CY Cergy Paris Université
DUDA 2023

Manipulation et prétraitement des données : UE 1

Auteur :
Francisco Martin-Gomez

Référent :
Pr. Mathieu CISEL

23 août 2023

Résumé

L'analyse des thèses référencées sur la plateforme officielle theses.fr pour la période s'étendant de 1984 à 2018 montre initialement une tendance singulière autour du 1er janvier : les thèses semblent majoritairement être soutenues ce jour précis, ce qui est matériellement impossible. Cependant, un retraitement préliminaire des données permet de contester cette observation et de dévoiler d'autres problèmes significatifs. Outre les incohérences dans les dates de soutenance dues à un artefact technique, le référencement des encadrants (directeurs et co-directeurs) présente aussi de nombreuses anomalies. Cela met en évidence un problème d'homonymie mal géré par les responsables de la saisie des informations dans la base de données de theses.fr. À l'issue de ces corrections préliminaires, il a été possible de constater l'évolution de la répartition des langues utilisées pour la rédaction des thèses en France. Si le français conserve sa position prédominante, son hégémonie est désormais mise à l'épreuve par l'anglais.

Table des matières

Table des figures

Liste des tableaux

Introduction

Cette étude s'est focalisée sur les soutenances de thèses entre 1984 et 2020, telles que répertoriées dans la base de données de theses.fr. À partir d'une méthode de collecte de données (scrapping) selon Cisel et al. (2020), qui a abouti à la création d'un jeu de données initial (PhDž.csv), nous avons entrepris des opérations de nettoyage des données pour les 18 variables disponibles. Des actions ont été entreprises pour néttoyer le jeu de données et traiter les valeurs manquantes, permettant la résolution de divers problèmes dans les données, tels que la répartition mensuelle des soutenances sur la période de référence, ainsi que des incohérences liées aux directeurs de thèse (présence d'outliers). Ces interventions ont permi l'obtention de résultats préliminaires concernant l'évolution des langues employées pour la rédaction de ces thèses..

1 Présentation des données

— *Le jeu de données PhD_ v2 (.csv) est disponible en ligne à l'adresse suivante : lien vers les données (drive.google)*
— *Le code et les graphiques sont disponibles en ligne à l'adresse suivante : lien vers le dépôt github.com*

Le Tableau 1 présente l'ensemble des variables et des observations du jeu de données PhD_v2 (.csv), issu d'un travail de scrapping à partir du site www.theses.fr (Cisel *et al.*, 2015). Ce jeu de données est composé de 18 variables (15 de type *charactère*, 2 de type *date*, 1 de type *facteur* et 1 de type *numérique*) avec 447644 observations (ici chaque ligne/observation correspond à une thèse répertoriée dans la base de données du site). Le jeu de données reflète le référencement officiel des thèses soutenues ou en cours dans les universités françaises, entre 1984 et 2018, indiquant les noms des auteurs, des encadrants (directrices et directeurs de thèses), les identifiants des universités de rattachement, la langue de rédaction ou encore la discipline scientifique de la thèse.

Table 1 – Vue synthétique du jeux de donénes PhD_v2 (nom de la variable, NA, type)

Variable	Valeurs manquantes	Type
Auteur	0	character
Identifiant auteur	129989	character
Titre	13	character
Directeur de these	17	character
Directeur de these (nom prenom)	17	character
Identifiant directeur	49172	character
Etablissement de soutenance	4	character
Identifiant etablissement	17085	character
Discipline	5	factor
Statut	0	character
Date de premiere inscription en doctorat	383995	date
Date de soutenance	221173	date
Year	56746	integer
Langue de la these	63765	character
Identifiant de la these	0	character
Accessible en ligne	0	character
Publication dans theses.fr	0	character
Mise a jour dans theses.fr	177	character

En plus de l'identification et la complétion des données manquantes exposées plus loin, divers traitements ont été simultanément entrepris (recodage de certaines variables comme les dates et les variables catégorielles, considérées comme des chaînes de caractères), le retrait des doublons ou le renommage des variables et des niveaux de la variable catégorielle afin de les rendre plus lisibles (pour plus de détails sur ces opérations de préparation et de nettoyage, voir le code R de notre notebook PhD.Rmd dans notre github).

1.1 Analyse des données manquantes

1.1.1 Répartition des données manquante

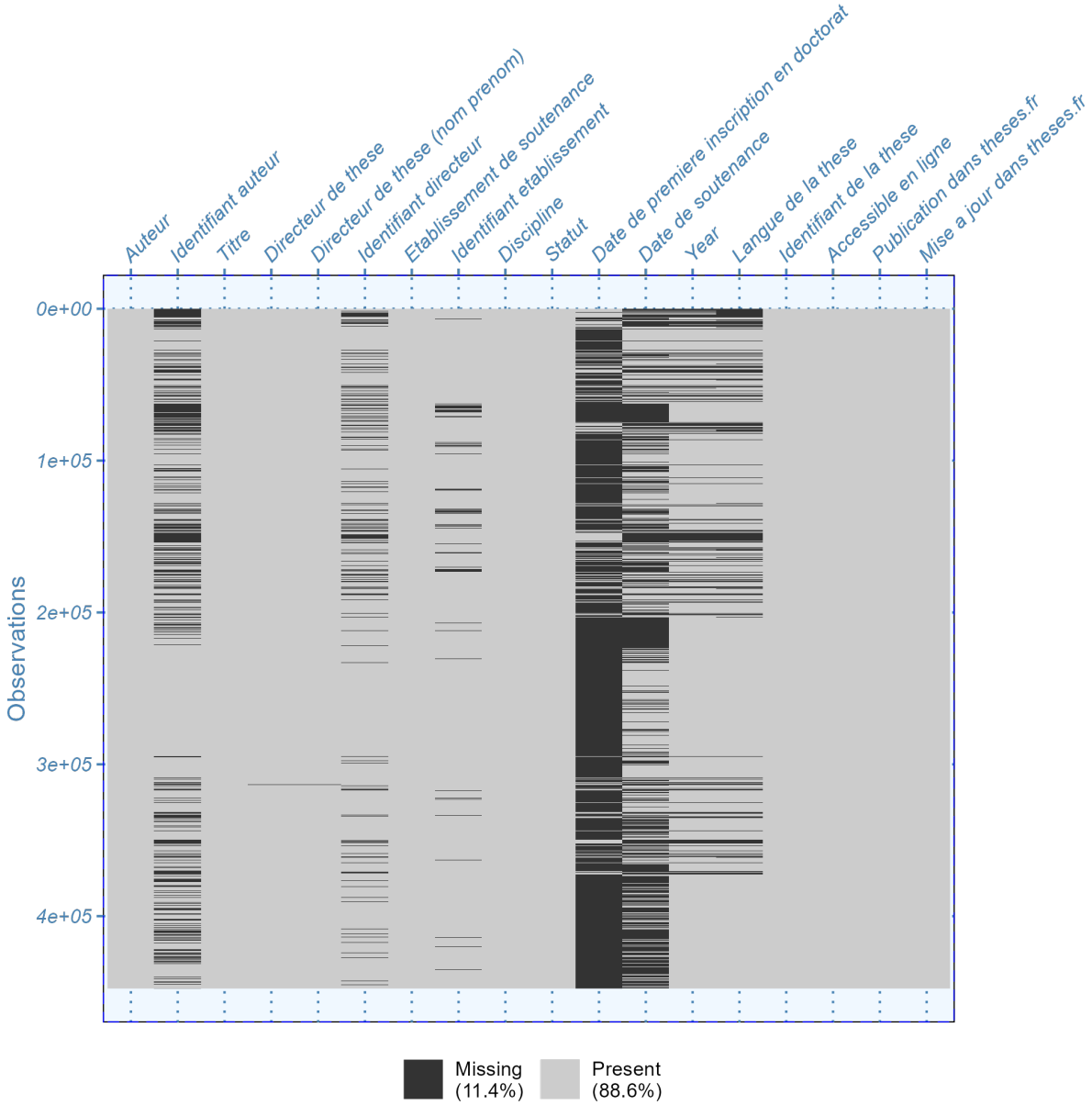


FIGURE 1 – Répartition des données manquantes à l'intérieur du jeu de données PhD_v2

