PhD$_v2 - THESISANALYSIS$

Emmanuel Humbert
Prétraitement manipulation de données

August 31, 2024

Contents

List of Figures

List of Tables

1 Introduction

It is not a secret PhDs have been considered a top academic rank among graduates from all horizons and is today well known for being as time consuming for the student as for the supervising director. An important step before graduating with a PhD is writing a thesis. And choosing the appropriate language assumed a greater significance years after years. Indeed, changes in reforms and ways of thinking supported by internet probably played a role in the evolution of thesis writing. We decided to study accurately the choice of a language through time trying to understand trying to give possible causes.

We also desire to salute research works carried out by Dr. Matthieu Cisel in the field of Learning Analytics. These works were a source of motivation to study this dataset thoroughly.

It is necessary to say we began this endeavour knowing Education is an important target of financing either through the State or thank to more private funding plans. Yearly effort from the State to improve the situation firstly convinced us detailed financial information were available somewhere. We were wrong. Consequently, without information on financial channels, studying the financing of PhD thesis or real causes behind language evolution became arduous. We also considered possible influences on thesis online accessibility but one more time the lack of information on Financial channels forced us to mere supposition.

Questions we asked ourselves are :

- How are distributed defenses months and how can we explain it ?
- How is organized the thesis supervision among directors and what could be the reasons ?
- What are the main languages chosen in writing reports and how did it evolved in time ?

To answer these questions we decided to proceed step by step verifying firstly missing data. Then we checked whether some months had more success than others and why. We also verified whether some mistakes had been done when registering authors focusing on namesakes. We also reckoned the quantity of thesis attributed to directors and for which reasons. Eventually, we evaluated what was the most successful languages when choosing to write a thesis.

Our hypothesis are :

- Mistakes are possible when registering thesis, author and directors in a software. That could lead to irregularity in the data
- Being in France, French probably gathers a large part of thesis

2 Material and Methods

In the next section, we will firtly introduce our sources and their reliability. Secondly we will present our methods to study this dataset.

2.1 Sources

To carry out this study we used the PhDv2 dataset, which was created from data available on the website Theses.fr. This website was created in 2011 and it is a Research browser gathering all french thesis. Managed by the Agence Bibliographique of higher education under the supervision of the Ministry of High Education, Research and Inovation, we considered it a first rank source to perform our mission.

2.2 Outliers

Counting the number of thesis supervised by directors we observed that the number of thesis attributed between 1984 and 2018 goes from only one thesis (Mr Gerald Panier, Mr Gerald Franz) to 208 thesis (Jean-Michel Scherrmann). We also found out 711 thesis have no thesis director.

We verified the distribution of thesis by month and by year for Mr. Jean-Michel Scherrmann (208 thesis), Mr. François-Paul Blanc (201 thesis) and Mr Philippe Delebecque (104 thesis) to compare possible outliers with a director having a ”normal” thesis distribution. Concerning Mr. Scherrmann and Blanc nearly all thesis were defended in January except one or two for each of them, from 1984 to 2018. For Mr Delebecque, thesis' defenses are distributed among every months of the year though January is also the prefered one.

In 1994, his most busy year, Mr Scherrmann supervised 40 thesis in January. Divided equally it means 10 thesis by week. For Mr. Blanc, the most busy year is 2004 with 33 thesis probably all in January. Dividing equally it means at least 8 thesis by week. We can suppose a mistake was committed making these two directors the outliers of the dataset. We can also suppose this two directors are hard worker and overachiever employees. We prefered to let you decide which one of these answers is true. However we haven't been able to find an explanation to the 711 thesis without a director. With more data on this, we will be able to go further.

2.3 Methods

To reach our results we firstly investigated missing data separating that kind of data from the dataset. It led us to firstly create a graph to locate columns the most concerned by missing data. We went further in this matrix because we also sorted data to verify whether data missing in a column is related to another column. Secondly we created an heatmap showing the proportion of missing data in these columns.

Afterwards we decided to focus on data from 1984 to 2018, to get the average number of thesis' defenses by month. Then we created a Facetgrid presentation focusing this time on 2005 to 2018 in order to develop our understanding of defenses by month. These operations led us to carry out other graphs and to exclude one month. We calculated again the proportion of thesis by month identifying consequently the real prefered month. We added error bars to get a more interesting result.

