Etude des soutenances des thèses en France

Sébastien Mertès

8 juin 2023

Table des matières

1. Introduction **1**

2. Présentation des données **1**
2.1 Recherche des variables 1
2.2 Nouveau dataset 2

3. Valeurs manquantes **2**
3.1 Heatmap 3

4. Principaux problèmes détectés **5**
4.1 Production des thèses 5
4.2 Répartition de la production des thèses 7
4.3 Le cas du mois de janvier 8
4.3.1 Le 1er janvier 8
4.3.2 Proportion des soutenances sur chaque année 9
4.4 Approche réelle de la distribution mensuelle des soutenances 9
4.5 La question des homonymes 10
4.5.1 Le cas d'un auteur 10
4.5.2 Résumé des interprétations 11
4.5.3 Hypothèse du nombre d'homonymes 11

5. Outliers et résultats anormaux **11**
5.1 Directeur de thèses 11
5.2 Le cas d'un directeur 12

6. Résultats préliminaires **13**
6.1 Evolution des langues d'écriture 13
6.2 Proportion des langues 13

Résumé

L'étude de la production de thèses depuis les années 2005 à 2020, montrerait qu'elles sont principalement soutenues au mois de décembre. La sélection de cette période a été choisie afin d'avoir les données les plus fiables en rapport aux dispositions légales sur l'obligation de transmettre la production de thèses par les établissements à partir de l'année 2006. Nous verrons que ces référencements contiennent des erreurs dans la distribution du nombre de thèses par auteur en remarquant la présence d'homonymes. Nous remarquerons également, que le nombre d'encadrements de thèse pour certains directeurs est anormalement élevé et d'en conclure qu'il est aberrant. Pour finir, nous observerons que la tendance du choix de la langue d'écriture évolue de manière significative par rapport à celle du français qui reste la langue encore préférée aujourd'hui.

1. Introduction

Cette étude est essentiellement axée sur les soutenance de thèses sur la période 1984-2020. Nous mettrons en lumière les difficultés d'extraction des données depuis un jeu de données comportant 22 variables où des techniques de nettoyage des données ont été appliquées. Nous étudierons les valeurs manquantes et les principaux problèmes détectés sur la question de la répartition mensuelle des soutenances et des auteurs. Nous nous pencherons également sur la détection de données aberrantes en prenant le cas des directeurs de thèse. Enfin, nous étudierons l'évolution des langues d'écriture au cours du temps.

2. Présentation des données

2.1 Recherche des variables

Nous observons, dans le Tableau 1, l'ensemble des variables du dataset PhD. La colonne 0 "Unnamed : 0" ne représente qu'un indexage et pourrait être supprimée du dataset. La nomination des variables nous indique le type de variables et ce qu'elles peuvent représenter. Nous sommes donc en présence d'un jeu de données composé de variables indiquant l'enregistrement des thèses par des établissements dans la base de données nationale dont le site www.theses.fr fait la référence. Principalement, nous connaissons les différents domaines de recherche par la discipline indiquée, mais également les noms des encadrants et des étudiants ainsi que la date de soutenance des thèses, qu'elles soient en préparation ou été soutenues (Statut), et la langue de rédaction. Nous remarquons la présence d'erreurs d'encodage et de typage de variable indiquant des données de type date, notamment pour les variables enregistrant les dates de publication et de mise à jour sur le site theses.fr.

	Variable	non nulle	Type
1	Auteur	448047	object
2	Identifiant auteur	317700	object
3	Titre	448040	object
4	Directeur de these	448034	object
5	Directeur de these (nom prenom)	448034	object
6	Identifiant directeur	448047	object
7	Etablissement de soutenance	448046	object
8	Identifiant etablissement	430965	object
9	Discipline	448047	object
10	Statut	448047	object
11	Date de premiere inscription en doctorat	64331	datetime64[ns]
12	Date de soutenance	390961	datetime64[ns]
13	Year	390961	float64
14	Langue de la these	448047	object
15	Identifiant de la these	448047	object
16	Accessible en ligne	448047	object
17	Publication dans theses.fr	448047	object
18	Mise a jour dans theses.fr	447870	object
19	Discipline_prÃ©di	448047	object
20	Genre	448047	object
21	etablissement_rec	444973	object
22	Langue_rec;;;;;;;;;;;;;;;;	383927	object

TABLEAU 1 – Variables du 1er dataset.

2.2 Nouveau dataset

Le Tableau 2, contient principalement les mêmes variables que la Tableau 1. Contrairement au tableau précédent, il sera nécessaire de changer le type des variables des dates de soutenance et de première inscription en thèse afin de représenter les différentes distributions et proportions selon les périodes dans la suite de l'étude.

	Variable	non nulle	Type
0	Auteur	448047	object
1	Identifiant auteur	317700	object
2	Titre	448040	object
3	Directeur de these	448034	object
4	Directeur de these (nom prenom)	448034	object
5	Identifiant directeur	448047	object
6	Etablissement de soutenance	448046	object
7	Identifiant etablissement	430965	object
8	Discipline	448047	object
9	Statut	448047	object
10	Date de premiere inscription en doctorat	64331	object
11	Date de soutenance	390961	object
12	Year	390961	float64
13	Langue de la these	448047	object
14	Identifiant de la these	448047	object
15	Accessible en ligne	448047	object
16	Publication dans theses.fr	448047	object
17	Mise a jour dans theses.fr	447870	object

TABLEAU 2 – Variables du 2ème dataset.

3. Valeurs manquantes

La Figure 1, ci dessous, représente visuellement les valeurs manquantes par variable, représentées par des lignes blanches. Ce graphique a été obtenu en triant les données par ordre chronologique des soutenances. Cela permet de révéler la régularité du manque de précision sur la date d'entrée en thèse de l'étudiant en rapport à sa soutenance. Le Tableau 3 permet de confirmer cette observation en comptabilisant le nombre de valeurs manquantes par variable.

