CY Cergy Paris Université

DU Data Analyst

UE Prétraitement et manipulation de données

Rapport

Quentin FOUCHÉ

26 avril 2022

Table des matières

Liste des figures

Liste des tableaux

1 Présentation des données

1.1 Première manipulation d'un jeu de données

Le premier jeu de données manipulé est un tableau comportant deux variables, l'âge et le genre, et 23 705 individus. La distribution de ces deux variables est représentée dans les Figures 1 et 2. La majorité des individus ont entre 0 et 60 ans. Parmi eux, la classe d'âge la plus représentée est celle des 20 à 40 ans (Figure 1). Le sex ratio est globalement équilibré (Figure 2).

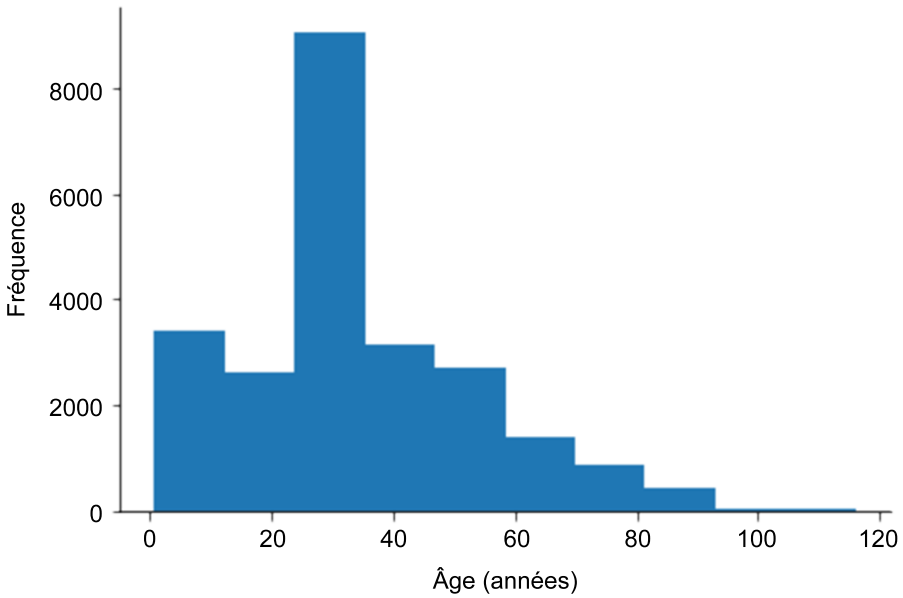

Figure 1. Distribution de la variable âge

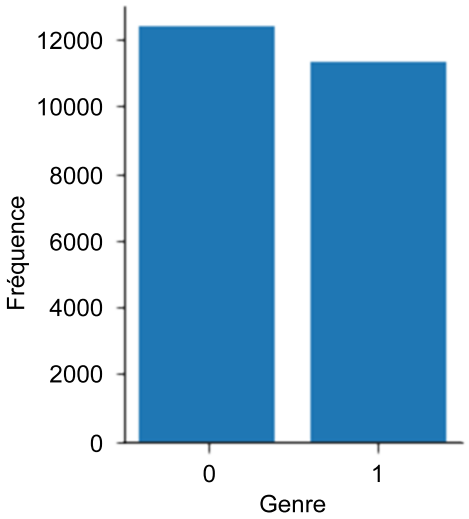

Figure 2. Distribution de la variable genre

1.2 Analyse d'un jeu de données réelles

Le jeu de données réelles est un tableau regroupant les métadonnées des thèses réalisées en France entre 1972 et 2020, extraites à partir du site web theses.fr. Deux jeux de données sont fournis : le premier ("PhD_v1") comporte manifestement des erreurs et est impossible à ouvrir sur la plateforme Jupyter, tandis que le second ("PhD_v2") s'ouvre sans problème. Toutes les analyses présentées dans la suite de ce rapport ont été réalisées sur ce deuxième jeu de données.

Le Tableau 1 résume le nom et le type des 18 variables du jeu de données, qui comprend 447 644 thèses. Les données des variables indiquant la date de première inscription en doctorat, la date de soutenance, l'année de soutenance et les dates de publication et de mise à jour dans theses.fr, sont des dates. La variable définissant si la thèse est accessible en ligne comporte des données logiques ("oui" ou "non"), tout comme la variable définissant le statut de la thèse ("en cours" ou "soutenue"). Toutes les autres variables comportent des données textuelles.

Tableau 1. Description des variables du jeu de données "PhD_v2". Le nombre de valeurs exclut les données manquantes, exprimées en pourcentage. "ID" : identifiant.

Variable	**Type**	**Nombre de valeurs**	**Données manquantes (%)**
Auteur	texte	447644	0
ID auteur	texte	317655	29.0
Titre	texte	447635	<0.1
Directeur	texte	447629	<0.1
Directeur (nom prénom)	texte	447629	<0.1
ID directeur	texte	447644	0
Établissement de soutenance	texte	447640	<0.1
ID établissement	texte	430559	3.8
Discipline	texte	447639	<0.1
Statut	logique	447644	0
Date d'inscription	date	63976	85.7
Date de soutenance	date	390898	12.7
Année	date	390898	12.7
Langue	texte	383879	14.2
ID thèse	texte	447644	0
Accessible en ligne	logique	447644	0
Publication dans theses.fr	date	447644	0
Mise à jour dans theses.fr	date	447467	<0.1

2 Données manquantes

2.1 Répartition des données manquantes

Plusieurs variables du jeu de données comportent un pourcentage élevé de valeurs manquantes : 4% pour l'identifiant établissement, entre 12 et 15% pour la date et l'année de soutenance et la langue, 29% pour l'identifiant auteur et jusqu'à 86% pour la date d'inscription (Tableau 1). Les autres variables contiennent moins de 0.1% de données manquantes ; pour cette raison, seules les 6 premières ont été prises en compte dans les analyses suivantes.

La répartition des données manquantes dans ces 6 variables est représentée dans la Figure 3. Dans cette figure, les thèses ont été triées de sorte à afficher en premier celles dont la date d'inscription en doctorat est présente. Cette répartition montre une régularité dans la localisation des données manquantes de certaines variables : les données manquantes dans la date et l'année de soutenance ainsi que dans la langue semblent concerner les mêmes thèses. En revanche, ces thèses contiennent toutes une date d'inscription en doctorat, qui est absente dans les autres thèses (Figure 3).