Conscious namesakes can lead to biased results we chose the name Cecile Martin to get a proof of the namesake problem problem in the dataset. We afterward managed to calculate the quantity of thesis baring the same author's name using the thesis ID. To identify possible outliers we compared the directors with the most thesis' defenses to a director having a "normal" quantity of thesis' defenses.

We finally studied the languages used to write thesis reports combining several lines in one graph. We did the same for reports' accessibility online. Both graphs helped us to understand how evolved preferences from 1984 to 2018.

2.4 Tools

The main Python packages we used to calculate and display our results are :

- Pandas (Data Manipulation)
- Numpy (Data Manipulation)
- Matplotlib (Visualization)
- Seaborn (Visualization)
- Datetime (To convert a column to datetime format)

3 Results

In this part we will firstly introduce our results. Afterwards, we will describe the tables and graphs who led us to these results.

3.1 Introduction

This dataset is composed of 447644 rows and 18 columns when beginning of our search. Data are of the object type. Most of them are string data except for the Year column whom are float data. It should be integer. Data are distributed between 1971 and 2020.

Working on this dataset we have been able to get the following results :

Investigating missing data we found out several columns are concerned by a large quantity of missing data. We also found that among these columns some behave together. Then looking for an issue in the dataset we found out an unequal distribution of defenses by month from 1984 to 2018. Removing a month from the data we have been able to find out the real prefered month for defenses.

Looking for namesakes we indeed found out the name Martin is often used. The Thesis ID was also critical in obtaining results. We have 3 possible outliers considering all their defenses date are scheduled in January. Consequently the number of defenses by week seems unrealistic. Many thesis do not have supervising directors too. Concerning languages, we determined which one is the most used and observed a change in writing habits.

3.2 Understanding missing data

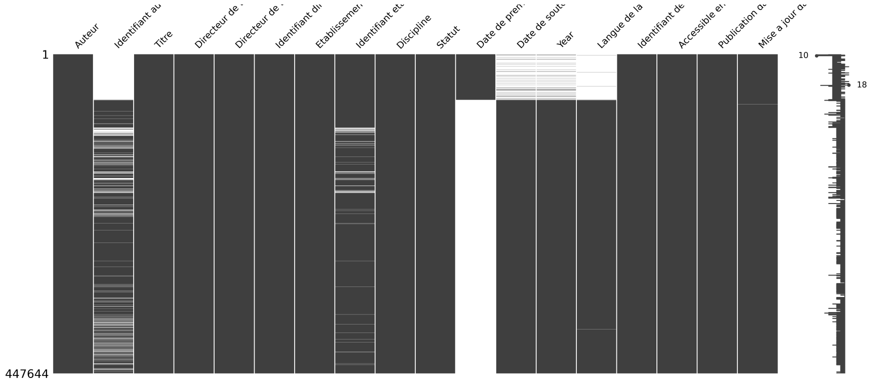

Figure 1: Distribution of missing data by column (sorted)

The figure 1 is a matrix of missingness. It shows missing data among other others. Columns concerned by missing data are "Identifiant Auteur", "Identifiant Etablissement", "Date de première inscription en doctorat", "Date de soutenance", "Year", "Langue de la these" et "Mise à jour dans these.fr". Sorting data helped to know the kind of missing data we have. "Identifiant Auteur" is missing at random (MAR), "Identifiant Etablissement" is missing completely at random (MCAR), "Mise à jour dans theses.fr" is missing completely at random. All the other columns we introduced here are not missing at random (MNAR)

Indeed as soon as we get data in the "Data de soutenance" column we lose data in the "Date de premiere inscription en doctorat" column. It is probably an administrative problem. As soon as the defense date is known it is entered in the software and this action delete the registration date. Columns "Year" and "Langue de la these" behave in opposition to "Date de premiere inscription en doctorat".

The following table shows the proportion of missing data for the columns the most concerned by missing data :

	En cours	Soutenue
Statut	0.0	0.0
Date de premiere inscription en doctorat	3.5	100
Date de soutenance	83.4	0
Year	85.4	0
Langue de la these	95.9	0
Identifiant auteur	99.7	16.7

