Data Analysis Report on National Doctoral Thesis Data

Ru WANG

SUPERVISED BY: DR. MATTHIEU CISEL

Wednesday 13th April, 2022

Abstract

Based on the data frame exported from website *theses.fr*, its characteristics: missing values, anomalies are investigated, it is fond out that digitization process helps to keep on tracking of the precise information, but more additional features need to be introduced in order to distinguish potentially errors such as homonym issue.

Keywords— data analysis - missing data - anomaly

Table of Contents

List of Figures

List of Tables

1 | Introduction

With the technical development and widely applied numerical methods, the broad use of internet and the increasing popularity of lean philosophy have increased the use and meaning of "digitizing" to describe improvements in the efficiency of organizational processes (Audrin, 2019). The digitization concerns all different domains: books, magazines, documents, images, paintings, sculptures, films, musics, devices, etc (Reis, Amorim, Melão, Cohen, & Rodrigues, 2020). All sorts of research work about digitization had been done in a wide way, especially in the digitization of library preservation (Bingham, 2010), the open access of the doctoral theses is one of the main topic (Rasuli, Solaimani, & Alipour-hafezi, 2019).

Today we will explore the most important website of the doctoral theses in France *http://www.theses.fr*, which was published in July 2011 and contains the doctoral theses dated from 1971 (l'ABES, 2011). This database is keeping updating not only for the defended theses, but also the theses in progress in all different disciplines regardless the original materials(paper version, digital version, commercial edition...)

In this report, we analyze part of the data, which comes from this french doctoral theses data frame. By detecting the missing information and its anomalies from the given data, the characteristic of this database is investigated. The remainder of this report is organised as follows:

Chapter 2 — Missing Data Exploration

Chapter 3 — Problem Detection And Analysis

Chapter 4 — Anomaly Detection And Analysis

Chapter 5 — Discussion

2 | Missing Data Exploration

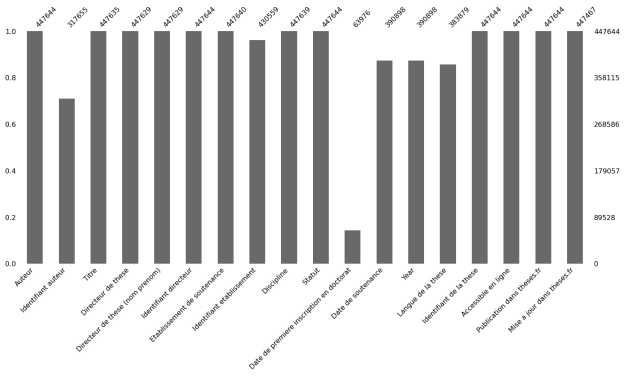
Based on the given database which is named as *PhD_v2.csv*, this file is partially exported from the website *http://www.theses.fr*. We firstly explore the information that this file contains. By checking the dimension of this data which has a matrix structure, it appears that it records totally 447644 theses from year 1971 to year 2020. The following Table 2.1 lists all the categories of information through the first thesis example.

Table 2.1: The category information of the file *PhD_v2.csv*

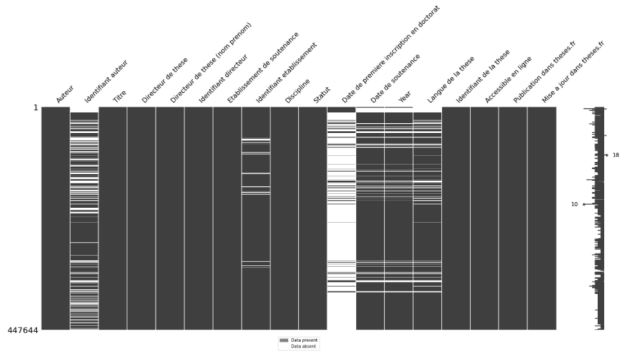
Category	**1**
Auteur	Saeed Al marri
Identifiant auteur	NaN
Titre	Le credit documentaire et l'onopposabilite des exceptions
Directeur de these	Philippe Delebecque
Directeur de these (nom prenom)	Delebecque Philippe
Identifiant directeur	029561248
Etablissement de soutenance	Paris 1
Identifiant etablissement	027361802
Discipline	Driot prive
Statut	enCours
Date de premiere inscription en doctorat	30-09-2011
Date de soutenance	NAN
Langue de la these	NAN
Identifiant de la these	s69480
Accessible en ligne	non
Publication dans theses.fr	26-01-2012
Mise a jour dans theses.fr	26-01-2012

From Table 2.1, We can count 17 categorical features concerning one thesis. Those are very typical information: The author, the title, the discipline etc. It also shows that some information is not registered, for example, for this Dr.Saeed,the

author identity, the date of defence and the language of thesis are missing. It is quite common that, the missing values appear in the database: even in a well-designed and controlled study, missing data occurs in the research. Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions. Therefore, we will study detailedly these missing values in this chapter.