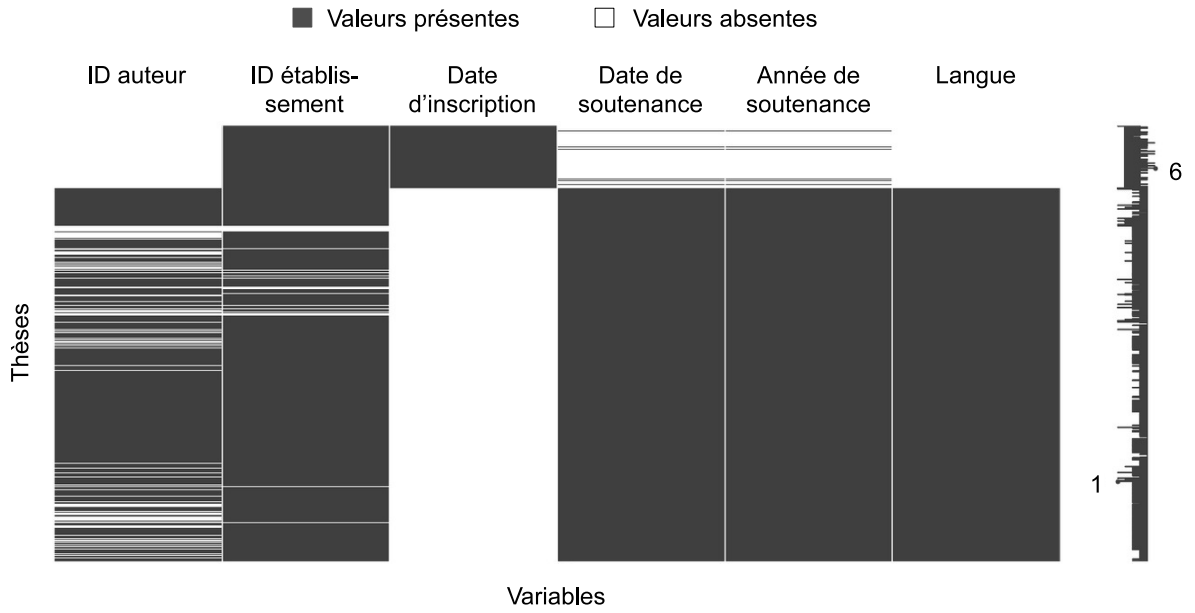

Figure 3. Répartition des valeurs présentes (en noir) et manquantes (en blanc) dans les 6 variables qui contiennent le plus de données manquantes, pour toutes les thèses du jeu de données "PhD_v2". L'histogramme vertical affiché à droite représente le nombre de valeurs manquantes par thèse. "ID" : identifiant.

La Figure 4 donne un aperçu plus détaillé de la répartition des valeurs manquantes. Trois patterns peuvent être distingués. Tout d'abord, la grande majorité des thèses (312 904 thèses, soit 69.9% du nombre total) contiennent une seule valeur manquante, située dans la date de première inscription en doctorat. Ensuite, un ensemble de 53 115 thèses (11.9%)

ont des valeurs manquantes à la fois dans la date d'inscription et dans l'identifiant auteur. Enfin, 55 868 thèses (12.5%) ont des valeurs manquantes dans 4 variables simultanément : la date et l'année de soutenance, la langue de la thèse et l'identifiant auteur (Figure 4). Ces observations confirment qu'il y a une séparation nette entre les valeurs manquantes de la date d'inscription et celles de la langue, la date et l'année de soutenance. Cette séparation s'observe aussi entre la date d'inscription et l'identifiant auteur, mais seulement pour une partie des données manquantes.

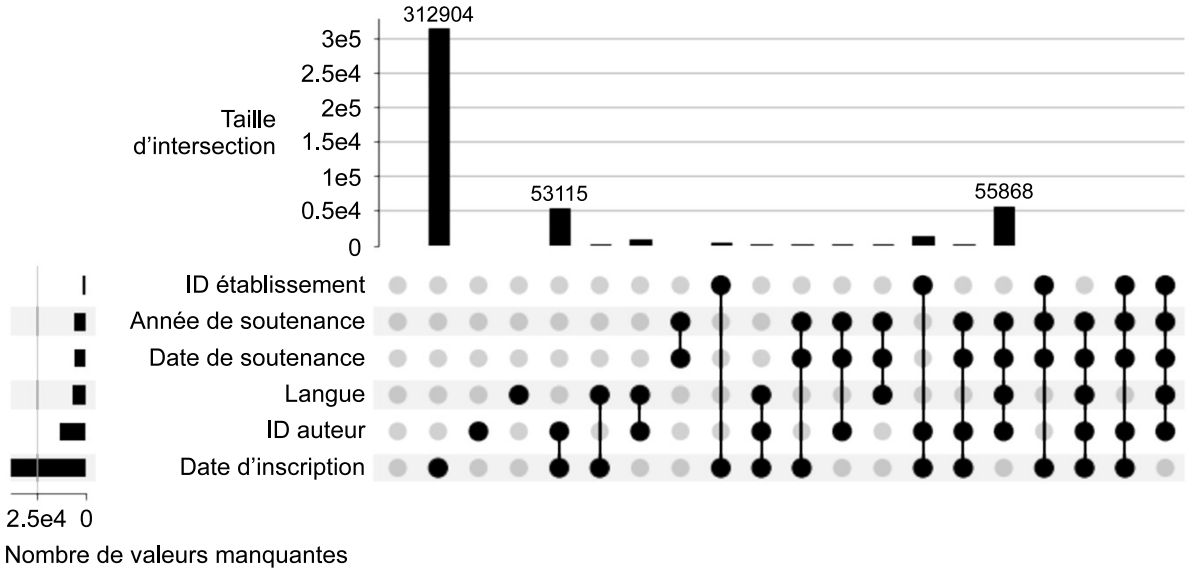

Figure 4. Diagramme UpSet représentant la nombre de valeurs manquantes dans chaque combinaison de variables, pour les 6 variables du jeu de données "PhD_v2" qui contiennent le plus de données manquantes. L'histogramme situé en bas à gauche montre le nombre total de valeurs manquantes dans chacune des variables. L'histogramme en haut à droite montre la taille d'intersection, i.e. le nombre de thèses qui contiennent soit aucune valeur manquante (1e colonne), soit des valeurs manquantes dans une seule variable (colonnes 2 à 4), soit des valeurs manquantes communes à 2 variables ou plus (colonnes 5 à 20). Les points colorés en noir indiquent dans quelle variable ou quelle combinaison de variables sont situées les valeurs manquantes. "ID" : identifiant.