Le Graphique 1 montre la répartition par variable des données manquantes dans le jeu de données telles qu'elles apparaissent après retraitement des valeurs vides (mais non affectées au NA) et des erreurs de typographie. Certaines variables, comme la variable Auteur, ayant très peu de valeurs manquantes, une procédure d'enquête humaine a été réalisée à partir de la base de données thèses.fr (détail dans le Notebook) et a permis le recollement des valeurs manquantes. Cependant, comme le montre l'illustre le Tableau 1, cette technique n'est hélas pas généralisable à toutes les variables, considérant le grand nombre de valeurs manquante sur lesquelles enquêtées. L'analyse de la répartition des valeurs manquante montrent que 11, 4% des données sont manquantes. Les valeurs manquantes se répartissent entre les variables : *Identifiant auteur*, *Identifiant directeur*, *Identifiants établissements*, *Year*, *Langue de la thèse*, *Date de première inscription en doctorat* et *Date de soutenance*, les deux dernières variables étant les plus touchées par le phénomène (voir Tableau 1 pour le détail).

1.1.2 Régularité/pattern dans les données manquantes

Pour approfondir notre analyse des données manquantes, nous nous sommes interrogés sur la possibilité de pattern ou de régularité dans le caractère manquant des données, pouvant alors indiquer des relations plus spécifiques entre variables. A cette fin, nous avons alors construit un corrélogramme sous la forme d'une heatmap (Graphique 2) sur lesprincipales variables pr"sentant des valeurs manquantes. Pour rappel, chaque cellule du corrélogramme affiche le coefficient de corrélation entre deux variables. Les valeurs possibles pour les coefficients de corrélation varient entre -1 et 1, tels que :

— Un coefficient de corrélation de 1 indique une corrélation positive parfaite, ce qui signifie que les deux variables augmentent ou diminuent ensemble de manière linéaire ;
— Un coefficient de corrélation de -1 indique une corrélation négative parfaite, ce qui signifie que lorsque l'une des variables augmente, l'autre diminue de manière linéaire.

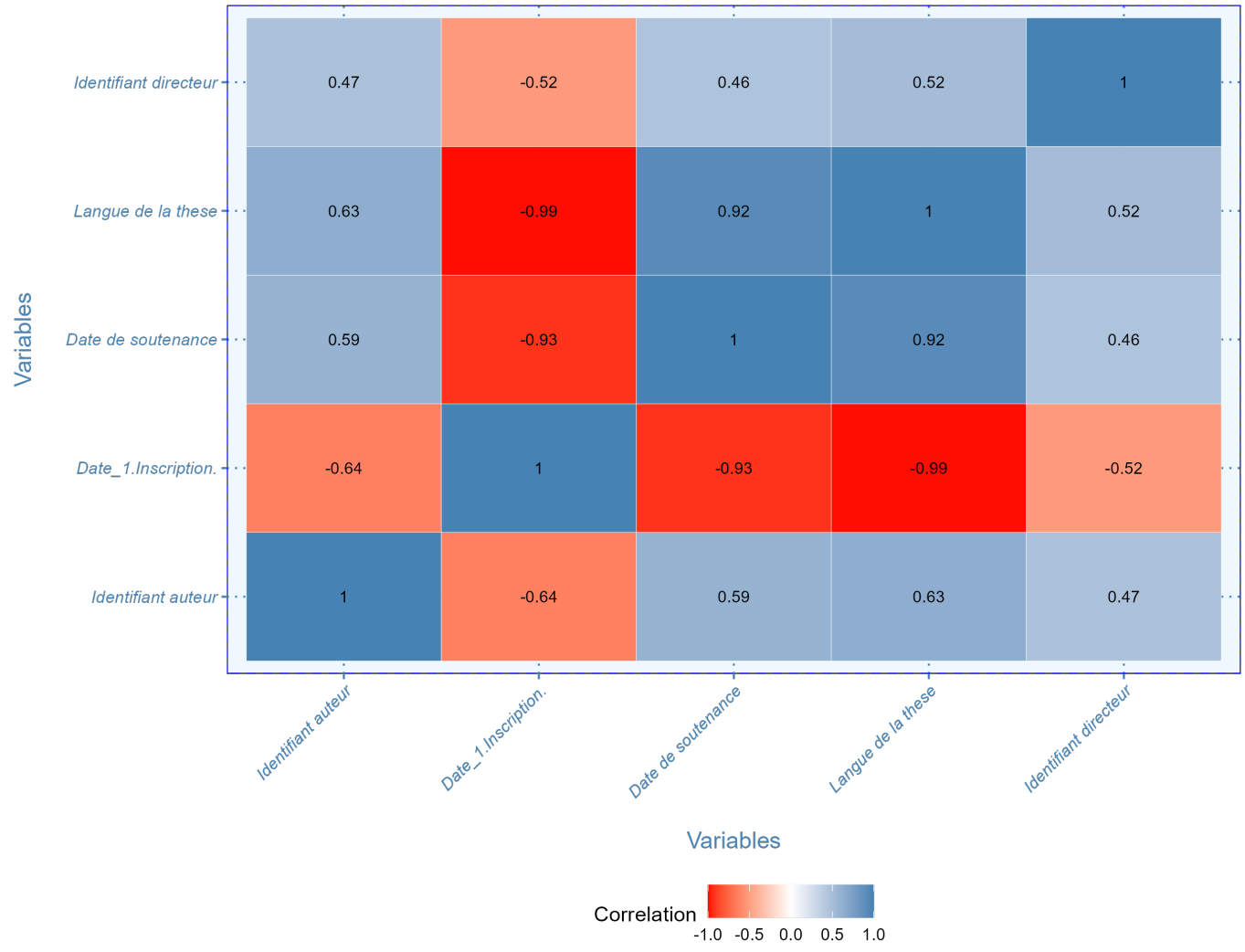


FIGURE 2 – Heatmap des données manquantes (PhD_v2)

Le Graphique 2 souligne l'existence de pattern particuliers dans la répartition des données manquantes, et tout particulièrement de fortes corrélations entre les variable *Date de soutenance de thèse* et la variable *Date de première inscription en doctorat* (corrélation négative de -0.93), entres les variables *Date de première inscription en doctorat* et *Langue de la thèse* (corrélation négative de -0.99) ou encore *Date de soutenance et Langue de thèse* (corrélation positive de 0.92). Ainsi, les données semblent manquantes dans *Date de soutenance de thèse* lorsque les données sont présentes dans *Date de première inscription*, et inversement (corrélation négative). La même relation se dégage entre *Date de première inscription* et *Langue de la thèse* (présence de données manquantes dans Date de première inscription lorsqu'il y a absence de données manquantes dans Langue de thèse). Enfin, les données sont manquantes simultanément

entre *Langue de thèse* et *Date de soutenance* (corrélation positive). L'hypothèse qu'il est possible de faire à ce stade des observations est qu'il existe vraisemblablement un lien entre ces NA en fonction du Statut de la thèse (en cours / soutenue), certaines informations présentes pendant la préparation de la thèse (i.e. avant la soutenance) comme la date de première inscription disparaissant après la soutenance de la thèse. A l'inverse, la Langue de these, qui est une variable descriptive d'un manuscrit de thèse déposé (donc à une thèse soutenue), est présente/absente en fonction de la variable Date de soutenance de thèse, ce qui apparaît cohérent avec notre hypothèse.

