DIPLOME UNIVERSITAIRE DATA ANALYST

UE PRETRAITEMENT ET MANIPULATION DE DONNEES

Devoir intermédiaire n° 2

MOISE EHIMIGAYE SENGHOR

Année Universitaire : 2024-2025

Table des matières

Liste des tableaux

Table des figures

I Présentation des données

Le jeu de données soumis à notre attention est relatif aux thèses en France . A la base le dataset à été extrait d' un moteur de recherche de thèses(`https://www.theses.fr`). Notre travail consistera à l'étude exploratoire de l' extraction qui a été faite et l' évaluation de la qualité des données. Pour se faire nous utiliserons le langage de Python et la plate-forme Anaconda.

Le jeu de données décrit 448047 en fonction de 23 variables , c' est donc un volume assez important d' informations . Ces 23 variables sont décrites dans le Table1 Le premier fait saillant est que la majeure partie des variables sont des variables textuelles, seules deux sont numériques (1 décimales (float) et une entier naturel (int64)). La variable "Unnamed :0" est l' une d' elle mais après réflexion on peut la considérer comme une indexation numérique des thèses .En effet on voit que c' est une suite de nombres entiers allant de 0 à 448047 . La variable "Year" est la deuxième variable numérique comme son nom l' indique elle semble plus correspondre à une variable temporelle qu'à une variable numérique.

On peut scinder le dataset en plusieurs axes de lecture l' auteur de la thèse, la thèse en elle même, l' encadreur, l' établissement de l' auteur et enfin l'implémentation dans le moteur de recherche. Les différentes variables sont rattachables aux thématiques précitées.

Un autre axe de lecture nous permets de retenir une autre structuration de l' analyse . En l'espèce, nous mettrons en exergue la priorisation d' une description exhaustive des thèses et des parties prenantes ('auteur', 'Etablissement de soutenance', 'directeur ...) d' une part , et de l' autre la temporalité de l'analyse. Cette dernière orientation est matérialisée par l' existence de plusieurs variables qui convoquent des dates('Year', 'Date de soutenance' , 'Publication dans these.fr'...).

En définitive, nous avons un jeu de données assez large que nous pouvons lire , analyser , et interpréter en fonction de nos objectifs . La réuissite de notre étude est assujettie à la perfection de la qualité de l' input informationnelle dont la première étape est la gestion des données manquantes.

TABLE 1 – Description des variables du jeu de donnée des thèses en France

N°	Nom Variable	Type Variable
1	Unnamed :0	int64
2	Auteur	object
3	Identifiant auteur	object
4	Titre	object
5	Directeur de thèse	object
6	Directeur de thèse (nom prénom)	object
7	Identifiant directeur	object
8	Établissement de soutenance	object
9	Identifiant établissement	object
10	Discipline	object
11	Statut	object
12	Date de première inscription en doctorat	object
13	Date de soutenance	object
14	Year	float64
15	Langue de la thèse	object
16	Identifiant de la thèse	object
17	Accessible en ligne	object
18	Publication dans theses.fr	object
19	Mise à jour dans theses.fr	object
20	Discipline	object
21	Genre	object
22	établissement_rec	object
23	langue_rec	object

II Données manquantes

Nous allons essayer de jauger la qualité des données qui nous sont soumises en nous focalisant sur l'étude des données manquantes.

Au préalable, nous avons utilisé le package Python Sweetviz afin de générer un rapport statistique des différentes variables du jeu de données(le

rapport est joint en annexe ('these.html'). Un lecture succincte de celui ci nous indique qu'il n' y a pas de doublon et que nous avons cinq variables catégorielles , deux numériques et 16 textuelles . Il existe une disparité du poids des données manquantes qui varie de 0 à 383 716 par variable.

TABLE 2 – Distribution des données manquantes (>1% par variable)

Variable	**Nombre total**	**Pourcentage**
Date de première inscription en doctorat	383716	85.64%
Identifiant auteur	130347	29.09%
Langue_rec	64120	14.31%
Year	57086	12.74%
Date de soutenance	57086	12.74%
Identifiant établissement	17082	3.81%

Dans une logique d' affinage nous avons choisi de considérer qu'un taux inférieur à 1% de données manquantes était acceptable. Ce qui nous a permis d' extraire une distribution illustrée par la Table 2. Les six variables retenues seront la base sur laquelle nous continuerons notre analyse pour plus d' efficacité.
Le table2 susmentionnée nous permets déjà de remarquer que le pourcentage de données manquantes est le même pour les variables "Year" et "Date de soutenance" . Cette égalité induit l' hypothèse d' un forte corrélation entre les deux mais les informations que nous avons jusque là ne nous permettent pas de l' affirmer. Concernant la variable "date de première inscription en doctorat" tenant compte du fait qu' il y a toujours plusieurs années entre cette ci et celle de la soutenance,en y ajoutant le temps de latence entre soutenance et le reporting dans le site dédié, on peut considérer ces faits comme la cause de l' existence des données manquantes . Cette explication n'est cependant pas prouvée de manière scientifique bien que les thèses les plus anciennes répertoriées , qui datent de la fin des années 70, confortent notre hypothèse.

