

Université Cergy Paris



CERGY PARIS

UNIVERSITÉ

Diplôme universitaire :
Data Analyst

UE 1 :
Data manipulation and preprocessing

January 6, 2023

Haury Fabien

Contents

1	Data presentation	3
1.1	PhD V2	3
1.1.1	Presentation	3
1.1.2	Nature of the variables	4
2	Missing data	5
2.1	Visualization of all data	5
2.2	Visualization of data for each value of the Status variable	6
3	Main problems detected	8
3.1	Exploration of data for defense dates	8
3.1.1	Distribution of defense by month	8
3.1.2	Distribution of defenses for each year	9
3.1.3	Percentage of defenses per month	9
3.1.4	Evolution of defenses on January 1st per year	11
3.2	Analysis of homonym issues	12
3.2.1	Cécile Martin namesake	12
3.3	Conclusion	13
4	Outliers	14
4.1	Presentation	14
4.2	Search for outliers	14
4.3	Outliers analysis	16
4.3.1	Analysis of outliers supervisors between 40 and 140 supervised theses	16
4.3.2	Analysis of outliers supervisors with more than 140 supervised theses	16
4.4	Conclusion	16

5	Preliminary results	18
5.1	Presentation	18
5.2	Evolution of languages over time	19
5.2.1	Entire dataset	19
5.2.2	Period from 2004 to 2018	19
5.3	Référence bibliographique	20
5.4	Conclusion	21
6	SQL	22
7	Bonus work	23
7.1	Presentation	23
7.2	Missing data	23
7.3	Problems with academic discipline	24
7.3.1	Problem of genres by disciplines	24
7.3.2	Problem of languages according to disciplines	26
7.4	Web scraping	27
	List of Figures	29
	List of Tables	31
	Bibliography	32

Part 1

Data presentation

1.1 PhD V2

1.1.1 Presentation

The PhD V2 dataset is based on information retrieved from the theses.fr website. It gathers a set of information centered on the theses such as the author, these director, language etc.

The variables are present separately each for a total of 447,644 rows and 18 columns.

Variable	Nb distinct	Variable	Nb distinct
Auteur	430277	Statut	2
Identifiant auteur	313775	Date premiere inscription soutenance	4010
Titre	446815	Date de soutenance	3992
Directeur de these	159019	Year	45
Directeur de these (nom prenom)	159021	Langue de la these	206
Identifiant directeur	98907	Identifiant de la these	447572
Etablissement de soutenance	568	Accessible en ligne	2
Identifiant etablissement	573	Publication dans theses.fr	2765
Discipline	24263	Mise a jour dans theses.fr	2634

Table 1: Distinct number by variables of the dataset based on the website theses.fr

Table 1 shows the distinct number of observations for each of the variables in the dataset. This allows us to have an overview of the data.

1.1.2 Nature of the variables

- Auteur : Class "Character"
- Identifiant auteur : Class "Character"
- Titre : Class "Character"
- Directeur de these : Class "Character"
- Directeur de these (nom prenom) : Class "Character"
- Identifiant directeur : Class "Character"
- Etablissement de soutenance : Class "Character"
- Identifiant etablissement : Class "Character"
- Discipline : Class "Character"
- Status : Class "Character"
- Date de premiere inscription en doctorat : Class "Character"
- Date de soutenance : Class Character
- Year : Class "Double"
- Langue de la these : Class "Character"
- Identifiant de la these : Class "Character"
- Accessible en ligne : Class "Character"
- Publication dans theses.fr : Class "Character"

We can see that all the variables except one are of Class "Character". It makes sense for the variable Year to be of Class "Double". However, for an ease of use, to convert the variables **Date de premiere inscription en doctorat**, **Date de soutenance**, **Year**, **Publication dans thèses.fr** et **Mise a jour dans theses.fr** in Class "Date" is preferable. In the same way, the variables **Status**, **Language of the theses**, **Accessible en ligne** can be changed in factor.

Part 2

Missing data

2.1 Visualization of all data

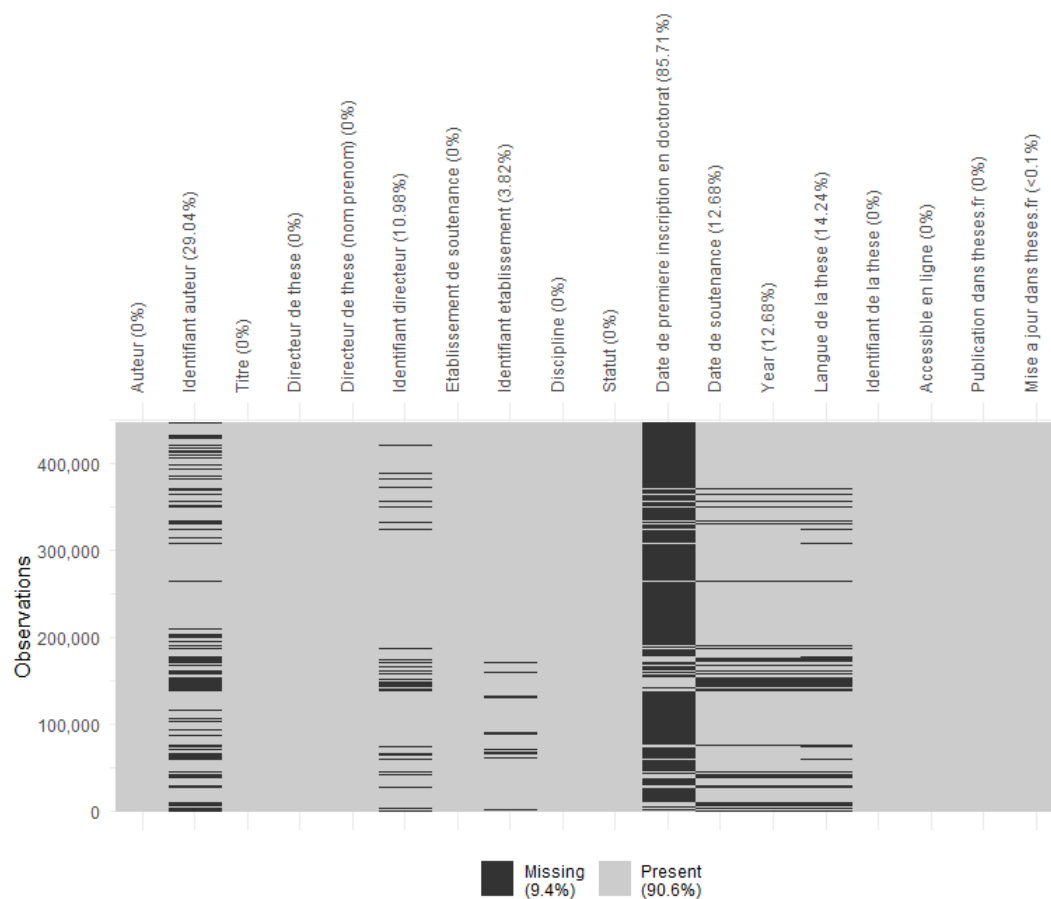


Figure 1: Visualization of missing data on the entire dataset

Figure 1 shows a visualization of missing data. There is 9,4% of missing data. The variable with the most missing data is Date de premiere inscription en doctorat with 85,71%. A pattern seems to emerge between Date de premiere inscription en doctorat and Date de soutenance, Year, Langue de la these. It seems that for a data missing in Date de premiere inscription en doctorat, the data of Date de soutenance, Year, Langue de la these are missing.

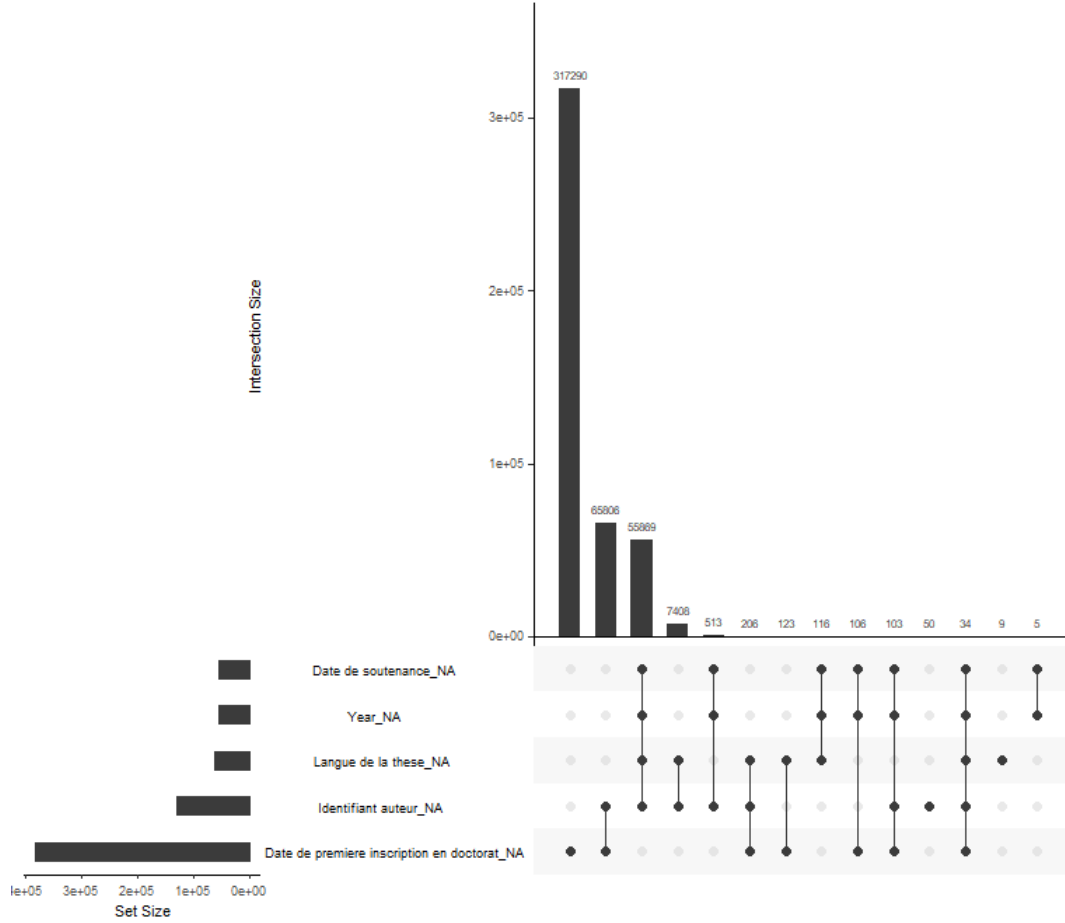


Figure 2: Missing data model

Figure 2 allows us to visualize the links between missing data of different variables. We can confirm the hypothesis of a link between Date de premiere inscription en doctorat, Date de defense, Year, Langue de la these.

2.2 Visualization of data for each value of the Status variable

As we have just seen, there is a link between Status and Date de premiere inscription en doctorat, Date de defense, Year, Langue de la these. We will separate Status into two: enCours (inProgress) et soutenue (Supported).

Figure 3 shows a visualization of missing data for Status: inProgress. 54% of the data are missing, the vast majority of which are attributable to the variables Date de defense, Year, Langue de la these. It can be seen that very little data is missing for Date de premiere inscription en doctorat, on the contrary for Date de defense, Year, Langue de la these for which very little data is present.