Pour confirmer cette hypothèse (sans recourt à un test statistique hors de propos dans ce rapport), nous avons construit un deuxième corrélogramme des données manquantes en fonction du Statut (Figure 3) sur l'ensemble des variables du jeu de données. Comme cette figure le souligne, il y a bien un lien entre le Statut de la thèse et la présence des NA entre les variables.

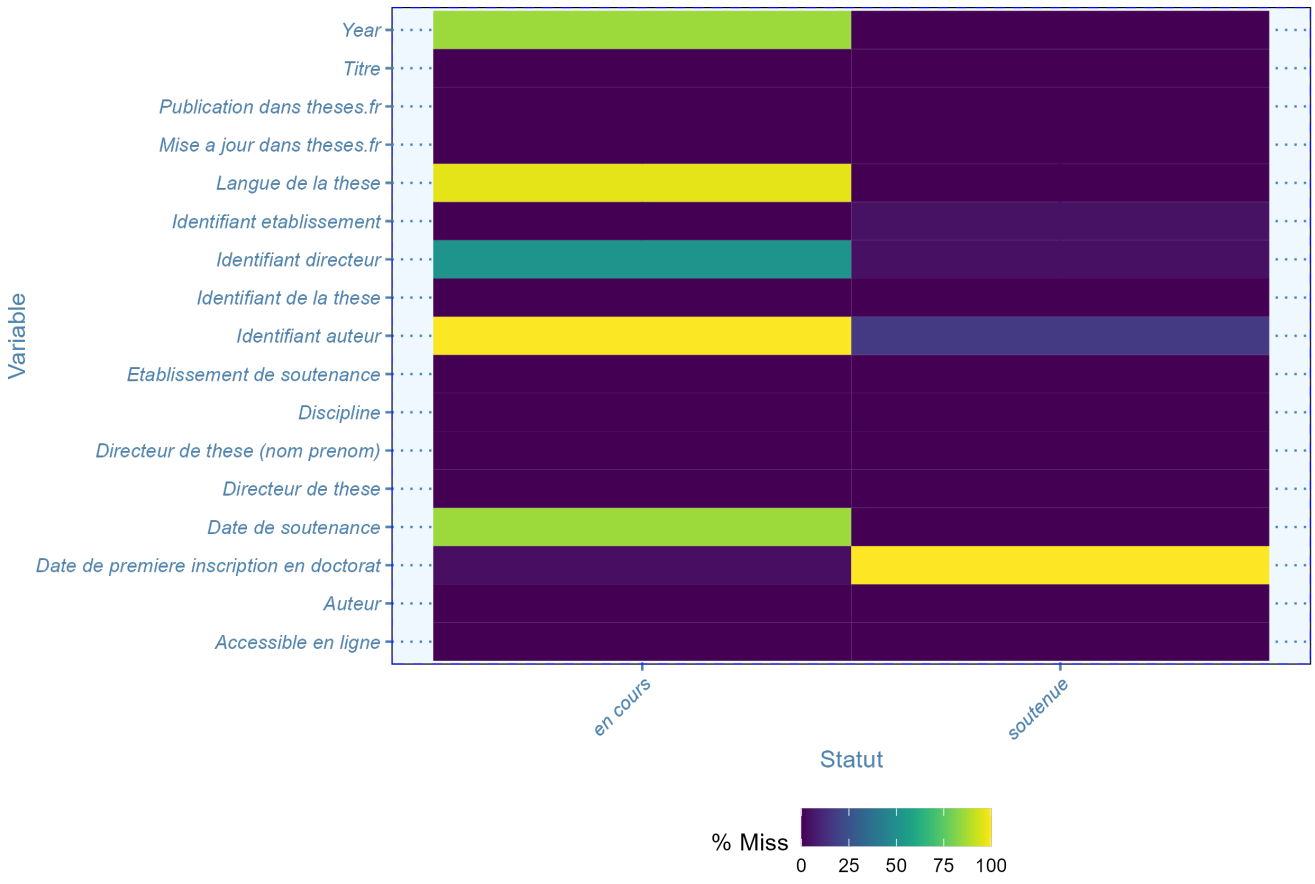


FIGURE 3 – Pourcentage de données manquantes selon le statut de la thèse

2 Détection de problèmes dans le jeu de données

2.1 Les mois de soutenance des thèses

2.1.1 Le problème du mois de janvier

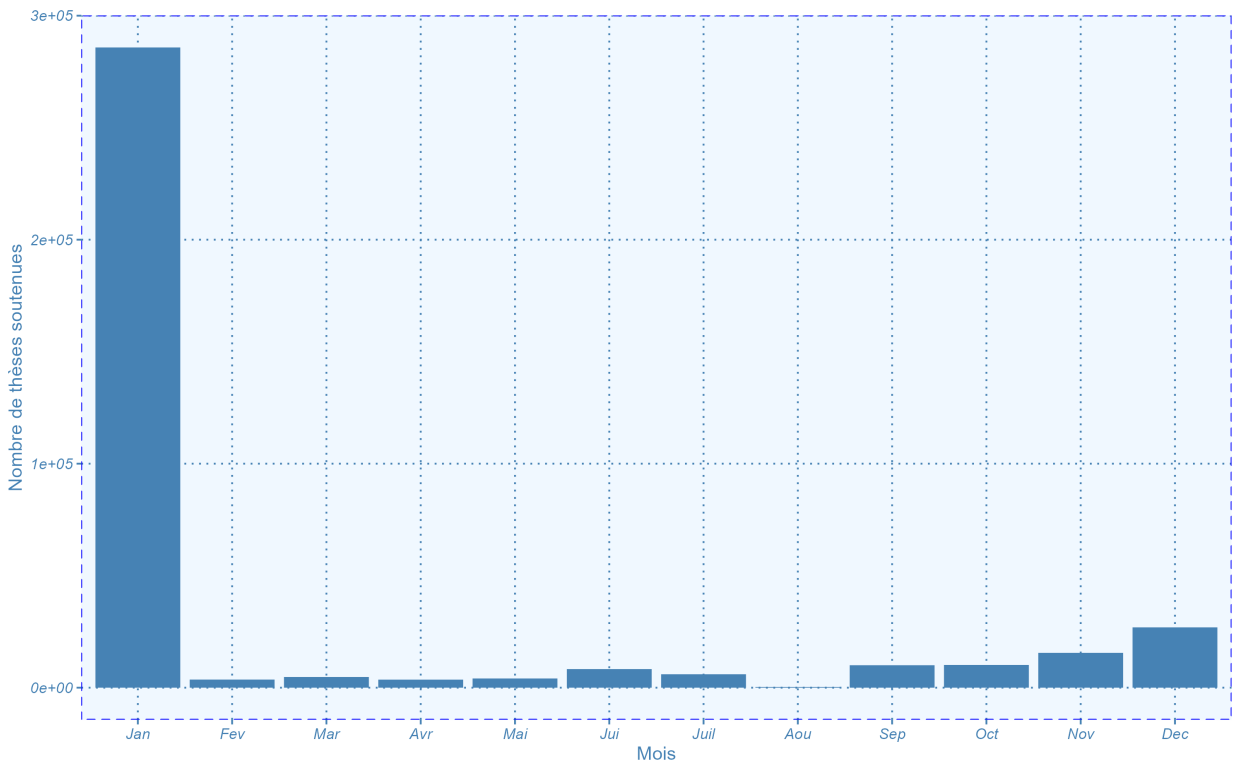


FIGURE 4 – Distribution des mois de soutenance (1984-2018)

