Manipulation et pré-traitement de données

Marion Mauran

Diplôme Universitaire DATA ANALYST

2024 - 2025

Résumé

Le nettoyage d'un jeu de données relatives aux soutenances de thèses entre 1984 et 2018 permet de mettre en lumière les problématiques d'homonymie et d'*outliers*, en se focalisant sur le cas des auteurs et des directeurs de thèses notamment. En ce qui concerne les auteurs de thèses, on peut relever des cas d'homonymie plutôt que des cas de thèses multiples en se focalisant sur l'exemple d'un auteur. *A contrario*, l'étude de la situation d'un directeur de thèses montre que ceux-ci peuvent être amenés à encadrer un grand nombre de soutenances.

Par ailleurs, l'analyse des données met en évidence une tendance croissante à l'utilisation de l'anglais dans les thèses depuis 2000, phénomène attribué à la mondialisation de la recherche et à la visibilité accrue des thèses en ligne. Ces résultats mettent en exergue l'importance de la qualité des données et de l'adaptation à des audiences internationales dans le domaine académique.

Summary

The cleaning of a dataset relating to thesis defenses between 1984 and 2018 sheds light on the problems of homonymy and outliers, focusing on the case of thesis authors and directors in particular. As far as thesis authors are concerned, we can identify cases of homonymy rather than multiple theses, by focusing on the example of one author. *A contrario*, the study of the situation of a thesis director shows that they may be required to supervise a large number of thesis defenses.

Analysis of the data also reveals a growing trend towards the use of English in theses since 2000, a phenomenon attributed to the globalization of research and the increased visibility of online theses. These results highlight the importance of data quality and adaptation to international audiences in the academic field.

Table des matières

1 Présentation des données

1.1 Présentation de la base de données

Ce jeu de données publiques contient les données des thèses de doctorat françaises soutenues entre 1984 et 2018. Les données sont saisies par les établissements habilités à délivrer le doctorat, de manière à répondre aux obligations réglementaires. Enfin l'ensemble des données recueillies est diffusé sur le moteur de recherche *theses.fr*.

L'objectif poursuivi à travers ce rapport est de nettoyer cette base, d'étudier la question des données manquantes et d'identifier des problèmes associés au jeu de données. A terme, cela permettra de tirer des conclusions sur les données analysées.

1.2 Taille et variables de la base de données

La base de données compte 23 colonnes et 448 047 lignes.

Les variables ainsi utilisées permettent de recenser les données suivantes (de manière globalisée) :

— le nom de l'auteur de la thèse, son genre et son identifiant,
— le titre de la thèse et son identifiant,
— le directeur de thèse et son identifiant,
— le nom de l'établissement de soutenance et son identifiant,
— la discipline concernée,
— le statut de la thèse (soutenue ou en cours),
— la date de première inscription en doctorat,
— la date de soutenance (comprenant l'année),
— la langue de la thèse,
— l'accessibilité de la thèse en ligne,
— la date de mise à jour et de publication dans *theses.fr*.

1.3 Typologie des données utilisées

Les variables rencontrées peuvent être de différents types :

— qualitatives (ne prennent pas de valeurs numériques),
— quantitatives (prennent des valeurs numériques et peuvent être mesurées).

Parmi les variables qualitatives, il existe plusieurs sous-types :

— nominales (n'ont pas d'ordre spécifique),
— ordinales (ont un ordre naturel).

Parmi les variables quantitatives, il existe plusieurs sous-types :

— discrètes (prennent des valeurs distinctes et souvent limitées),
— continues (prennent n'importe quelle valeur dans un intervalle donné).

La synthèse des typologies de variables est présentée dans la Table 1.

En analysant la Table 1, on s'aperçoit qu'il existe une majorité de données qualitatives nominales, liées :

Nom de variable	Type de variable
Auteur	qualitative nominale
Identifiant auteur	qualitative ordinale
Titre	qualitative nominale
Directeur de thèse	qualitative nominale
Directeur de thèse (nom et prénom)	qualitative nominale
Identifiant directeur	qualitative ordinale
Etablissement de soutenance	qualitative nominale
Identifiant établissement	qualitative ordinale
Discipline	qualitative nominale
Statut	qualitative nominale
Date de première inscription en doctorat	quantitative continue
Date de soutenance	quantitative continue
Year	quantitative continue
Langue de la thèse	qualitative nominale
Identifiant de la thèse	qualitative ordinale
Accessible en ligne	qualitative nominale
Publication dans theses.fr	quantitative continue
Mise à jour dans theses.fr	quantitative continue
Discipline - prédi	qualitative nominale
Genre	qualitative nominale
Etablissement - rec	qualitative nominale
Langue - rec	qualitative nominale

TABLE 1 – Noms et types de variables rencontrées dans la base de données

— aux noms d'auteurs,
— aux noms de directeurs,
— aux noms d'établissements,
— aux titres de thèses,
— aux noms de disciplines,
— au genre,
— au statut de la thèse,
— aux langues d'écriture,
— à l'accessibilité en ligne.