Figure 4 shows a visualization of missing data for Status: Supported. 20% of the data are missing, the vast majority of which are attributable to the Date de premiere inscription en doctorat variables.

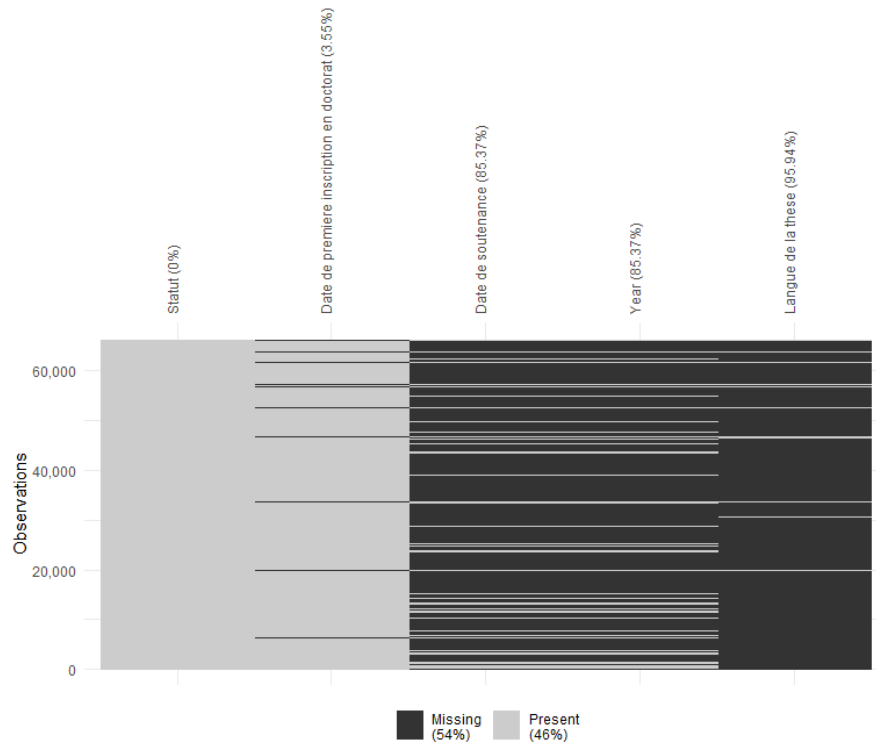


Figure 3: Visualization of missing data for the inProgress status value

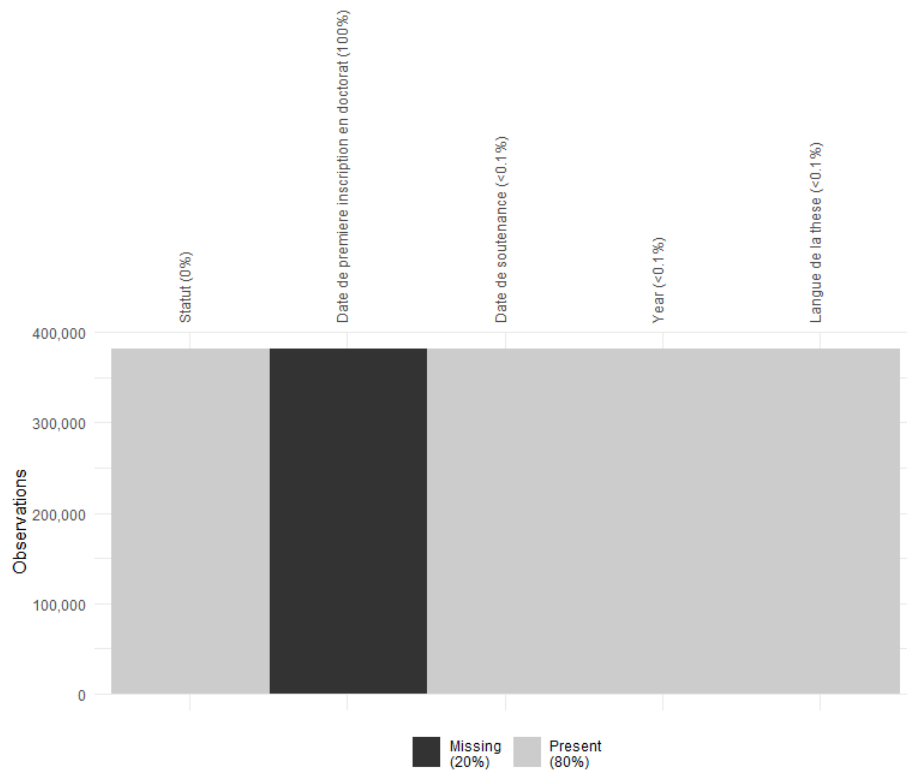


Figure 4: Visualization of missing data for the Supported status value

Part 3

Main problems detected

3.1 Exploration of data for defense dates

3.1.1 Distribution of defense by month

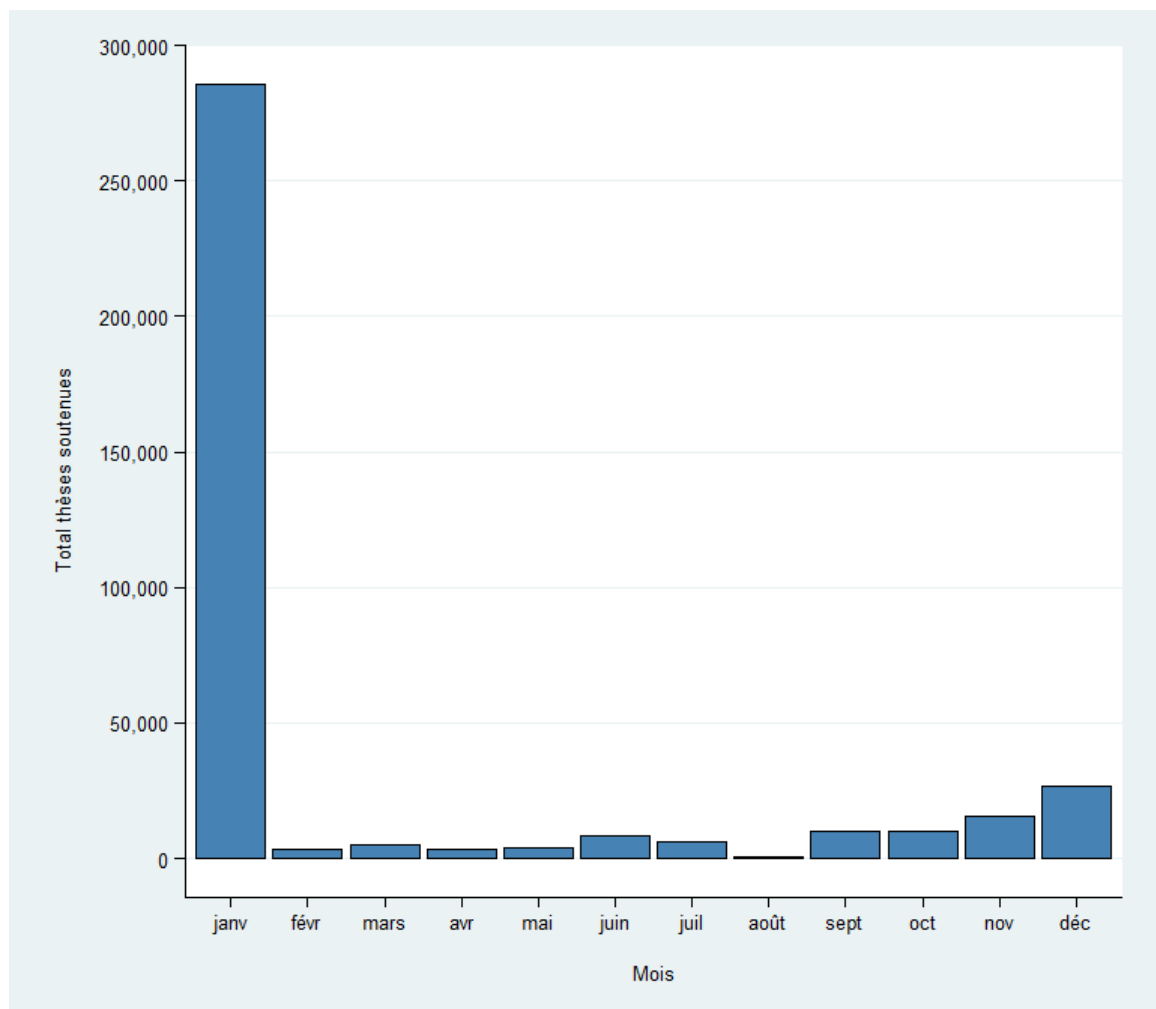


Figure 5: Distribution of defense by month, period 1984-2018.

Figure 5 represents the distribution of theses defenses per month for the period from 1984 to 2018. The month January is the most represented defense month, with around 275,000 theses defended for this month. The period from February to July is almost uniform, with several thousands theses defended. The month of August represent a drought, finally the period from September to December shows an increase in

defenses with approximately 12,000 theses defended in September and approximately 25,000 theses defended in December.