La Figure 4 montre une très nette prédominance du mois de janvier dans les mois de soutenance de thèse entre 1984 et 2018 [1] et une "absence" de soutenance au mois d'août. Ce dernier point est cependant tout à fait cohérent avec les périodes de vacances estivales des universités françaises et ne semble pas être une erreur. Ce qui n'est pas le cas du mois de janvier. Rien n'explique une telle répartition à priori des soutenances, et la situation représentée par la Figure 4 semble devoir être investiguées plus en profondeur, reflétant selon nous un problème dans les données.

En suivant les recommandations des travaux de Cisel et al. (2021) qui ont déjà porter sur l'analyse de ce jeu de données issues du référencement des theses.fr, une première hypothèse peut être faite concernant la qualité du jeu de données. En effet, l'obligation légale de référencement des thèses n'a été faite aux université qu'à partir de l'année 2006 (article 10 de l'arrêté du 07 août 2006 portant sur les métadonénes de thèses). Avant cela, le référencement était local (i.e. relatif à chaque université), réalisé suivant des modalités d'indexation propres aux établissements. Ainsi, rien ne garantit que les dates puissent être exactes et précises. Cisel et al. ont déjà soulevé ce problème de date de soutenance sur la partie la plus ancienne de la base theses.fr. Le graphique de la Figure 5 montre l'évolution du nombre de thèse par mois de soutenances (période 2005-2018). On y constate une distribution nettement plus homogène des mois de soutenance, bien que le mois de janvier reste le mois le plus fréquent. Cela va dans le sens de notre hypothèse d'un mois de janvier « artefact » d'une indexation tardive des thèses anciennes. Toutefois, la présence d'écarts-types très important questionne encore la qualité des données, et doit inciter à une certaine prudence dans nos conclusions.

1. L'année 2018 a été choisi afin de conserver le maximum d'informations exactes dans la base de données lors de l'extraction en 2020, permettant de considérer le temps de traitement et d'indexation post-soutenance dans les universités française.

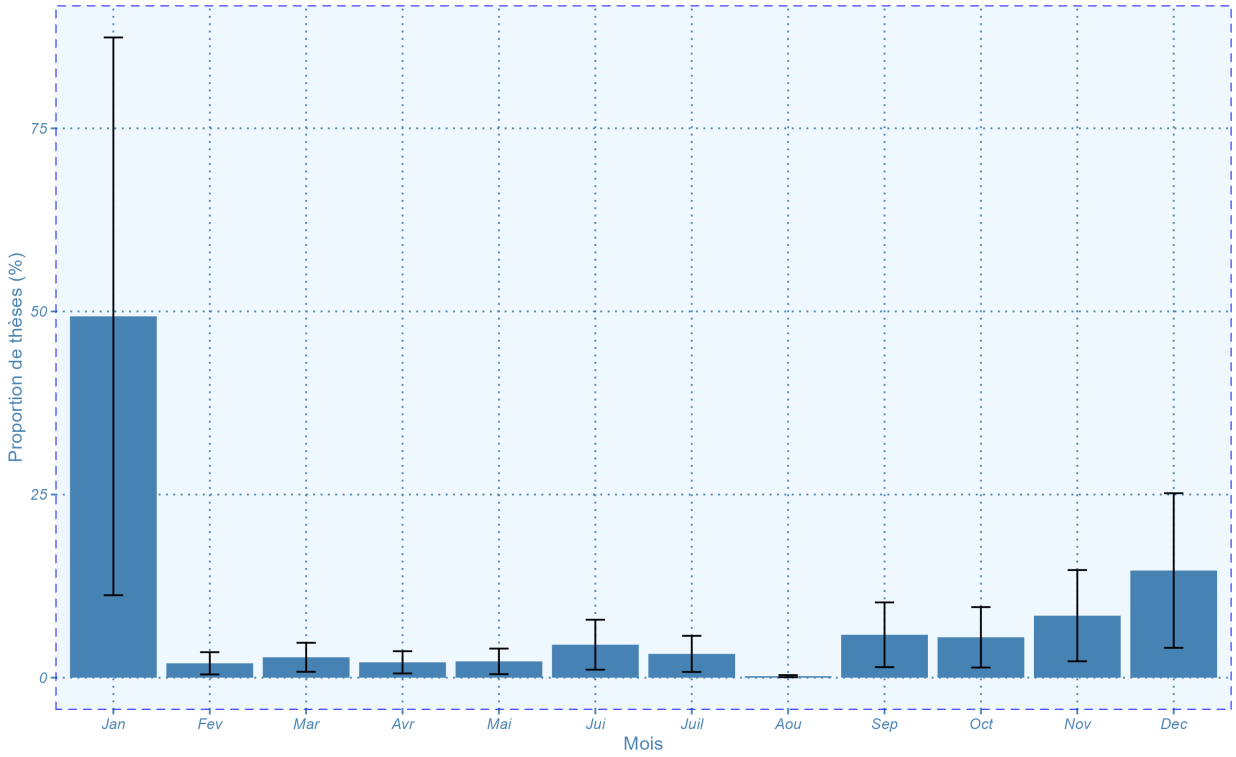


FIGURE 5 – Distribution des mois de soutenance (2005-2018) avec écart-type

En réalité, la valeur affichée dans la Figure 5 pour mois de janvier semble reposer sur un effet « volume », i.e. un nombre de thèses important encore référencées en janvier durant la période 2005-2018. La Figure 6, qui représente l'évolution de la proportion des mois de soutenance sur la même période 2005-2015 montre le très net déclin de l'usage de cette date de soutenance, confirmant selon nous que les nouveaux critères d'indexation, après une période de latence liée à la diffusion des usages, ont conduit à des données plus précises. Cette évolution est confirmée par l'analyse du graphique de la figure 8 (Annexes), où l'on peut constater qque la diminution du recourt au mois de janvier commence à se percevoir à partir de 2007.