(a) Presentation of missing data in the form of bar



(b) Presentation of missing data in the form of matrix

Figure 2.1: Presentation of missing data in the form of bar(a) and matrix(b)

In order to have a general view on the missing data, the missing values bar is shown in Figure 2.1 (a). It plots the present value by a bar chart, the x-axis below each bar is the name of each category and the x-axis on top indicates the total number of existing values, while left y-axis is the percentage of present values

and the right y-axis is the total number of present one. Figure 2.1(b) shows matrix plot with the missing values represented by white lines, and the non-empty values drawn as black lines.The y-axis are the observations, the x-axis are the categories of data. From the Figure 2.1 (a), we can see that beside the columns *Auteur*, *Identifiant directeur*, *Statut*, *Identifiant de la these*, *Accessible en ligne* and *Publication dans theses.fr* which have no missing values, all the rest are not completed, especially the columns of *Date de premiere inscription en doctorat* and *Identifiant auteur*. It is logical that, the columns *Auteur* has no missing data, since it is the crucial information of one thesis. Furthermore, once this website is being published, we can verify easily if the thesis can be or can not be traced online and inside this website. So, these columns for the accessibility online and visibility on the theses website will not have missing data. On the other hand, in France, after one successful defense, the date of the registration for the doctorate will be eliminated. That explains the large portions of missing data of column *Date de premiere inscription en doctorat* in the Figure, and this is consistent with the observations of the two corresponding columns in the Figure(b). To investigate more quantitatively these correlations, the Figure 2.2 is given.

Theses additional information about missing values can be gained by looking at correlations between features of their missing values. Figure 2.2 shows a heatmap of the file. As the figure indicates, this database has both positive and negative correlations of missing values between different categories. Here the negative correlation presents the places where the missing values present in one category is absent in the other category, vice versa, the positive correlation means two category have same missing value present places. As the figure indicates, the correlation between the PhD registration date and the dissertation defense date/year is -0.9, and this number is close to -1 when comes to column *Langue de la these*. These correlation coefficients are coherent with the doctoral regulation in

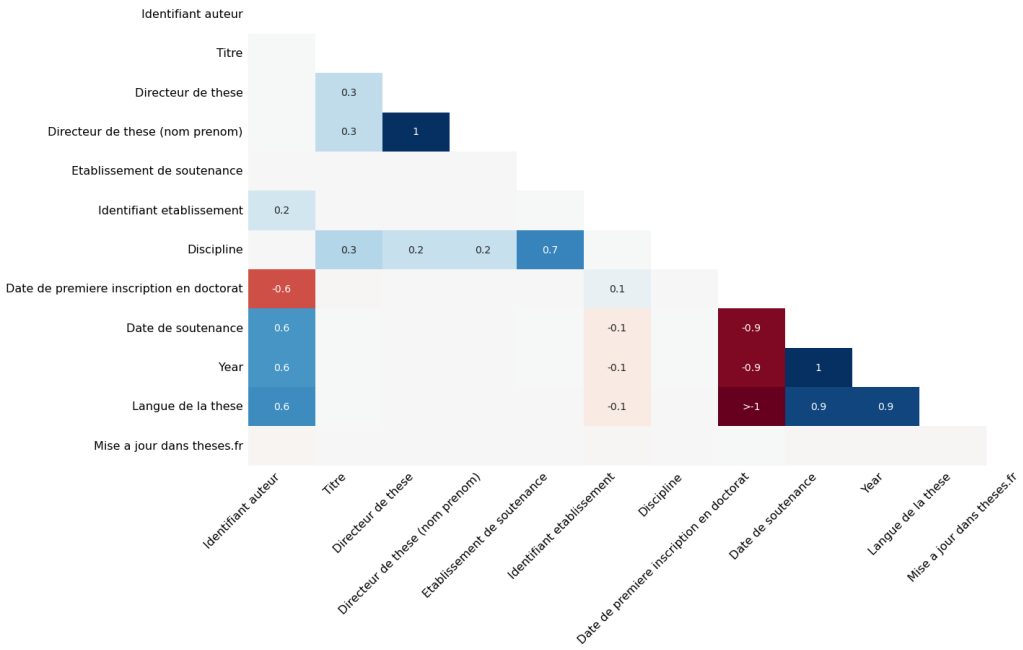


Figure 2.2: Heatmap based on the file *PhD_v2.csv*

France. As well as the number 0.9 between column *Date de soutenance* and column *Langue de la these* reprove this regulation, because only after successfully defended a dissertation, the thesis will be hand out and recorded. By that time, its language can be detected and registered. In addition of negative correlation, there also have some important positive correlations, such as *Directeur de these(nom prenom)* and *Directeur de these* which partly proves the accuracy of this database. The correlation between the columns *Discipline* and *Etablissement de soutenance* is 0.7, which can be explained as that since the doctoral diploma is the highly specialized one, the specialists who work on the research need to group together in order to advance on their projects, so, these are the places where cultivate the doctors and become more influential on their domain, step by step, they grow into the specialty of the certain establishments. How the discipline or number of doctors within establishments could influence on its reputation or ranking, how many PhD students will become a supervisor later, such as the same person ap-

pears both in the author and supervisor column, whether they will stay in the same establishment or not, could be the interesting exploration for the further work.