La matrice de nullité de la figure1 nous permets d' aller plus loin dans notre étude . Nous distinguons de fortes similitudes entre les colonnes 'Year' et 'Date de soutenance' . En faisant une petite comparaison entre les 50 premières valeurs des deux variables il est clair que ce sont les mêmes lignes de données qui sont manquantes . De plus, on peut voir que l' année correspond à l' année de la date soutenance on peut donc en déduire que la variable 'Year' correspond à l' année de soutenance . Ce qui expliquerait la quasi

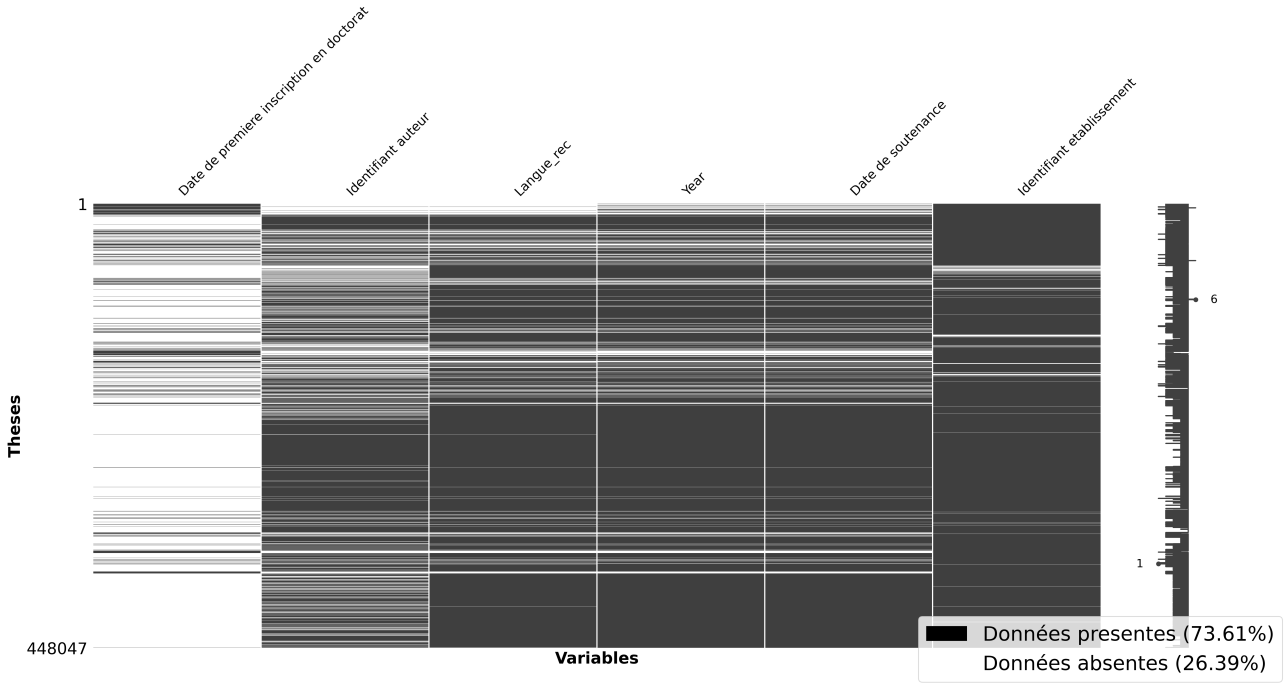

FIGURE 1 – Matrice des données manquantes des variables sélectionnées

égalité entre la 4ém et la 5ém colonne dans la Fig1.

La heatmap de la Figure2 montre une certaine relation négative entre les variables date de soutenance et date de première inscription Ce paradigme peut aisément être compris, il est logique que pour les thèses soutenues nous ayons certaines données qui soient exhaustives telles que la langue , l' année de soutenance , la date de soutenance , et l' établissements de soutenance . En effet la soutenance marque la fin de d' un processus de plusieurs années, il va de soit que plus d' informations sont disponibles . En revanche les thèses en cours sont dans un processus dynamique susceptible de changer (changement de sujet , d' encadreur ou d' établissement ou même abandon...). Il est donc plus efficient d' attendre que la soutenance soit actée pour faire la mise à jour des informations . Le fait que nous ayons 85% de données manquantes pour les dates de première inscription des thèses soutenues peut aussi se comprendre si les personnes qui ont fait le reporting ont considéré que l' information était peu pertinente . On peut aussi penser cela concerne des thèses très anciennes ce qui rend l'accès au données à mettre à jour difficile (une étude plus approfondie dans ce sens serait intéressante).

Il aurait été aussi pertinent de voir comment le pourcentage de données manquantes a évolué dans le temps étant donné que la mise en oeuvre effec-

tive de la gouvernance des données dans le domaine est assez récente (alors que nous étudions les thèses sur plus de 40 ans). La seule variable où les pourcentages sont relativement proches quelque soit le statut des thèses est la variable 'identifiant d' établissements', ce qui semble couler de source car tout les doctorants doivent être étudiant d' une université pour pouvoir faire une thèse (l' information est assez facile à obtenir).