3.1.2 Distribution of defenses for each year

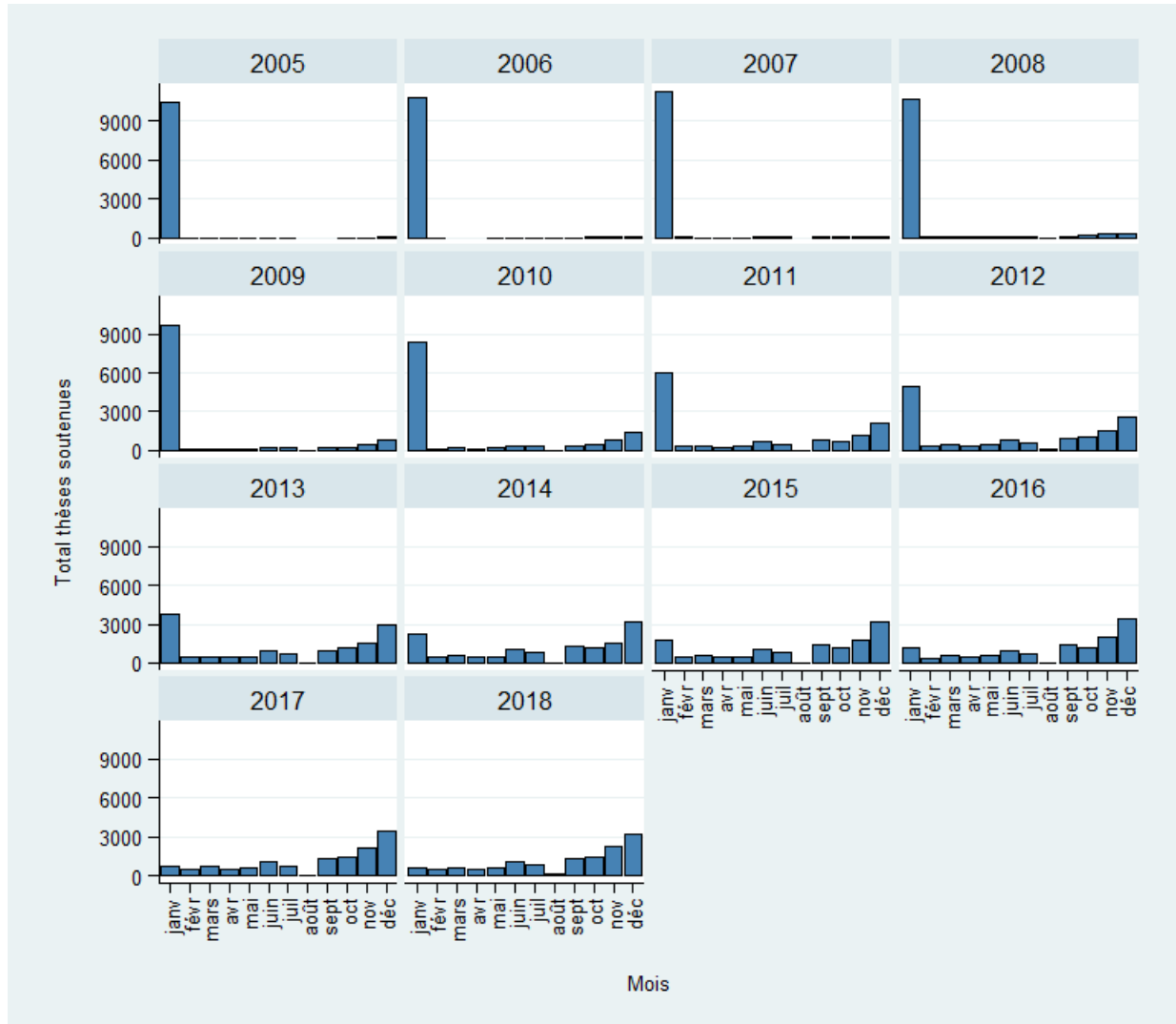


Figure 6: Distribution of defenses per month per respective year, period 2005-2018

Figure 6 represents the distribution of defenses per month per respective year for the period from 2005 to 2018. The period from 2005 to 2010 is strongly centered on the month of January with, ± 9000 theses on average. the following period from 2011 to 2018 shows the fall of the month January to ± 800 theses while the other months all increase, especially the month of December which goes from ± 1500 theses to more than 3000 theses defended.

3.1.3 Percentage of defenses per month

Figure 7 represents the defense percentage by month with its standard deviation without filtering the January 1st for the period from 2005 to 2018. The month of January represent nearly 50% of defenses dates with a standard deviation of 32, 5%. This standard deviation is none other than the representation of the evolution seen in Figure 6.

Since January 1st is a public holiday in France and the day after New Year's Eve, it is plausible that there is a problem with this particular day. Let's filter that particular day and look at the result.

Figure 8 represents the defense percentage by month with its standard deviation, with filtering the

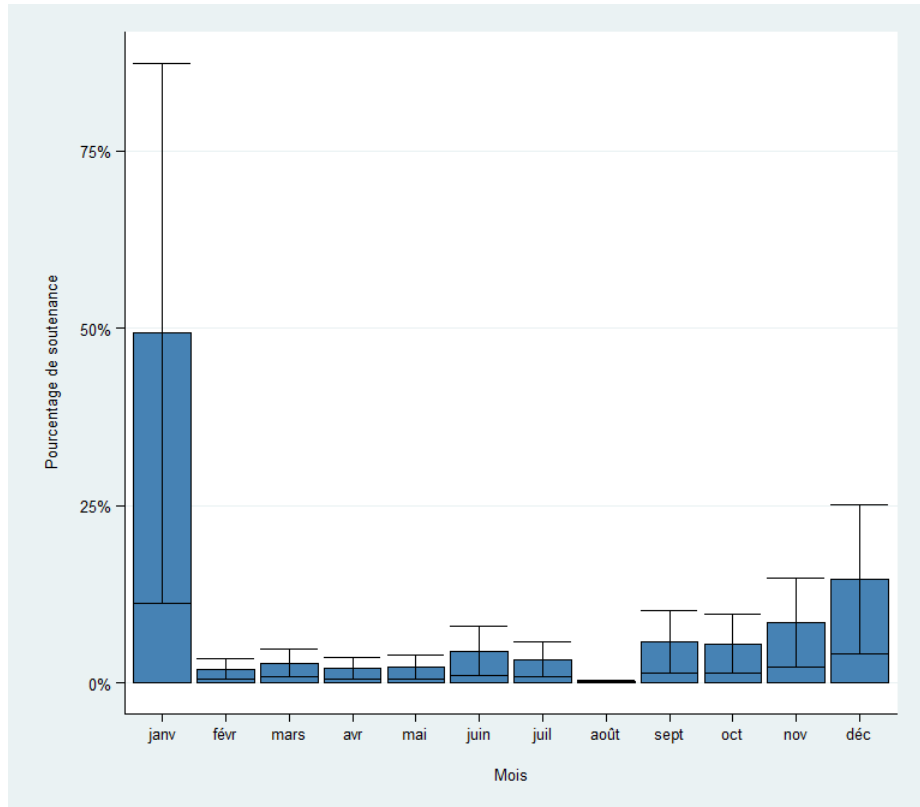


Figure 7: Percentage of defenses per month with standard deviation, period 2005-2018, without filtering out the first of January

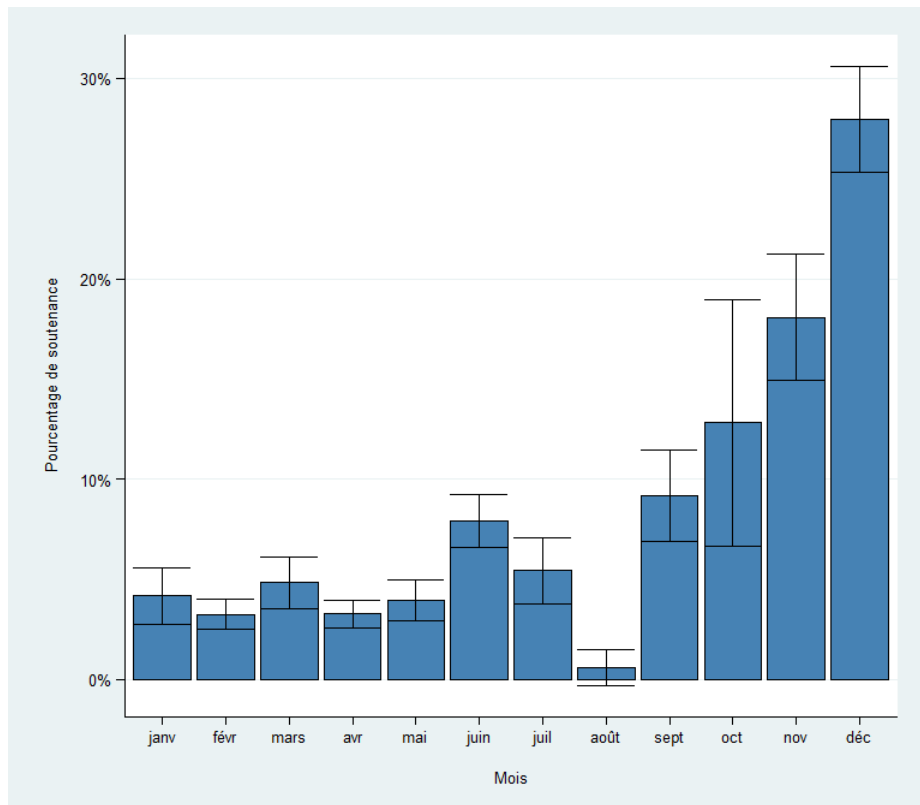


Figure 8: Percentage of defenses per month with standard deviation, period 2005-2018, with filtering the first of January

January 1st for the period from 2005 to 2018. After filtering the January 1st, it clearly appears that a majority of defenses are held in December with nearly 30% of the total. The month of June, July and the Fall period cover a good chunk of the remaining. Several hypotheses can be envisaged to explain those behaviors:

- For the month of June and July, it is possible that doctoral students want either to finish their theses before the holidays or finish their theses to be able to be assigned to a teaching post for the following academic year.
- For the Fall period, it seems to be divided in two parts: Fall and December.
 - As we can see, the defenses in August are few in numbers, it is possible to imagine that this month serves as a rest period or as a proofreading/finishing period for the defense for the Fall.
 - For the month of December, a possible hypothesis is simply to validate your thesis in order to be able to announce it during the end-of-year festivities or to finish your thesis before the end of the calendar year to be eligible for prizes/awards during the following year.

3.1.4 Evolution of defenses on January 1st per year

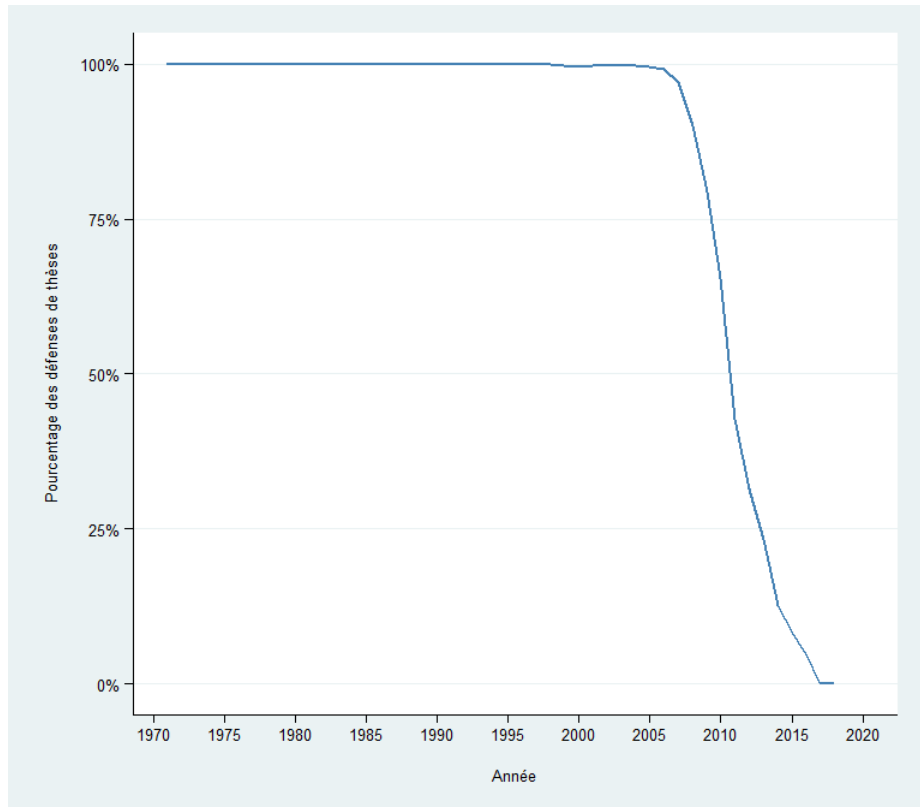


Figure 9: Evolution of defenses on January 1st per year over the entire set of data

Figure 9 represents the evolution of defenses held on January 1st over the year for the entire dataset. For the period from 1970 to 2005 almost all defenses took place on that day. From 2005 to 2018 the curve decreases from nearly 100% to close to 0%. This is also shown in Figure 6

3.2 Analysis of homonym issues

3.2.1 Cécile Martin namesake

Analysis of the data set

Variable	Nb distinct	Variable	Nb distinct
Auteur	1	Statut	1
Identifiant auteur	4	Date de premiere inscription en soutenance	1
Titre	7	Date de soutenance	7
Directeur de these	7	Year	7
Directeur de these (nom prenom)	7	Langue de la these	2
Identifiant directeur	7	Identifiant de la these	7
Etablissement de soutenance	7	Accessible en ligne	2
Identifiant etablissement	7	Publication dans theses.fr	3
Discipline	7	Mise a jour dans theses.fr	5

Table 2: Distinct number of PhD v2 variables

Table 2 shows the distinct number of observations for each of the variables in the dataset. That allows us to have an overview of the data.