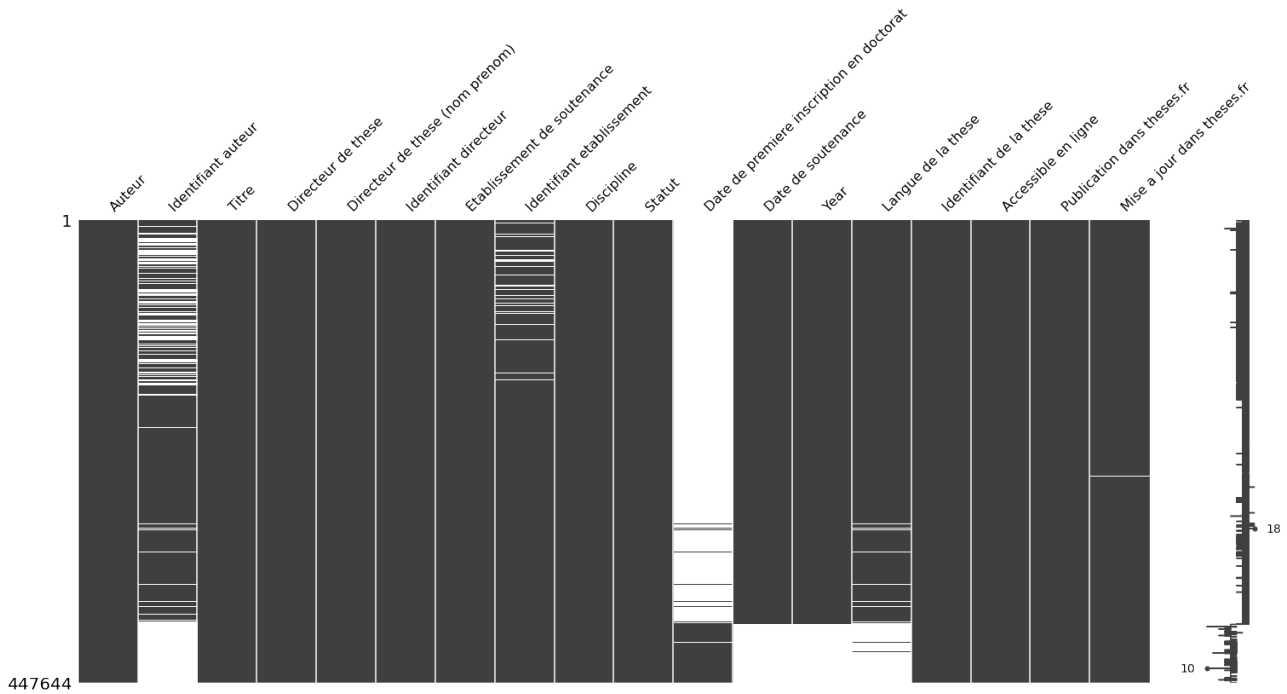

FIGURE 1 – Valeurs manquantes

Variable	valeur manquante
Auteur	0
Identifiant auteur	129989
Titre	9
Directeur de these	15
Directeur de these (nom prenom)	15
Identifiant directeur	0
Etablissement de soutenance	4
Identifiant etablissement	17085
Discipline	5
Statut	0
Date de premiere inscription en doctorat	383668
Date de soutenance	56746
Year	56746
Langue de la these	63765
Identifiant de la these	0
Accessible en ligne	0
Publication dans theses.fr	0
Mise a jour dans theses.fr	177

TABLEAU 3 – Nombre de valeurs manquantes.

3.1 Heatmap

Nous visualisons sur la Figure 2 la proportion des valeurs manquantes en pourcentage de trois variables selon que les thèses soient en préparation ou soutenues. En rapport à la Figure 1, les trois variables dont nous observons la proportion de valeurs manquantes, ont été choisies selon le heatmap de corrélation entre variables de la figue 3. Il ressort que l'indication des dates de première inscription en doctorat est corrélée avec les dates de soutenance et la langue de la thèse. La figure 2, permet d'apporter davantage de précisions sur la nature de cette corrélation. Ainsi nous constatons que, systématiquement, les données des dates de soutenance sont fournies lorsque les thèses ont été soutenues. L'article de Martin I. intitulé *"Le signalement des thèses de doctorat"* et la recherche sur le site de

theses.fr, précisent que les établissements ont une obligation légale d'enregistrer les thèses produites dans les répertoires nationaux Sudoc et STAR depuis l'année 2006. Cela supposerait qu'avant cette date, il n'était pas obligatoire d'enregistrer les thèses soutenues. Cela s'observe avec les résultats obtenus dans la figure 1 de la répartition des valeurs manquantes où l'on remarque un manque important d'enregistrements des dates de première inscription en thèse au profit des dates de soutenance qui sont systématiquement indiquées.