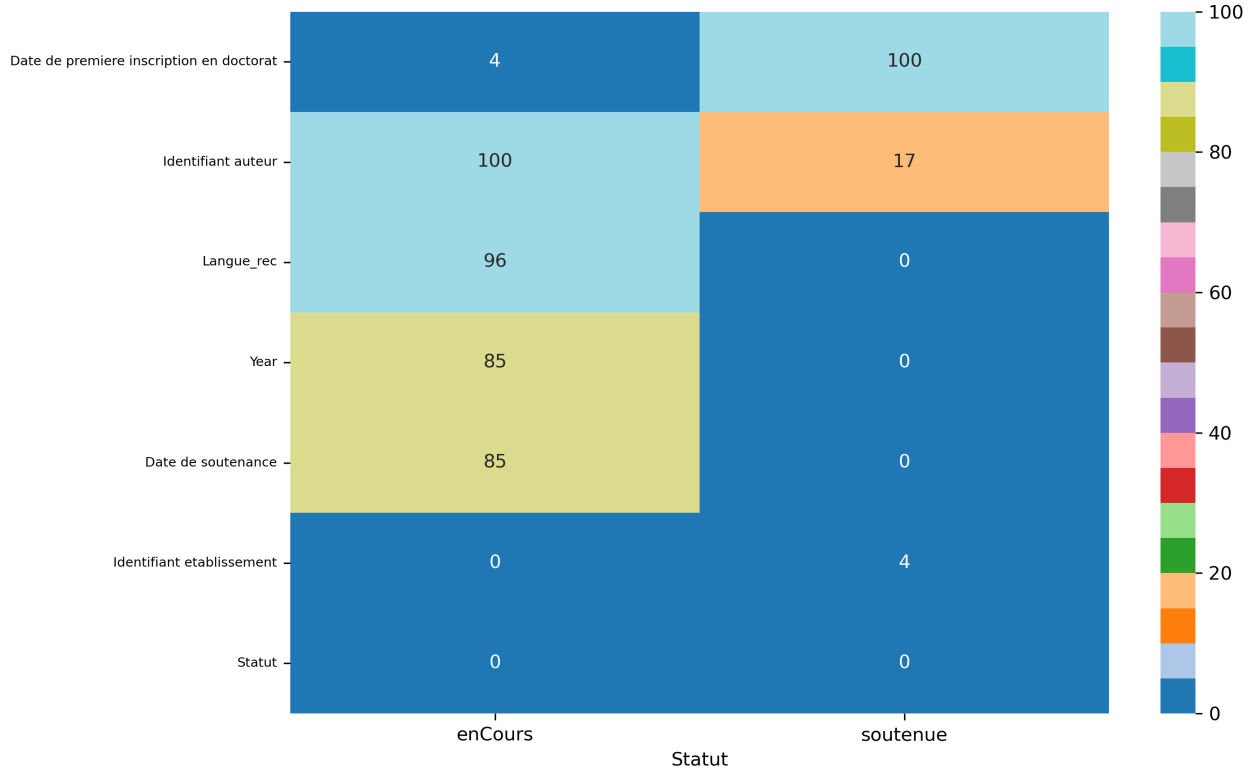

FIGURE 2 – Carte thermique du pourcentage de données manquantes en fonction du statut de la thèse

Il existerait donc un certain rapport négatif dans la manière d' intégrer les informations des thèses soutenues et des thèses en cours d' écriture . Nous avons pu émettre quelques suggestions pour l' expliquer mais nous ne pouvons affirmer ou infirmer nos hypothèses

Constant dans notre logique d' assainissement et de compréhension des données , nous allons maintenant nous concentrer sur les possibles anomalies rencontrées dans notre base de données.

III Principaux problèmes détectés

Nous allons maintenant concentrer notre étude sur les dates de soutenances des thèses composant le dataset soumis à notre attention.

Il nous est demandé de nous limiter à la période de 1984 à 2018 . On peut supposer que celle-ci a été choisi pour plusieurs raisons . Tout d' abord , (la loi Savary) (`https://fr.wikipedia.org/wiki/Loi_Savary`) a modifié la nomenclature de l' enseignement supérieur en général et celle de la recherche doctorale en particulier. Antérieurement, le doctorat pouvait être un doctorat d' état plus long et plus reconnu au niveau académique ou un doctorat du troisième cycle. Il faut aussi préciser que dans la base de donnée , les années antérieures à 1984 présentes sont les années 76,79,80,82 et pour chacune d'elle une seule thèse par année à été prise en compte (soit 4 thèses pour les années inférieures à 84 sont répertoriées). Le nombre de thèses avant 1984 est donc résiduel.

Les périodes postérieures à 2018 concernent les années 2019, 2020, 2071, 2072, 2073 les trois dernières années peuvent être considérées comme des outliers . Les discussions sur une nouvelle refonte du parcours doctoral se sont tenues en 2019. La fin de cette année et l' année 2020 ont été le théâtre de la pandémie du Covid-19 . Ces deux conjonctures ont du impacter sur le volume de thèses soutenues. A notre sens , la deuxième supposition nous semble avoir un impact plus significatif sur le volume de thèses soutenues.