Since the heatmap above considers the missing values, we can go further to see the correlation of missing values based on the textual information. Such as the missing correlation of two states of PhD defense (successful passed or under preparation) with the date of PhD registration and the date of dissertation. As the Figure 2.1(b) indicates, if the missing value of PhD registration date is totally due to the large proportion of successful dissertation in this database, then, this column should be completely positively related to the state of success of the PhD defense.

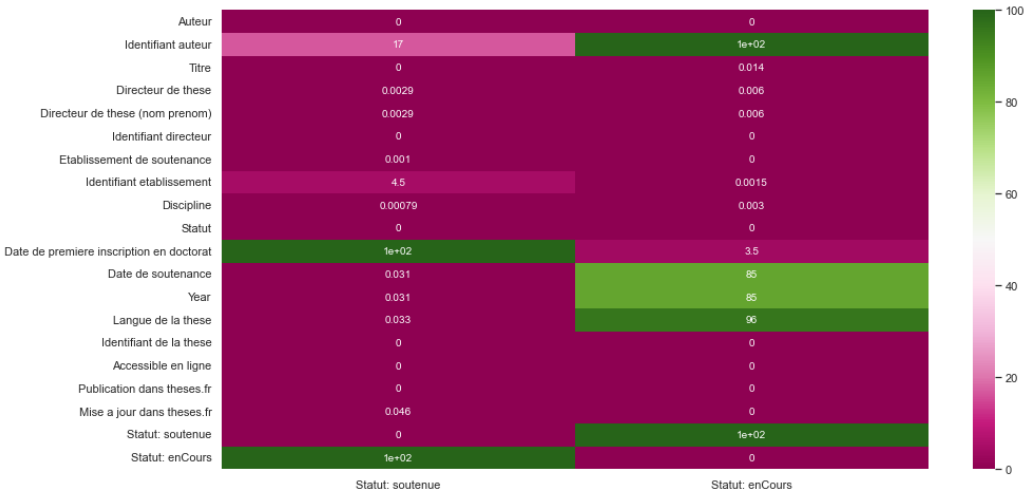


Figure 2.3: Percentage of missing value between PhD states with others categories features

Figure 2.3 shows the percentage of missing value between two PhD states with other categories features. We can see that when the PhD defense is successful passed, the inscription date information is totally removed, which leads to 100% of missing values. While there are still non negligible 17% of missing for the identity author and identity establishment. While, as to the PhD is in progress, the identity

of author is 100% unknown, it could be explained as that the author ID is given once the defense is successfully passed. This could be the same explanation for the almost zero missing information for the title, the information of the director, since once the PhD student is registered, those information is already recorded into system. But we also should notice that the date defense and language of theses is not 100% missing, but only 85% unknown. Theoretically, once the PhD is still in progress, those information should be unknown, unless there has delay for the update this online system or there include the abandoned PhD, or failed PhD defense could be due to the students who abandon the PhD study. More detail work can be done considering those missing data, such as the link between abandoned these with its discipline or the time period of historical social events, etc. Meanwhile, we can also use some imputation methods to replace the missing values.

3 | Problem Detection and Analysis

In this chapter, we will explore the number related features of this data frame, which is the date/year of dissertation defense. Firstly, we go through the evolution of number of theses from year 1971 to 2020 in Figure 3.1. This plot shows that