2.2 Corrélations entre les données manquantes

Le corrélogramme présenté en Figure 5 confirme les informations révélées par les Figures 3 et 4 : il montre que les données manquantes dans la langue, la date et l'année de soutenance sont très positivement corrélées entre elles (coefficient situé entre 0.9 et 1), et négativement corrélées avec celles dans la date d'inscription (-0.9 à -1). Les valeurs manquantes de ces quatre variables sont donc de type "MNAR" (Missing Not At Random). Cela semble être aussi le cas pour une partie des données manquantes dans l'identifiant auteur, qui concernent les mêmes thèses que celles où manquent la langue, la date et l'année de soutenance (coefficient à 0.6 ; Figures 3 et 5). En revanche, l'autre partie des données manquantes dans l'identifiant auteur, ainsi que la quasi-totalité des données

manquantes de l'identifiant établissement, ne sont pas corrélées à celles d'autres variables (Figure 5). Ces valeurs sont donc de type "MAR" (Missing At Random).

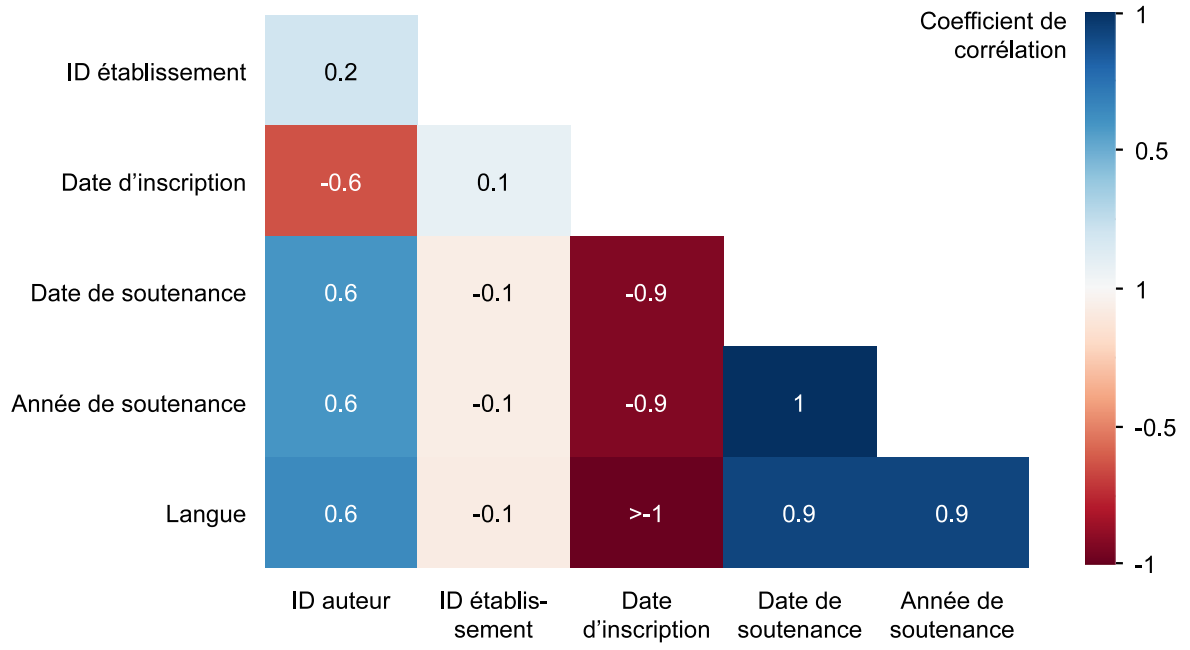

Figure 5. Corrélogramme des valeurs manquantes entre deux variables, pour chaque paire de variables parmi les 6 qui contiennent le plus de données manquantes. Le gradient de couleur indique le degré de corrélation entre les variables (bleu : corrélation positive ; blanc : pas de corrélation ; rouge : corrélation négative). "ID" : identifiant.

Comme le révèle la Figure 6, la corrélation négative entre la date d'inscription et la date de soutenance est liée au statut de la thèse. Lorsque la thèse n'a pas encore été soutenue, la date d'inscription est présente et la date de soutenance absente, et inversement lorsque la thèse a été soutenue (Figure 6). Cela pourrait signifier que la date de première inscription en thèse est automatiquement effacée lorsque la date de soutenance est ajoutée sur theses.fr. Le lien entre la date, l'année de soutenance et la langue semble plus évident : l'année de soutenance est ajouté en même temps que la date, et la langue ne peut probablement être connue qu'une fois le manuscrit de thèse soumis, donc seulement après la soutenance. Il semble aussi que l'identifiant auteur n'est donné qu'une fois la thèse soutenue, puisqu'aucun identifiant n'est donné pour les thèses encore en cours (Figure 3).