There are seven Cécile Martin in the dataset, however, only four are unique. Each of theses was done in a different establishment with a different supervisor as well as in a different discipline. Only one status exist: Supported. A single theses was defended in a language different from the others/

Distribution of years of defense for the unique author identifier

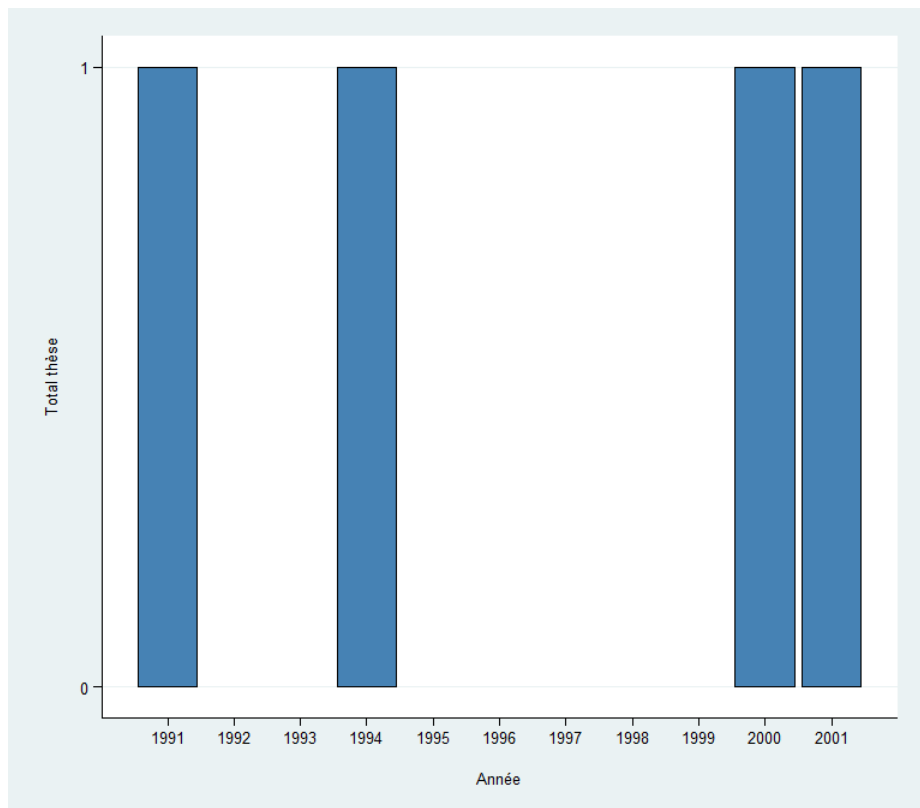


Figure 10: Histogram of years of defense for the unique identifier

Figure 10 represents the distribution of years of defenses for the unique identifier. We note that the data is separated into three periods, 1991, 1994 and 2000/2001. If the graph does not allow us to conclude, an

exploration "by hand" can allow us to draw some conclusions.

Discipline	Date de soutenance
Neurosciences	1991-01-01
Sciences biologiques et fondamentales appliquees. Psychologie	1994-01-01
Sciences biologiques fondamentales et appliquees. Sciences medicales	2000-01-01
Genie des procedes industriels	2001-01-01

Table 3: Discipline defense date for unique identifier

Table 3 contains four different disciplines as well the respective defense dates. We can conclude that the discipline related to industrial process engineering (Genie des procedes industriels) has no connection with the other three.

We also see that two of the three remaining theses were defended in the same field: Fundamental and Applied Biological Sciences (Sciences biologiques et fondamentales appliquees). Their defense date is sufficiently far apart to make it possible to assume that this two theses were, possibly, defended by the same person.

3.3 Conclusion

We set ourselves the problem of studyingg the namesakes Cécile Martin. During this analysis, we found a duplicate issue and pushed the analysis further.

We have potentially solved the problem of Cécile Martin namesakes duplicates by showing that it is plausible that one of these duplicates is indeed the same person based on the discipline of the theses and their date of defense. But even more, it showed us the usefulness of doing these analyses while keeping the other variable to be able to discern patterns.

Part 4

Outliers

4.1 Presentation

The dataset contains the information for the period from 1984 to 2018. There are a total of 308,587 rows and 66,148 distinct directors, 56,680 distinct director IDs, and a total of 1,9% missing data. It therefore appears that duplicates exist, given the number of directors in relation to the number of rows. Similarly, it appears that certain director IDs are duplicate too.

4.2 Search for outliers

Let's find out from how many directed theses we can consider a director as an outlier. Let's start with a histogram to see the distribution.

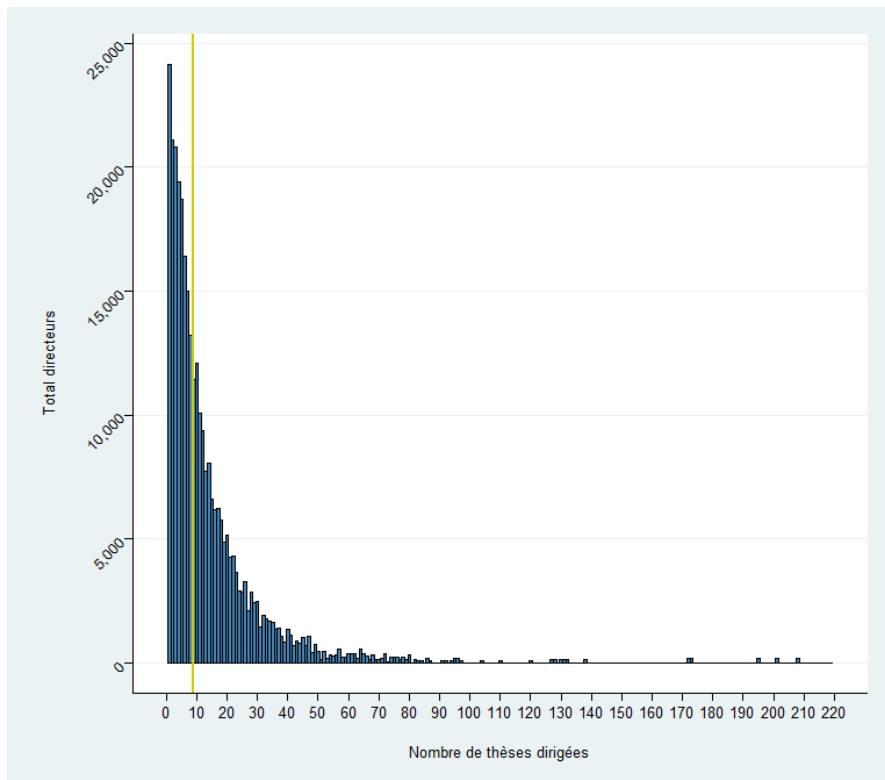


Figure 11: Distribution of total supervisors by number of theses supervised, with median

Figure 11 represent the distribution of total supervisors by number of theses supervised over the entire

dataset with median line. The median line allows us to see that 50% of supervisors have surveyed between one and nine theses. It is clear that there are outliers, a number of supervisors have surveyed more than 100 theses!

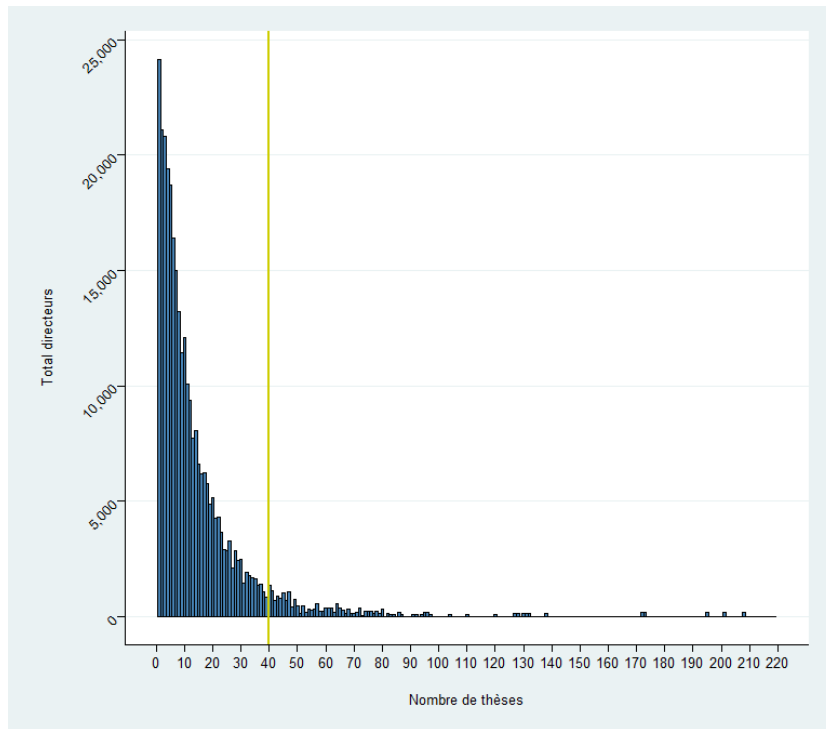


Figure 12: Distribution of total supervisors by number of theses supervised, with line representing the start of outliers

Figure 12 represents a histogram of the number of directed theses on the entire dataset, with a representation by a vertical line marking the separation between the non-outlier data and the outlier data. This line was calculated by the Interquartile Range (IQR) method.