Par ailleurs, les identifiants associés apparaissent comme des données qualitatives ordinales :

— identifiants des auteurs,
— identifiants des établissements,
— identifiants des directeurs,
— identifiants des thèses.

Enfin, un certain nombre de valeurs sont rattachées à des variables quantitatives continues, ce sont principalement les dates :

— date de première inscription en doctorat,
— date de soutenance,
— date de publication dans *theses.fr*,
— date de mise à jour dans *theses.fr*.

Pour les variables qualitatives, il est nécessaire d'avoir un aperçu du nombre de valeurs non-nulles et uniques pour chacune des variables (voir 2).

D'après la Table 2, il ressort que certaines variables semblent avoir des valeurs nulles, comme par exemple :

— identifiant auteur,
— identifiant établissement,
— établissement - rec,
— langue - rec.

Il est également à noter que toutes les variables présentent des données redondantes.

Ainsi, il apparaît nécessaire de pouvoir quantifier les proportions de valeurs manquantes ou aberrantes de la base de données.

Variables qualitatives	Nombre de valeurs non nulles	Nombre de valeurs uniques
Auteur	448047	430273
Identifiant auteur	317700	313771
Titre	448040	446816
Directeur de these	448034	159019
Directeur de these (nom et prénom)	448034	159021
Identifiant directeur	448047	98906
Etablissement de soutenance	448046	567
Identifiant etablissement	430965	572
Statut	448047	2
Discipline	448047	24262
Langue de la these	448047	206
Identifiant de la thèse	448047	447567
Accessible en ligne	448047	2
Discipline - prédi	448047	15
Genre	448047	6
Etablissement - rec	444973	110
Langue - rec	383927	4

TABLE 2 – Nombre de valeurs non nulles et uniques par variables qualitatives présentes dans la base de données

2 Données manquantes

2.1 Matrice de nullité

Il s'agit de déterminer le taux de valeurs manquantes pour chaque variable afin de pouvoir s'assurer de la qualité des données présentes dans la base.

Les résultats sont présentés dans la Figure 2, qui est une représentation graphique des emplacements de valeurs manquantes dans l'ensemble de données.

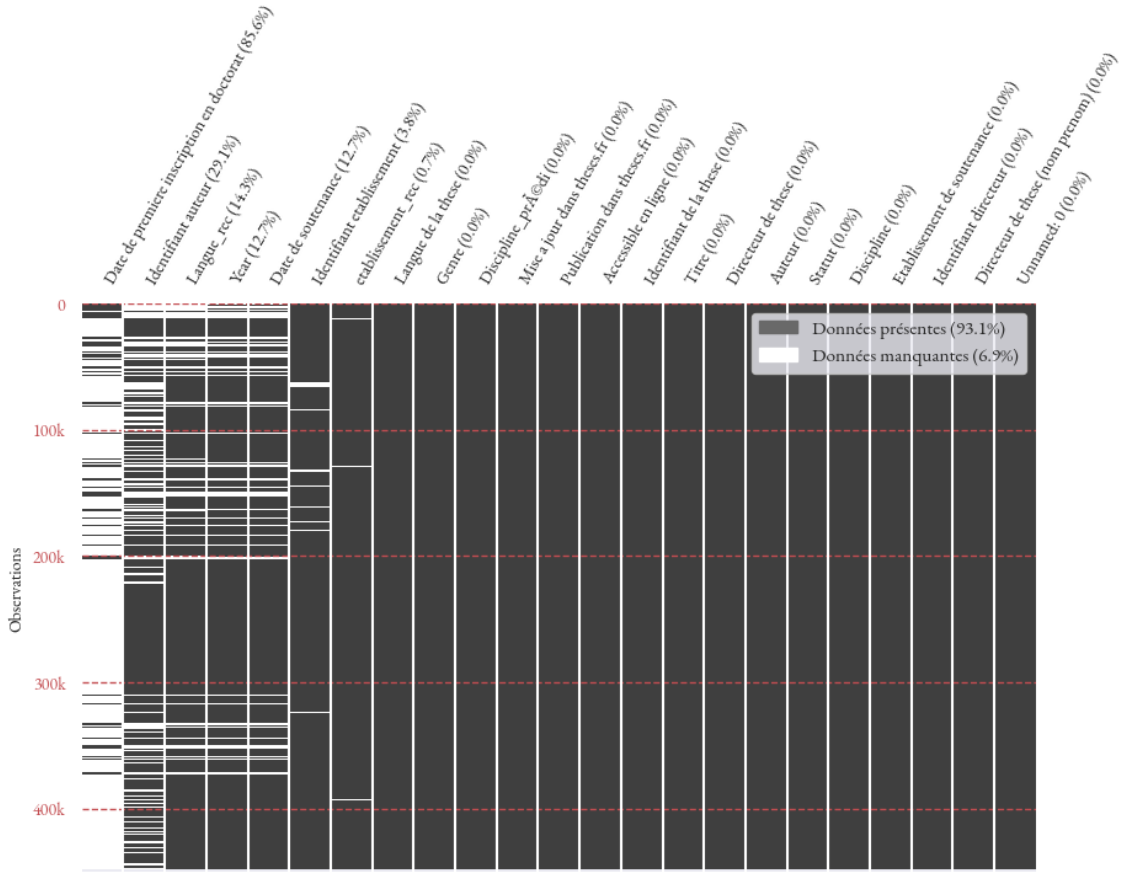

FIGURE 2 – Représentation des données manquantes par variables (matrice de nullité)