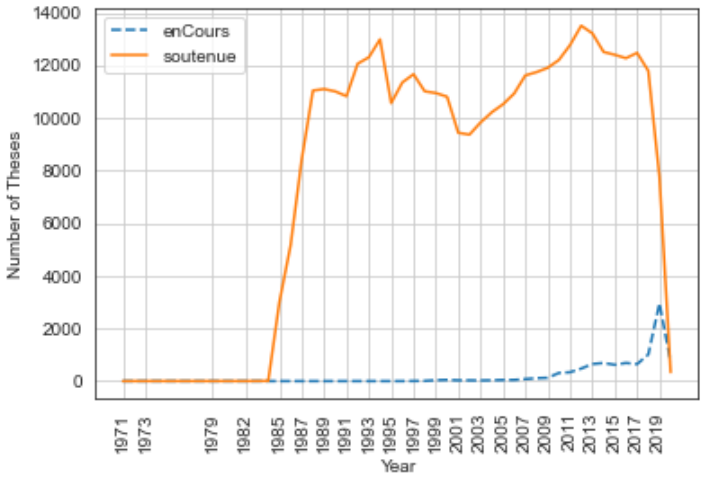


Figure 3.1: The evolution of number of theses under two status

for the statute is in progress, it almost 0 until after 2009 and then reach a peak in 2019 around 2000, then drop to 1000 in 2020. While, the number of theses, which are successfully defensed, is almost 0 until 1985, then increases to 11000 around 1989 and fluctuate around this number until it goes down vertically to 0 in 2020. This can be explained as that these changes reflect some typical time periods. Year 1985 is the end of the the Cold War, so there are more people able to do the high level study. Year 2008 is the global financial crisis, to continue study to avoid the recessional job market is a good option, While year 2020 starts the propagation of Covid-19 internationally. Face to this historical unrecorded disaster, it takes time for students, universities, government to reconsider a new system to continue the study, research, PhD defense etc.

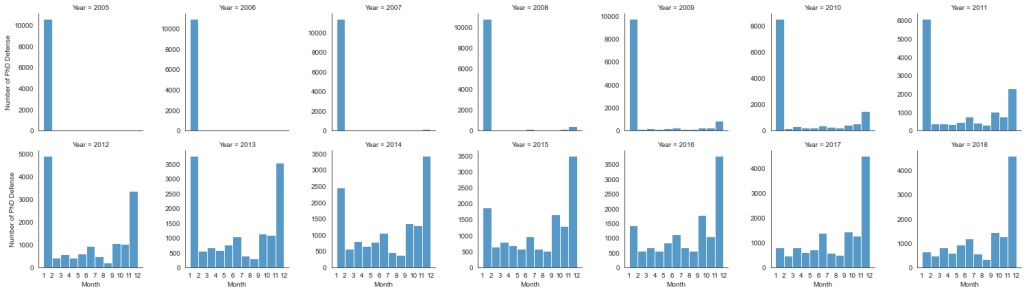


Figure 3.2: The number of PhD defense in each month from 2005 to 2018

Beside the evolution in the large scale, we also explore its monthly evolution. we find that the number of PhD defense not only link to the historical, human society changing, but also varies with the school year agenda. For example, the new school year always starts in September and end in June, with the two main holidays: summer holiday and winter holiday. So the plot the number of PhD defense in each month from 2005 to 2018 as shown in Figure 3.2. It indicates that until 2008, there start to have some PhD defense record in other months than January, part of this result is that this website is published in 2011, so, when the dissertation defense date is unknown, it is estimated as January. We can see that after the digitization of PhD defense, there are more and more record take place in December, and quite fewest ones register in July and August. This tendency is in accord with the school year agenda, the job market for PhD, the annual budget planning of companies, etc.

Furthermore, we find that there are a lot of PhD defense date is documented as 1st January. Since that day is a international holiday, it is impossible to have any defense presentation, and these abnormal records may due to unknown defense date but the defense year information is registered, so, the 1st January is set as a default data when the precise data information is unavailable. But this default setting will give statistical bias when considering the monthly defense result. In order to see more details on this issue, we plot the Figure 3.3. Since each year there are dif-

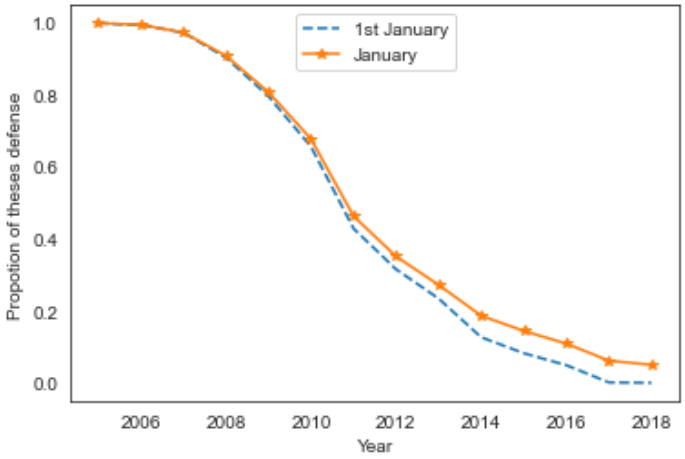


Figure 3.3: The propotion of PhD defense from 2005 to 2018 considering only the 1st January and January completely

ferent total number of defense, as indicated in Figure 3.1, due to these different bases, we cannot compare the absolute value for the conclusion, so, we compare the relative value-percentage in order to reach the same comparable level. Figure 3.3 is the proportion of PhD defense evolution from 2005 to 2018, considering only 1st January (blue) and the whole January completely (orange). We can see that both of them start from 1 in 2005 and then decrease to around 0.1 in 2018. And they are starting to have slightly difference around 2011 which is the year that published the *theses.fr* website, since this helps to restore more detail information, the defense date precisely registered. And it is also shown that less and less PhD dissertation presentation is chosen in January.