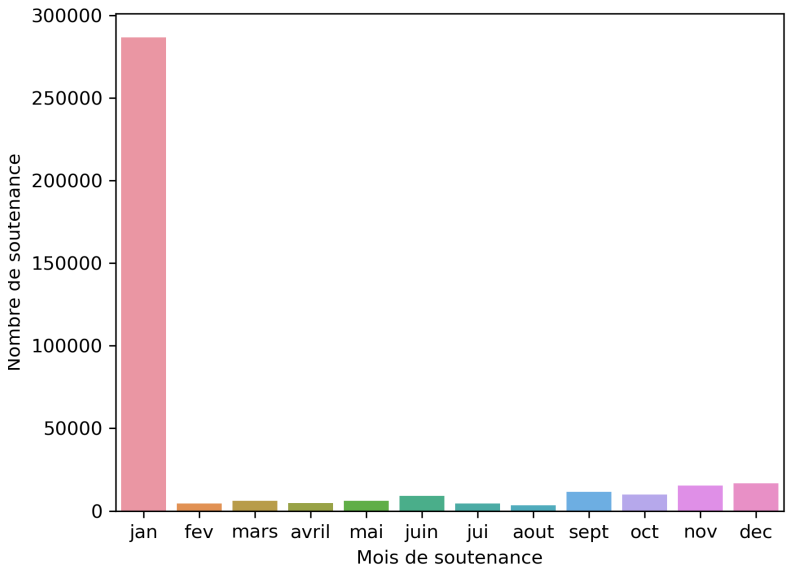

FIGURE 3 – Distribution du nombre de thèses soutenues par mois de 1984 à 2018

La figure 3 ci dessus mets en lumière une très grande disparité entre le nombre de soutenances au mois de janvier et les autres mois . Le pourcentage de thèses soutenues au mois de janvier dans cette période est de 75,6% . Il nous est très difficile de pouvoir donner une raison à ce paradigme nous allons essayer d' approfondir notre raisonnement . Dans cette optique, essayons de disséquer la distribution de la date de soutenance année par année.

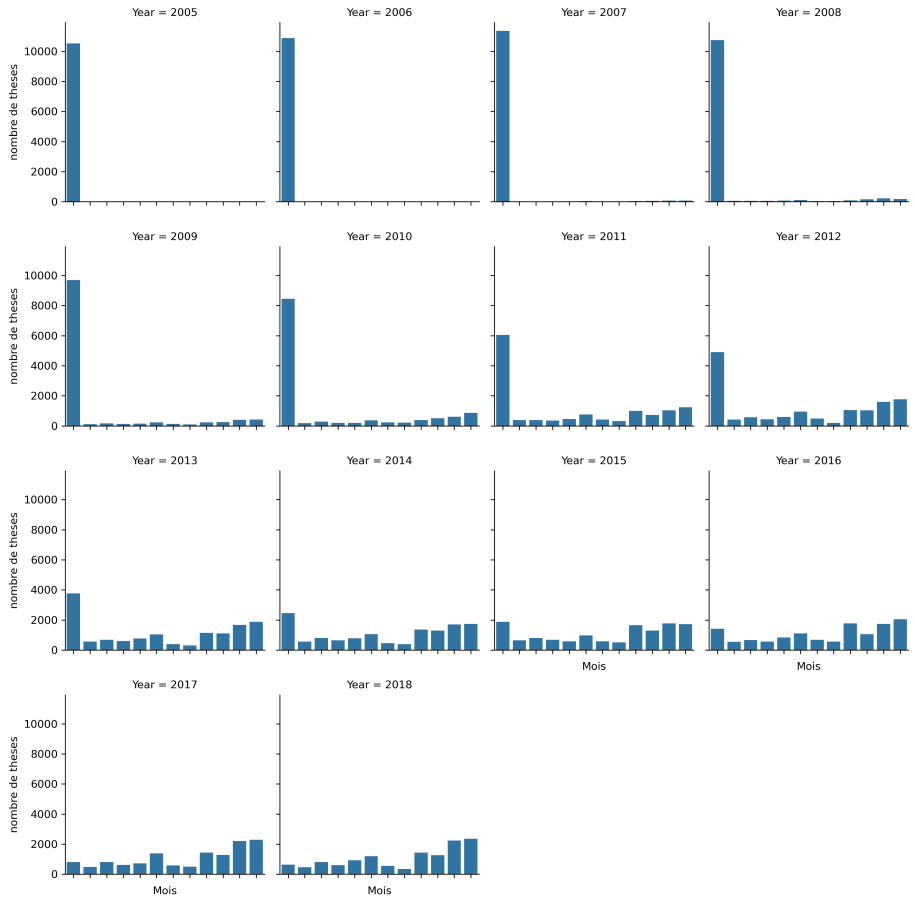

FIGURE 4 – Distribution du nombre de thèses soutenues par mois de chaque année de 2005 à 2018

La figure 4 nous permets de visualiser cette distribution mais cette fois entre 2005 et 2018. On observe une certaine prépondérance de 2005 à 2009 ensuite une deuxième phase où le nombre de thèses soutenues en janvier diminue progressivement au profit des autres mois (jusqu' en 2016) . Enfin une phase de stabilisation de la répartition à partir de 2017 où on en vient progressivement à une certaine homogénéité et renversement en faveur des trois derniers mois .