L'analyse de la Table 2 montre que 86 % des données de la variable "date de première inscription en doctorat" sont manquantes.

D'autres variables présentent entre 10 et 30 % de données manquantes. Il s'agit des variables suivantes :

— identifiant de l'auteur,
— langue - rec,
— année de soutenance,

— date de soutenance.

Enfin, d'après la matrice de nullité (Figure 2), on peut souligner que deux variables présentent un faible taux de données manquantes :

— identifiant établissement,
— établissement - rec.

Les autres variables présentent un taux de nullité proche de zéro :

— langue de la thèse,
— genre,
— discipline - prédi,
— mise à jour dans theses.fr,
— publication dans theses.fr,
— accessible en ligne,
— identifiant de la thèse,
— titre,
— directeur de thèse,
— auteur,
— statut,
— discipline,
— établissement de soutenance,
— identifiant directeur,
— directeur de thèse.

Cela permet de conclure que la qualité des données de la base est suffisante pour construire l'analyse. En effet, on observe un taux de données manquantes total de 6,9 %.

2.2 Carte thermique des absences en fonction du statut de la thèse

À ce stade, l'objectif est de se concentrer sur une variable spécifique afin d'examiner sa relation avec les autres en termes d'absence de données. Il est pertinent d'approfondir l'analyse des valeurs manquantes pour les variables concernées en fonction du statut de la thèse ("soutenue" ou "en cours") par exemple.

Les variables retenues pour l'analyse des taux d'absences sont celles qui ont un taux de nullité différent de zéro :

— date de première inscription en doctorat,
— identifiant auteur,
— langue - rec,
— year,
— date de soutenance,
— identifiant établissement,
— établissement - rec.

Les taux d'absences des variables en fonction du statut de la thèse sont affichées dans la Figure 3.

Cette carte thermique permet de mettre en évidence les variables pour lesquelles les valeurs manquantes coïncident au maximum, indiquant ainsi une forte relation entre ces variables en termes d'absences de données.

Les données présentant une forte similarité en termes d'absences apparaissent en bleu foncé, avec un taux proche de 1. À l'inverse, les variables ayant peu de lien en ce qui concerne l'absence de leurs données montrent un faible taux et des couleurs plus proches du jaune. La carte thermique (3) montre que l'absence de données sur

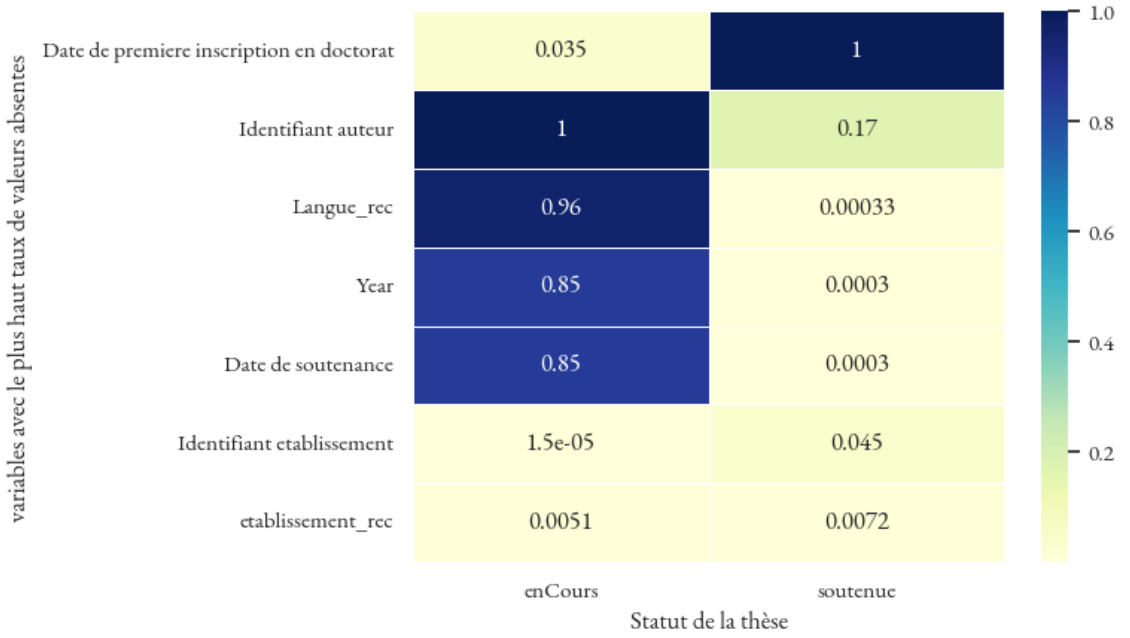

FIGURE 3 – Représentation des taux d'absences des différentes variables en fonction du statut de la thèse (carte thermique)

la variable "date de première inscription en thèse" est très fortement corrélée au statut de la thèse "soutenue".