3 Principaux problèmes détectés

3.1 Anomalie dans le mois de soutenance

La première investigation sur les problèmes potentiellement présents dans le jeu de donnés a porté sur le mois de soutenance. La période analysée s'étale de début 1984 à fin 2018, en raison du faible nombre de thèses soutenues avant 1984 et d'un manque de données pour

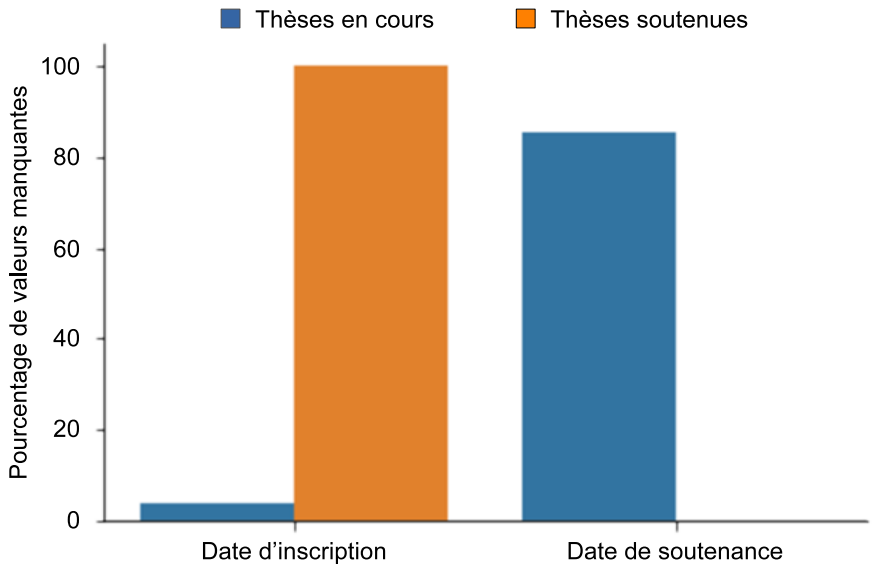

Figure 6. Pourcentage de valeurs manquantes dans la date d'inscription et la date de soutenance en fonction du statut de la thèse (bleu : en cours, n = 2690 ; orange : soutenue, n = 381189).

les thèses soutenues en 2019, 2020 et 2021 (ce manque de données étant probablement lié à une latence parfois longue entre la soutenance de la thèse et la rentrée des données sur theses.fr).

Le graphique présenté en Figure 7 montre le nombre de thèses soutenues par mois entre 1984 et 2018. Le nombre de thèses soutenues en janvier (environ 275 000) s'avère être nettement plus élevé que celui de tous les autres mois (moins de 30 000), ce qui semble peu réaliste. Cette différence pourrait s'expliquer par un manque de données relatives au

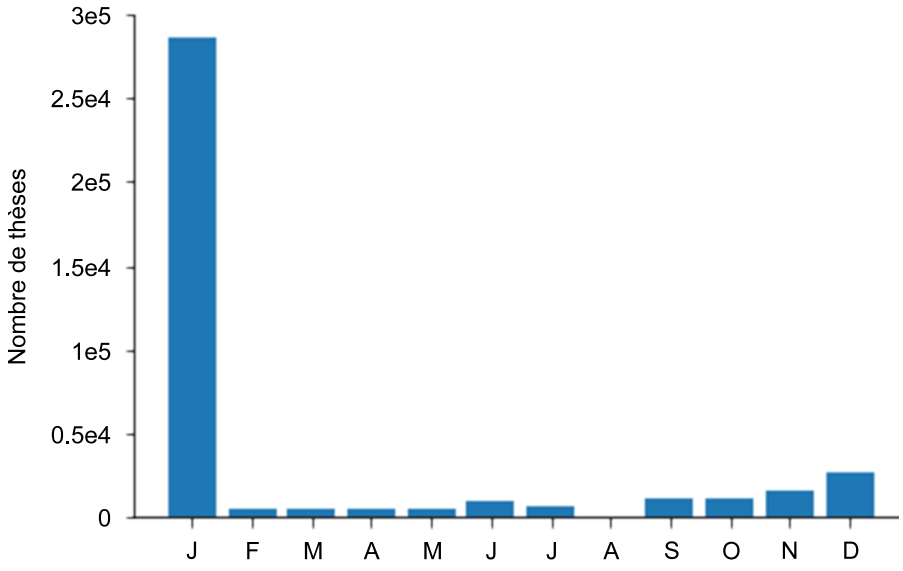

Figure 7. Nombre de thèses soutenues en fonction du mois, entre 1984 et 2018 inclus. Les mois sont indiqués par leur initiale, dans l'ordre chronologique.

mois et au jour de soutenance pour la grande majorité des thèses lorsqu'elles sont rentrées sur theses.fr. Le jour et le mois seraient alors automatiquement fixés au premier janvier.

La Figure 8 représente la proportion de thèses soutenues par mois pour chaque année prise séparément, de 2005 à 2018. La proportion élevée du mois de janvier chute progressivement à partir de 2008, jusqu'à atteindre un niveau proche de celui des autres mois à partir de 2014. La Figure 9 confirme cette observation : le pourcentage de thèses soutenues le 1er janvier diminue à partir des années 2008 jusqu'à devenir quasi-nul en 2017. Cela suggère qu'à partir des années 2000, un nombre croissant de thèses ajoutées sur theses.fr contenaient des informations concernant le mois et le jour de soutenance. A partir de 2017, la quasi-totalité des thèses ajoutées contenaient ces informations.