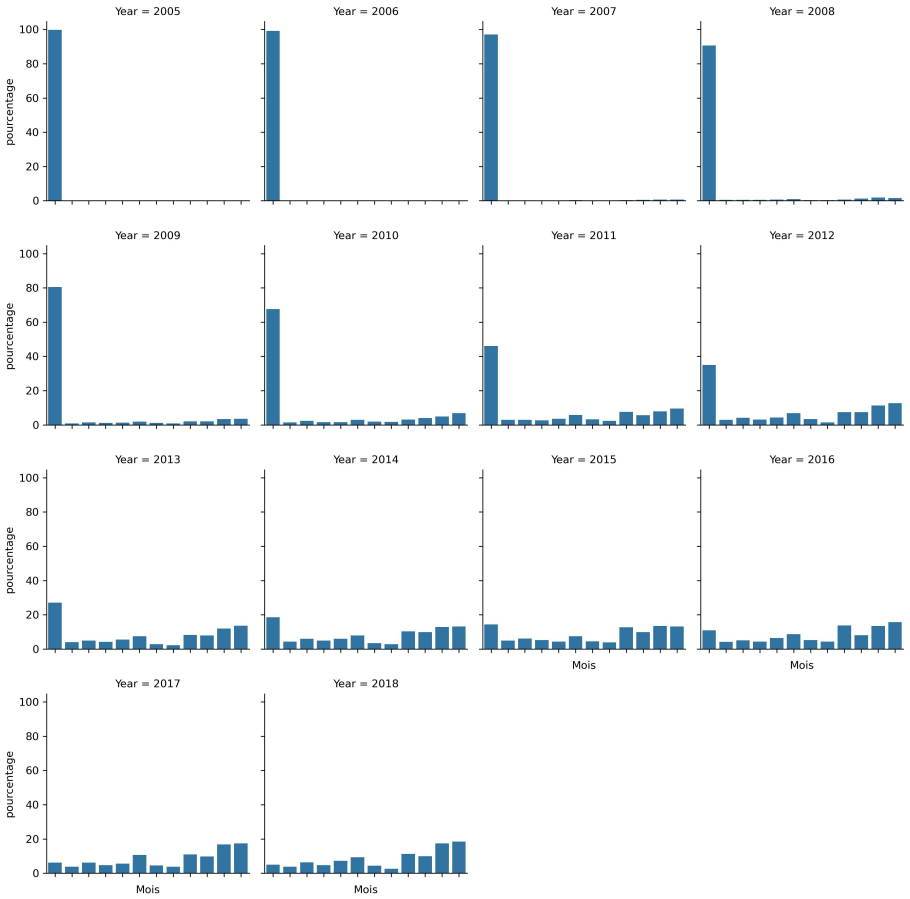

FIGURE 5 – Distribution du pourcentage de thèses soutenues par mois de chaque année de 2005 à 2018

Ces observations sont confortées par la figure5 ci dessous qui reprend la distribution précédente mais cette fois avec des pourcentages . On passe de presque 100% en 2005 de thèses soutenues en janvier à 80% en 2010 , ensuite les baisses s' accentuent significativement chaque année pour arrivée à a peu prés 25% en 2017. A partir de 2018 , le mois de janvier n' est plus le mois de soutenance majoritaire au profit du dernier quart annuel.

Toujours dans le but d' approfondir notre réflexion, nous pouvons aussi revenir à la figure 6 qui nous permet d'étayer nos constats précédents . Sur l' intervalle 2005-2018 , plus de 45% des thèses ont été soutenues en janvier. Pour être plus clair , un peu moins de la moitié des thèses est soutenue sur 1 douzième de l' année . Ce ratio est caractéristique d' une anomalie ou encore d' une explication à trouver.

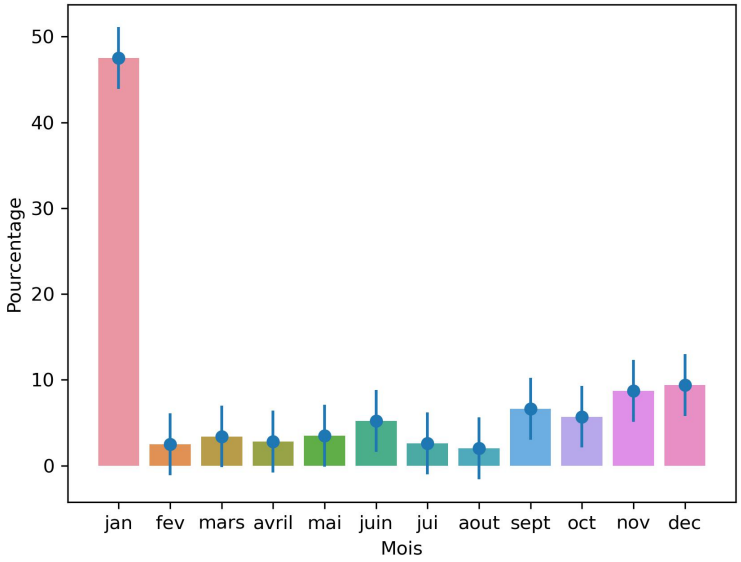

FIGURE 6 – Distribution du pourcentage de thèses soutenues par mois de 2005 à 2018 (écart-type)

Cette distorsion est d' autant plus criarde quand on observe la fig7 , on se rend compte que de 1985 à 2005 toutes les thèses sont soutenues le premier janvier de chaque année . Ce qui voudrait dire que le premier janvier, jour férié quand on connaît la difficulté à rassembler un jury de thèse, que toutes les soutenances au niveau national se seraient tenues. Cela semble très peu plausible . Essayons de voir le poids du premier janvier sur l' intégralité de l' année .
La figure8 représentant la propension de thèses soutenues à l' exception du jour de l' an montre que l' anomalie ne vient pas du mois de janvier mais plutôt de la date du 1er janvier . En effet en éludant cette date, se retrouve avec une répartition mensuelle des soutenances plus homogène (variant de 4% a peu près à 16%) . En plus , en adéquation avec les conclusions que nous avions faites précédemment, on voit que les mois dont les pourcentages sont les plus élevés sont les mois de septembre , d' octobre et de décembre.