Cela peut s'expliquer par un changement dans le renseignement de la base de données. Ce changement pourrait ainsi ne pas affecter les thèses plus récentes dont le statut est "en cours".

Par ailleurs, il est aussi notable que l'absence de données au niveau des variables suivantes :

— identifiant auteur,
— langue - rec,
— year,
— date de soutenance,

est très fortement liée à l'absence de données lorsque la thèse est "en cours".

Cela pourrait être lié aux processus administratifs établis, tels que l'octroi d'un "identifiant auteur" uniquement lorsque la thèse est déjà soutenue.

Il en est de même pour la langue de rédaction de la thèse qui peut ne pas être définie ou déclarée tant que la thèse n'est pas achevée ou proche de l'achèvement.

La date de soutenance semble naturellement absente pour les thèses en cours car la soutenance n'a pas encore eu lieu, et l'année de soutenance ne peut être connue que lorsque la date de soutenance est fixée.

2.3 Étude des corrélations entre données manquantes

Cette analyse centrée sur le statut de la soutenance, permet de s'interroger sur les corrélations d'absences entre variables à plus large échelle.

Sur la Figure 4, la visualisation des corrélations d'absences filtrée sur la première variable rend le phénomène plus visuel.

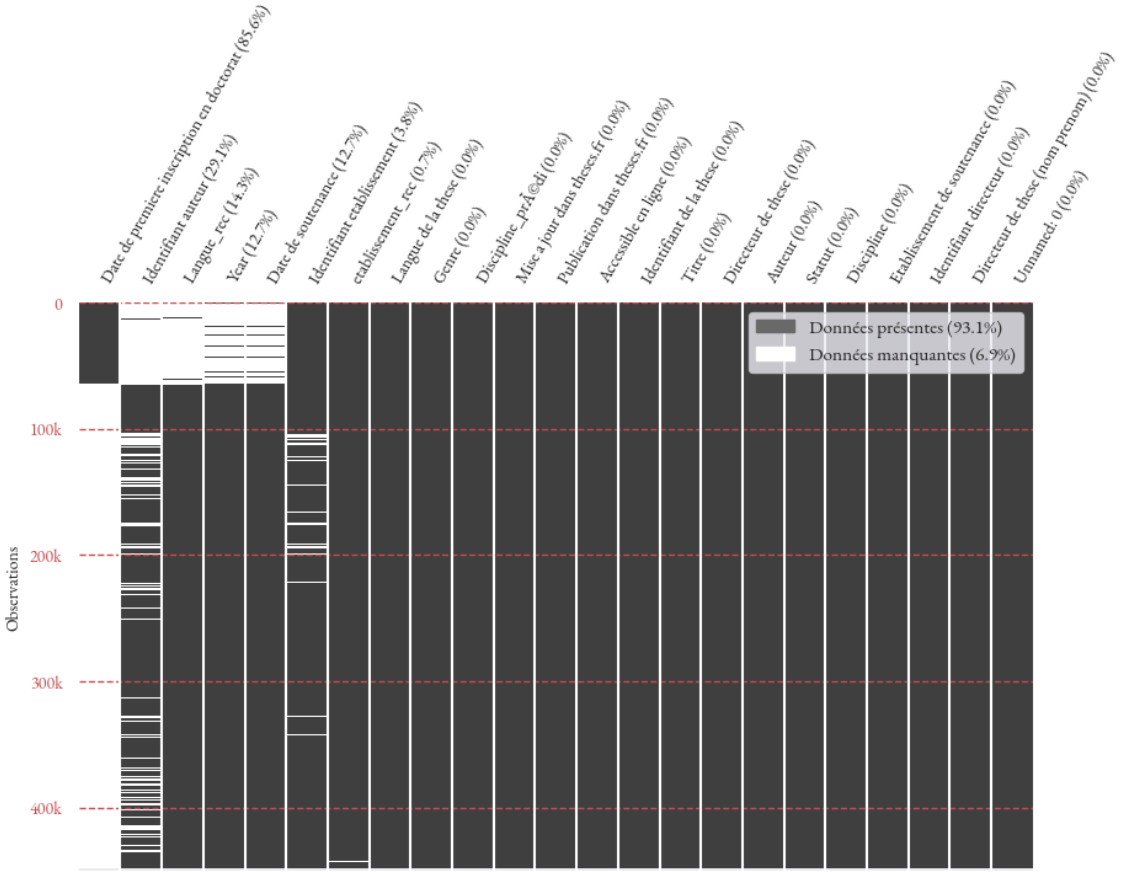

FIGURE 4 – Représentation des données manquantes par variables, filtrées sur la "date de première inscription en doctorat"