Table 1: Missing data in percentage

3.3 Distribution of defenses by month and year

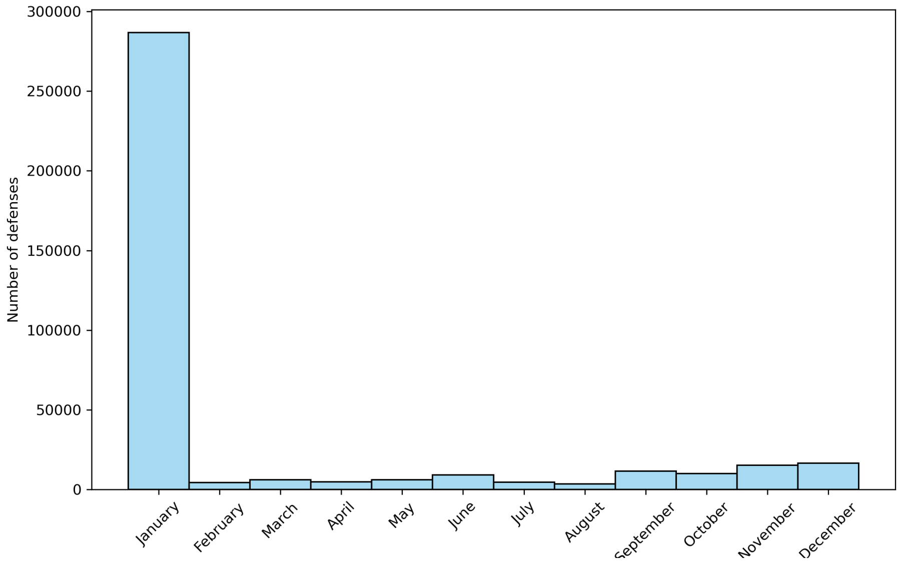

Figure 2: Distribution of defenses months (1984-2018)

We created this histogram to understand how are distributed defenses in time. From 1984 to 2018. It is clear that January is the month during which most of defenses were planned. However, it seems that all other months where also selected to schedule defenses. We consequently decided to shorten the time period focusing on data from 2005 to 2018.

The following figure shows the average number of defenses by month from 2005-2008. We added an error bar to highlight confidence intervals :

In Figure 3 below, we can observe January is still the most selected month in proportion with an average number of 5968.50 thesis' defenses. We also see the error bars are very long. A very long error bar in a bar chart indicates a high level of variability or uncertainty in the data. Here are several possible interpretations:

1. High Variability: If the error bars represent standard deviation, a long error bar means that the individual data points are spread out widely

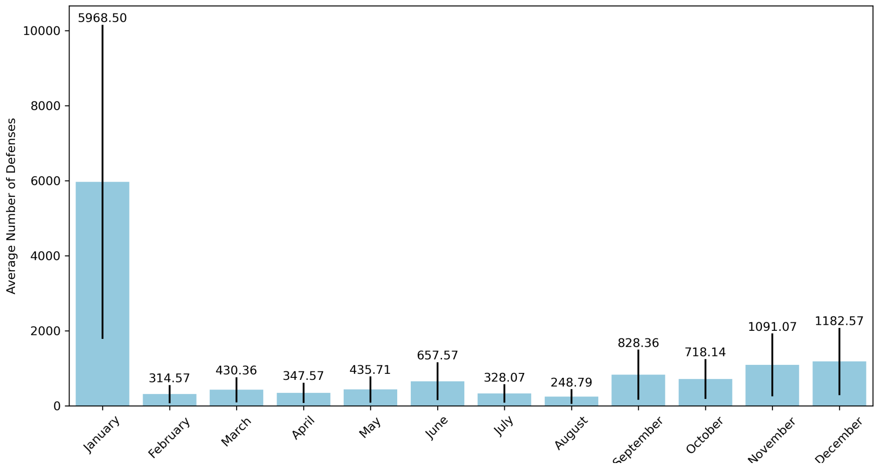

Figure 3: Average Number of Defenses per Month (2005-2018)

around the mean, indicating high variability within the dataset.

2. Low Precision: If the error bars represent standard error or confidence intervals, a long error bar suggests that the sample mean is a less precise estimate of the population mean. This can occur with smaller sample sizes or highly variable data.

3. Measurement Uncertainty: Long error bars can indicate that the measurements themselves are subject to significant uncertainty or error, implying that the true value could vary wide

4. Potential Outliers: Large error bars might be a sign of outliers or extreme values in the dataset, which increase the overall range of the data.

Due to the presence of these long error bars, any conclusions drawn from the data should be made with caution. It suggests that the true values could be quite different from the observed values.