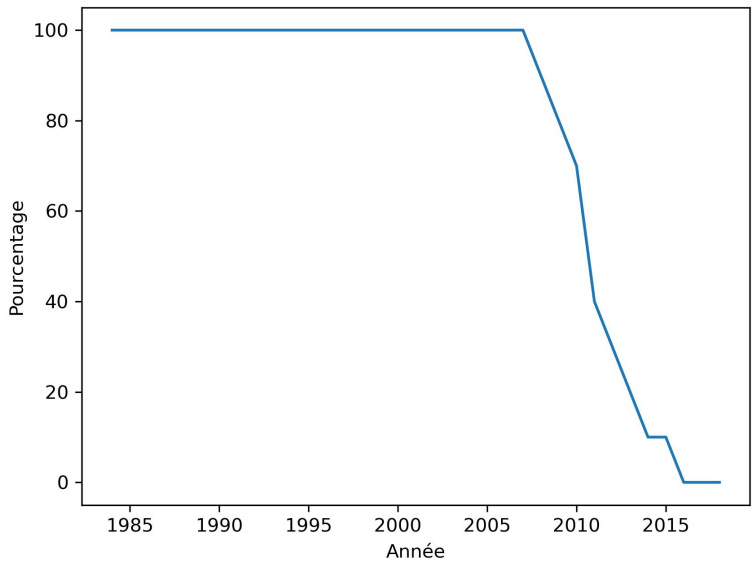

FIGURE 7 – Proportion des thèses soutenues au premier janvier

Nous n' avons pas l' explication scientifique de cette singularité , on va partir du postulat qu'elle relève de l' arbitraire . Tenant compte du fait que notre étude porte sur un intervalle de 23 ans et de la difficulté d' accéder à des données anciennes , nous présumons que lors de lors du reporting sur thèse.fr , il a été convenu que toute les thèses soutenues à des dates inconnues seraient considérées comme soutenues le jour de l' an Notre déduction est hypothétique mais elle nous semble la plus logique bien que difficile à prouver.

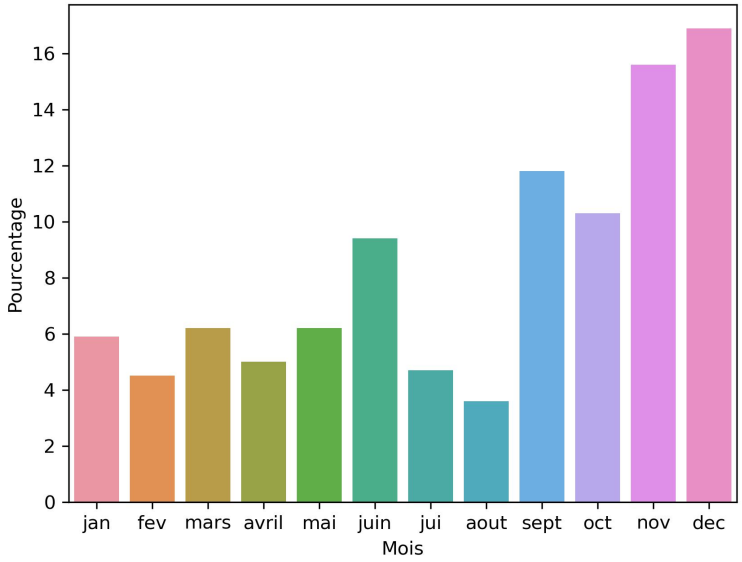

FIGURE 8 – Proportion des thèses soutenues les thèses en dehors du premier janvier de 2004 à 2018

Dans ce type de base de données les cas d' homonymie peuvent être problématiques à titre d' exemple nous allons essayons d' en examiner un . Il s' agit du cas de Cecile Martin.

TABLE 3 – Résumé des informations relatives aux thèses soutenues par un auteur du nom de Cécile Martin

ID Auteur	Etablissement	Année	ID these	Discipline
203208145	SORBONNE PARIS CITE	2017	2017USPCA018	SHS
81323557	Inst. Nl AGR. PARIS-GRIGNON	2000	2000INAP0034	BIO...
179423568	PARIS 9	2014	2014PA090003	ECO-GEST
81323557	COMPIEGNE	2001	2001COMP1380	Sc. De L'INGE
81323557	BORDEAUX 2	1991	1991BOR22005	BIO...
81323557	CLERMONT-FERRAND 2	1994	1994CLF21651	PSYCHO...
182118703	PARIS 11	1989	1989PA112163	Mat , mil, chimie

Le tableau 3 ci-dessus nous récapitulatif le cas précité.