2.3.1 Différents types de corrélations

En analysant la Figure 4, on peut s'apercevoir de deux types de corrélations :
— des corrélations positives (lorsque les valeurs manquantes pour une variable coïncident avec les valeurs manquantes pour une autre variable),
— des corrélations négatives (lorsque la présence de valeurs manquantes pour une variable est associée à l'absence de données chez une autre variable).

On retrouve ainsi des corrélations positives entre :
— identifiant auteur,
— langue - rec,
— year,
— date de soutenance.

On retrouve des corrélations négatives entre "date de première inscription en doctorat" et :
— identifiant auteur,
— langue - rec,
— year,
— date de soutenance.

D'autres corrélations négatives peuvent être relevées avec les variables ci-dessous :
— identifiant établissement,
— établissement - rec,

mais le taux d'absences de données étant inférieur à 5 %, cela peut ne pas être révélateur d'une réelle piste d'analyse.

2.3.2 Différentes forces de corrélations

La Figure 5 nous permet d'étudier la force des corrélations entre données manquantes (qu'elles soient positives ou négatives).

Il y a confirmation des corrélations entre absences que nous avons détectées précédemment :
— "date de soutenance" et "year" ont les mêmes absences,
— les absences de "langue - rec" sont également très proches de "date de soutenance" et "year".

Il semble qu'il y ait également une forte corrélation (négative) entre ces variables et :
— établissement - rec,
— identifiant établissement.

Cependant, ces deux variables ayant un faible taux de données absentes, il ne parait pas pertinent à ce stade de les maintenir pour l'analyse.

Le noeud de de jonction de la branche "identifiant auteur" est également relativement proche du haut du graphique, ce qui montre un certain niveau de proximité avec les absences des variables :
— date de soutenance et year,

— langue - rec.

La "date de première inscription en doctorat" présente aussi une corrélation (négative) avec :

— identifiant auteur,
— date de soutenance,
— langue - rec.

Ces observations concordent avec les analyses effectuées sur la carte thermique des absences.

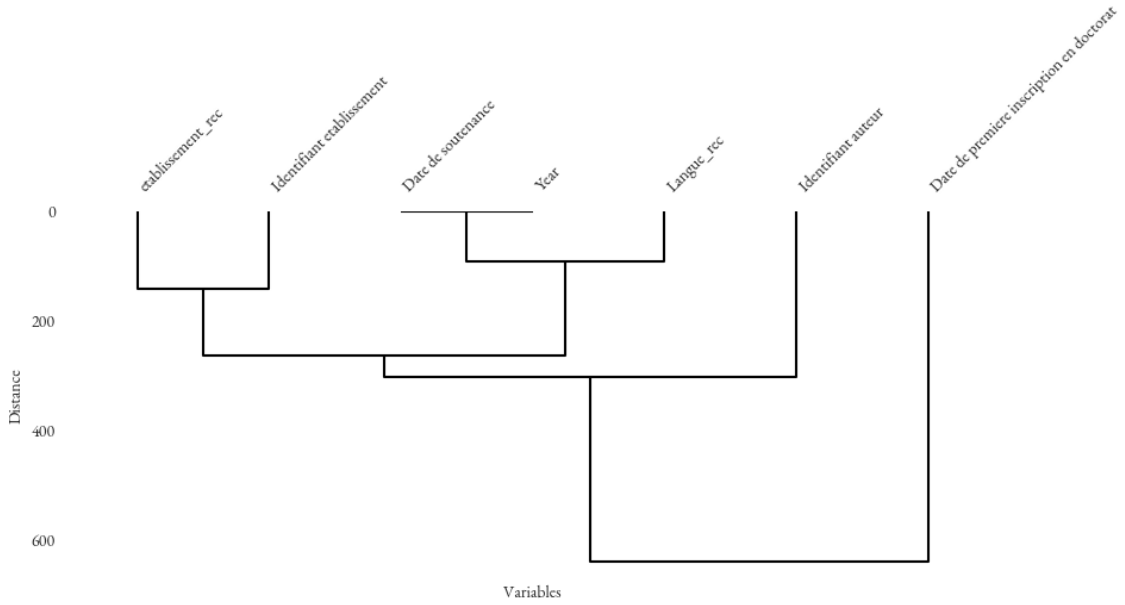

FIGURE 5 – Représentation des données manquantes par niveau de proximité (dendrogramme)

En conclusion, les hypothèses pouvant être formulées à ce stade sont les suivantes :

— "year" est un extrait de la "date de soutenance",
— il existe un lien de corrélation élevé positif entre les absences des données des variables "date de soutenance", "year", "langue - rec" et "identifiant auteur",
— les données absentes pour ces variables sont étroitement liées au statut de la thèse "en cours",
— il existe un lien de corrélation négatif entre les données manquantes pour ces variables et les données absentes pour la variable "Date de première inscription en doctorat",
— il existe un lien de corrélation très élevé entre l'absence de données pour la variable "Date de première inscription en doctorat" et le statut de la thèse "soutenue".