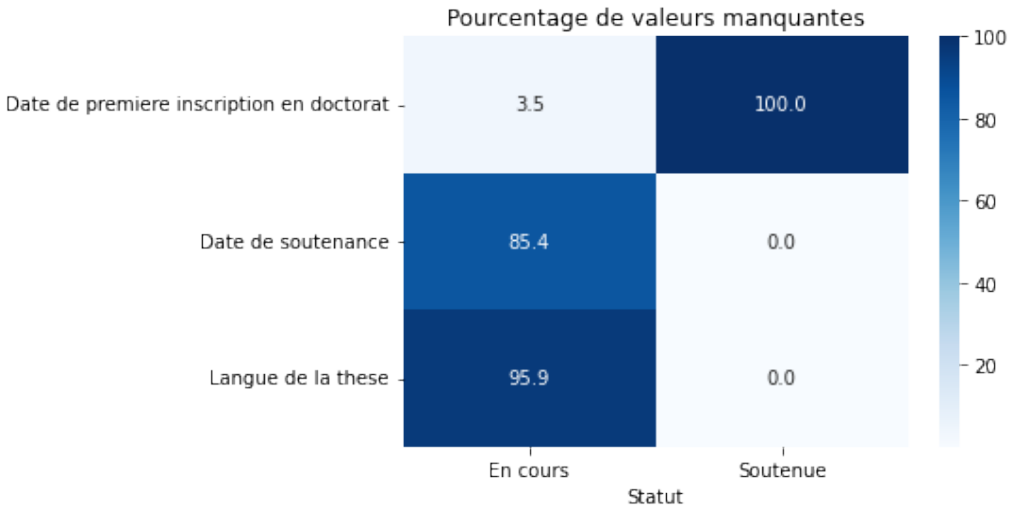

FIGURE 2 – Heatmap de la proportion des valeurs manquantes

En observant plus précisément la Figure 3, ci-dessous, nous remarquons que la proportion de valeurs manquantes des dates de soutenances et la langue choisie sont inversement corrélées aux dates de première inscription en thèse et aux langues. Cette observation corrobore les résultats obtenus par le calcul du nombre de valeurs manquantes et leurs représentations dans le Tableau 3 et la Figure 1 où l'on observe que leurs distributions sont inversement proportionnelles.

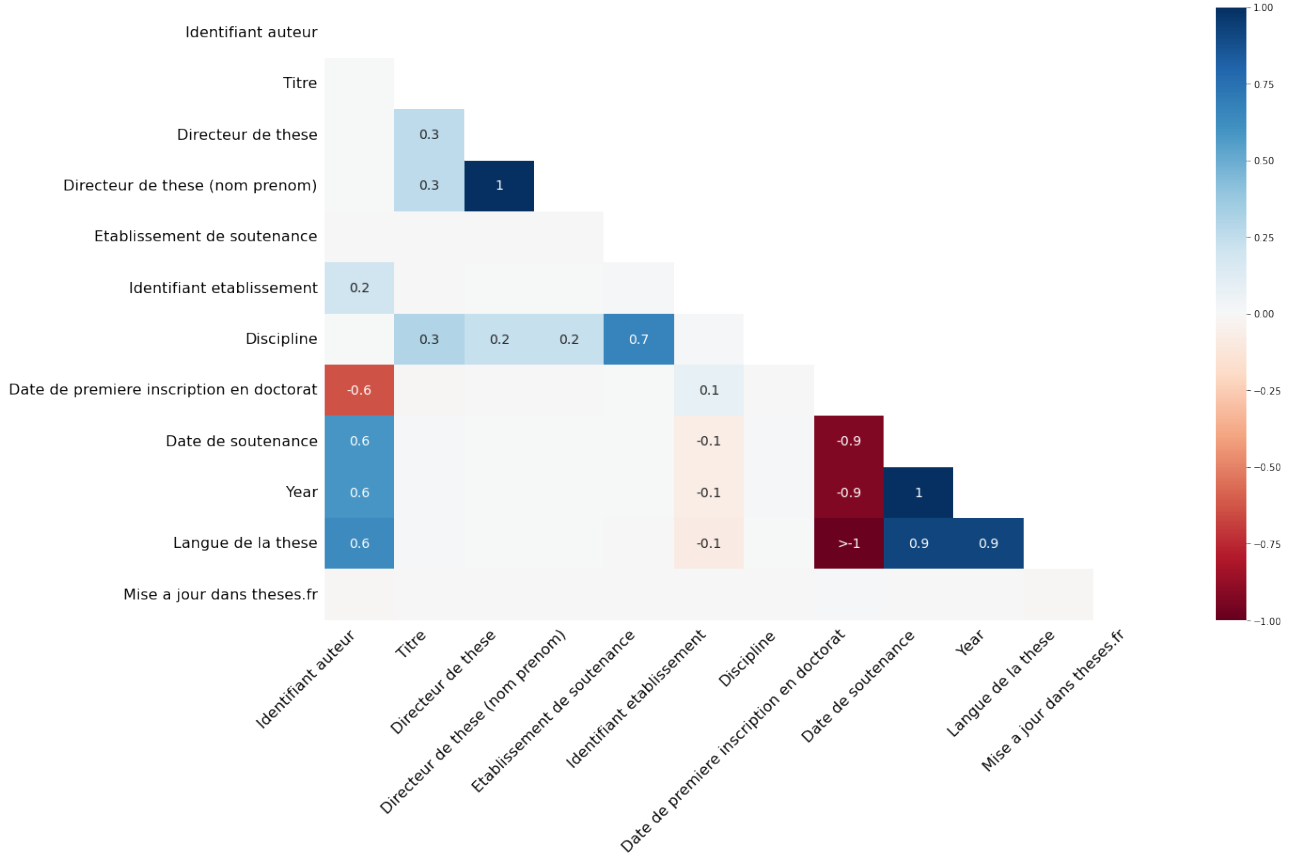

FIGURE 3 – Heatmap de corrélations

4. Principaux problèmes détectés

4.1 Production des thèses

La Figure 4 permet de visualiser la période à partir de laquelle le nombre de thèses soutenues a fortement chuté. Nous constatons qu'elle commence approximativement à partir de l'année 2018.