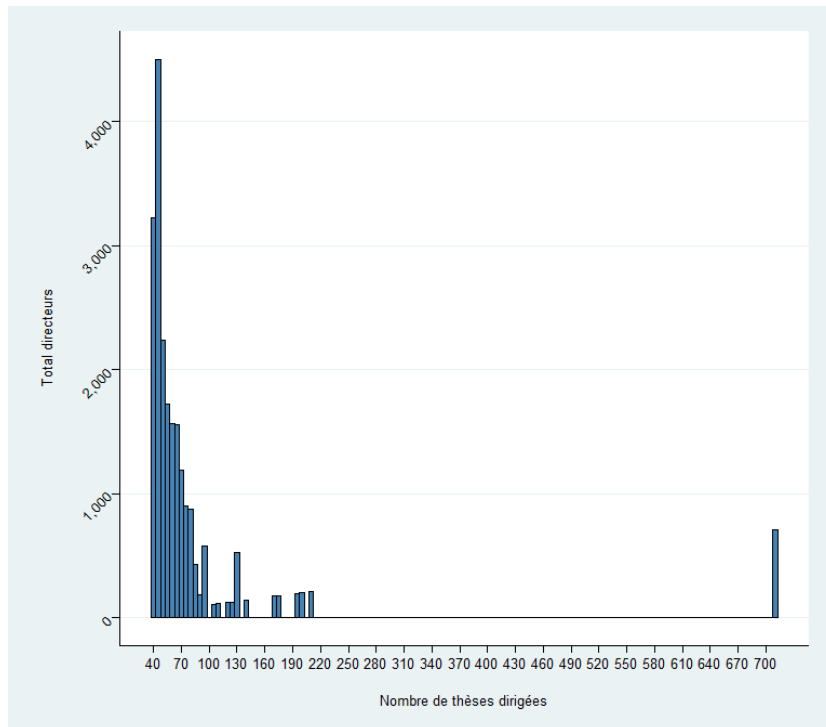


Figure 13: Distribution of total outlier supervisors by number of supervised theses

Figure 13 represents the distribution of the total of outlier supervisors by the number of supervised theses. Now that we have targeted the anomalies, let's take a closer look at them. For a greater clarity, the outliers

will be divided into two parts: from 40 to 140 supervised theses and from 140 to 250 supervised these.

4.3 Outliers analysis

4.3.1 Analysis of outliers supervisors between 40 and 140 supervised theses

There are 20,068 rows for 367 distinct directors as well as 447 distinct director IDs. This allows us to see that numerous directors are duplicated, but also that a certain number of directors share the same identifier, which is not possible.

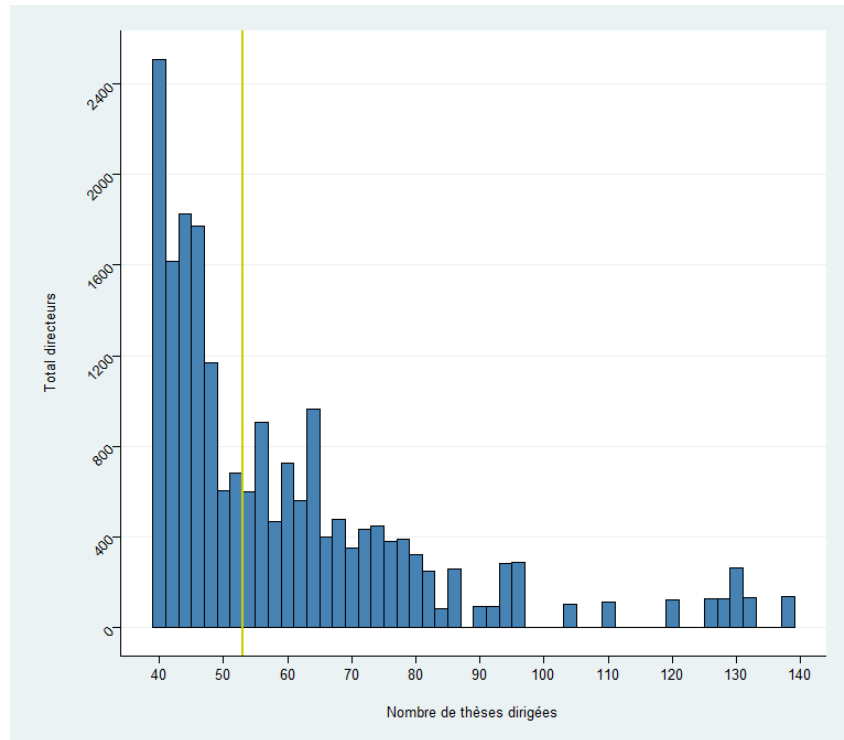


Figure 14: Distribution of outliers between 40 and 140 supervised theses, with median

Figure 14 represents the distribution of outliers between 40 and 140 supervised theses with a median line.

4.3.2 Analysis of outliers supervisors with more than 140 supervised theses

There are 1660 rows for six separate directors as well eight separate director IDs. This allows us to see that numerous directors are duplicated, but also that a certain number of directors share the same identifiers, which is not possible. There is a group of extreme anomaly with a total of supervised these over 700. Let's take a closer look at what it contains. There are 711 lines for a unique director. It seems that this anomaly is the result of a single director! The name of the single director is: "Unknown theses director".

By combining the two pieces of information cited above, it is obvious that the correlation between the two variables is strong, but it is also plausible to hypothesize that this single director is in fact a multitude of different directors. When entering information, either this was unknown or it is the default name, or an input error was made.

4.4 Conclusion

We set ourselves the problems of studying theses supervisors in order to find the outliers. During this analysis, we found many duplicates within the chosen variables, as well as differences in the outliers themselves.

We have just seen the outliers only in relation to the variables: thesis director (surname first name) (directeur de these (nom prenom)), Director identifier (identifiant directeur). If this allowed us to better see the distribution and the problems related to those variables, whatever variable we choose to analyze, in the future, should be done using the entire dataset and not just the variables of interest in order to be able to discern pattern through it. (i.e - link between the extreme outliers and their establishment identifier to determine if the problem comes from a particular establishment?)

Part 5

Preliminary results

5.1 Presentation

this chapter will focus mainly on the evolution of the languages used in the theses. To do this, the variable `Langue de la these` is renamed `Langue` and the observations are grouped together and put change into factor as follows:

- "fr" is renamed "Français"
- "en" is renamed "Anglais"
- "enfr" and "fren" are renamed to "Bilingue"
- missing data are renamed to "NA" (note that we are deleting their status of N/A and count them as separate language)
- all the remaining languages are renamed "Autres"

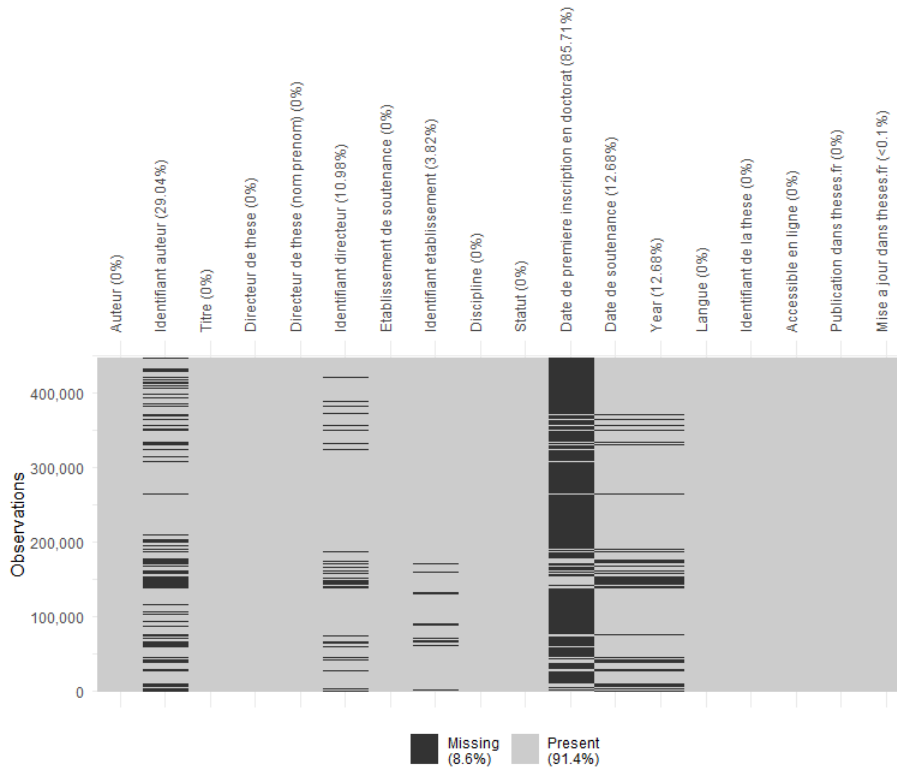


Figure 15: Visualization of missing data

5.2 Evolution of languages over time

5.2.1 Entire dataset

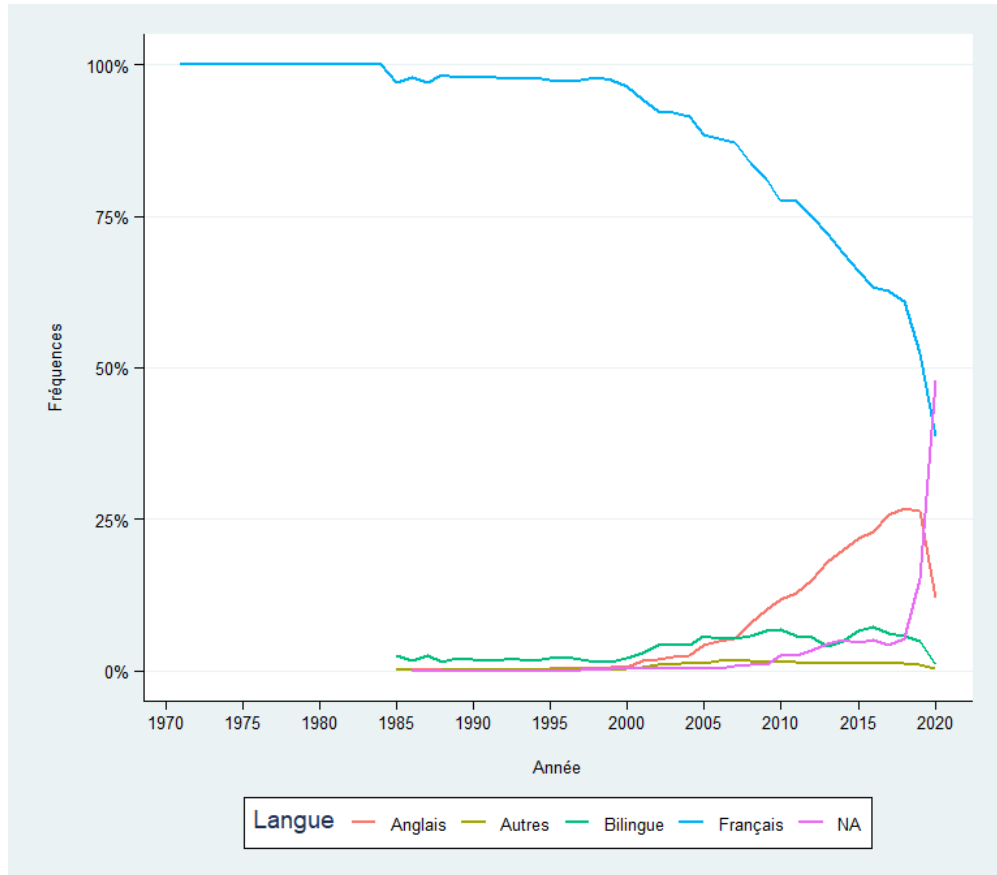


Figure 16: Evolution of the frequencies of the different languages use over the years

Figure 16 represents the evolution of the use of different languages over the year for the entire dataset. We can see four periods, the first from 1971 to 1985 allows us to note the absence of non-french language. The second from 1985 to 2000 shows the appearances of other languages, with French evolving at 95%. the third period from 2000 to 2018 shows a growth of English going from nearly 0% to more than 25% and a fall of French from 95% to 60%, finally from 2019 to 2020 shows a fall of languages and increase in missing data.