Sur notre jeu de données nous repérerons sept thèses soutenues par un auteur du nom de Cecile Martin. Les auteurs sont tous de sexe féminin . Nous remarquons aussi que quatre d' entre elles ont le même identifiant auteur (81323557) les trois autres en ont des différents (203208145 , 179423568

,182118703). Pour les quatre premières nous avons deux thèses de Biologie, une de sciences de l' ingénieur et une de psychologie . les deux thèses de biologie sont soutenues en 2000 à Paris et en 1991 à Bordeaux . Les deux autres sont soutenues à Compiègne et à Clermont-Ferrand respectivement en 2001 et 1994. Si l' intervalle de 1991 à 2001 donnerait 10 ans pour rédiger 3 théses à la même personne (celle de 91 écrite antérieurement) au vu des spécialités et aussi des sujets qui sont assez variés on ne peut pas affirmer sans doute que le l' identifiant 81323557 correspond à la même personne, il nous est aussi impossible de l' infirmer . Relativement aux trois autres Cécile Martin dont les indentifiants sont discordants , nous avons aussi des distinctions pour les années de soutenance (intervalle de 18 ans), les spécialités et les établissements . Nous sommes plus enclins à affirmer l' homonymie de minimum quatre personnes aux identifiants dissemblables et nous opterons pour la prudence pour les doctorantes ayant le même ID. Même si cette assertion nous semble cohérente , nous sommes obligés d' être relatif sur sa force probante.
En dehors de ces cas spécifiques, il peut aussi arriver que certaines informations soient incompréhensibles ou erronées . Ces données aberrantes seront l' objet de la suite de notre étude.

IV Outliers et résultats anormaux

Les deux participants à l' élaboration d 'une thèse sont le doctorant et son encadreur . Après nous être appesanti sur les thèses et leurs auteurs, nous allons maintenant nous concentrer sur les directeurs de thèse. Au préalable nous commencerons par essayer d' avoir une idée sur le nombre de thèses dirigées par chaque encadreur (au vu du nombre pléthorique nous choisissons les 10 directeurs les plus prolifiques table 4).

TABLE 4 – Récapitulatif des 10 directeurs de thèse ayant encadré le plus de thèses de 1984 à 2018

Directeur de These	**Nombre de thèses encadrées**
Directeur de these inconnu	711
Jean-Michel Scherrmann	208
Francois-Paul Blanc	201
Pierre Brunel	195
Michel Bertucat	173
Guy Pujolle	172
Bernard Teyssie	138
Henry de Lumley	132
Jean-Claude Chaumeil	131
Bruno Foucart	130

Le directeur le plus productif est le directeur de thèses est nommé inconnu . Il couvre à lui tout seul 711 mentorats en plus de 20 ans soit à peu près 30 doctorats par an . Il est hautement improbable qu'une seule personne aie pu abattre tout ce travail. D' ailleurs le second de la liste est à 208 thèses encadrées . Par ailleurs , il est établi 90% des encadreurs ont accompagnées 7 thèses ou moins , on peut donc considérer le directeur de thèse inconnu soit comme une donnée aberrante ou encore comme une agrégation de plusieurs personnes dont les données ne sont pas disponibles .

Nous allons donc allons parler du cas du second à savoir Mr. Jean-Michel Scheerman. Ce dernier à mentoré plus de 200 thèses en l' espace de 20 ans(table 5). Il dispose de deux identifiants (26404788 et 27787087) respectivement liées à l' université Paris 5 et Paris 6. Les périodes couvertes sont de 1989 à 2012 pour le premier identifiant et de 1993 à 2000 pour le second. On remarque les périodes de Paris 6 sont inclues dans celle Paris et que le nombre de thèses de paris est relativement faibles (5) . Il est donc possible que les deux identifiants représentent la même personne.

TABLE 5 – Thèses dirigées par Jean-Michel Scherrmann

Id directeur	Année soutenance	Etablissement	Nombre de thèses
26404788	1989	Paris 5	11
26404788	1990	Paris 5	13
26404788	1991	Paris 5	11
26404788	1992	Paris 5	22
26404788	1993	Paris 5	27
26404788	1994	Paris 5	39
26404788	1995	Paris 5	1
26404788	1995	Paris 5	26
26404788	1996	Paris 5	1
26404788	1996	Paris 5	23
26404788	1997	Paris 5	6
26404788	1998	Paris 5	1
26404788	1998	Paris 5	1
26404788	1999	Paris 5	3
26404788	2003	Paris 5	1
26404788	2003	Paris 5	1
26404788	2003	Paris 5	1
26404788	2004	Paris 5	1
26404788	2005	Paris 5	1
26404788	2007	Paris 5	1
26404788	2008	Paris 5	2
26404788	2009	Paris 5	1
26404788	2009	Paris 5	1
26404788	2011	Paris 5	1
26404788	2012	Paris 5	1
27787087	1993	Paris 6	1
27787087	1994	Paris 6	1
27787087	1997	Paris 6	1
27787087	1999	Paris 6	1
27787087	2000	Paris 6	1

Nous voyons dans le tableau 6 ci dessus que la plupart des thèses concernées ont pour spécialité la pharmacie. Il est aussi notable que les thèses restantes (un nombre résiduel) ont toutes pour domaine des sous spécialités de la pharmacie (pharmacie clinique, pharmacocinétique , radiopharmacie...) ou encore des disciplines incluses dans le domaine médical (ou connexes). Cette deuxième grille de lecture renforce notre hypothèse . Nous aurions un professeur Mr Scherrman qui serait Professeur de pharmacie , il ayraitdirigé des thèses de doctorat à Paris 5 et Paris 6 de 1989 à 2012 aussi bien en pharmacie que de dans des branches annexes. Cette assertion semble judicieuse

mais il est difficile d' être catégorique.