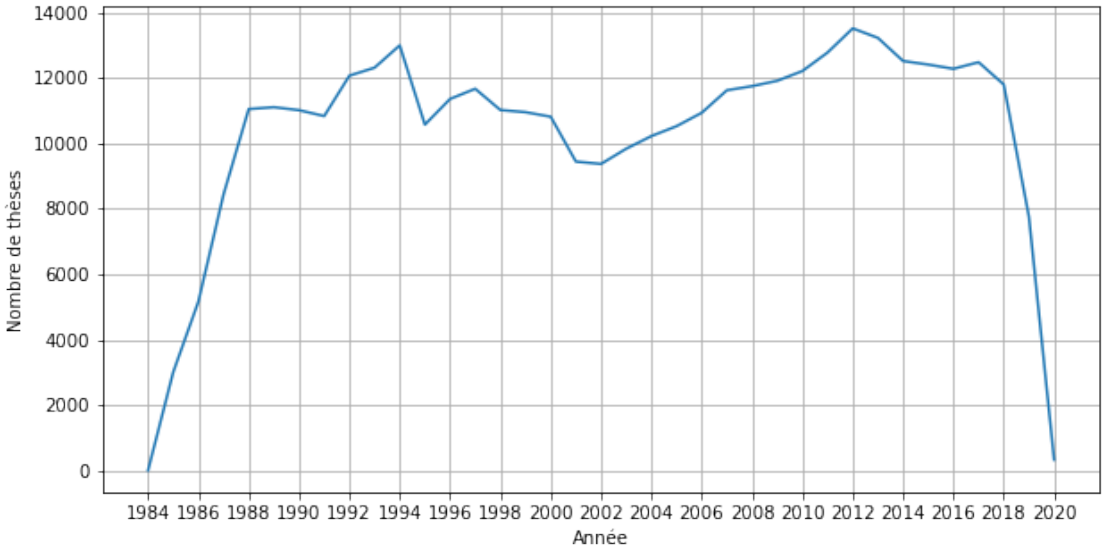

FIGURE 4 – Production de thèses par année

Les Figures 5 et 6 représentent la distribution et la proportion de thèses soutenues. Nous observons une concentration anormale de la production de thèses au mois de janvier par rapport aux autres mois de l'année sur la période entière de 2005 à 2018. Nous pourrions interpréter cette observation par le fait qu'il y a un décalage entre la date de soutenance de l'étudiant dans son établissement et son enregistrement effectif dans la base nationale. On peut envisager, en observant la Figure 5, que cet enregistrement se produit essentiellement au mois de janvier. Il n'est pas exclu qu'à ce décalage, s'ajoute également le rattrapage des enregistrements des dates de soutenances de l'époque précédant l'obligation d'utiliser les répertoires nationaux Sudoc et STAR et que le choix arbitraire en début du mois de janvier ait été retenu.

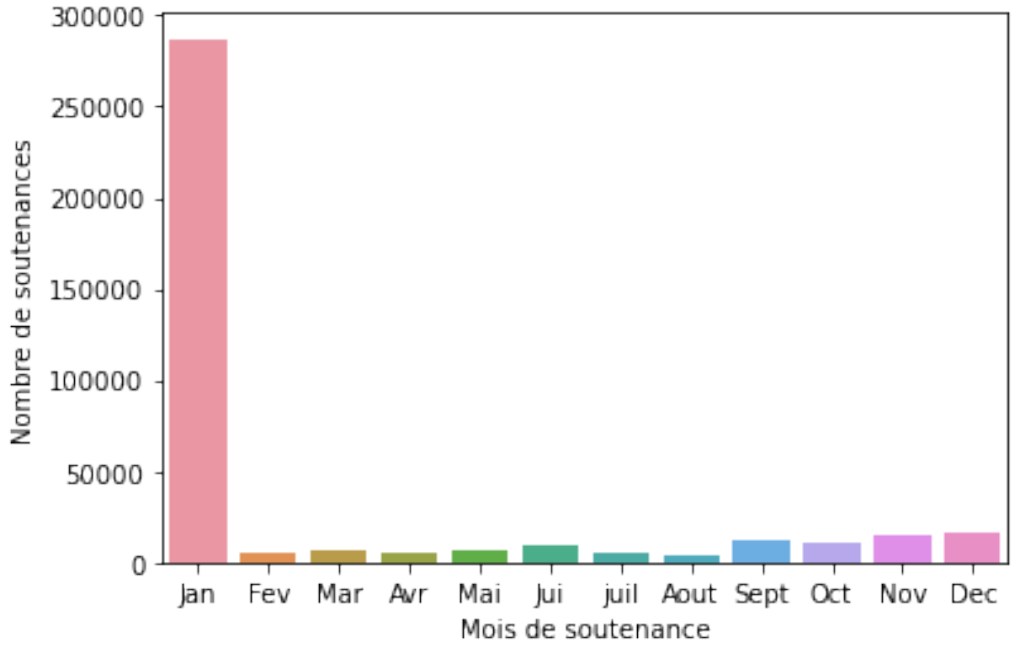

FIGURE 5 – Distribution des thèses par mois entre 2005-2018

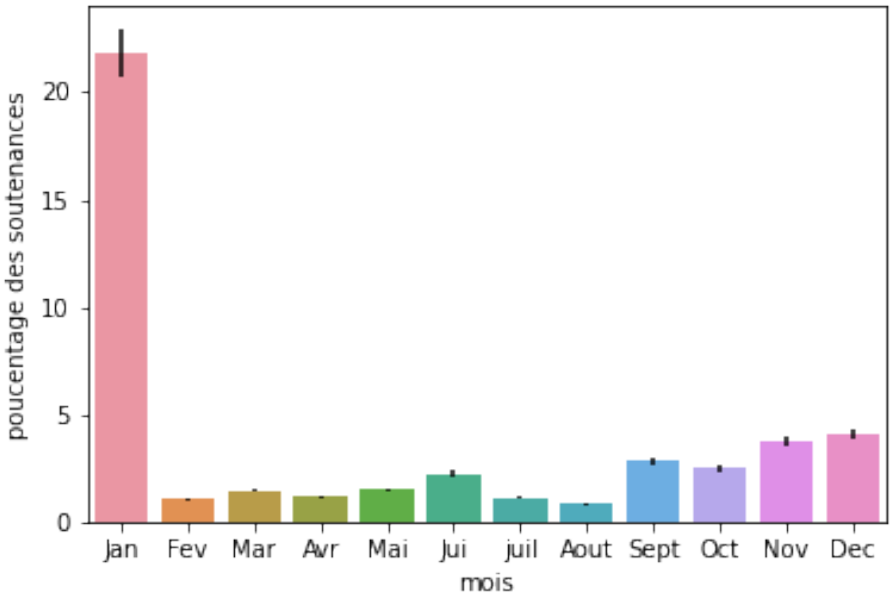

FIGURE 6 – Proportion des thèses soutenues entre 2005-2018

4.2 Répartition de la production des thèses

La Figure 7 montre la répartition par année et par mois de cette production sur toute la période. Cela permet de comprendre que le nombre de thèses tend à se répartir de manière plus homogène au cours du temps sur l'ensemble des mois de l'année. Nous remarquons que la proportion des soutenances au mois de janvier diminue progressivement au fil des ans.