To see the statically monthly result during this period, the figure 3.4 is given. The left on is the subplot in which the 1st January is including, while the right one excludes this international holiday. The error bar is the sanded deviation. We can see the big difference with/without considering that day. For the left, January takes half of the proportion of yearly theses defense, and the December is the second largest with proportion 0.2. Beside August is almost 0, all the remains is around

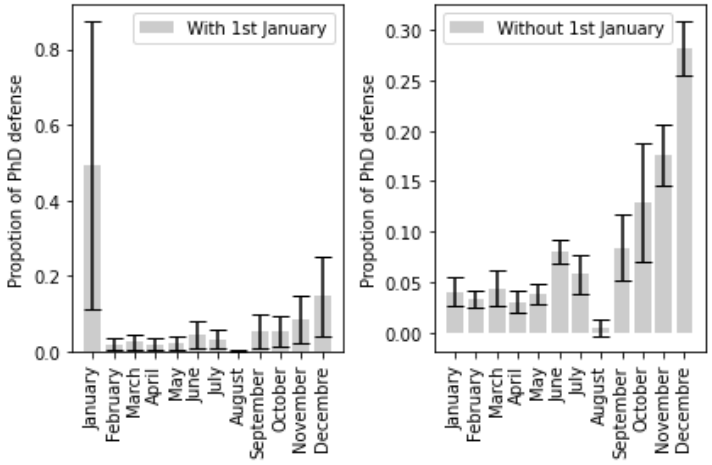


Figure 3.4: The bar chart of monthly PhD dissertation defense from 2005 to 2018 with 1st January (left) and without 1st January (right)

0.1. And the error bar is also proportional to the proportion of defense, such as January has the largest error bar. While, when comes to the right subplot, December has the largest proportion around 0.3, then comes to November, October, the September and June has around 0.1, the the rest are around 0.05 except August is around 0. We can see that if considering date of 1st January as a normal date, due to the default setting, January is the month gathering half of the yearly defense while December is the second month which has second largest proportion. That explains both the large percentage and range of error-bar in January during those 14 years. But if we remove this abnormal setting information, which means that we don't take the 1st January into account, the right figure shows quite differently the monthly PhD defense. The months which are close to the holiday is largely being chosen for the defense: November and June. This is coherent with Figure3.2 that after 2011, more and more defense hold in December, and after 2014, the number of defense drop quickly less than those in January. Then, comparably in December, a lot of defense also happen in June, due to the job opportunities in September, the teacher position offers in university, post-doc offers, industry

employments,etc. As we can see that after the publishing of *theses.fr* website, a lot of precise information can be resisted and those records are consistent with the reality and can help us to provide more useful guidelines, furthermore, we need to carefully check the database in order to avoid the statistical result bias.

4 | Anomaly Detection and Analysis

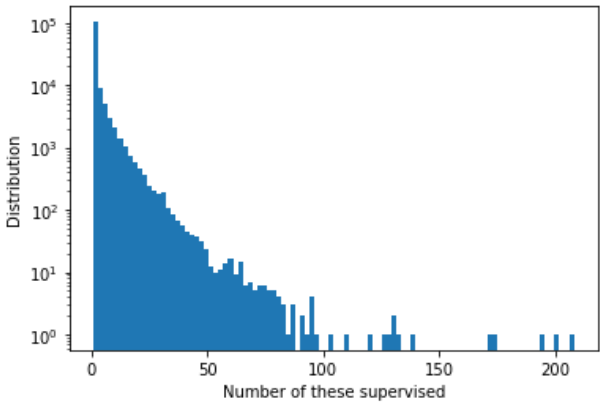
In this chapter, more complex text-based features will be investigated. Firstly, we look into the author name which is one essential category. Here, we take name of *Cecile Martin* as a example. From the database, this name is recorded 7 times with 5 different author identities, and 6 different disciplines, as the following Table 4.1 shows. Since there may exist the PhD students with the same name, so, the author ID is designed to distinguish them. But from this result, it seems that this distinguishing method is not reliable: four PhD dissertation defenses with the same author ID in 3 different disciplines. But on the other hand, form this table, it is possible that Cecile Martin get her PhD in domain of neruosciences in 1991, then, another PhD in domain of biology science in 1994, and the third one in genie des procedes industriels in 2001. Since all of the three theses are recorded as written in French, we need more information to find if these three theses are done by the same person. From this table, it indicates that we need more features about the author to differentiate them, birth day, birth place for example.

The similar question can be asked for the supervisors. Due to reasons such as the research work refers to different domains, the share of the research work, etc, there exists co-supervisors. Generally, when there are more than one supervisors, those

Table 4.1: The category information on the author name *Cecile Martin*.

