = 0` or `VP_DISPO> 0` - `VP_DISPO` is the number of available cars in household)
- Variable X: **taillemng** - household size
- Variable Z: **Permis_mng**: at least 1 driving licence in the household
- Specific focus W: influence of **sociodemographic household** on your decision variable Y

5. Engine Types

Study the Grenoble car fleet. What are the determinants of choosing a diesel engine? The question can focus on the 1st vehicle of the household.

[1] French for TTB: Travel Time Budget is the sum of duration of all trips performed in a day by the individual

- Variable of interest Y: **diesel** (variable to be created according to `ENERGIE1 = 3` or `ENERGIE1 <> 3` - `ENERGIE1`is the type of engine of the 1rst car of the household)
- Variable X: **TYPE_HAB** - type of housing
- Variable Z: **taillemng** - household size
- Specific focus W: influence of **occupation** on your decision variable Y

6. Frequence of use of urban public transport

Study the factors associated with frequent (daily) use of urban public transport (`freqtcu`)

- Variable of interest Y: **freqtcu** (variable to be created according to `freqtcu = 1` or not)
- Variable X: **permis** - driving licence holder
- Variable Z: **taillemng** - household size
- Specific focus W: influence of **car ownership** on your decision variable Y

7. Trips during rush hours

Study the factors associated with time of a trip during the rush hours (daily) (`heuredep` and `mindep`)

- Variable of interest Y: **`heuredep` and `mindep`**
- Variable X: **motifor** - trip purpose at origin
- Variable Z: **D12** - traveled distance (distance as the crow flies)
- Specific focus W: influence of **household mobility equipment** on your decision variable Y

8. Traveled distance

Study the factors associated with traveled distance

- Variable of interest Y: `D13`
- Variable X: **motifor** - trip purpose at origin
- Variable Z: **TYPE_HAB** - type of housing
- Specific focus W: influence of **mobility accessibility and speed** on your decision variable Y

Data operations are in two parts.

1. First part describes data preprocessing (off the project) that were performed to create the dataframe. It shows you the operation, the scripts applied before the exercise. *You can not run these instructions*
2. Second part describes processing (on the projet) that are to be performed to manage the data for the project. *You have to run these operations* and eventualy modify them.

Part One - Preprocessing

Database loading and exploration

```
library(tidyverse)
```

Data loading This section describes former data manipulation that lead to the actual database.

The data are extract of half of the real household in the survey

```
#  echo=TRUE, eval=FALSE
# Loading of the original database
load("allgre_TR.RData")
# Extracting half of the household for confidentiality issue
set.seed(123)
allgre$PB <- round( ave(  sample( x= c(0,1), prob= c(0.5, 0.5), size =
         length(allgre$id_men),replace=T) , allgre$id_men, FUN= mean) )
table(allgre$PB)
allgre.PB <- allgre[allgre$PB ==1,]
# Saving
# Exclusion of non interesting variables
allgre.PB_V2 <- allgre.PB[, -c( 5 ,10 ,47 ,49 ,51 ,64 ,81 ,97 ,98 ,101 ,124:136, 139:148, 150, 151)]
save(allgre.PB_V2, file = "allgre.PB_V2.RData")
```

Part Two - Processing introduction

Here starts the real data work

- Here we load the actual database
- We built some count table to check the correct loading of the data
- We built ID variable: `id_depl` to point each unique trip. `id_depl` is composed of the individual ID concatenated with the trip number

```
#  echo=TRUE, eval=FALSE
# Loading of the resulting DF
load("allgre.PB_V2.RData")
table(allgre.PB_V2$nbd)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 1277   91 4023 1967 7037 3649 4063 2386 2063 1413  901  618  412  251  197   75
##   16   17   18   22
##  146   53   36   44
```

```r
# Check of NO_DEPL - number of trips
table(allgre.PB_V2$NO_DEPL)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 1277 7025 6795 4745 4152 2392 1665  990  648  394  237  144   91   55   36   21
##   16   17   18   19   20   21   22
##   16    7    4    2    2    2    2
```

```r
# Check of id_depl - id of trips
table(is.na(allgre.PB_V2$id_depl))
```

```
##
## FALSE  TRUE
## 29425  1277
```

```r
# Creation of new id_depl
allgre.PB_V2$id_depl <- allgre.PB_V2$id_pers * 10  + allgre.PB_V2$NO_DEPL
table(is.na(allgre.PB_V2$id_depl))
```

```
##
## FALSE
## 30702
```

```r
table(allgre.PB_V2$nbd)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 1277   91 4023 1967 7037 3649 4063 2386 2063 1413  901  618  412  251  197   75
##   16   17   18   22
##  146   53   36   44
```