Cela pourrait s'expliquer par les procédures administratives :

— certaines données ne sont pas accessibles tant que la thèse n'a pas eu lieu,
— la requête de la date de première inscription en doctorat n'était peut être pas renseignée pour les thèses qui ont été soutenues, alors que pour celles en cours (plus récentes), elle l'est (changement de processus administratifs).

3 Principaux problèmes détectés

3.1 Détection de données aberrantes dans les dates de soutenance

3.1.1 Nombre de soutenances par mois

La Figure 6 permet de mettre en lumière la distribution des soutenances en fonction des mois de l'année. Les données sont prises en compte entre 1984 et 2018, dernière année de collecte des données de la base.

Le choix de 2018 comme dernière année retenue pour la représentation des données permet également d'éviter la visualisation de données aberrantes sur l'intervalle de collecte (*exemple* : une année saisie non écoulée comme 2075).

En effet, il peut y avoir des anomalies de saisies de dates mais aussi des retards dans les traitement des thèses plus récentes, ce qui peut diminuer la qualité des valeurs affichées.

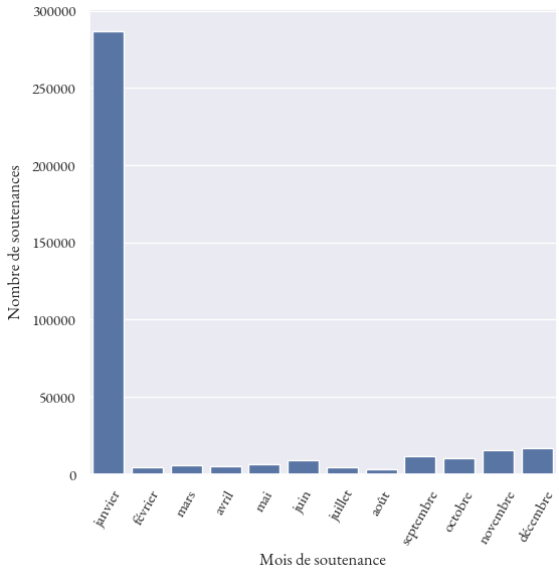

FIGURE 6 – Distribution des soutenances par mois, sur la période 1984 - 2018

A la lecture de la Figure 6 , on peut observer un nombre anormalement élevé de soutenances en janvier.

Cela peut être dû au fait que certaines données ont été mises à jour soit *a posteriori*, soit de manière imprécise ou automatique en validant la date du 1er janvier en lieu et place de la date réelle. Cela peut être en lien avec le paramétrage de l'application de recueil qui remplit ce champ de manière automatique (en fonction des données accessibles), lorsque celui-ci n'est pas complété par l'utilisateur.

3.1.2 Nombre de soutenances par mois, en fonction des années

La Figure 7 nous permet de visualiser de manière plus précise l'évolution du nombre de soutenances mensuelles au fil des ans, entre 2005 et 2018.