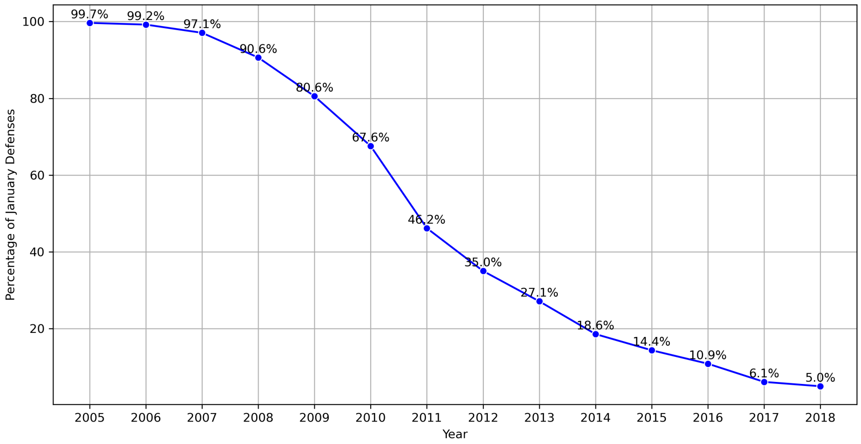

Figure 4: Proportion of defended thesis in January (2005-2018)

In Figure 4 the percentage of defended thesis in january goes from 99.7% in 2005 to 67.6% in 2010. From here, it diminishes from 46.2% in 2011 to 5% in 2018. This graph, added to the error bars of the previous graph, confirms january month biased our results until now.

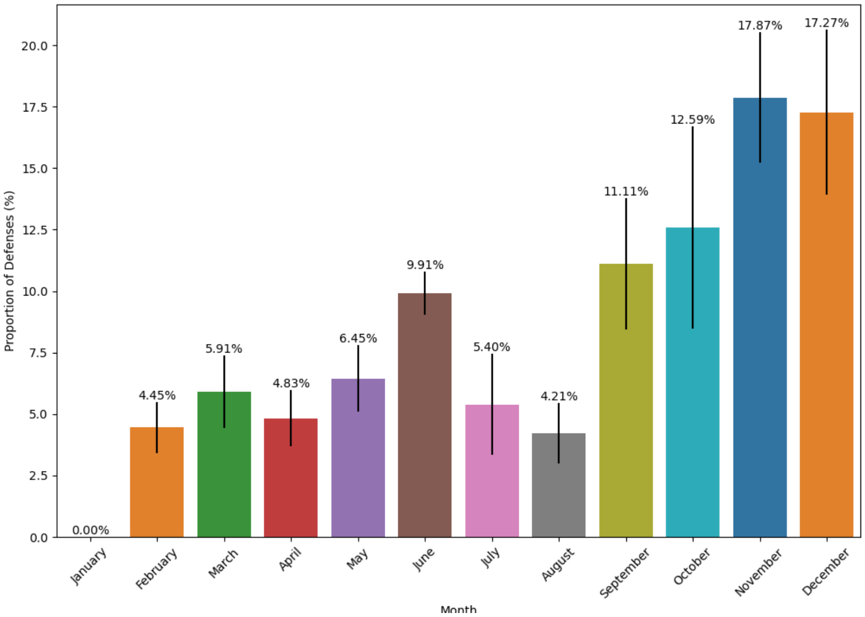

Figure 5: Proportion of thesis defenses by month (Excluding January) from 2005 to 2018

Excluding January we got the real proportion of thesis defenses by month. The most selected month is November with 17.87%. The least selected month is February with 4.45%. Error bars are shorter than in Figure 3. We can consequently say that removing January helped making our results more reliable.

Author	Thesis Count
Bruno Martin	10
Celine Martin	11
Franck Martin	12
Guillaume Martin	16

Table 2: Sample of thesis count by name sake.

We found out the namesake thesis count goes from 2 to 16. To get this result we counted unique thesis IDs. We can consequently say though we have 16 thesis registered to Guillaume Martin, all of them are unique. There is no mistake waiting to be corrected.