5.2.2 Period from 2004 to 2018

Why start in 2014? Starting in 2004 seems a relevant choice due to the implementation of the BMC (Bachelor-Master-Doctorate) system in France, which consists of harmonizing the university system at a European level. Therefore, studying this period makes it possible to see the evolution of languages after its establishment and therefore to note change that result from it. The chosen period ends in 2018 to avoid side effects due to missing data after this year.

Let's calculate the percentage change between 2004 and 2018 : Table 4

We can see in connection with Figure 17 the evolution of languages.

Langue	count 2004	count 2018	% changement
Français	9371	7807	-16.7
Anglais	267	3429	1184.3
Bilingue	435	741	70.3
Autres	137	155	13.1
NA	40	464	1060

Table 4: Evolution of languages for the period from 2004 à 2018

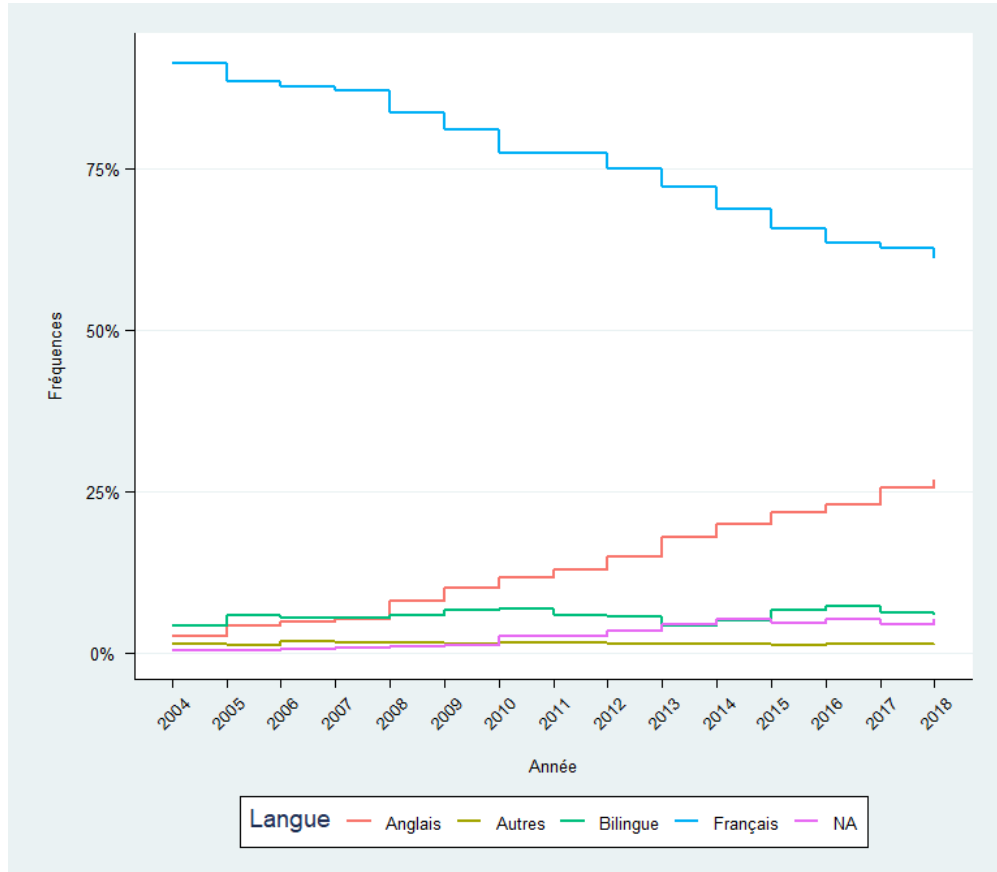


Figure 17: Evolution of the frequencies of use of different languages over the year from 2004 to 2018

Figure 17 shows the evolution of the frequencies of use of different languages over the year. We can see that French drop from 80% to 63%. English has increased from 5% to 27%. The other languages as well as the theses defended in bilingual remain constant with 2% and, 6% respectively. An increased of 6% of missing data is to be noted.

5.3 Référence bibliographique

As Martin (2015) points out, the reporting of doctoral theses: «Ce sont 10 000 doctorats environ qui sont délivrés en France chaque année. this figure uncreases continuously, by 32% between 2005 and 2012».[1]. We will verify this figure.

Table 5 represents the total languages for the year 2005 and 2012 and will be used to calculate the percentage increase.

to calculate the percentage increase, you must:

1. Calculate the coefficient of increase:

$$\frac{13985}{10561} = 1.32$$

Langue	count 2005	count 2012
Français	9352	10477
Anglais	437	2089
Bilingue	611	771
Autres	121	184
NA	40	464
Total	10561	13985

Table 5: Total language for the year 2005 and 2012

2. Multiply by 100 to get a percentage

$$1.32 * 100 = 132\%$$

3. Subtract 100% because we assume that the initial value was 100%

$$132\% - 100\% = 32\%$$

We find the same result of 32% for the same period.

5.4 Conclusion

We proposed to study the evolution of the languages used for the defense of theses. during this analysis, we focused in particular on the period 2004 to 2018 to study the language changes due to the BMD system. If this analysis does not allow us to affirm that the BMD system influenced the choices of languages for theses defense, it did allow us to observe that English has been constantly increasing since the beginning of the 2000s.

Part 6

SQL

This part is an introduction to simple SQL request and was done through DataCamp SQL courses.

Question 1 : What is the duration of the shortest film produced between 1942 and 1968 inclusive?
Open Secret in 1948 with a running time of 68 minutes.

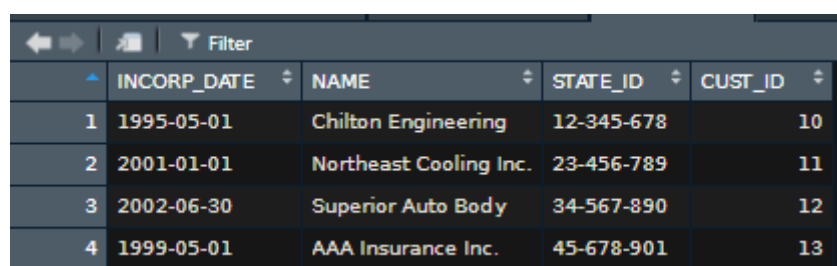
Question 2 : What is the average duration of films made between 1954 and 1967 inclusive?
The average duration of films made between 1954 and 1967 is 132.80 minutes.

Question 3 : Over the period 1960-1970 (inclusive), how many distinct languages were used in films?
There are four different languages. (French, German, Italian, English)

Question 4 : Count how many films were produced after the 2000s in French or Spanish.
There were 100 films produced.

Question 5 : Identify the French-language film that brought in the most money between 1990 and 1999 inclusive.
The Red Violin, 1998, grossing \$9,473,382

Question 6 : Give 5 films in English starting with the letter Z, all eras combined.
Zero Effect, Zoolander, Zoom, Zodiac, Zombieland



The image shows a screenshot of a database interface, likely Wampserver, displaying a table with 5 columns: INCORP_DATE, NAME, STATE_ID, and CUST_ID. The table contains 4 rows of data. The interface includes a 'Filter' button and a table with alternating light and dark gray rows.

	INCORP_DATE	NAME	STATE_ID	CUST_ID
1	1995-05-01	Chilton Engineering	12-345-678	10
2	2001-01-01	Northeast Cooling Inc.	23-456-789	11
3	2002-06-30	Superior Auto Body	34-567-890	12
4	1999-05-01	AAA Insurance Inc.	45-678-901	13

Figure 18: Wampserver manipulation result table

Part 7

Bonus work

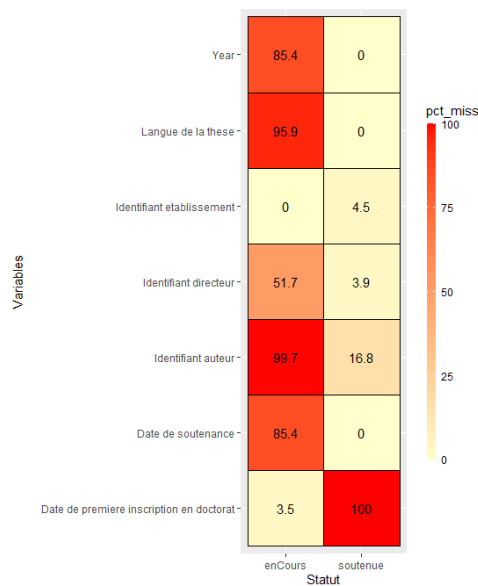
7.1 Presentation

This chapter will be dedicated to missing data with a different approach from chapter 2, to the evolution of male/female relationship and languages within university disciplines and finally to web scrapping to reproduce the data set.

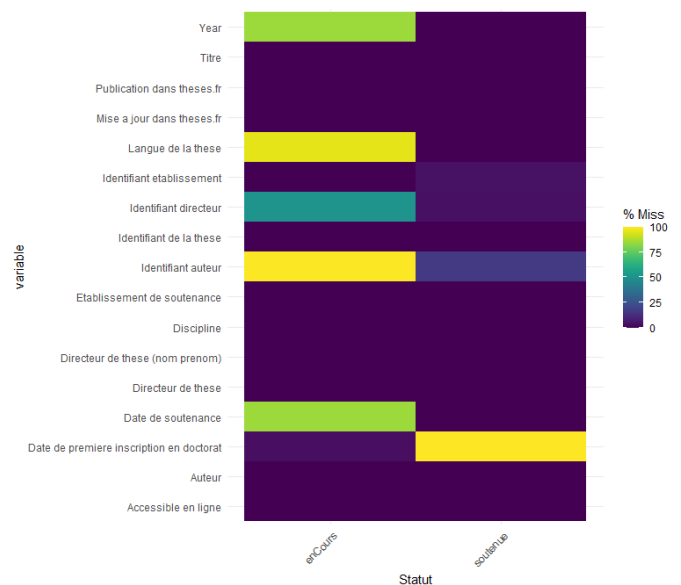
We will also use a new dataset named PhD_V3. This is a cleanup version that also contains new variables such as gender or a recoding of disciplines variables.

7.2 Missing data

In this section, we will represent missing data using heatmap



(a) Heatmap of missing data by Status



(b) Heatmap of missing data by Status for the entire dataset

Figure 19: Heatmap des données manquantes

7.3 Problems with academic discipline

7.3.1 Problem of genres by disciplines

PhD V2

since the author's gender was not present in PhD_V2 dataset, it was added using the Genderguesser library in Python. Note that for clarity, only the five most represented disciplines are used below. The following two graphs will serve as an example to show that it is possible to present the same data differently, resulting in a different reading of it.