Author	Identifiant auteur	Year	Staut	Langue de la these
Cecile Martin	203208145	2017	soutenue	fr
Cecile Martin	81323557	2000	soutenue	fr
Cecile Martin	179423568	2014	soutenue	fr
Cecile Martin	81323557	2001	soutenue	fr
Cecile Martin	81323557	1991	soutenue	fr
Cecile Martin	81323557	1994	soutenue	fr
Cecile Martin	182118703	1989	soutenue	enfr

Author	Identifiant auteur	Discipline
Cecile Martin	203208145	Etudes cinematographiques et audiovisuelles
Cecile Martin	81323557	Sciences biologiques fondamentales et appliquees
Cecile Martin	179423568	Sciences economiques
Cecile Martin	81323557	Genie des procedes industriels
Cecile Martin	81323557	Neurosciences
Cecile Martin	81323557	Sciences biologiques fondamentales et appliquees
Cecile Martin	182118703	Physique



Name	Number Of Supervised Theses
Directeur de these inconnu	711
Blanc Francois-Paul	227
Scherrmann Jean-Michel	209
Brunel Pierre	206
Pujolle Guy	196
Bertucat Michel	173
Teyssie Bernard	140
Lumley Henry de	139
Maffesoli Michel	136
Foucart Bruno	135
Chaumeil Jean-Claude	132
Bessiere Jean	128
Boulu Roger G.	127

Figure 4.1: The distribution of number of theses by one supervisor (left) and the name list(right).

supervisors are more dynamic and have more influence in the research, which also results to have more PhD students. Here, we consider all the name of supervisors, including the name of co-supervisor. According to this database, there are totally 101252 different name of supervisors who directed 379109 theses from 1984 to 2019. Figure 4.1 left shows the distribution of the number of theses directed under one name of supervisor, while the right table list the ranking name of the supervisor by the number of theses supervised, from 1984 to 2019. We find that there are 711 number of these with unknown name of the supervisors, and then the supervisor *Blanc Francois-Paul*, *Scherrmann Jean-Michel* and *Brunel Pierre* are in the list of top 3 who directed more than 200 theses. We can see that there are 711 unknown name supervisor, which could be explained as that the missing information too long time ago or they are dated before the year of publishing *theses.fr* . It is not surprising that one professors can have more than 2 PhD in the same grade, especially for the discipline like society, literature, which can last more than 4 years. For this professor, Blanc Francois-Paul, he directed 227 theses during 36 years, which is around 6 these by year in the continues 36 years. It is a impressive work if there is no homonym issue. We go further with this name in the

database, it turns out that there are only one identification under this name, which range from 10 disciplines, and there are two disciplines with the same name 'Droit prive', and the other one is 'Droit prive.', Due to the typing errors with one extra point. The Table 4.3 list the disciplines that he refers to for the research work.

Table 4.3: Number of theses in different disciplines supervised by Blanc Francois-Paul From 1984 to 2019

Discipline	**Identifiant directeur: 26730774 Number of theses**
Droit	36
Droit compare	3
Droit et sciences politiques	1
Droit prive	81
Droit prive et sciences criminelles	28
Droit prive.	1
Droit musulman compare Droit public.	24
Histoire du droit.	10
Histoire du droit et des institutions	15
Science politique	2

Besides the productive PhD students of the professors who contribute to the ranking, but also there exist some confusion due to the homonyms problems. For example, under the name of *Brunel Pierre*n it has 3 different supervisor identities. Then, the question comes out if these identities represent 3 different people or 2 different people or it is the same person, only due to the typing mistake. To answer this question, we need to delve into more information, such as the working establishment, the disciplines, etc.

On the other hand, once a professor starts to supervise a PhD, which means this professor is important in academic circle and he stars the academic career more maturely. Without exception situations, he will has more and more PhD students under his name. But from this database, we find that 1/3 professors supervised no more than 3 theses during these period, which is the opposite situation like we

discussed before. This could because those professors are newly recorded into this system or they just star their career as a PhD supervisor yet. For the same reason, other information is needed in order to explain it. The total number of supervised theses by each supervisor, by collaboration with other supervisor, all those statistics can reflect not only the dynamism of the supervisors, but also the tendency of the research area, which could be the future work.

Table 4.4: Number of theses supervised

Number Of Theses Surprised	**Number of Supervisors**
1	41150
2	15071
3	9495
4	6645
5	5079
6	3902
7	3180
8	2518
9	2056
10	1679

Beside the name context information, there exists categorical context variable: disciplines and language of the theses. We firstly go through the disciplines, here, with the algorithm, the disciplines are divided into 15 categories. So, the evolution of the 15 disciplines is ploted as Figure 4.2. From this figure, we can see that half of the disciplines are studied by more and more people, and the biology, science are the top domains,while another half have less and less PhD defenses, especially the number of theses in medicine reaches the peak around 4000 in 1990 and then drop to 800 within 5 years.