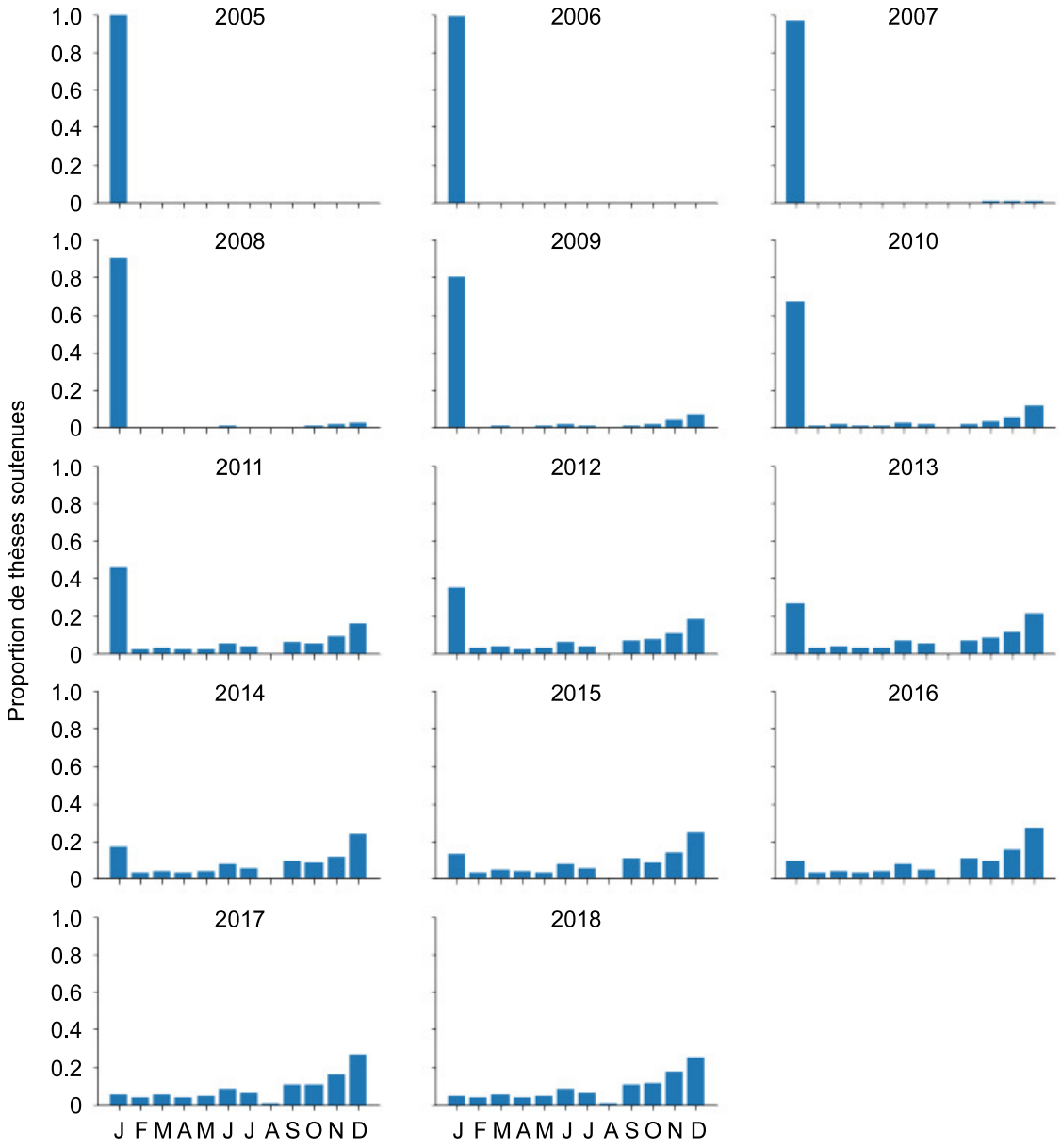

Figure 8. Proportion de thèses soutenues en fonction du mois (indiqués par leur initiale dans l'ordre chronologique), pour chaque année prise séparément entre 2005 et 2018 inclus.

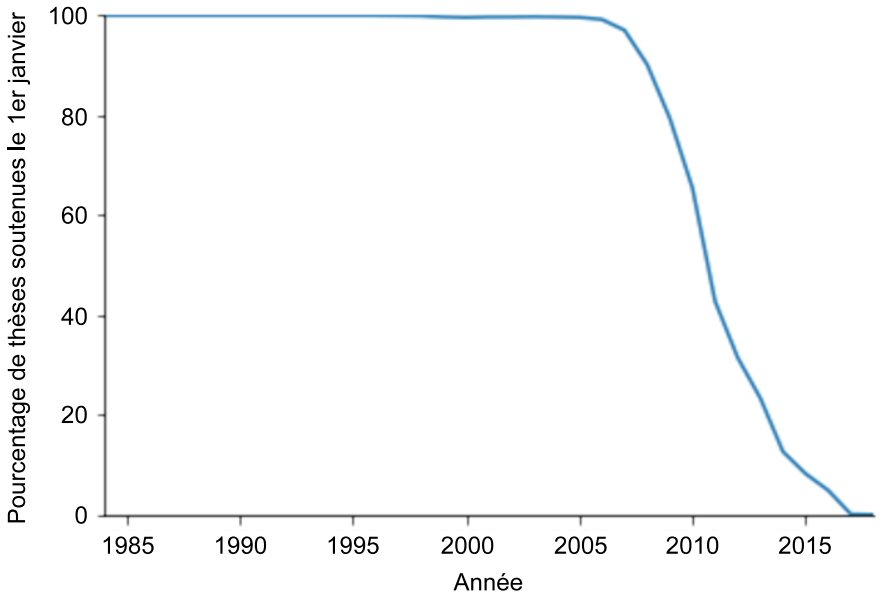

Figure 9. Pourcentage de thèses soutenues le 1er janvier en fonction de l'année, entre 1984 et 2018.

Puisque les dates de soutenance du 1er janvier semblent erronées, les thèses indiquées comme ayant été soutenues à cette date peuvent être retirées du jeu de données avant d'analyser la dynamique temporelle des soutenances. C'est ce qui a été fait dans la Figure 10 : cette figure représente le pourcentage de thèses soutenues par mois sur la période 1984-2018, sans les thèses soutenues le 1er janvier. Elle montre que les thèses ont été le plus souvent soutenues en fin d'année à partir de septembre (9 à 16% de septembre à novembre), avec un pic à 27 $\pm$ 3.5% de soutenances au mois de décembre. Ce pic peut s'expliquer par le fait que certaines écoles doctorales autorisent le doctorant à soutenir avant la fin de l'année civile sans qu'ils n'aient besoin de s'inscrire à l'université pour une année supplémentaire. Une fois passé le nouvel an, l'inscription devient obligatoire, expliquant le pourcentage plus faible de thèses soutenues à partir de janvier (entre 3 et 7% de janvier à juillet). Par ailleurs, peu de soutenances ont eu lieu en août (moins de 1%), période à laquelle les écoles doctorales et universités sont généralement fermées.

3.2 Erreurs liées aux homonymies : le cas de Cécile Martin

Un autre type d'erreurs qu'il est possible d'observer dans un jeu de données comportant des noms d'individus est celui des homonymies. Dans le cadre des thèses, des homonymes peuvent se trouver soit dans parmi les auteurs, soit parmi les directeurs. Pour illustrer ce problème, le cas de l'auteure ayant pour nom Cécile Martin a été analysé.

Le jeu de données contient 7 thèses dont Cécile Martin est l'auteure. Parmi ces thèses, trois ont un identifiant auteur unique (203208145, 179423568 et 182118703), ce qui permet d'emblée d'en déduire qu'elles ont été réalisées par des personnes différentes, ayant toutes le même nom et prénom. En revanche, les quatre autres thèses ont le même identifiant auteur (81323557) : une analyse plus approfondie est ici nécessaire pour déterminer s'il s'agit ou non d'une même personne.