TABLE 6 – Nombre total de thèses encadrées par discipline(Jean-Michel Scherrmann)

Discipline	**Total thèses**
Médecine	2
Pharmacie	186
Pharmacie clinique	1
Pharmacie. Pharmacie clinique et pharmacocinétique clinique	1
Pharmacie. Pharmacocinétique	2
Pharmacie. Toxicologie	1
Pharmacocinétique	2
Pharmacocinétique, Radiopharmacie	1
Pharmacologie cellulaire et moléculaire	1
Sciences biologiques et fondamentales appliquées. Psychologie	1
Sciences médicales	4

Une recherche sur la personne du Professeur Schermann a confirmé notre point de vue (`https://u-paris.fr/deces-du-pr-jean-michel-scherrmann/`, `https://www.inserm.fr/portrait/histoire/jean-michel-scherrmann/`, `https://fr.wikipedia.org/wiki/Facult%C3%A9_de_pharmacie_de_Paris`...) . Les renseignements collectés prouvent l' existence d' un Professeur Schermann émérite de Pharmacie qui a eu à être le doyen de la chaire de Pharmacie à l' université de Paris 5 et qui a aussi eu à intervenir à Paris6. La période d' activité coïncide avec notre période de référence . Nous pouvons considérer que les deux identifiants correspondent à une même personne .

V Résultats préliminaires

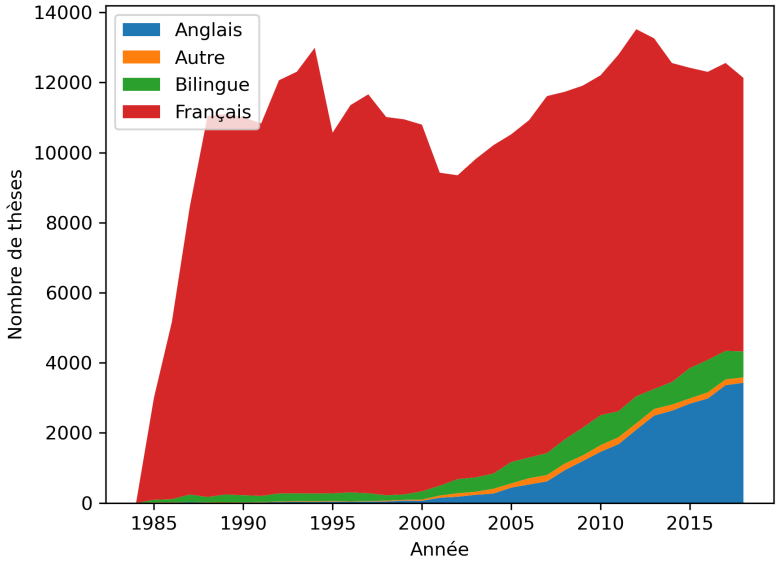

FIGURE 9 – Évolution de la langue d' écriture des thèses en France de 1984 à 2018

Il nous reste à nous intéresser aux langues d' écritures des thèses objet de notre étude . La table 9 illustre cette thématique . Nous notons une prédominance certaine du français . Mais relevons des mutations qui se dessinent au fil des années . D' abord une introduction timide des thèses bilingues (francais–Anglais et Anglais francais) et des thèses en Anglais (les thèses en autres langues restes subsidiaires) . Cette tendance va crescendo et commence à devenir vraiment prégnante à partir de la fin des années 90. Le mouvement s' accélère à partir de 2015, les thèses autres que le français deviennent beaucoup plus présentes avec une primeur pour les doctorats en anglais. L' augmentation des thèses en anglais et bilingues se comprend avec la globalisation de la recherche scientifique et la place des pays anglophones en la matière spécialement dans le domaine scientifique. Le but de la démarche étant d' offrir une meilleure visibilité à l' international. Le bilinguisme est plus observé dans les sciences sociales cependant le paradigme reste le même. Il serait aussi opportun de rappeler aussi que la co-tutelle , mécanisme très usité dans la recherche doctorale moderne, a un impact sur les langues d' écritures est aussi à qu' il serait judicieux de quantifier. L' arrivée des autres

langues d' écriture s' explique aussi par les mutations géopolitiques internationales , par exemple le cas de la chine devenu une superpuissance mondiale ou le Brésil qui sont devenus des pays attractifs en matière de recherche . Il faut par ailleurs considérer la forte immigration ciblée (universitaires et chercheurs) qui permets aujourd'hui à beaucoup de personnes d' horizons divers de venir étudier faire de la recherche en France.

Afin d' étayer nos propos nous convoquerons une citation suivante "En 1987, moins de 3 % des thèses de doctorat étaient rédigées dans une autre langue que le français. Cette situation est restée inchangée jusqu'à l'an 2000, six ans après la promulgation de la loi Toubon, année à partir de laquelle la situation a évolué rapidement. Cette évolution s'est poursuivie après l'adoption de la Loi Fioraso. En 15 ans, le nombre de thèses rédigées en anglais a augmenté régulièrement pour atteindre un tiers des thèses en 2015. Les thèses rédigées dans d'autres langues que le français ou l'anglais restent rares (moins de 3 %), et il s'agit systématiquement de co-tutelles internationales de thèse.

Sur la même période, les effectifs de docteurs étrangers ont fortement augmenté, passant de 1 745 en 2000 à 4 815 en 2010 (+175 %), tandis que celui des docteurs français a varié de 6 163 à 7 269 (+17 %) (ENS Paris-Saclay, 2023)." (ENS Paris-Saclay. (2023). Le français perd-il du terrain dans les thèses de doctorat ? *ENS Paris-Saclay.*)