COMPARING HOSPITALS WITH TRAUMABASE ®

*Under the supervision of **Professor Julie Josse***
*With the help of **Doctor Sophie Hamada** invested in trauma research and organization*

Main goal : **Using *R*, we wanted to explore in a systematic way the Traumabase dataset in order to find differences as well as similarities between hospitals** for both procedures and patient oriented variables.

I. MISSING VALUES

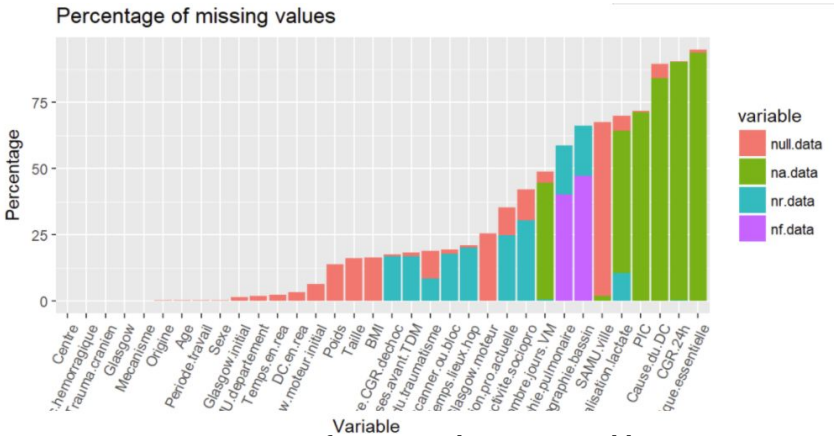
We are working on **real-life data**, therefore it is of the utmost importance to **assess the soundness** of the data in order for our results to be credible. The first order of business is therefore : **missing values**.

- **To start,** we **analysed the missing values** in the dataset
  - **Null** : The **missing** data when we imported the dataset, actually an empty string "" in it ;
  - **NA** : **Not Applicable**, actually as "NA" in the dataset because when the form was completed by doctors it wasn't relevant ;
  - **NR** : **Not Specified**, actually as "NR" in the dataset. We transform this in Null for quantitative variables and keep it for categorical variables ;
  - **NF** : **Not Done**, actually as "NF" in the dataset. Means a medical treatment was not done. We keep it because it is informative.

- **Then :** We **imputed missing (null) values** with the function imputeFAMD from the package missMDA (whose author is Professor Josse).

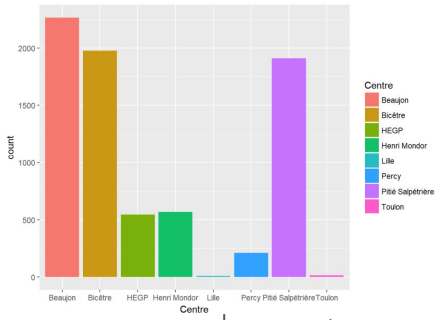
Patients	Hospital	Age	Days in coma	...	Nature of Trauma
Patient 1	Hospital A	85	NA	...	Fall
Patient 2	Hopistal B	54	5	...	Car accident
...	...	...	...	...	
Patient N	Hospital A	33	NA	...	Gun wound

Mock sample of the dataset



Percentage of missing values per variable

Observations per hospitals?



Now that we have imputed missing values, **we may wonder how much observations do we have for each hospital.** Indeed, if we want to do statistics on our dataset, we need to have enough data for each hospital so that our results are statistically sound.

We therefore decide to **remove** the hospitals of Lille and Toulon because we have **too little data on them.**

COMPARING HOSPITALS WITH TRAUMABASE ®

*Under the supervision of **Professor Julie Josse***
*With the help of **Doctor Sophie Hamada** invested in trauma research and organization*

II. SELECTING AND EXPLORING VARIABLES

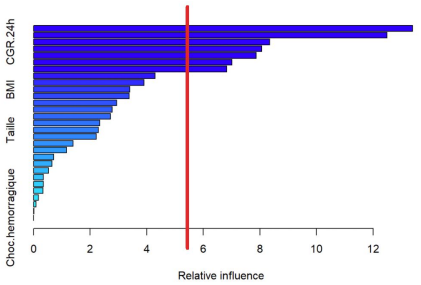
1) SELECTING INSIGHTFUL VARIABLES

Problem : There are many variables. Which one should we study in depth to compare hospitals?

1. Using the package GBM, we try to predict, using non-geographical variables, the hospital an observation was made at ;
2. The idea behind this is to select variables with the biggest relative influence, meaning they were the variables most likely to give us insight on the differences between hospitals.