Since the gender also plays a important role in the different disciplines, so, the evolution of the 15 disciplines by the different gender, male is the up figure and female is the is the down figure. As Figure 4.3 shows, there are generally more

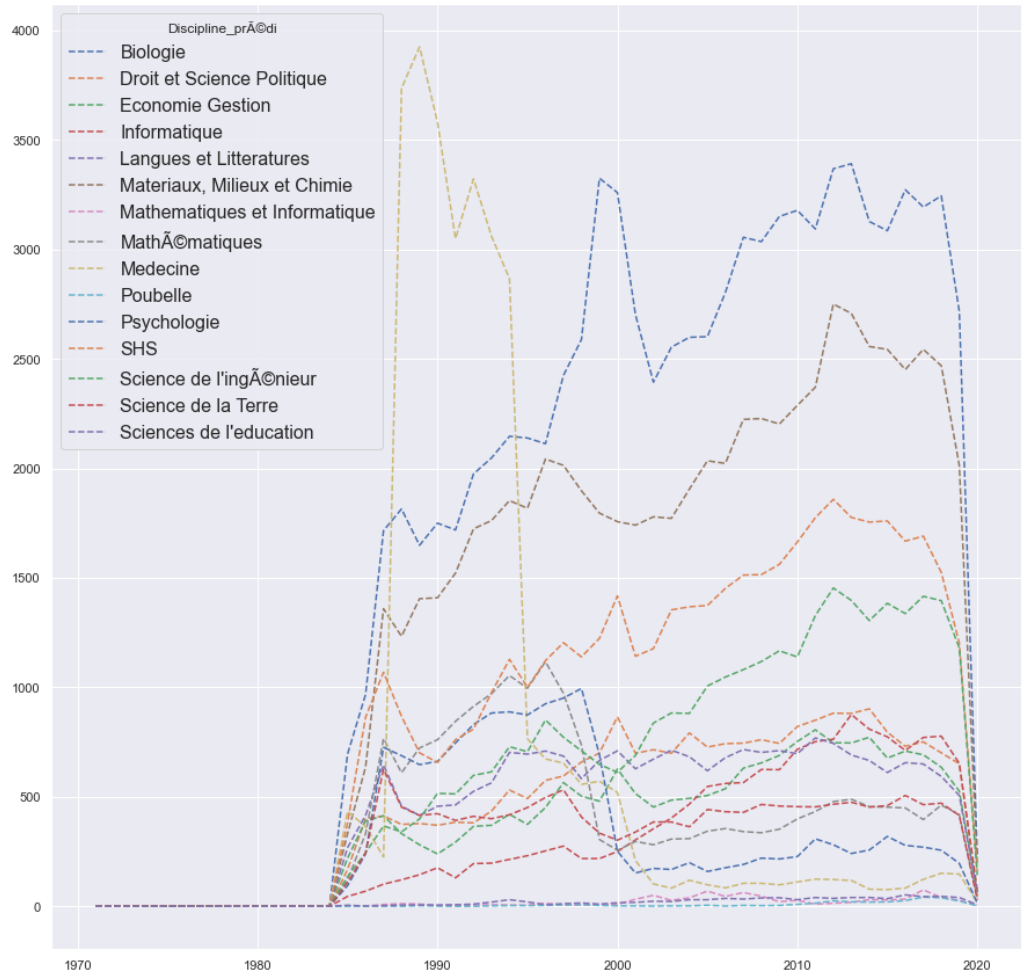


Figure 4.2: The evolution of number of theses in 15 disciplines

male in the PhD defense than female. And there exists domains which has more gender than other, such as in physics and chemistry, computer science, engineering, mathematics have more male than female, while the biology, psychology, language and literature are the opposite situation.

Then, the evolution of number of theses by different language is also examined. We find that there are totally 206 different languages record, for simplify, we classify into four categories: English, French, bilangue which represents both English and French, Others which is the language besides what we mentioned above, then

the nan for the language is unrecorded. Figure 4.4 shows the number of theses in different languages change with year by different gender (up) for male and (down) for female. We can see the increasing of English theses around 2000 for both gender, but always more male in this language, while more female in french than male, this could be explained as there are more inscription of international students due to the global crisis 2008, and female francophone prefer to write in french. The drop for all of the languages around 2020 is because of the Covid-19.