3.4 Details on outliers

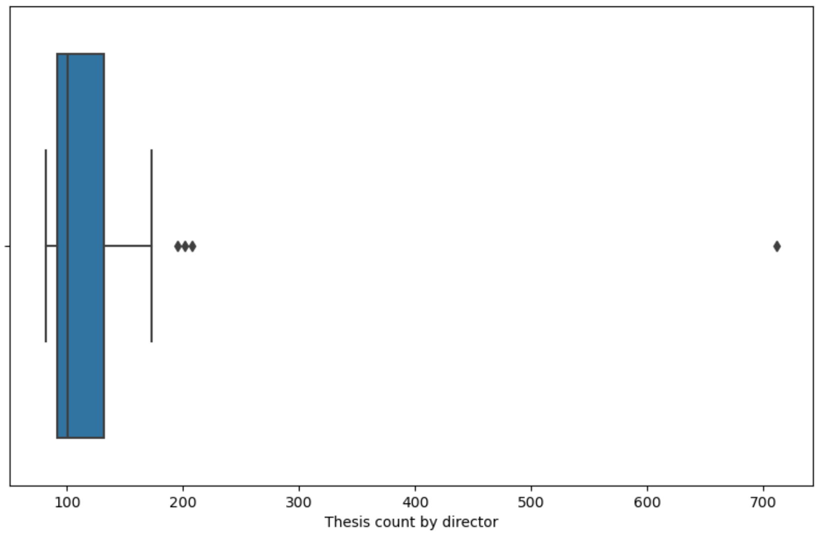

Figure 6: Identifying outliers among supervising directors (1984-2020)

We used a graph named boxplot to look for outliers. As you could see, 4 points are located outside the figure on the left. These are the possible outliers. Starting by the biggest we found 711 thesis with an unknown supervising director. After, there is M. Jean-Michel Scherrmann with 208 thesis, M. François-Paul Blanc with 201 thesis and M. Pierre Brunel with 195 thesis.

3.5 Thesis Language and Online Accessibility

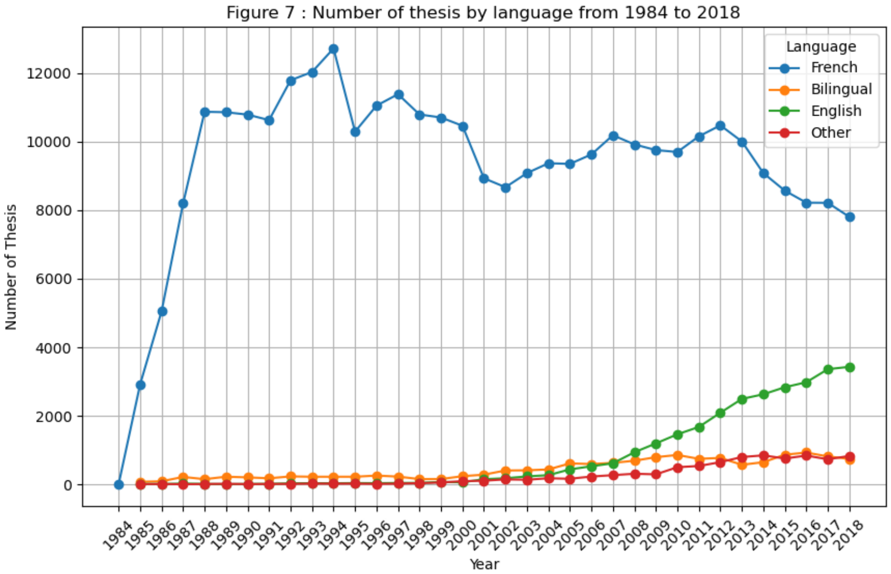

Figure 7: Number of thesis by language from 1984 to 2018

We created a line plot with multiple lines to make the comparison more clear. We observed French is the language the most used. From 1984 to 1994 we have a steep rise in french thesis while other languages are nearly completely ignored. From 1994, french thesis began slowing down. Starting from more than 12000 in 1994 it slowly diminished to a little less than 8000 in 2018.

Bilingual, English or Other language thesis remained unattractive until 2000. From here we observed a low rise of bilingue and other language thesis to nearly 1500 a year in 2018. English begins a slow rise from 2000 and a more dynamic one from 2007 to reach nearly 3500 thesis in 2018.

Several reasons are possible to explain these changes.

The rise of french thesis may be due to educational reforms in the 1980's and an increase access to higher education. Its slowing down may be due to a saturation of the market for academic position, an economic recession or some policy changes.

The rise of English from 2000 may be due to the globalization of education and the academic prestige when publishing in this language.