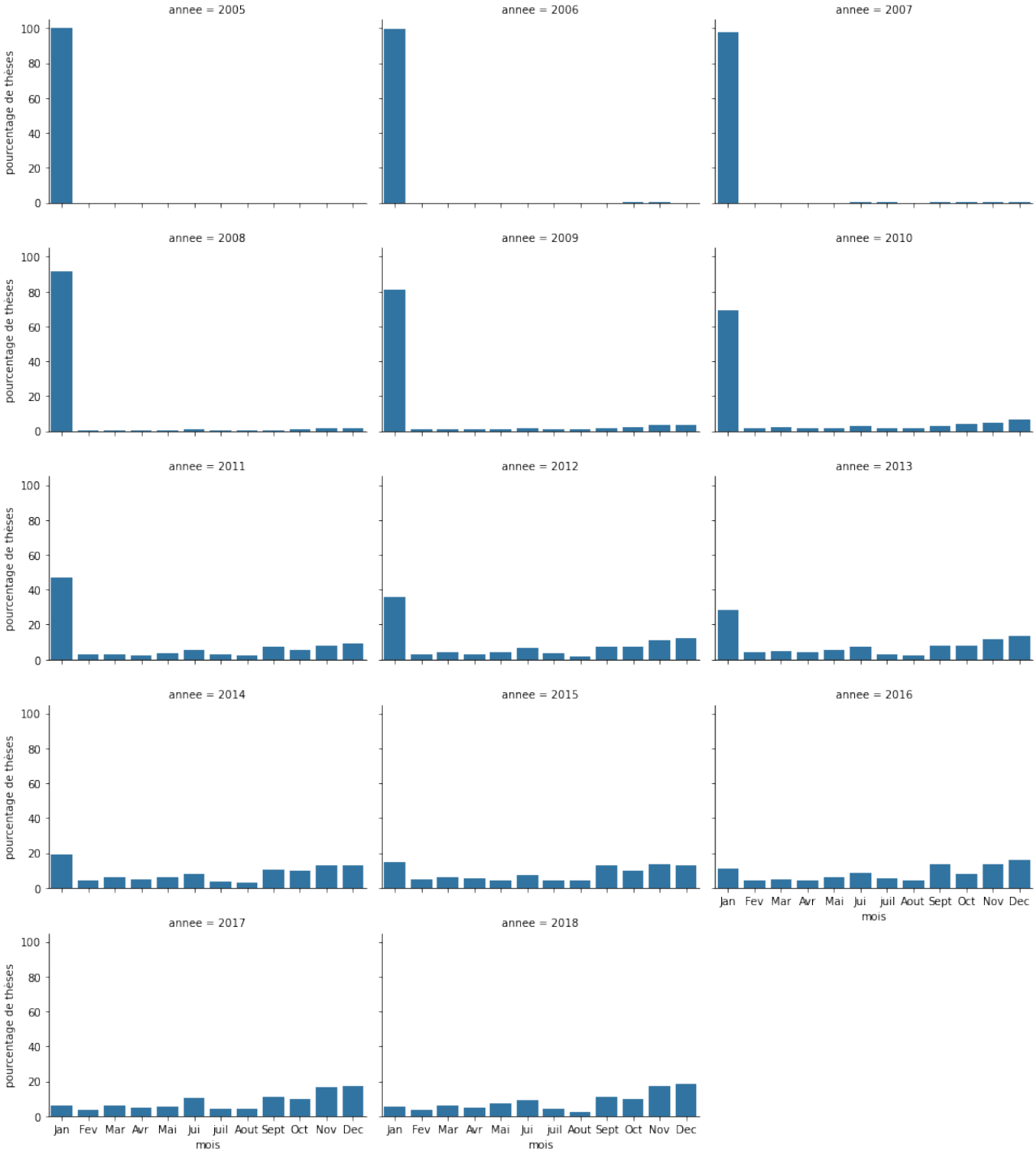

FIGURE 7 – Proportion des thèses soutenues par année entre 2005-2018

4.3 Le cas du mois de janvier

4.3.1 Le 1er janvier

Par les observations précédentes, le mois de janvier concentre un nombre curieusement élevé de thèses produites par rapport aux autres mois de l'année. Le Tableau 4 permet de nous intéresser à la distribution des soutenances par jour selon les mois de l'année.

Mois	Jour	Pourcentage de thèses
1	1	20.41
1	2	0.03
1	3	0.02
1	4	0.04
1	5	0.00
...	...	...
12	27	0.00
12	28	0.00
12	29	0.00
12	30	0.00
12	31	0.00

TABLEAU 4 – Nombre de thèses par mois et par jour entre 2005-2018.

Nous constatons que le 1er janvier concentre l'essentielle de la production des thèses, en comparaison avec la Figure 6, avec plus de 20 %, entre 2005 et 2018.

4.3.2 Proportion des soutenances sur chaque année

La Figure 8, montre une diminution importante des soutenances au mois de janvier au fil des années. Cela permet de synthétiser l'observation de la Figure 7, montrant la progression de la répartition des soutenances sur l'ensemble des mois de l'année pour la période étudiée.