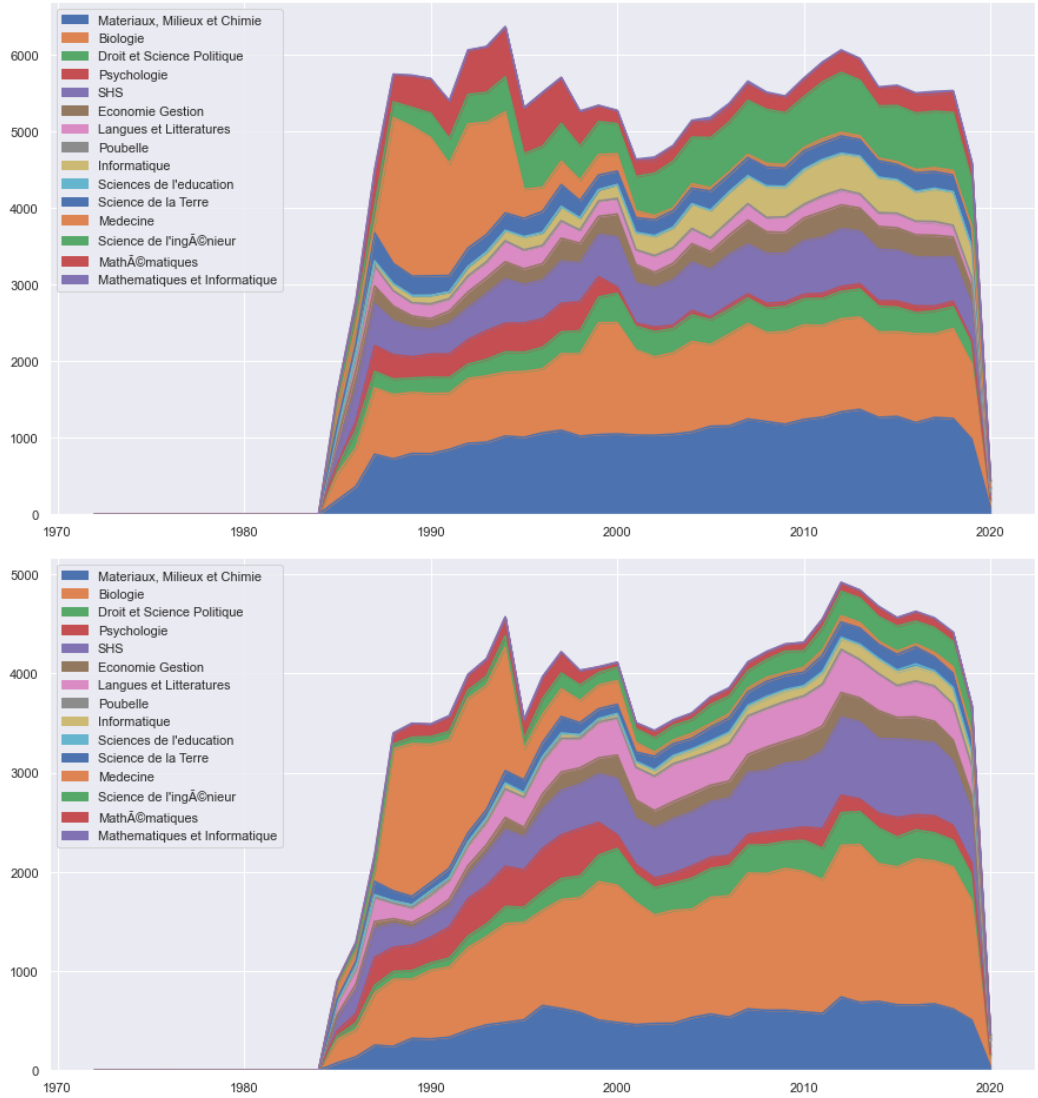


Figure 4.3: The evolution of number of theses in different displines by gender (up) male and (down) female

Figure 4.4: The evolution of number of theses in different language by gender (up) male and (down) female

5 | Discussion

In this report, we analyze the data frame from *theses.fr*, which contains 17 category features of one thesis, and there has a record of 447644 theses dated from 1971 to 2020. We investigate the missing values in this data frame, and find out the missing value like the PhD registration date is due to the regulation in France, while other missing values like year/date of the PhD dissertation defense, language, supervisor etc, are due to the incompleteness of the data. By going through the context based information, we find that the number of PhD theses increases with year and the dissertation date matches with school agenda, especially after the publishing website *theses.fr*, which provides more precise information. Meanwhile, we find that the homonym issue on the author and supervisor doesn't been solved by the additional Identity feature, and the category of disciplines needs to be further examined for clarification.

Bibliography

Audrin, B. (2019). *Making sense of digitization: three studies on the digitization concept and its implementation in organizations* (Doctoral dissertation). University of Fribourg.

Bingham, A. (2010). The digitization of newspaper archives: Opportunities and challenges for historians. *20TH CENTURY BRITISH HISTORY*, *21*(2), 225-231. Retrieved from https://academic.oup.com/tcbh/article-abstract/21/2/225/1703826?redirectedFrom=fulltext#no-access-message doi: https://doi.org/10.1093/tcbh/hwq007

l'ABES. (2011). *A propos.* http://www.theses.fr/apropos.html. (Online)

Rasuli, B., Solaimani, S., & Alipour-hafezi, M. (2019). Electronic theses and dissertations programs: A review of the critical success factors. *College & Research Libraries*.

Reis, J., Amorim, M., Melão, N., Cohen, Y., & Rodrigues, M. (2020). Digitalization: A literature review and research agenda. In Z. Anisic, B. Lalic, & D. Gracanin (Eds.), *Proceedings on 25th international joint conference on industrial engineering and operations management – ijcieom* (pp. 443–456). Cham: Springer International Publishing.