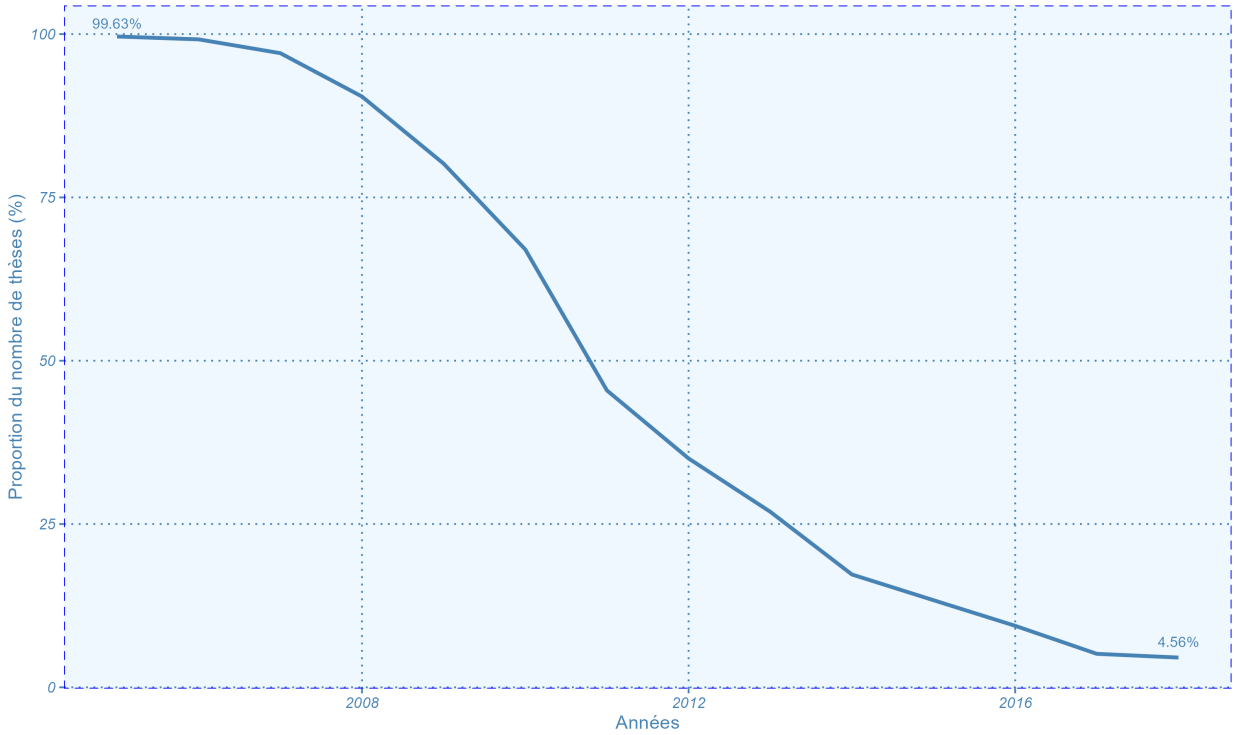


FIGURE 6 – Distribution de la proportion du mois de janvier dans les date de soutenance (2005-2018)

2.1.2 Isoler l'artefact des date de soutenance : le cas des "01-01"

En y regardant de plus près (Tableau 2), il semble que la surreprésentation du mois de janvier provienne d'un nombre anormal de thèse soutenue le « 01-01 », date impossible puisque jour férié (1er de l'an). Selon nous, il s'agit ici d'un artefact informatique, une entrée crée de manière ad hoc lors de la construction de la base de données ou de son extraction. En effet, par défaut, dans de nombreux logiciel de manipulation/traitement de données (par exemple R), la reconstruction d'un format date à partir d'une année seule produit un format générique de type « 01-01- année considérée ». Cette reconstruction artificielle permet par la suite les manipulations des variables au format date et/ou facilite certains types de visualisation temporelle, là ou l'année seule est souvent de type entier voire caractère. En étudiant les résultats du Tableau 2, on note une très forte et très constante diminution dans le temps des occurrences "01-01", signe selon nous de la diffusion des pratiques normées d'indexation des thèses et d'une collecte plus précise des dates de soutenances, correspondant aux exigences administratives.

2.1.3 La vraie distribution des moins de soutenance

Une fois retirée les date commençant le 01-01 (car ne pouvant être de vraies dates), nous obtenons une distribution corrigée (Figure 7), permettant de souligner que le mois de soutenance préféré des doctorant est le mois de décembre

TABLE 2 – Nombre de thèse soutenue le 01-01, par année (2005-2018)

Annee	Totale these en 01-01	Total these an.
2005	10522	10562
2006	10885	10975
2007	11349	11697
2008	10686	11854
2009	9554	12033
2010	8190	12516
2011	5605	13110
2012	4398	13985
2013	3237	13868
2014	1666	13202
2015	1069	13023
2016	633	12965
2017	15	13123
2018	1	12805

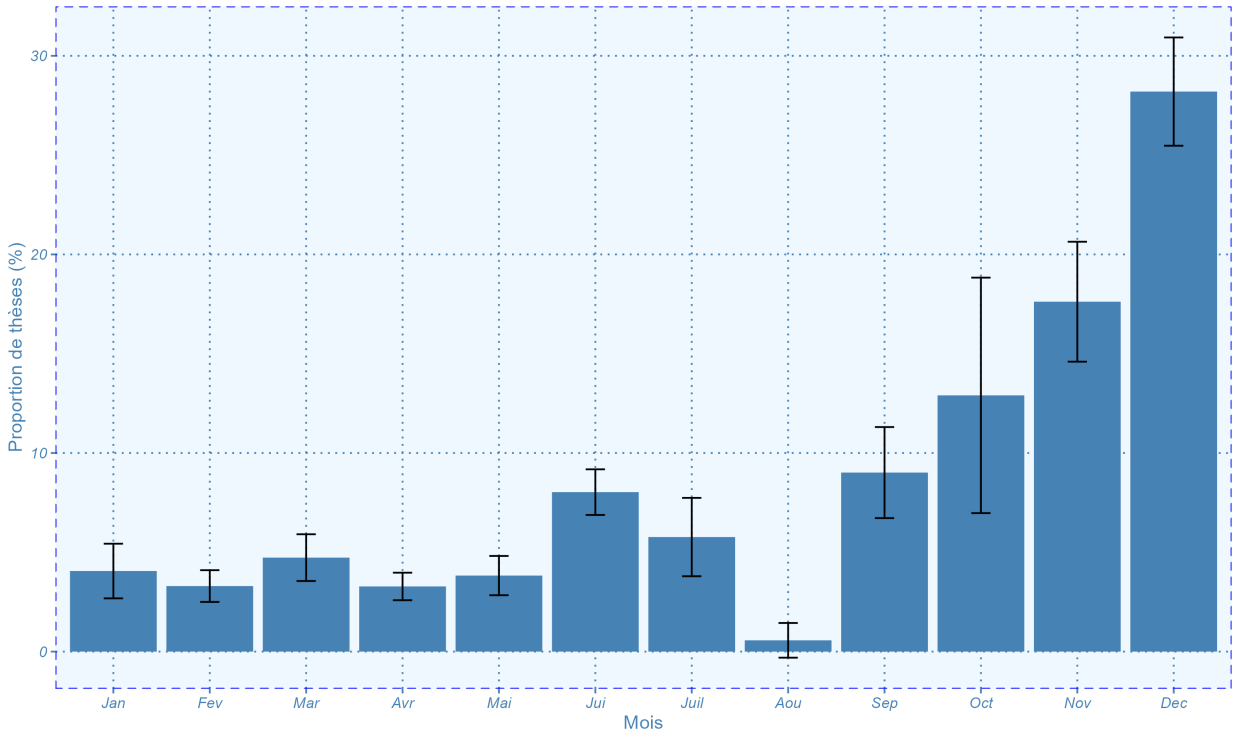


FIGURE 7 – Distribution corrigée (retrait de l'artefact du 01-01) des mois de soutenance de thèse par année (2005-2018)

2.2 Hyper-docteurs et homonymie

Un autre problème est apparu à l'exploration du jeu de données. Celui d'individus présentant un nombre anormalement élevé de thèses, dont tous ne sont pas nécessairement des hyper-docteurs (cas d'individus ayant soutenus plusieurs thèses dans leur vie) et signale l'existence d'homonymes mal gérés dans la base sur lesquels exercés des traitements correctifs. Ainsi, le cas de Cécile Martin (Tableau 3) est apparu comme caractérisant ce problème, et servira ainsi d'illustration au travail à réaliser.