Solution :

1. By optimizing on the number of trees, we can reach a success rate of **33%** which is twice as much as what we would get by predicting randomly. This means that **the prediction is meaningful**, which is what we were aiming for, more than precision. Given the nature of the data, it is unlikely that we could do much better.
2. Given the drop in relative influence (see red line), we can argue that **there are 7 variables significantly more interesting to compare/distinguish hospitals**.

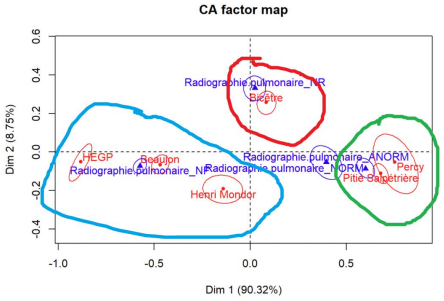


Relative influence of each variable in the GBM algorithm

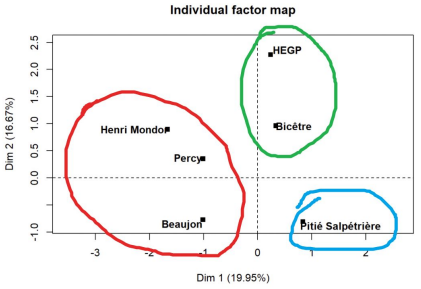
2) EXPLORING SELECTED VARIABLES

Problem : Given a variable, we can plot it, but how to get an insight on which hospitals are similar or different for it?

Solution : For **categorical variables**, we use **Correspondence Analysis** between the hospitals and the variable, and for **quantitative variables** we do a **FAMD** (~mixed PCA) between hospitals and the variable, finally we look at the **coordinates of hospitals** resulting from one or the other method. We do this for the **7 variables**.



CA to analyse correspondence between X-RAYs and hospitals



Hospital factor map in the FAMD between time before going to the Operating room and hospitals

COMPARING HOSPITALS WITH TRAUMABASE ®

*Under the supervision of **Professor Julie Josse***
*With the help of **Doctor Sophie Hamada** invested in trauma research and organization*

III. CLUSTERING HOSPITALS GIVEN CLUSTER FOR EACH VARIABLE

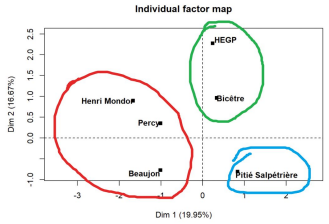
Remark : When looking at the plots from the previous slide, it seems natural to try and **summarize them by doing clustering on hospitals coordinates** in those plots using **K-Means** with the elbow method.
Problem : For each hospital, for a given variable we now have a cluster name. How can we summarize that information?
Solution : We do one last MCA and apply hierarchical clustering on the coordinates of hospitals in that plot :

COMPLETE PIPELINE :

Dataset

CA or FAMD

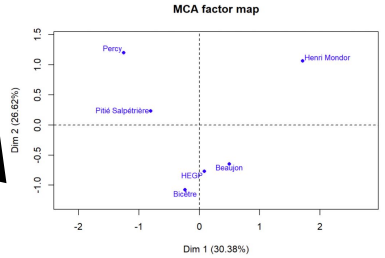
For each of the 7 variables



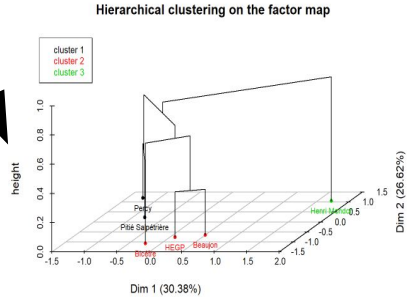
K-Means

Hospital	Var 1	...	Var 7
Hospital 1	Cluster 2	...	Cluster 3
…	...	...	...
Hospital 7	Cluster 1	...	Cluster 3

Multiple Correspondence Analysis



Hierarchical clustering



Cluster of hospitals

Conclusion :

1. Thanks to our analysis, we are able to say : **we have 3 clusters of hospitals who "behave" significantly in a similar fashion** ;
2. We are able to say **for which variables they are similar** and **for which they are different from other hospitals** ;
3. We are also able to say **why** and **how for each of the 7 selected variables**.

Our work will be used by Doctor Sophie Hamada to steer discussions in regular meetings between hospitals' directors aiming at improving major trauma care.