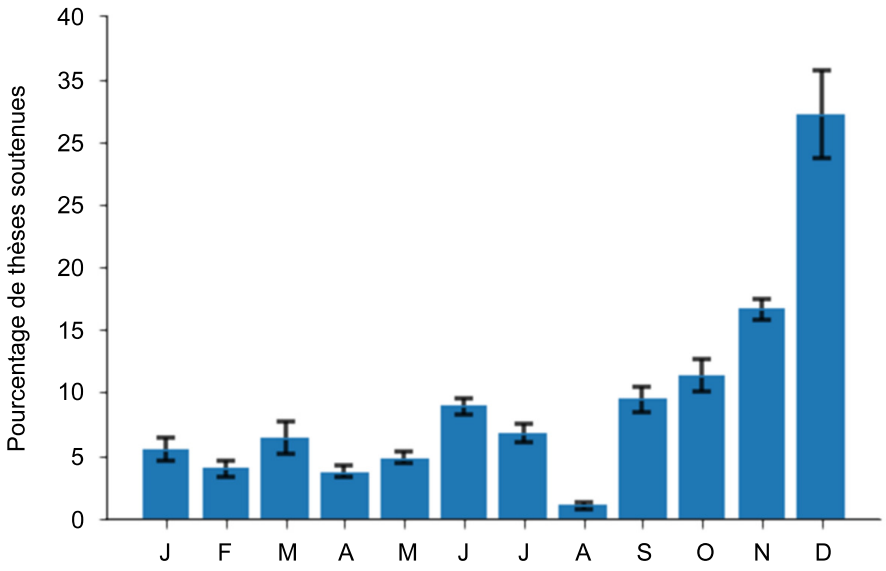

Figure 10. Pourcentage moyen de thèses soutenues par mois sur l'ensemble des années entre 1984 et 2018. Les barres d'erreurs représentent l'erreur standard à la moyenne. Les mois sont indiqués par leur initiale dans l'ordre chronologique.

Le titre des thèses soutenues par Cécile Martin dont l'identifiant est 81323557 sont cités ci-dessous :

1. Système laitier et filière lait au Mexique : contraintes de développement, stratégies d'acteurs, enjeux de leur coévolution. Cas de la Ciénega de Chapala, Jalisco.

2. Modélisation et critères de combustibilité en incinération combinée de déchets ménagers et de déchets industriels banals.

3. Caractérisation électrophysiologique et pharmacologique des canaux ioniques : sodium, calcium, activés par l'ATP, des cellules myométriales, effets de la gestation et de l'ocytocine.

4. Influence du ph ruminal sur la digestion des parois végétales, en relation avec les modifications de l'activité fibrolytique de l'écosystème microbien.

Ces quatre titres étant différents, l'hypothèse de doublons lors de la rentrée des informations sur theses.fr peut être rejetée.

Le Tableau 2 présente les principales informations discriminantes concernant ces quatre thèses : l'établissement de soutenance, le directeur de thèse, la discipline, la date de soutenance et la date de mise à jour des données sur theses.fr. D'après ces informations, les quatre thèses ont été réalisées dans des villes différentes, dans des disciplines différentes et encadrées par des directeurs différents. Elles ont également été soutenues à des dates différentes, avec 3 ans d'écart entre les thèses n°3 et n°4, 6 ans entre les thèses n°4 et n°1 et seulement 1 an entre les thèses n°1 et n°2. Il semble donc peu probable que les thèses n°1 et n°2 soient l'œuvre d'une même personne. Par ailleurs, les thèses n°2, 3 et 4 ont été mises à jour en même temps sur theses.fr, à un jour près. Il se pourrait donc que ces trois

thèses aient été réalisées par une même personne.

Tableau 2. Description des thèses soutenues par Cécile Martin dont l'identifiant auteur est 81323557. Le numéro de la thèse correspond à l'ordre de citation des titres dans le texte.

	Cécile Martin - ID : 81323557			
	Thèse n°1	**Thèse n°2**	**Thèse n°3**	**Thèse n°4**
Établissement	Agro Paris-Grignon	Compiègne	Bordeaux 2	Clermont-Ferrand 2
Directeur	Jean Lossouarn	Gerard Antonini	Jean Mironneau	Yves Briand
Discipline	Sciences médicales	Génie des procédés	Neurosciences	Psychologie
Soutenance	01/01/2000	01/01/2001	01/01/1991	01/01/1994
Mise à jour	10/12/2019	08/07/2020	07/07/2020	07/07/2020

Pour résumer, les informations contenues dans le jeu de données suggèrent fortement que le nom Cécile Martin désigne au moins 5 personnes différentes. Une même personne pourrait avoir réalisé trois thèses à elle seule, entre 1994 et 2001. Des recherches au-delà du jeu de données restent nécessaires pour confirmer ces hypothèses.

4 Outliers

4.1 Directeurs ayant encadré un nombre relativement anormal de thèses

Les outliers représentent les valeurs extrêmes des variables dans un jeu de données. S'ils correspondent parfois à des erreurs dans la collecte des données, leur traitement (i.e. le choix de les garder ou de les supprimer) peut considérablement impacter l'analyse et nécessite donc d'être envisagé avec précaution. Dans le jeu de données de theses.fr, l'analyse des outliers a porté sur les directeurs et directrices de thèses.

Un aperçu de la distribution du nombre de thèses encadrées par directeur (incluant les cas de co-direction) est donné dans la Figure 11. Cette figure montre que la grande majorité des directeurs et directrices ont encadré moins d'une dizaine de thèses entre 1984 et 2018, mais que quelques uns en ont dirigées entre 100 et 230, jusqu'à environ 700 pour l'un d'eux.