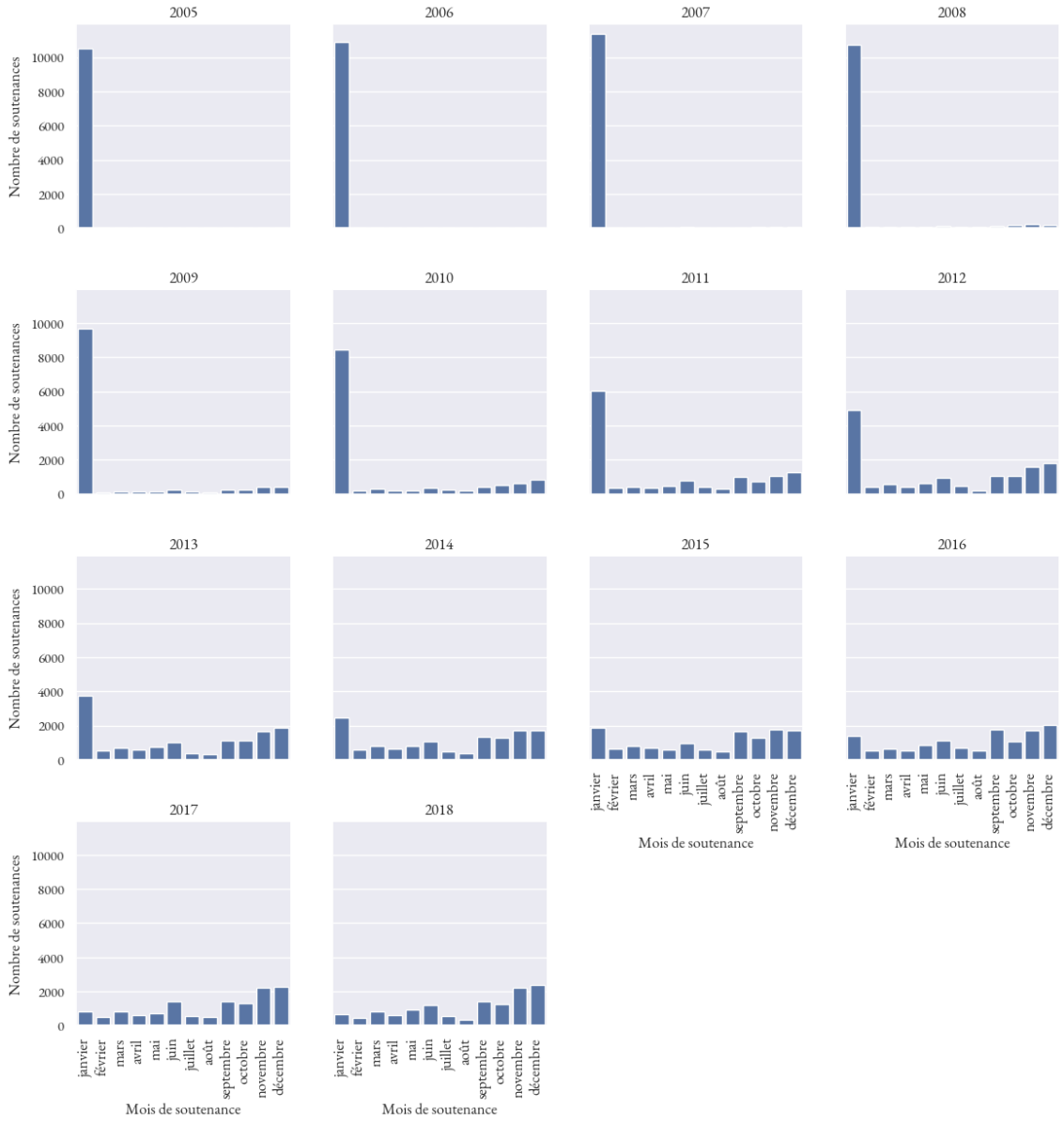

FIGURE 7 – Distribution des soutenances de thèses en fonction des mois, au fil des années, entre 2005 et 2018

Sur ce graphique, on peut observer que la distribution des thèses soutenues en janvier diminue au fil du temps, entre 2005 et 2018. Cette évolution peut être divisée en 3 périodes :

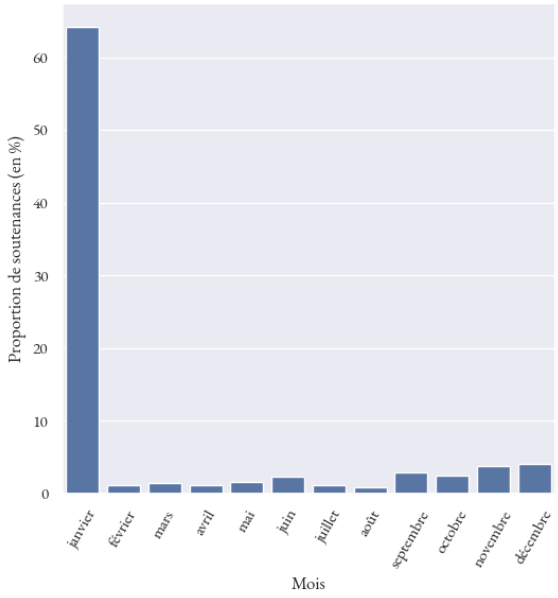

FIGURE 8 – Proportion de soutenances par mois, entre 1984 et 2018

— entre 2005 et 2007, les soutenances de thèses sont toutes enregistrées en janvier,
— entre 2008 et 2014, le nombre de soutenances enregistrées en janvier diminue fortement, tout en restant supérieur au nombre de soutenances des autres mois,
— entre 2015 et 2018, le nombre de soutenances enregistrées en janvier diminue progressivement, tout en étant inférieur au nombre de soutenances d'autres mois de l'année.

3.1.3 Proportion de soutenences par mois

La Figure 8 met en évidence la proportion de thèses soutenues en fonction des mois entre 1984 et 2018.

Il ressort ainsi nettement que plus de 60 % des thèses sont soutenues en janvier sur la période sélectionnée.

3.1.4 Proportions de soutenances le 1er janvier

La Figure 9 permet quant à elle de préciser l'évolution de la proportion de thèses soutenues au 1er janvier en fonction des années, entre 1984 et 2018.