Please find below the line plot on online accessibility :

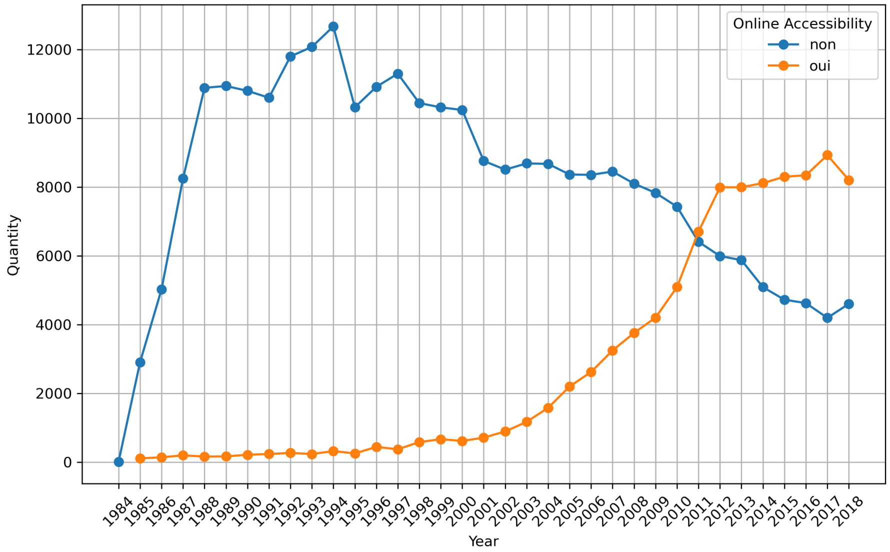

Figure 8: Thesis online accessibility from 1984 to 2018

In this line plot we have one line for YES and another for NO. We observed a steep rising line of offline thesis from 1984 to 1994 and a regularly rising line of online thesis from 1994 to 2018. To compare, we had more than 12000 offline thesis and less than 500 online thesis in 1994. In 2018, we had nearly 5000 offline thesis and 8000 online thesis.

We explain this evolution by the commercialization and accessibility of internet starting between 1993 and 1995. The explosion of web content from

1995 to 2000 and the development of social media in the 2000s had a great impact on academic habits.

4 Discussion

Research key points

- Columns showing large quantity of missing not at random data work together
- Due to the presence of outliers and possible mistakes January appeared wrongly as the prefered month for defenses.
- November is the prefered defense month
- Namesakes are numerous but adding the thesis ID avoid counting twice a thesis.
- Prefered language for writing thesis is French
- More thesis are accessible online than staying offline in 2018.

4.1 Structure of the discussion

We decided to focus the discussion section on the distribution of defenses month because it required us to ignore a part of the dataset to reach the correct result.

4.2 Interpretation of Defenses by Month and year

As developed earlier our first distribution graph showed January right away appeared as the prefered month for defenses. Creating a facetgrid focusing on years from 2005 to 2018, we found defenses were nearly always performed in January from 2005 to 2008. Respectively January defenses were 10526 in 2005, 10889 in 2006, 11355 in 2007 and 10744 in 2008. In 2009 January defenses began to slow down from 9693 to 640 in 2018.

The difference with other months was indeed visible. The real prefered month for defense, November, had 8 defenses in 2005, 19 in 2006, 72 in 2007 and 230 in 2008. The maximum for November being reached in 2018 with 2229 defenses. Error bars used in Figure 4 gave us the confirmation data at hand could be unreliable. The quantity of defenses supervised by 3 directors (208, 201 and 195), almost all in January, helped us also in deciding to remove January from the calculation.

The new proportion of defenses was more realistic. November, December and October with respectively 17.87%, 17.27% and 12.59% are the most selected month for defenses. We also observed that the distribution of defenses all along the year presented no extremes. The remaining months are between 4.45% and 11.11%.

Without evoking outliers and the possibility of committing mistakes we can also suppose the concentration of thesis defenses in January and the assignment of many theses to a few directors is explained by a combination of academic calendar considerations, administrative efficiencies, director expertise and availability, departmental policies, and student preferences. These factors interplay to create patterns in thesis scheduling and supervision assignments.

Understanding these dynamics requires a look at both institutional policies and individual choices within the academic environment. Each factor contributes to the observed trends, highlighting the complexity of academic planning and management.

4.3 Conclusion

Our results, though giving a detailed picture of the situation and corroborating our hypothesis, are uncomplete. Indeed our data concerned thesis defended in France. With data from several countries we would have been able to provide a more global view of the situations we encountered.

In order to give a more interesting and up to date account of these data, it would be also interesting to study the gender among phd students. Added to the fund attributed to student and subjects, getting to know whether we have a difference based on gender and fields of study would be an interesting addition to our research.