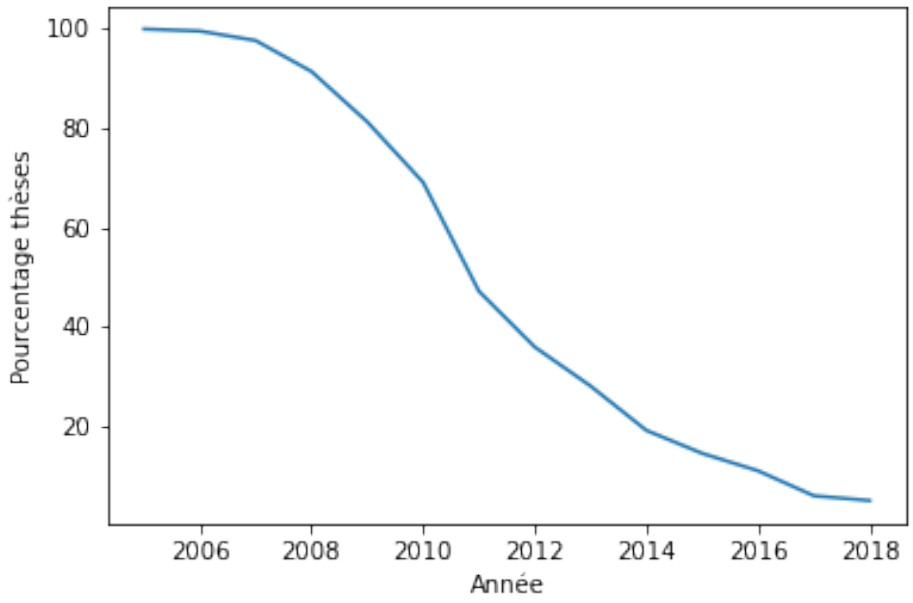

FIGURE 8 – Proportion des thèses soutenues en janvier entre 2005-2018

4.4 Approche réelle de la distribution mensuelle des soutenances

La prédominance des soutenances au 1er janvier suggère, depuis la loi exigeant que les établissements de formation transmettent les thèses produites, qu'une partie des thèses précédent la loi ait été enregistrée de manière arbitraire à cette date. Une hypothèse, expliquant cette observation, serait une récupération des thèses soutenues antérieures à l'obligation légale et enregistrées dans le répertoire nationale au 1er janvier selon l'année, à défaut de connaître la date précise. En supprimant les thèses soutenues au premier janvier à partir de l'année 2006, nous observons sur la Figure 9, une répartition

mensuelle des soutenances ces dernières années plus fidèle à la réalité. On remarque que le mois de décembre, serait le mois où le nombre de soutenances est le plus important dans l'année en approchant les 17%.

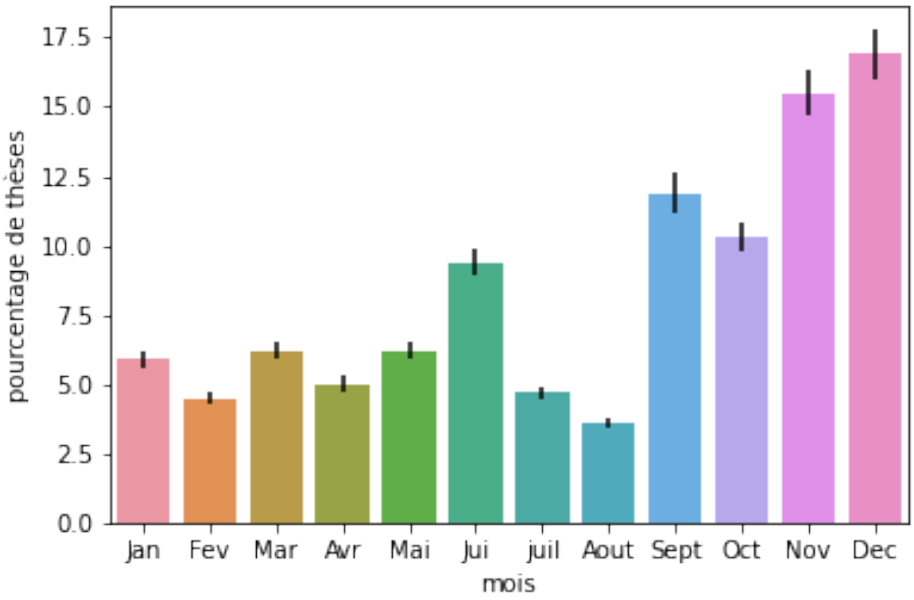

FIGURE 9 – Proportion des thèses soutenues par mois entre 2005-2018

4.5 La question des homonymes

4.5.1 Le cas d'un auteur

Prenons maintenant le cas du nombre de thèses par auteur. Nous pouvons soumettre l'hypothèse qu'il est rare pour un étudiant de soutenir plus d'une thèse dans sa carrière. Pourtant, certains auteurs recensent plus d'une thèse. Penchons-nous sur le cas de Cécile Martin dont le Tableau 5 liste les soutenances depuis 1984 à 2018.

Identifiant	Auteur	Etablissement	Soutenance	Discipline
182118703	Cecile Martin	Paris 11	1989-01-01	Physique
81323557	Cecile Martin	Bordeaux 2	1991-01-01	Neurosciences
81323557	Cecile Martin	Clermont-Ferrand 2	1994-01-01	Sciences biologiques et fondamentales applique...
81323557	Cecile Martin	Institut national agronomique Paris-Grignon	2000-01-01	Sciences biologiques fondamentales et applique...
81323557	Cecile Martin	Compiegne	2001-01-01	Genie des procedes industriels
179423568	Cecile Martin	Paris 9	2014-01-24	Sciences economiques
203208145	Cecile Martin	Sorbonne Paris Cite	2017-01-16	Etudes cinematographiques et audiovisuelles

TABLEAU 5 – Liste des thèses soutenues par Cécile Martin.

Nous constatons un nombre important de thèses soutenues pour un auteur. On remarque que pour certaines soutenances, les domaines de recherche sont très éloignés les uns des autres. En effet, la physique, la neuroscience, les sciences biologiques fondamentales et appliquées ainsi que les génies des procédés industriels, sciences économiques et études cinématographiques et audiovisuelles sont des sciences très différentes. Cependant, la neuroscience et les sciences biologiques fondamentales sont des sciences voisines. Mais certaines soutenances sont rapprochées à moins de 2 ou 3 ans, ce qui est impossible pour des thèses demandant 3 ans au minimum de préparation et d'un nombre conséquent d'années d'études pour acquérir les connaissances théoriques. On pourrait croire que les

identifiants des auteurs donneraient une indication précise sur l'identité d'une personne. Mais selon les remarques précédentes, il semble difficile qu'un même auteur puisse soutenir à la suite plusieurs thèses, à trois, et pire, à moins d'un an d'intervalle dans des domaines très différents. Il semblerait que nous sommes en présence d'homonymes, composés de 5 ou 6 personnes ayant des identités différentes.