Table 3 – Liste des thèses référencées pour l'auteur Cécile Martin

Auteur	Id. auteur	Année	Discipline	Id. thèse
cecile martin	203208145	2017	Etudes cinématographiques et audiovisuelles	2017USPCA018
cecile martin	81323557	2000	Sciences biologiques fondamentales et appliquées. Sciences médicales	2000INAP0034
cecile martin	179423568	2014	Sciences économiques	2014PA090003
cecile martin	81323557	2001	Génie des procédés industriels	2001COMP1380
cecile martin	81323557	1991	Neurosciences	1991BOR22005
cecile martin	81323557	1994	Sciences biologiques et fondamentales appliquées. Psychologie	1994CLF21651
cecile martin	182118703	1989	Physique	1989PA112163

2.2.1 Le cas Cecile Martin

Le tableau 3 montre le nombre important de thèse attribuée à l'auteur Cécile Martin dans la base de thèses.fr, dans des disciplines en outre très diverses. La considération de la variable identifiant auteur permet dans un premier temps de réduire le nombre de thèses soutenues par individu unique pour ne considérer que le seul identifiant 81323557 comme problématique. Cet individu sera la base du travail d'enquête et de correction.

Toutefois, même en écartant les autres auteurs sur leurs identifiants, nous constatons toujours un nombre anormalement élevé de thèses soutenues pour un seul individu (Id 81323557). Les domaines de recherches semblent en outre relativement éloignés les uns des autres (Neurosciences versus Sciences biologiques fondamentales versus Génie des procédés industrielles) [2] ce qui sous-entend à minima des reprises d'études et/ou des années de formation avant de pouvoir soutenir une thèse dans ces disciplines. Or, certaines soutenances sont rapprochées à moins de 2 ou 3 ans, ce qui parait suspect. Il paraît difficile qu'un même auteur puisse réussir à soutenir plusieurs thèses dans des disciplines si différentes à 3 ou encore 1 an d'intervalle. L'hypothèse la plus vraisemblable est que nous soyons encore en présence d'homonymes, malgré cet identifiant unique.

Pour compléter cette première analyse du problème "Cécile Martin", nous avons ensuite commencé notre enquête en concentrant nos efforts sur la situation la plus crédible à nos yeux : le cas d'un individu ayant soutenu deux thèses à 6 ans d'intervalle en Sciences biologiques fondamentales (supposant que les différences dans les labels soient une nouvelle fois issus de l'indexation non contrôlée prévalant avant 2005). A partir d'une recherche par les identifiants chercheurs officiels ORCID de l'auteur et de ses directeurs de thèses, complétée par une recherche Google/Google Scholar, nous avons pu trouver trois Cécile Martin, dont une spécialiste du monde bovin (spécialisée dans les émissions de méthane et l'analyse des effets des rejets écologiques des élevages, https ://orcid.org/0000-0002-2265-2048), thème centrale des deux thèses de 1994 et 2000. Ces données suffisent, en l'état, à confirmer notre hypothèse d'une homonymie, et confirme l'intérêt des facteurs retenus (identifiant auteur, proximité disciplinaire et temps).

2. En effet, si les Neurosciences et les Sciences biologiques fondamentales semblent de prime abord des sciences voisines, se concentrant sur l'étude du vivant, elles possèdent de nombreuses différences épistémologiques nécessitant des reprises d'études et ou de formation professionnelle. Les neurosciences étudient le système nerveux et tout particulièrement le cerveau pour analyser et comprendre les phénomènes humains de type comportements, émotions ou perceptions ; les Sciences biologiques fondamentales couvrent un large éventail d'études concernant tous les aspects de la biologie, comme la génétique, la physiologie, la biologie cellulaire ou l'écologie. Les deux disciplines se construisent par le même tronc commun initial mais représentent des domaines de recherche séparés avec des corpus théoriques et des objectifs très différents : comprendre comment le système nerveux fonctionne, se développe, change avec l'expérience et dysfonctionne pour les neurosciences ; comprendre les principes de base qui régissent la vie sous toutes ses formes, de la molécule unique aux écosystèmes complexes pour les Sciences biologiques fondamentales.

2.2.2 Enquête et corrections possible : synthèse

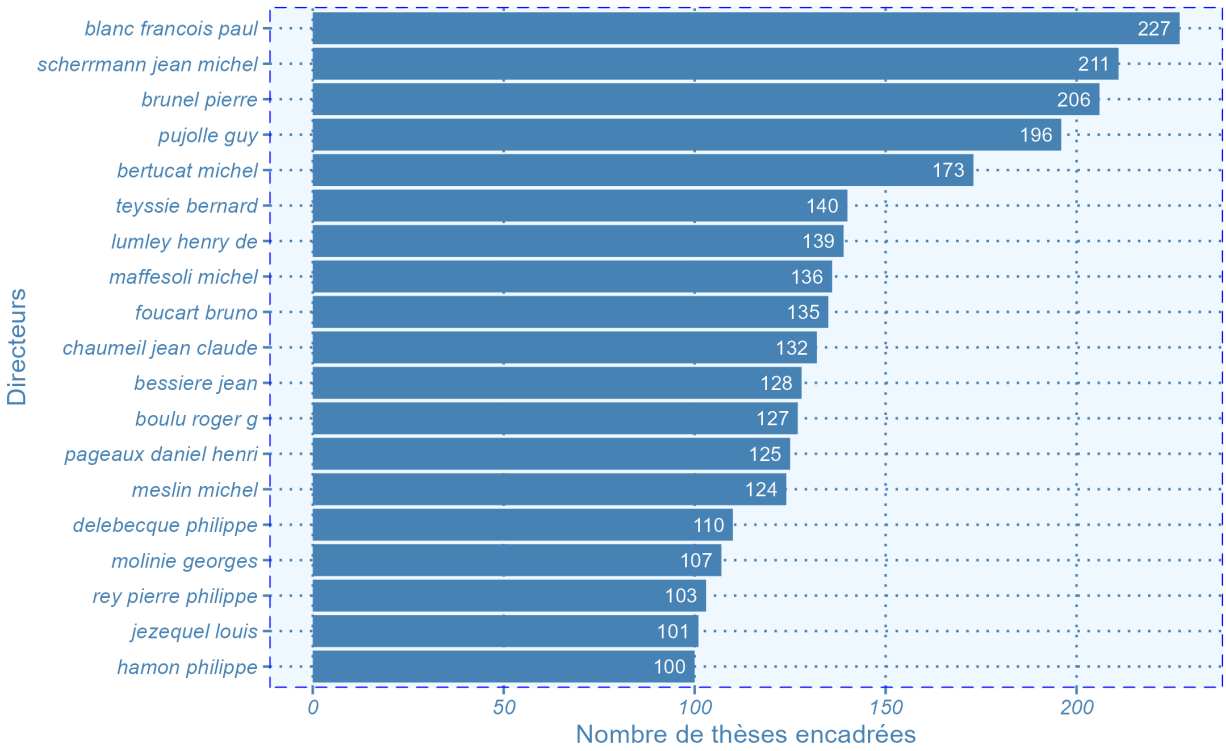
Table 4 – Synthèses des éléments utilisés pour l'enquête