Description of the database content

We handle the variables names to built a 7 columns table containing all variables names and the list and details of each variables in the data.

```r
Ncol <- 7;
v <- Ncol* (Nlin <-  ceiling(length(names(allgre.PB_V2))/Ncol )) - length(names(allgre.PB_V2))
mat <- matrix(c(names(allgre.PB_V2), rep("",v)), ncol = Ncol,
              dimnames = list(1:Nlin,c("col1", "col2", "col3", "col4", "col5", "col6", "col7"))
              )
# content
knitr::kable( mat, digits = 2, caption = "Table: Variable names in *allgre.PB*")
```

Variables Description and labels

See variables dictionary (file: **DICO_VAR_EMD_GRE2010_v22.xls**)

Here follows the column number, the variable name and the variable description

Col #	Var name	Description
1	tir	Drawing number of the observation (often close to the residence zone number)
2	NO_MEN	Household number
3	NO_PERS	Number of the person in the household
4	NO_DEPL	Number of the person's trip
5	zoneres.x	Number of the area of residence (see correspondence file)
6	jourdepl	Day of the move
7	TYPE_HAB	Type of residence of the household
8	TYPE_OCU	Type of occupation of the person
9	Gare2	Department number of the reference sncf station
10	Gare5	Postal code of the reference sncf station
11	telefon	Availability of a telephone
12	annuaire	Present in the telephon directory
13	internet	Availability of an internet connection
14	VP_DISPO	Number of private car available in the household
15	GENRE1	Type of car for the 1st car of the household
17	AN_VP1	Year of entry into service of the 1st househol car
18	PUIS_VP1	Power of the 1st car in the household
19	POSSES1	Type of ownership of the 1st car in the household
20	LIEU_STAT1	Parking place of the 1st car of the household
21	TYPE_STAT1	Type of parking for the 1st car of the household
22	GENRE2, ENERGIE2, AN_VP2, PUIS_VP2, POSSES2, LIEU_STAT2, TYPE_STAT2, GENRE3, ENERGIE3, AN_VP3, PUIS_VP3, POSSES3, LIEU_STAT3, TYPE_STAT3, GENRE4, ENERGIE4, AN_VP4, PUIS_VP4, POSSES4, LIEU_STAT4, TYPE_STAT4	idem for cars n ° 2, 3 and 4 of the household
43	NB_velo	Number of bikes in the household
44	NB_2Rm	Number of 2 motorized wheels in the household
45	COEF_MNG	adjustment coefficient associated with the household
46	zoneres.y	Residence zone number
47	sexe	Gender
48	lien	Link with the household reference person
50	TEL_PORT	Possession of a cell phone
51	mail	Possession of an email address
52	permis	Possession of driving license
53	etabscol	Last school attended
54	OCCU1	Main occupation
55	OCCU2	Other occupation
56	csp	Socio-professional category (PCS)
57	ABO_TC	Possession of a Public Transport subscription in general
58	VAL_ABO	Validity of the TC subscription yesterday