4.5.2 Résumé des interprétations

Interprétation	Signification	Données
Données cohérentes	Domaines de recherche similaires	Neuroscience Science biologique fondementales
	Soutenances chronologiquement éloignées Identifiants auteurs communs	1994-01-01 / 2000-01-01 81323557
Données incohérentes	Fréquences des soutenances < à 3 ans	1989-01-01 / 1991-01-01 2000-01-01 / 2001-01-01
	Domaines de recherche différents	Physique Genie des procedes industriels Sciences economiques Etudes cinematographiques et audiovisuelles
	Identifiants auteur différents	182118703 81323557 179423568 203208145

TABLEAU 6 – Synthèse des données cohérentes et non cohérentes.

4.5.3 Hypothèse du nombre d'homonymes

Il serait plausible que l'auteur, indiqué comme homonyme C dans le Tableau 7, soit la même personne.

Homonyme	Auteur	Soutenance	Discipline
A	Cecile Martin	1989-01-01	Physique
B	Cecile Martin	1991-01-01	Neurosciences
C	Cecile Martin	1994-01-01	Sciences biologiques et fondamentales applique...
C	Cecile Martin	2000-01-01	Sciences biologiques fondamentales et applique...
D	Cecile Martin	2001-01-01	Genie des procedes industriels
E	Cecile Martin	2014-01-24	Sciences economiques
F	Cecile Martin	2017-01-16	Etudes cinematographiques et audiovisuelles

TABLEAU 7 – Liste hypothètique des homonymes.

5. Outliers et résultats anormaux

5.1 Directeur de thèses

Intéressons nous dorénavant aux directeurs de thèses et par conséquent à ceux qui ont le plus grand nombre de thèses dirigées. Le Tableau 8 nous renseigne sur le nombre de thèses encadrées par directeur pour les dix plus prolifiques entre 1984 et 2018. Nous remarquons que pour certains, leur nombre est très élevé. En effet, il est plus qu'improbable qu'un directeur ait dirigé ou codirigé plus d'une centaine de thèses dans toute sa carrière.

Directeur	Thèse
Scherrmann Jean-Michel	208
Blanc Francois-Paul	199
Brunel Pierre	195
Bertucat Michel	173
Pujolle Guy	170
Teyssie Bernard	134
Lumley Henry de	132
Chaumeil Jean-Claude	131
Foucart Bruno	130
Maffesoli Michel	128

TABLEAU 8 – Nombre de thèses dirigées par directeur les plus prolifiques (1984-2018)

Le Tableau 9 permet de comptabiliser leurs nombres de thèses dirigées par année en affinant la recherche selon les domaines scientifiques. On remarque, par le nombre important de thèses dirigées en une seule année, que nous serions en présence de valeurs aberrantes pour les directeurs les plus prolifiques. L'indication des domaines scientifiques permettent d'évaluer si le nom des directeurs peut représenter une même personne, comme dans l'exemple de Cécile Martin (Tableau 5).

Directeur	Discipine	Etablissement	Date de soutenance	Thèse
Scherrmann Jean-Michel	Pharmacie	Paris 5	1994-01-01	39
Bertucat Michel	Pharmacie	Bordeaux 2	1995-01-01	28
Scherrmann Jean-Michel	Pharmacie	Paris 5	1993-01-01	27
Scherrmann Jean-Michel	Pharmacie	Paris 5	1995-01-01	27
Bertucat Michel	Pharmacie	Bordeaux 2	1993-01-01	26
...	...	...	...	...
Fabre Genevieve	Etudes latino-americaines	Paris 7	1998-01-01	1
Fabre Genevieve	Etudes anglophones	Paris 7	1995-01-01	1
Fabre Genevieve	Etudes anglophones	Paris 7	1991-01-01	1
Fabre Genevieve	Etudes anglaises	Paris 7	1993-01-01	1
van Isacker Pieter	Physique	Caen	1999-01-01	1

TABLEAU 9 – Nombre de thèses dirigées par directeur par date entre 1984-2018

5.2 Le cas d'un directeur

En prenant pour exemple le directeur Michel Bertucat, les résultats du Tableau 10 confirmeraient que ce dernier soit la même personne en remarquant que les domaines scientifiques et les établissements de soutenance sont récurrents. Cette observation confirmerait que le nombre de thèses indiqué en début de chaque année est aberrant.

Discipine	Etablissement	Date de soutenance	Thèse
Pharmacie	Bordeaux 2	1995-01-01	28
Pharmacie	Bordeaux 2	1993-01-01	26
Pharmacie	Bordeaux 2	1992-01-01	25
Pharmacie	Bordeaux 2	1994-01-01	23
Pharmacie	Bordeaux 2	1997-01-01	22
Pharmacie	Bordeaux 2	1996-01-01	16
Pharmacie	Bordeaux 2	1991-01-01	14
Pharmacie	Bordeaux 2	1998-01-01	12
Pharmacie	Bordeaux 2	2000-01-01	3
Sciences pharmaceutiques	Bordeaux 2	1992-01-01	1
Sciences pharmaceutiques	Bordeaux 2	1990-01-01	1
Pharmacie	Bordeaux 2	1999-01-01	1
Pharmacie	Bordeaux 2	1988-01-01	1

TABLEAU 10 – Nombre de thèses dirigées par Michel Bertucat entre 1984-2018

6. Résultats préliminaires

6.1 Evolution des langues d'écriture

En étudiant la proportion des langues d'écriture par année ces deux dernières décennies, nous obtenons le graphique de la Figure 10. Il indique clairement une croissance pratiquement linéaire de la langue anglaise, avec une baisse proportionnelle du français puisque l'anglais croît de 24,11 % tandis que le français décroît de 23,65 %. Mais le français reste la langue préférée d'écriture avec plus de 64 % de la production des thèses contre approximativement 28 % pour la seconde. Par contre les autres langues restent assez confidentielles avec une proportion quasi constante de 1.15 % en 2005 et 1.28 % en 2018.