Interprétation	Signification	Données
Cohérent	Identifiants auteur multiples	203208145 | 81323557 | 179423568 | 182118703
	Domaines de recherche proches	Sciences biologiques fondamentales et appliquées. Sciences médicales | Sciences biologiques et fondamentales appliquées. Psychologie
	Dates de soutenance compatibles	2000 | 2006
Incohérent	Identifiant auteur unique	81323557
	Domaines de recherche éloignée	Etudes cinématographiques et audiovisuelles | Sciences économiques | Génie des procédés industriels | Neurosciences | Physique (versus Sciences biologiques fondamentales et appliquées)
	Dates de soutenance incompatibles	2017 - 2014 | 2001 - 2000 | 1991-1994 | 1989 - 1991
SOLUTIONS	Recherche externe	ORCID, Google/Google Scholar, Research Gate (+ Thèses.fr pour le noms des encadrants permettant de vérifier la communauté autour de l'auteur)

3 Détection d'outliers

3.1 Le problèmes de l'encadrement des thèses

Nous nous intéressons maintenant aux directeurs et directrices de thèses. L'analyse des variables et données relatives aux encadrants (directrices et directeurs de thèses référencés dans le jeu de données PhD_-v2) permet de relever un problème assez similaire à celui exposé pour les hyper-docteurs. En classant les encadrants par nombre de thèses soutenues, nous constatons en effet que les 19 premiers ont eu à en diriger chacun plus de 100 sur la période 1984-2018 (Figure 9). Ce chiffre apparait très difficilement soutenable et semble plus qu'improbable : 100 thèses (le minimum des 19 premiers) en 34 ans cela revient à mener à la soutenance environ 3 thésardes et thésards par année universitaire. Or une thèse se déroule sur une période minimale de 3 à 5 ans, ce qui déplace la fourchette d'étudiants encadrés entre 9 et 15 simultanément par tranche de 3 à 5 ans. Ces valeurs semblent à priori incohérentes et semblent soulignées des problème d'homonymie proche de ceux déjà soulevés dans la partie précédente, mais dont l'ampleur crée ici des risques d'outliers, pouvant biaisées les analyses futures.



'il s'agit bien de la même personne Liste des 19 directrices/directeurs ayant encadrés plus de 100 thèses

Reprenant la même méthodologie d'enquête qu'avec l'analyse de l'auteur Cécile Martin (enquête reposant sur trois facteurs : identifiant, discipline, temps) afin d'enquêter sur ces valeurs extrêmes et confirmer notre hypothèse d'homonymie, nous avons constaté l'impossibilité d'utilisé la variable de l'identifiant pour les directeurs. En effet, cette variable affiche de trop nombreuses valeurs manquantes. Afin de remplacer ce facteur, nous avons donc choisi la variable Directeur de these (nom prenom) qui n'affiche pas de valeurs manquantes (Tableau 5).

Table 5 – Nombre de valeurs manquantes par variables utiles à l'enquête

	Nombre de NA	Pourcentage de NA
Identifiant directeur	19137	5.05
Directeur de these (nom prenom)	12	0.00

3.2 Outliers ou hyper-encadrement : élément de l'enquête et conclusion

Comme dans la partie précédente, concentrons l'analyse sur un de ces hyper-directeurs, ici « blanc francois paul ». Le Tableau 6 ci-dessous, qui résume les 15 principales années en nombre de thèses soutenues pour ce directeur, indiquent que nous sommes vraisemblablement en présence de données aberrantes, issues d'un référencement des thèses inapproprié ou erroné. Il ne nous semble en effet pas envisageable qu'il s'agisse du même individu, bien qu'à prioiri les disciplines référencées soient toutes juridiques. Il est incohérent, par exemple, qu'en 2006, « blanc francois paul » ait encadré 5 thèses soutenues en droit public, et 12 thèses en droit privé tant ces disciplines restent profondément différentes, représentant des spécialités et des domaines éloignés. En outre, il aurait encadré 10 thèses supplémentaire en droit privé et sciences criminelles (une autre discipline), pour un total de 27 thèses en une année ! Si nous considérons l'échantillon du jeu de données et les quartiles associés, nous pouvons observer que 75% des directrices et directeurs de thèses ont encadré 5 thèse ou moins sur la période 1984-2018 (Tableau 7). En observant une règle standard de détection des outliers basés sur ces derniers, la limite haute au-delà de laquelle des valeurs sont considérés comme outliers est 11. Ils semblent que ce jeu de données manifeste la présence de problèmes de données aberrantes à considérer avec attention. Ces outliers pourraient, comme dans le cas des hyper-docteurs reposer sur un problème d'homonymie.

TABLE 6 – Les 15 années les plus importantes en terme de thèses encadrées pour "blanc francois paul"

Année	Discipline	Identifiant etablissement	Nombre de theses
2004	Droit prive	26403692	17
2005	Droit prive	26403692	13
2006	Droit prive	26403692	12
2007	Droit prive et sciences criminelles	26403692	12
2006	Droit prive et sciences criminelles	26403692	10
2003	Droit prive	26403692	9
2009	Droit prive	26403692	9
2002	Droit prive	26403692	8
2004	Droit	26403692	8
2001	Droit prive	26403692	7
2005	Droit public	26403692	7
2000	Droit	26403692	6
2008	Droit	26403692	6
2001	Droit	26403692	5
2006	Droit public	26403692	5

TABLE 7 – Analyse inter-quartile, à 25% et 75% de l'échantillon (valeur limite haute des outliers calculée par IQR à 1,5)

Q1	Q3	IQR	Limite Haute
1.00	5.00	4.00	11.00

4 Résultats préliminaires : analyses des langues de rédaction

La Figure 10 illustre l'évolution entre 2001 et 2018 des langues de rédaction utilisées pour les thèses soutenues.