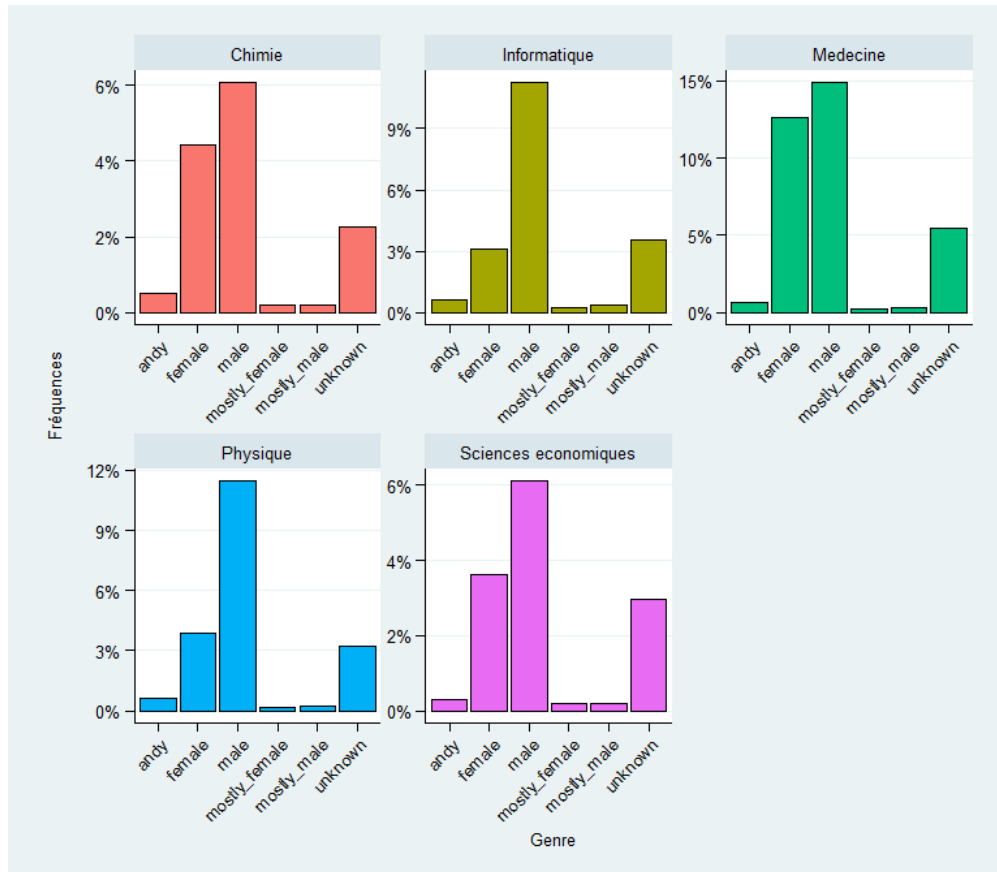


Figure 20: Facet wrap of percentages of genres by disciplines

Figure 20 represents the percentage of genres by discipline. We can see a male dominance in all disciplines. Note the near gender parity in medicine and chemistry, this type of graph allows you to observe the distributions of the target variable by category.

Figure 21 represents the relative frequencies of each discipline by gender. We can see that about 45% of women are in medicine. This type of graph allows you to see the distribution within the same category.

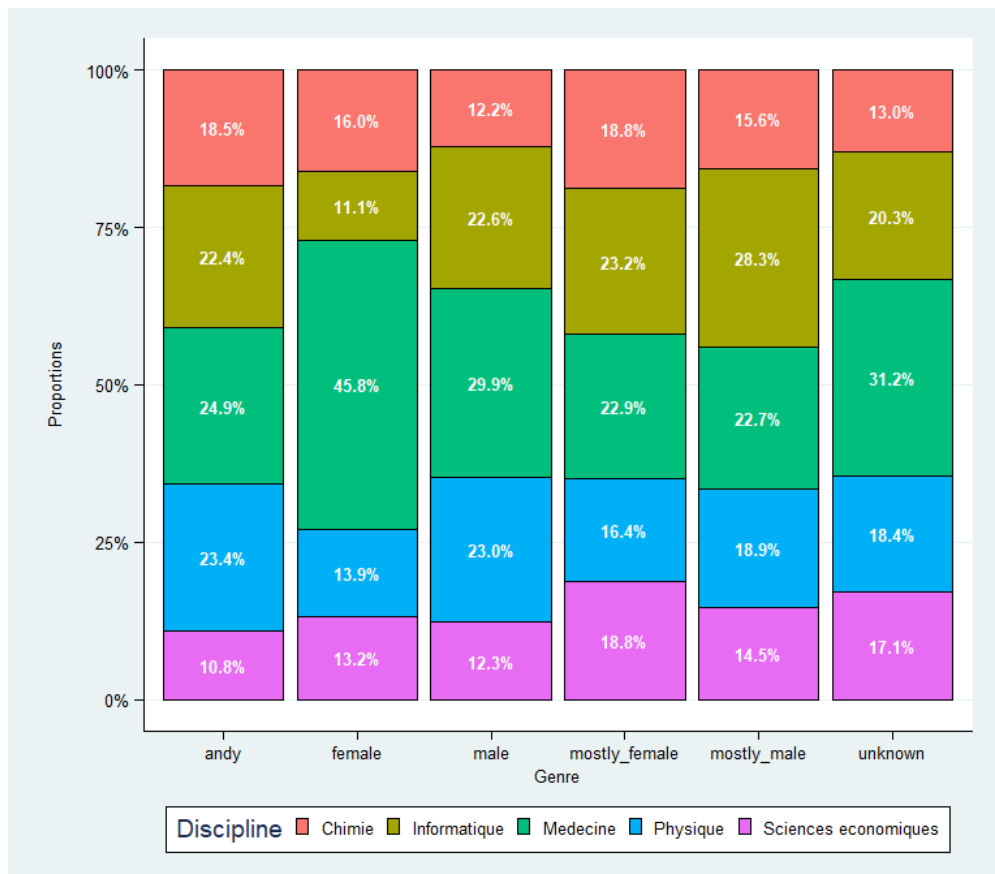


Figure 21: Relative frequencies of each discipline by gender

PhD V3

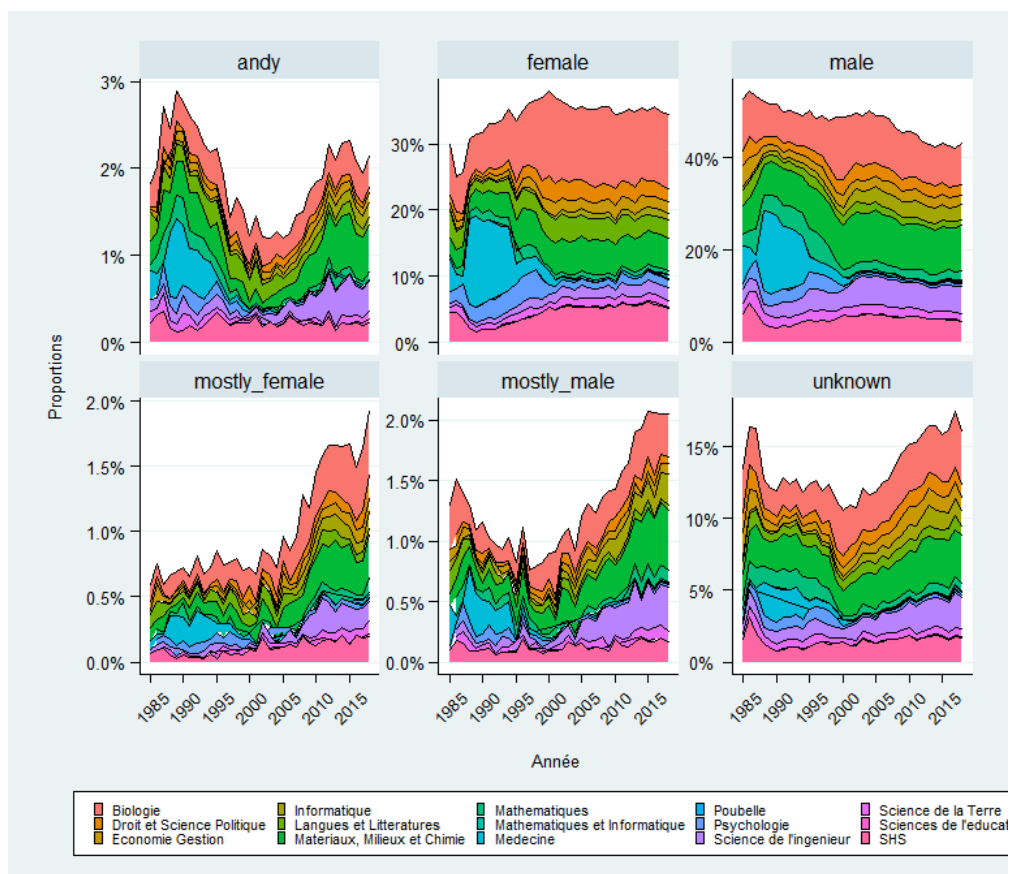


Figure 22: Evolution of genres by disciplines over the year, period from 1985 to 2018

Figure 22 represents the evolution of the proportions of genres by disciplines over the years. We can see that from 1990 to 1995, women in medicine represented 12% of the total. We can also see after 1995, women are mostly represented in biology.

7.3.2 Problem of languages according to disciplines

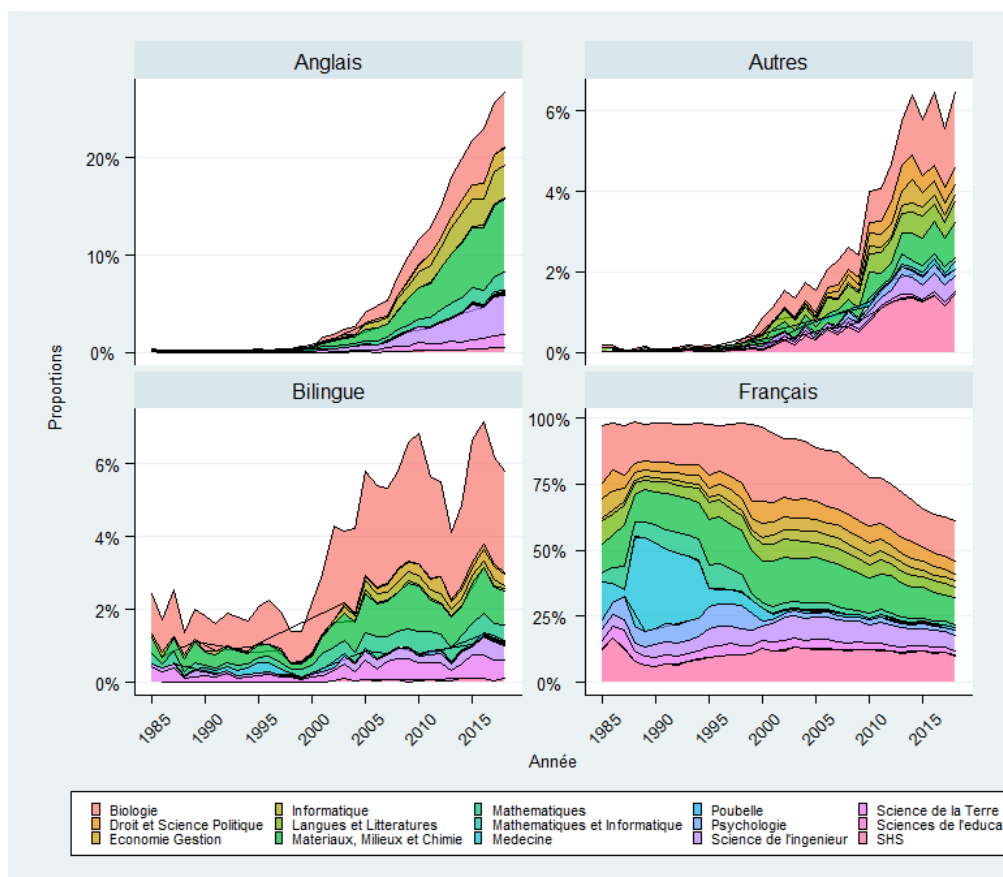


Figure 23: Evolution of languages by disciplines over the year, period from 1985 to 2018

Figure 23 depict the evolution of language use by topic over the years. We can see the fall of French as seen in chapter five. The French language is mainly used in biology (12,5%), materials, media and chemistry (10%) and SHS (12,5%) after the 2000s.