Le Tableau 3 liste les prénoms et noms des 10 directeurs ayant encadré le plus de thèses, par ordre décroissant. La première ligne indique que pour 711 thèses, le nom du directeur de thèse est inconnu : cet oulier est une erreur, qui peut être supprimée sans risque de

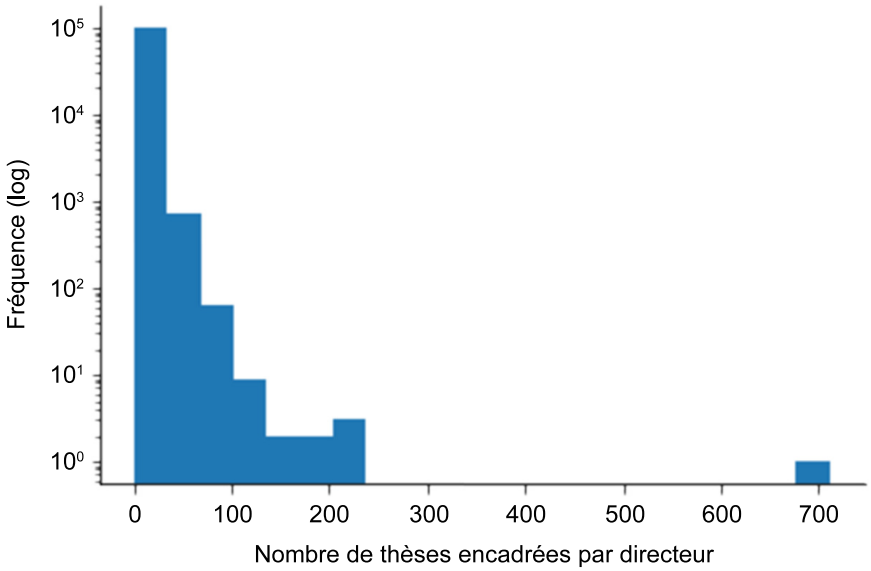

Figure 11. Distribution du nombre de thèses encadrées par directeur, parmi celles soutenues entre 1984 et 2018. La fréquence est exprimée en logarithme décimal.

biaiser l'analyse. Toutes les lignes suivantes indiquent des noms de directeurs réels. Il n'est donc pas possible, sur la seule base de ces données, de déterminer si le nombre élevé de thèses encadrées par ces directeurs reflète ou non une erreur dans le jeu de données.

Tableau 3. Nombre de thèses encadrées par les dix directeurs les plus prolifiques.

Directeurs	Nombre de thèses
Directeur de thèse inconnu	711
Francois-Paul Blanc	227
Jean-Michel Scherrmann	209
Pierre Brunel	206
Guy Pujolle	196
Michel Bertucat	173
Bernard Teyssie	140
Henry de Lumley	139
Michel Maffesoli	136
Bruno Foucart	135

La section suivante présente une analyse plus détaillée des informations disponibles sur François-Paul Blanc, le directeur ayant encadré le plus de thèses (227 entre 1984 et 2018).

4.2 Enquête sur le directeur le plus prolifique : François-Paul Blanc

Le Tableau 4 regroupe les principales informations sur les thèses encadrées par François-Paul Blanc entre 1984 et 2018. La première information concerne l'identifiant directeur : celui-ci est le même pour toutes les thèses dont Mr Blanc était le seul directeur, mais varie lorsque plusieurs directeurs ont co-encadré la thèse. Au vu des différences dans leur syntaxe, ces identifiants de co-directeurs semblent erronés et ne peuvent donc pas permettre d'identifier des personnes différentes.

Tableau 4. Description des thèses encadrées par François-Paul Blanc entre 1984 et 2018. Pour l'auteur, le nombre de thèses indique le nombre de doctorants différents. "ID" : identifiant.

Variable	Valeur	Nombre de thèses
ID directeur (direction solo)	26730774	201
ID directeur (co-direction)	267,307,740	11
	267,307,741	7
	26730774	4
	112,501,095	1
	112,299,172	1
	3	1
	973,903,640	1
Établissement	Perpignan	227
Auteur	(Nombre de doctorants)	227
Langue	Français	220
	Français-Arabe	5
	Inconnue	2
Discipline	Droit	195
	Histoire du droit	29
	Sciences politiques	3

Selon le Tableau 4, toutes les thèses dirigées par François-Paul Blanc ont été soutenues dans le même établissement, à Perpignan. Elles ont toutes été réalisées par des doctorants différents, ce qui réfute l'hypothèse de doublons lors de la saisie des données. De plus, la quasi-totalité des thèses ont été rédigées dans une même langue, le français, et 5 d'entre elles sont des thèses bilingues français-arabe. Enfin, toutes les thèses relèvent des sciences humaines et sociales (droit, histoire du droit et sciences politiques ; Tableau 4). Toutes ces informations tendent à confirmer l'hypothèse que François-Paul Blanc est bien une

seule et même personne et qu'il a bel et bien encadré 227 thèses entre 1984 et 2018.

L'évolution temporelle du taux d'encadrement de François-Paul Blanc renforce cette hypothèse. Cette évolution est représentée dans la Figure 12, en fonction de l'année de soutenance. Le taux d'encadrement tourne autour d'une à deux thèses par an entre 1995 et 1999, puis augmente progressivement jusqu'en 2004, année où Mr Blanc a encadré le plus de thèses (35 au total). Le nombre de thèses encadrées reste supérieur à 20 thèses par an jusqu'en 2009, puis devient nul en 2011. Dans l'ensemble, ce taux d'encadrement annuel semble réaliste : il reflète une carrière professionnelle assez classique en recherche, avec une augmentation progressive de la notoriété et des responsabilités de François-Paul Blanc dans son domaine d'expertise. Toutefois, un encadrement de 20 à 35 thèses par an sur six ans reste très élevé et soulève la question du temps que Mr Blanc a réellement pu investir dans l'encadrement de chacune de ces thèses.