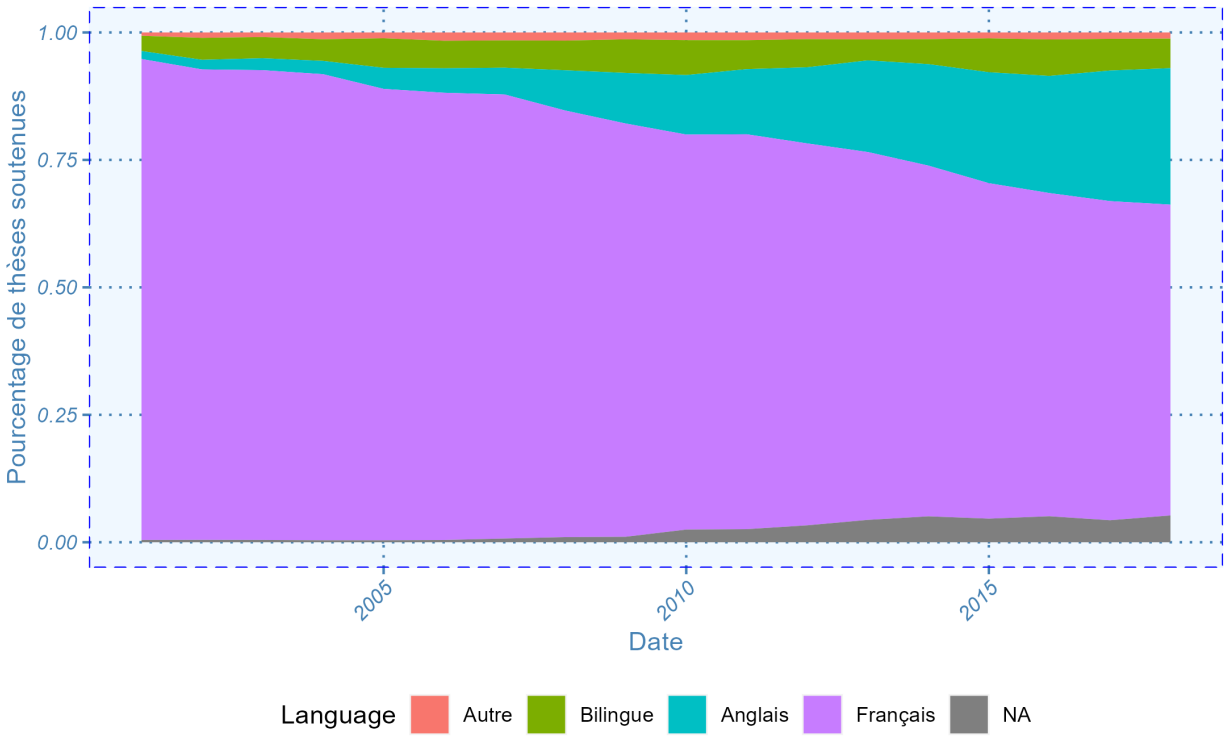


FIGURE 8 – Evolution de la proportion des langues de rédactions par thèses soutenues entre 2001 et 2018

Nous pouvons observer sur le graphique une progression constante et forte des thèses rédigée en anglais (149 en 2001, 1460 en 2010 et 3429 en 2018 ; Tableau 8) pouvant s'expliquer selon nous par l'ouverture internationale de la recherche française. Les financements des projets de recherche sont par exemple fortement dépendant de la dimension collaborative à l'échelle européenne et mondiale, faisant de l'anglais une seconde langue de référence pour la rédaction des travaux de recherche en cotutelle internationale. Il est également possible de voir un effet conjoint de deux phénomènes à l'œuvre dans les milieux académiques : une montée en compétences linguistique des étudiants français (qui pratiquent davantage et mieux l'anglais que leurs prédécesseurs) et la nécessité de plus en plus forte de publier qui pèsent sur les doctorants désireux de devenir chercheurs. Comme le souligne Martin (2015), ces derniers sont souvent incités à rédiger de nombreux manuscrits en anglais (des articles basés sur leurs travaux de thèses), et il est possible de penser que rédiger une thèse en anglais représente un gain de temps et d'efforts. Malgré cette croissance, le Français reste encore la langue dominante

Table 8 – Nombre de thèses soutenue par langue de rédaction, en 2001, 2010 et 2018

Année	Language de rédaction	Nb. de thèses
2001	Autre	62
2001	Bilingue	282
2001	Anglais	149
2001	Français	8932
2001		43
2010	Autre	189
2010	Bilingue	857
2010	Anglais	1460
2010	Français	9699
2010		311
2018	Autre	155
2018	Bilingue	741
2018	Anglais	3429
2018	Français	7807
2018		673

5 Annexes

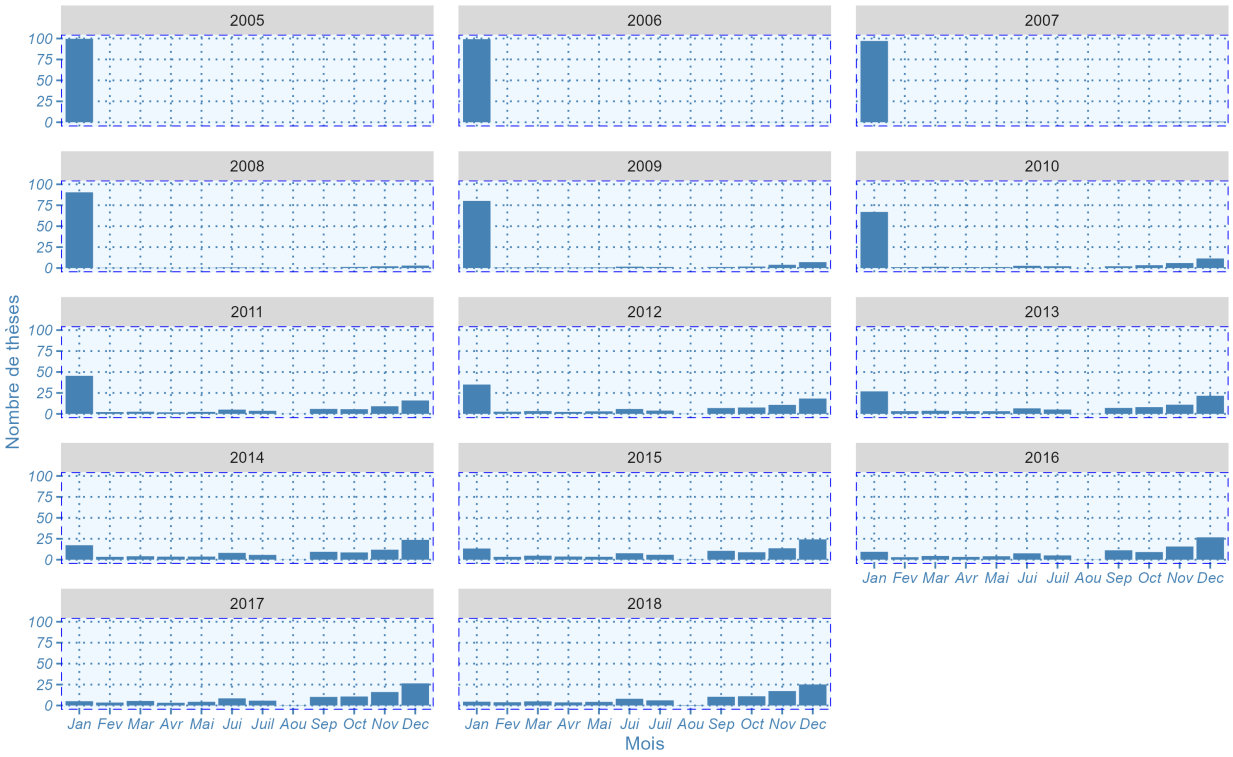


FIGURE 9 – Répartition du nombre de thèses soutenues en janvier par année (2005-2018)

6 Références

Jones, L. (2018). *Deep Learning for Natural Language Processing* (thèse de doct.). University of Techland.

Miller, A. (2020). *Introduction to Data Science*. Tech Publishers.

Smith, J., & Doe, J. (2021). Data visualization techniques in modern research. *Journal of Modern Data Science*, *1*(1), 1-12.

Williams, R., & Thompson, M. (2019). Neural networks and their applications in image recognition. *Proceedings of the 5th International Conference on Machine Learning*, 224-230.