7.4 Web scraping

Only the resulting web scraping will be shown below. It was done using harvestr library in R

	Auteur	Titre	Discipline	Directeur
1	Djelloul Kab	La responsabilité médicale en...	Droit privé	.
2	Hamza Tarin	Caractérisation et optimisatio...	Sciences de l'ingénieur	Abdellah Arhaliass,
3	Farah Bibi SHAIK DAWOOD	Formulation et procédés de p...	Sciences de l'ingénieur	Abdellah Arhaliass, Raphaëlle Savoie.
4	Riad EL HAMOUD	Réduction de la fatigue des é...	Sciences de l'ingénieur	Abdul-Hamid Soubra, Mourad Ait-ahmed.
5	Simon Husser	Privé et public en droit pénal	Droit pénal	Agathe Lepage
6	Marthe Cachard-Chastel	Synergie d'effets neurochimi...	Pharmacologie expérimentale et clinique	Alain Gardier
7	Alexandre Derre	Douleurs chroniques : implica...	Biologie Santé	Alexandre Pattyn
8	Lara Aldaou	Etude expérimentale et numé...	Sciences de l'ingénieur	Ali-Nordine Leklou,
9	Elise Madec	Etude multiparamétrique de l...	Physique	Amanda Silva brun
10	Diane Letourneur	Impact combiné de toxines d...	Sciences de la vie et de la sante	Amel Mettouchi
11	Clara Gandrez	Modèle comportemental d'id...	Conception (AM)	Améziene Aoussat
12	Adèle Kauffmann	Les effets du brexit sur la cito...	Droit de l'Union européenne	Anastasia Iliopoulou
13	Jeanne-Valérie Hell	Histoire, imaginaire dans "les...	Littérature française	André Daspre
14	Trang Nguyen Vinh	Analogie entre le courant éle...	Géographie et aménagement	André Dauphiné
15	Lucette Heller-Goldenberg	Histoire des auberges de jeun...	Histoire	André Nouschi
16	Alia Lakhoua	Le tissage de la soie à tunis d...	Histoire et civilisation	André Nouschi
17	Pierre Bicaba Nanye	La crise économique de 1929...	Histoire	André Nouschi
18	Monique Jacomino-Laborieux	L'algérie coloniale 1830-1962...	Histoire	André Nouschi
19	Angga Perima	Combinatorial antibiotic scre...	Chimie physique et chimie analytique	Andrew D. Griffiths
20	Patie Cendra Rakotoarimanana	Nanoscale surface engineerin...	Chimie	Anne-Marie Gonçalves
21	Renata Andrade (Da silva andrade)	Le cannibalisme dans l'art co...	Arts	Anne Creissels
22	Amelie Francois	Une sexualisation virtuelle de...	Arts	Anne Creissels
23	Clemence Canet	La visite guidée comme perfo...	Arts	Anne Creissels
24	Hiam Dahanni	Conception environnemental...	Sciences de l'ingénieur	Anne Ventura, André Orcesi.
25	Athul Kaitheri	Caractérisation des variations...	Sciences de la Planète et de l'Univers	Anthony Mémin, Frédérique Rémy.
26	Geoffroy Laurin	La rente et le droit des sûreté...	Histoire du Droit	Anthony Mergey
27	Lucas Tuduri	Croyances motivées et théori...	Sciences économiques	Antoine Billot
28	Martin Odoh	La détermination de la respo...	Droit Public	Antoine Delblond, Dodzi Kokoroko.
29	Jalal Elmir	Le secret bancaire, une insti...	Droit des affaires	Antoine Gaudemet
30	Benjamin Fontaine	Intérêt social et activisme act...	Droit des affaires	Antoine Gaudemet
31	Jérôme Beaumont	Rôle prépondérant des cellule...	Droit des affaires	Antoine Gaudemet
32	Alessandro Lauro	La "justiciabilité" du système ...	Droit public	Armel Le Divellec,
33	Mainak Sarkar	Trois articles sur l'analyse des...	Science de gestion - EM2PSI	Arnaud De Bruyn, Arnaud De Bruyn.
34	Coline Fonderflick	L'accord collectif de travail à ...	Droit social	Arnaud Martinon
35	Baptiste Bataille	Création et institutionnalisati...	Sciences de l'information et de la communication	Arnaud Mercier
36	Guillaume Perissat	Les enjeux de la communicati...	Sciences de l'information et de la communication	Arnaud Mercier
37	Clarisse NYNGONE MAYAZA	Consommateurs de soins et s...	Droit	Augustin Emame
38	Leslie-anne Merleau	Effets à l'échelle physiologi...	Dynamique des milieux naturels et anthropises pass...	Aurélie Goutte, Olivier Lourdais.
39	Malak Dia	Nanoparticules et matière or...	Science de la Terre et de l'Environnement	Béatrice Bechet,
40	Simon Gouzy	Conditions de formation de la...	Sciences de la Terre et des planètes	Benjamin Rondeau, Vassilissa Vinogradoff.
41	Fang Zhao	Sous-spécification et analyse ...	Sciences du langage - linguistique	Benoit Crabbe
42	Clara Lahiani (Coudert)	Les datacenters à l'épreuve d...	Droit fiscal	Benoît Delaunay

Figure 24: Data frame résultant du web scraping partie une

Etablissement	Statut	Date_inscription	Date_soutenance_ymd	Date_soutenance_y	Link_to_pdf	Langue
Catherine Puigellier, Corinne Pizzio-Delaporte.	en_cours	NA	2022-02-03	NA	NA	NA
Ahmed Rhallabi, Jamal Fajoui.	en_cours	2022-02-03	NA	NA	NA	NA
Nantes	en_cours	2022-02-07	NA	NA	NA	NA
Nantes	en_cours	2022-01-06	NA	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	NA	2022-03-15	NA	NA	NA
Université Paris XI	soutenue	NA	NA	2007	NA	NA
Montpellier	en_cours	NA	2022-04-08	NA	NA	fr
Nabil Issaadi, Ouali Amiri.	en_cours	2022-02-25	NA	NA	NA	NA
Université Paris Cité	en_cours	2021-07-13	NA	NA	NA	NA
Université Paris Cité	en_cours	2021-07-19	NA	NA	NA	NA
Paris, HESAM	en_cours	NA	2022-03-31	NA	NA	fr
Université Paris-Panthéon-Assas	en_cours	2021-10-08	NA	NA	NA	NA
Nice	soutenue	NA	NA	1988	NA	fr
Nice	soutenue	NA	NA	1985	NA	fr
Nice	soutenue	NA	NA	1985	NA	fr
Nice	soutenue	NA	NA	1987	NA	fr
Nice	soutenue	NA	NA	1988	NA	fr
Nice	soutenue	NA	NA	1985	NA	fr
Paris 6	soutenue	NA	2017-12-11	NA	NA	en
université Paris-Saclay	en_cours	NA	2021-11-16	NA	NA	fr
Paris 8	en_cours	2021-10-16	NA	NA	NA	NA
Paris 8	en_cours	2021-11-26	NA	NA	NA	NA
Paris 8	en_cours	2021-10-27	NA	NA	NA	NA
Nantes	en_cours	2021-12-03	NA	NA	NA	NA
Université Côte d'Azur	soutenue	NA	2021-12-02	NA	theses.fr/2021COAZ41...	en
Université Paris-Panthéon-Assas	en_cours	2022-01-26	NA	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	2021-10-28	NA	NA	NA	NA
Nantes	en_cours	2020-01-13	NA	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	NA	2022-01-21	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	2021-10-28	NA	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	2019-09-30	NA	NA	NA	NA
Luigi Benvenuti, Marco Mancini.	en_cours	2019-09-30	NA	NA	NA	NA
CY Cergy Paris Université	en_cours	2017-09-01	NA	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	2021-11-18	NA	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	2021-10-12	NA	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	2022-01-06	NA	NA	NA	NA
Nantes	en_cours	2021-12-16	NA	NA	NA	NA
Université Paris sciences et lettres	en_cours	2021-09-01	NA	NA	NA	NA
Denis Courtier-muras, Pierre-Emmanuel Peyneau.	en_cours	2021-12-16	NA	NA	NA	NA
Nantes	en_cours	2021-10-20	NA	NA	NA	NA
Université Paris Cité	en_cours	2021-09-08	NA	NA	NA	NA
Université Paris-Panthéon-Assas	en_cours	NA	2022-01-15	NA	NA	NA

Figure 25: Data frame résultant du web scraping partie deux

List of Figures

1	Visualization of missing data on the entire dataset	5
2	Missing data model	6
3	Visualization of missing data for the inProgress status value	7
4	Visualization of missing data for the Supported status value	7
5	Distribution of defense by month, period 1984-2018.	8
6	Distribution of defenses per month per respective year, period 2005-2018	9
7	Percentage of defenses per month with standard deviation, period 2005-2018, without filtering out the first of January	10
8	Percentage of defenses per month with standard deviation, period 2005-2018, with filtering the first of January	10
9	Evolution of defenses on January 1st per year over the entire set of data	11
10	Histogram of years of defense for the unique identifier	12
11	Distribution of total supervisors by number of theses supervised, with median	14
12	Distribution of total supervisors by number of theses supervised, with line representing the start of outliers	15
13	Distribution of total outlier supervisors by number of supervised theses	15
14	Distribution of outliers between 40 and 140 supervised theses, with median	16
15	Visualization of missing data	18
16	Evolution of the frequencies of the different languages use over the years	19
17	Evolution of the frequencies of use of different languages over the year from 2004 to 2018	20
18	Wampserver manipulation result table	22
19	Heatmap des données manquantes	23
20	Facet wrap of percentages of genres by disciplines	24
21	Relative frequencies of each discipline by gender	25

22	Evolution of genres by disciplines over the year, period from 1985 to 2018	25
23	Evolution of languages by disciplines over the year, period from 1985 to 2018	26
24	Data frame résultant du web scraping partie une	27
25	Data frame résultant du web scraping partie deux	28

List of Tables

1	Distinct number by variables of the dataset based on the website theses.fr	3
2	Distinct number of PhD v2 variables	12
3	Discipline defense date for unique identifier	13
4	Evolution of languages for the period from 2004 à 2018	20
5	Total language for the year 2005 and 2012	21

Bibliography

- [1] I Martin. “Le signalement des thèses de doctorat”.
In: *I2D - Information, données & documents* (2015), pp. 46–47. DOI: 10.3917/i2d.151.0046.