Col #	Var name	Description
59	travdom	Work or home study
60	zonetrav	Zone of the place of work or study (Main occupation)
61	dispovp	Arrangement of a car in general (Travel home-work or study)
62	PBM_STAT	Parking problems in general (at work or study)
63	STAT_TRAV	Parking difficulties at work or study place
64	fqvelo	Frequency of bicycle use
65	FQ2R1	Frequency of use 2 wheels with motors (type 1)
66	FQ2R2	Frequency of use 2 wheels with motors (type 2)
67	fqvpcond	Frequency of use Car with driver
68	fqvppass	Frequency of use Car as a passenger
69	freqtcu	Frequency of use Urban network tag
70	freqtram	Tramway use frequency only
71	freqrurb	Frequency of use Other urban network
72	freqtransisere	Frequency of use of the Cars Transisère network
73	freqter	TER use frequency
74	situveil	Situation of the person the day before
75	zoneres.x.1	Number of the area of residence
76	motifor	Purpose at the origin
77	motoracc	Purpose for the origin of the accompanied person
78	zoneorig	Number of the area origin of the displacement
79	heuredep	Start time (hour)
80	mindep	Start minute
81	motifdes	Purpose at destination
82	motdeacc	Purpose at destination of the accompanied person
83	nbarret	Number of stops in the trip
84	zonedest	Destination area number
85	heurearr	Arrival time (hour)
86	minarr	Arrival minute
87	duree	trip duration (declared)
88	nbmodemec	Number of mechanized mode of transport used in the trip
89	prisecharge	Transport cost are covered
90	D12	traveled distance (distance as the crow flies)
91	D13	traveled distance
92	zoneres.y.1	number of the residential area (should be the same as `zoneres.x.1`)
93	NO_TRAJ	number of the element in the trip
94	TPS_MAP_DEP	Walk time at origin
95	mode	Transport mode
96	ZONE_D_TRAJ	Area at origin of a stop
97	ZONE_A_TRAJ	Area at destination of a stop
98	TPS_MAP_ARV	Walk time at destination
99	NUM_VEH	Number of the vehicle
100	NB_OCCU	Number of occupant
101	LIEU_STAT	Parking place
102	NAT_STAT	Type of parking place
103	durstat	Parking duration
104	autoroute	Use of the highway
105	abonpeage	Public transport pass holder
106	id_men	Household id
107	id_pers	Person id
108	id_depl	Trips id
109	id_traj	Stop id

Col #	Var name	Description
110	nb_pers	Number of persons
111	nbd	Number of trips
112	ntraj	Number of stops
113	btt	Daily travel time
115	Couteff	Transport cost estimation
116	mode_depl_ag	Transport mode aggregated version

Data Base Structure

This is a so-called 'hierarchical' database. It initially consists of 4 files:

1. File *Household*
2. File *Person*
3. File *Trip*
4. File *Route*

In surveys, trips are made up of a series of trips or stages made for a certain reason. For a trip, there may be several routes described, when the trip is made in several stages (for example: a bicycle trip from home to the station, a train trip, then a bicycle trip from the station instead of job). This is represented in the database provided by several lines for the same trip.

On the same hierarchical principle: an individual can make several trips during the day, and a household can be made up of several individuals.

The database provided is the 'finest', since it describes all the household, individual and travel information for each trip. There are therefore as many lines as there are trips.

The structure is managed by the identifying numbers of households, individuals, trips and route:

- `id_men`: Unique Household identifier
- `id_pers`: Unique identifier Person
- `id_depl`: Unique identifier Displacement
- `id_traj`: Unique identifier of the route

The work requested in this project may require work, for example on an individual basis instead of trips. For this, it will be necessary to reduce the database (reduce the number of lines) to keep only information at the individual level (one line per individual).

Definitely MongoDB format will help understanding the data organisation and make your data treatment easier and more efficient

---

Statement for the first class:

- Your first step in this project is to be able to reproduce the statement in pdf format kniting the Rmd file
- This implies
  - to organize your working directories (in your OS)
  - to organize your files (in your OS and report paths in Rmd files)
  - to install all needed libraries in R (RStudio - specifically `knitr` and `tinytex`)
  - to explore the Rmd file chunck by chunck to ensure everything is ok
  - to knit the Rmd in pdf / solve the bugs
  - to enjoy your first reproducible pdf document

References

Cerema. 2013. *Enquêtes Ménages Déplacements « Standard Certu »*. Cerema (ex-Certu). https://books.google.fr/books?id=nhtpmAEACAAJ.

Chang, Winston. 2013. *R Graphics Cookbook*. O'Reilly Media, Inc.

Harrell, F. E. 2013. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer New York. https://books.google.fr/books?id=7D0mBQAAQBAJ.

Munzner, Tamara. 2015. *Visualization Analysis and Design*. CRC Press, Routledge.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.name/knitr/.

———. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Boca Raton, Florida: Chapman; Hall/CRC. https://github.com/rstudio/bookdown.