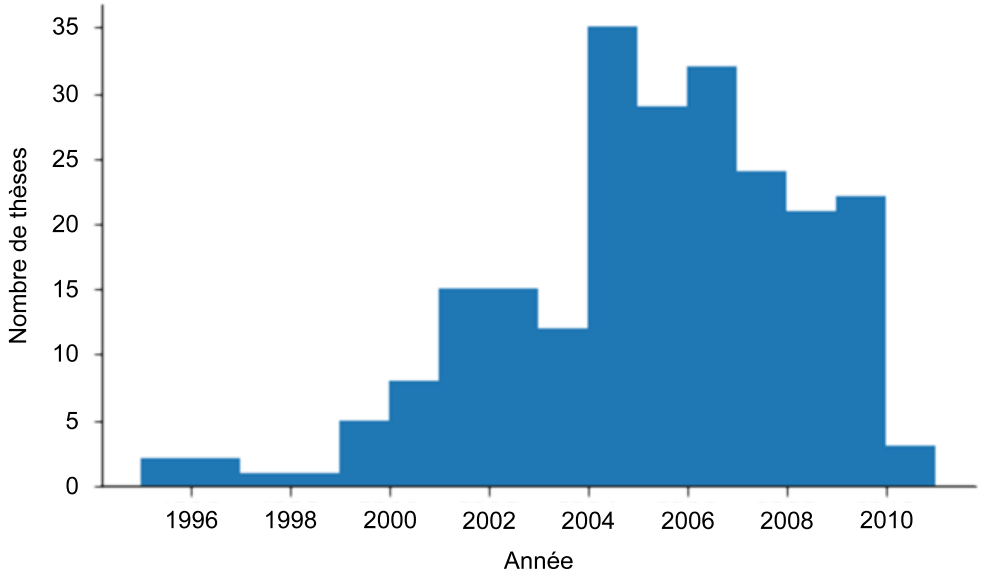

Figure 12. Nombre de thèses encadrées par François-Paul Blanc en fonction de l'année de soutenance.

Au-delà de l'identification d'erreurs, les informations sur les directeurs les plus prolifiques soulèvent plusieurs questions. Par exemple, quelle est la discipline la plus représentée dans les thèses encadrées par ces directeurs ? Les sciences humaines et sociales prédominent-elles sur les sciences dures ? Par ailleurs, les 9 directeurs ayant encadré le plus de thèses sont tous des hommes (Tableau 3) : ce biais dans le sex ratio est-il généralisable à toutes les thèses ? Des analyses plus approfondies sont nécessaires pour répondre à ces questions.

5 Résultats préliminaires

Une fois passé l'étape de nettoyage, les données issues de theses.fr offrent de nombreuses possibilités d'analyse. Le choix de ces analyses va dépendre des problématiques auxquelles

l'analyste souhaite répondre. Par exemple, quelles sont les langues les plus souvent choisies pour rédiger la thèse ? Comment le choix de la langue a-t-il varié au cours du temps ?

Pour répondre à ces deux questions, l'évolution du choix de la langue de la thèse a été analysée au fil des ans. Elle est représentée dans la Figure 13, sur la période 1985-2018. Cette figure montre que le français est la langue prédominante quelle que soit l'année, représentant plus de 97% des thèses soutenues dans les années 1985 à 2000. Sa proportion commence toutefois à diminuer en 2000. A partir de cette date, la part de la langue anglaise augmente progressivement jusqu'en 2018, passant de moins de 1% à plus de 28%, ce qui en fait alors la deuxième langue la plus choisie. Le pourcentage de thèses bilingues tourne autour de 2% entre 1985 et 2000, puis augmente légèrement (entre 3% et 7.5% jusqu'en 2018). Les autres langues sont quant à elles largement sous-représentées (moins de 0.2% avant 2000, entre 0.6% et 1.7% après).

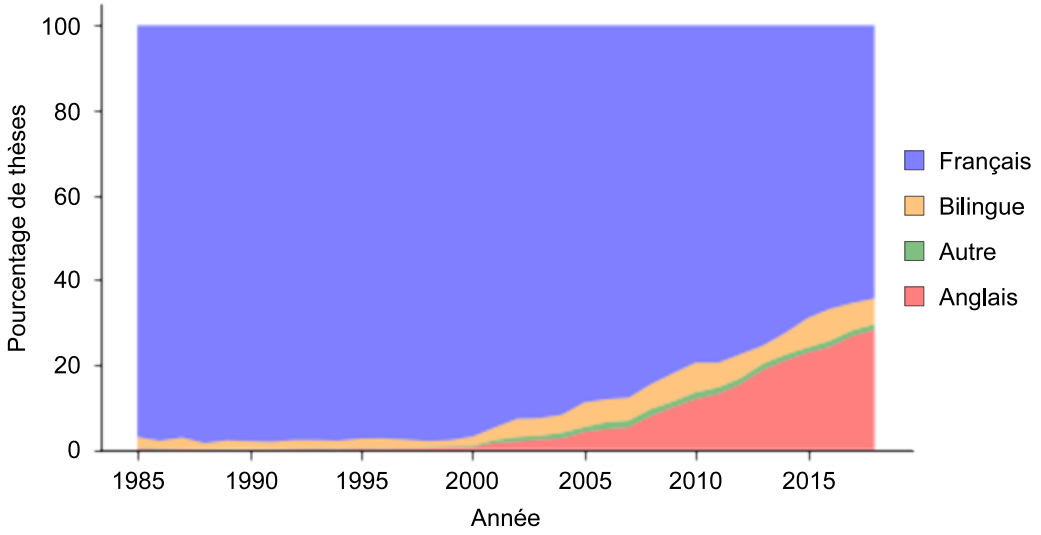

Figure 13. Pourcentage de thèses en fonction de l'année de soutenance (entre 1985 et 2018) et de la langue de la thèse (bleu : français ; orange : bilingue anglais-français ; vert : langues autres que l'anglais et le français ; rouge : anglais).

Une interprétation possible de l'augmentation du choix de la langue anglaise est qu'il y aurait eu de plus en plus d'étudiants étrangers venus faire leur thèse en France à partir de 2000. Il se pourrait aussi que cette augmentation soit liée au mode de diffusion de la thèse : selon Martin (2015), les thèses ont été de plus en plus diffusées sous forme numérique ces deux dernières décennies, ce qui facilite grandement leur accès par rapport au format papier, y compris depuis l'étranger. De plus en plus d'étudiants français pourraient donc avoir choisi de rédiger leur thèse en anglais, afin d'augmenter sa visibilité à l'échelle internationale. Une analyse de la corrélation entre la langue de la thèse et son accessibilité en ligne permettrait de donner une première indication sur la vraisemblance de cette hypothèse. Toutefois, si cette corrélation existe bien, elle pourrait aussi vouloir dire que seule la langue des thèses dont le manuscrit est en libre accès peut être déterminée avec précision. Dans ce cas, les données concernant toutes les autres thèses nécessiteraient une vérification avant de pouvoir être interprétées.