Année	Langue	Distribution	Total	Proportion(%)
2005	Français	9352	10521	88.89
2018	Français	7807	12132	64.35
2005	Anglais	437	10521	4.15
2018	Anglais	3429	12132	28.26
2005	Bilingue	611	10521	5.81
2018	Bilingue	741	12132	6.11
2005	Autres	121	10521	1.15
2018	Autres	155	12132	1.28

TABLEAU 11 – Distribution et proportion des langues d'écriture en début et fin de période

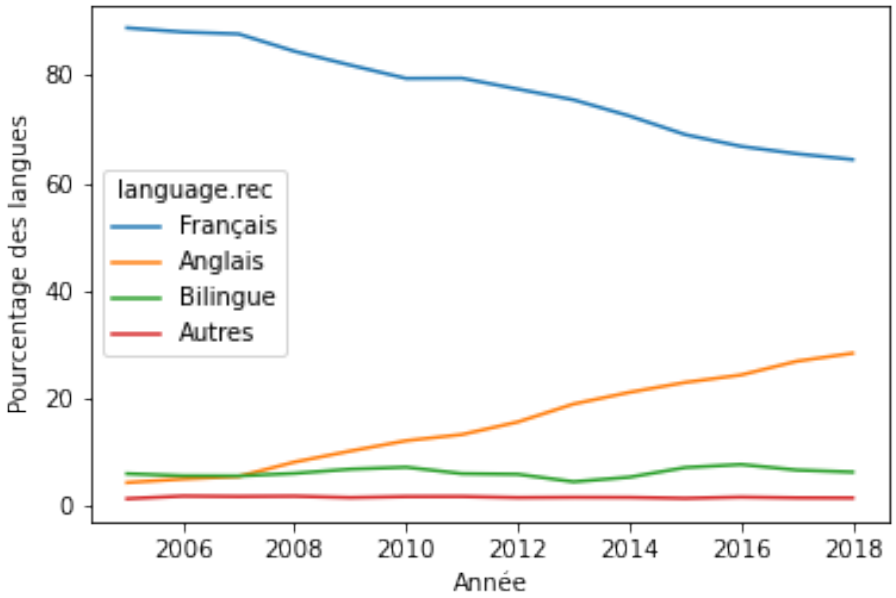

FIGURE 10 – Proportion des langues d'écriture entre 2005-2018

6.2 Proportion des langues

Le Figure 11, permet de mieux représenter la proportion d'une langue par rapport aux autres au fil du temps puisque la totalité des langues affichées représente 100 % des données analysées.

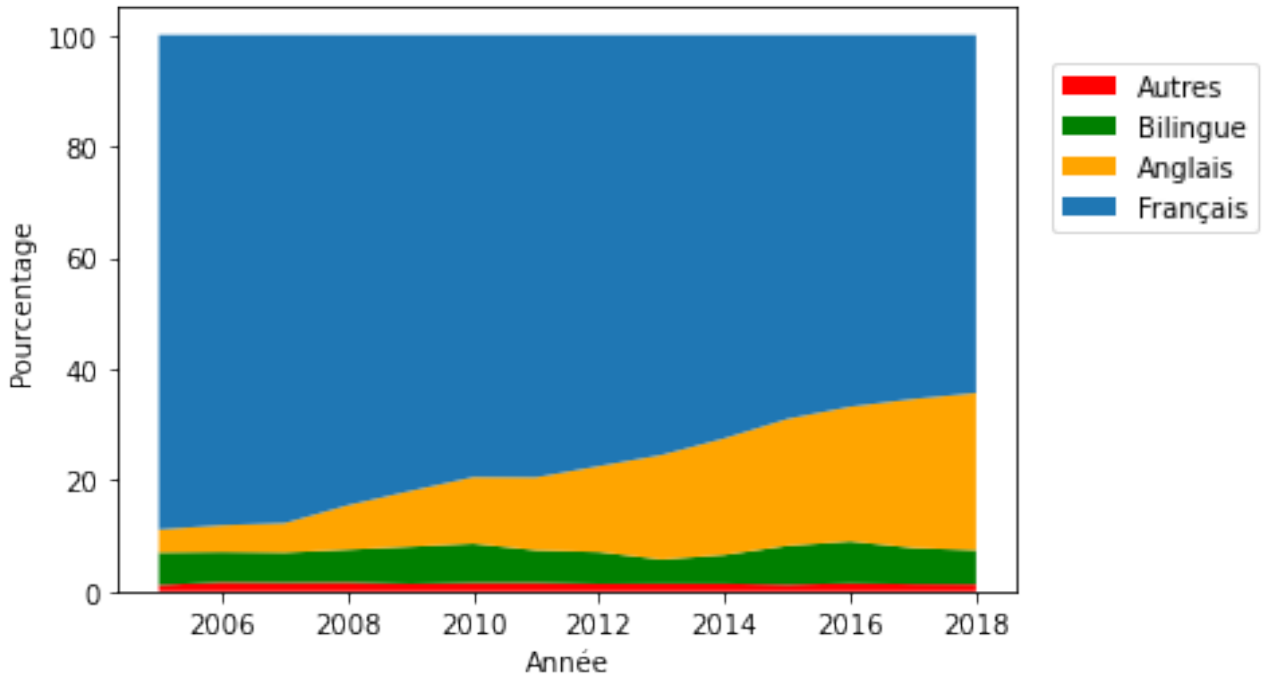

FIGURE 11 – Part des langues d'écriture entre 2005-2018