Deux tendances apparaissent :

— entre 1984 et 2005, les thèses sont toutes soutenues le 1er janvier,
— entre 2005 et 2017, la proportion de thèses soutenues le 1er janvier chute violemment, pour atteindre une valeur proche de 0 en 2018.

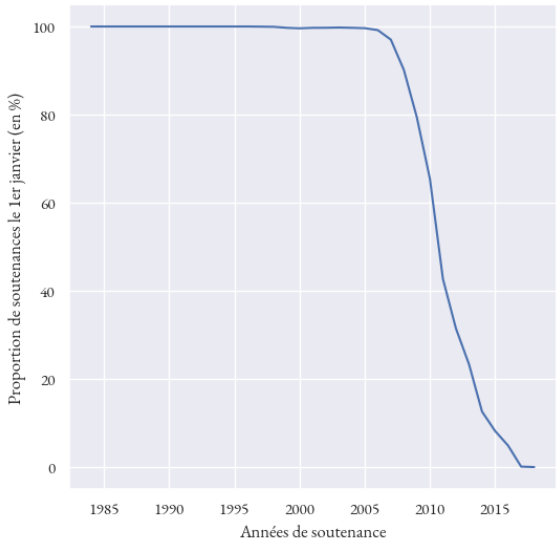

FIGURE 9 – Proportion de soutenances le 1er janvier, entre 1984 et 2018

3.1.5 Proportions de soutenances par mois, pour les soutenances en dehors du 1er janvier

La donnée temporelle "01-01 de l'année X" semble être une donnée aberrante, qui peut être due par exemple à des rattrapages ou à des corrections de saisies comme vu précédemment. Il convient ainsi d'analyser les données en la supprimant. La Figure 10 nous donne ainsi un aperçu plus fidèle de la réalité.

On observe au final une préférence pour des dates de soutenances en fin d'année, et notamment au mois de décembre. Les mois de juillet et août sont au contraire les moins attractifs.

3.2 Cas d'homonymies chez les noms d'auteurs de thèses

Afin de poursuivre l'a démarche de nettoyage des données, il semble nécessaire de s'attacher à l'étude des noms d'auteurs de thèses, afin de déterminer la présence ou non d'homonymies.

Pour analyser les cas d'homonymies chez les auteurs de thèses, un tri rapide des valeurs les plus courantes montre un nombre anormalement élevé de thèses pour des auteurs portant le nom de Martin. Par exemple :

— Nicolas Martin a soutenu 16 thèses,
— Franck Martin, 12 thèses,
— Philippe Martin, 12 thèses,
— Cécile Martin, 7 thèses,
— etc.

Focalisons-nous sur le cas de Cécile Martin. La Table 3 permet d'orienter l'ana-

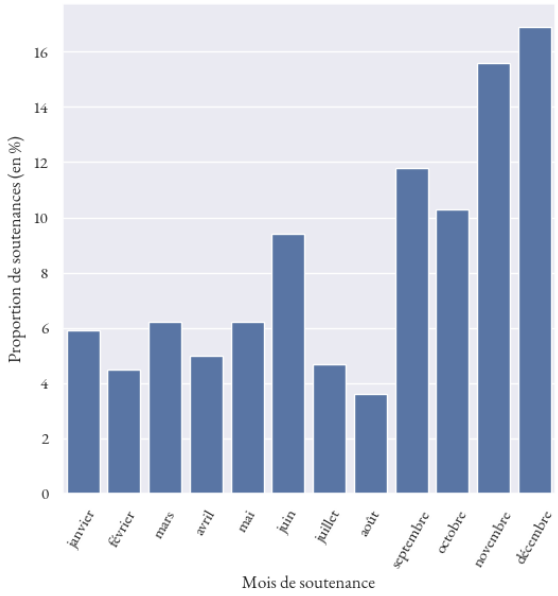

FIGURE 10 – Proportion de soutenances en fonction du mois de l'année, sans les soutenances enregistrées au 1er janvier, entre 1984 et 2018

lyse :

— soit vers un cas d'homonymie,
— soit vers un cas de publications multiples.

3.2.1 Données non utiles à l'analyse

Une sélection des variables à prendre en compte pour l'analyse est nécessaire afin de mener à bien cet objectif.

Les variables redondantes n'apportent pas d'élément à prendre en compte pour l'analyse du cas :

— year,
— discipline.

Les variables sans données n'apportent aucun élément à la réflexion portée sur le cas d'homonymie de Cécile Martin :

— date de première inscription en doctorat.

Les variables suivantes ne donnent pas d'information sur le caractère d'homonymie de Cécile Martin :

— accessible en ligne,
— statut ("en cours" ou "soutenue"),

Les variables qui ne comptent que des valeurs distinctes ne sont pas à prendre en compte pour l'analyse :

— directeur de thèses,
— code directeur,

Identifiant auteur	Publication dans theses.fr	Mise à jour dans theses.fr	Discipline	Etablissement de soutenance
203208145	26-09-11	03-10-17	Cinéma et audiovisuel	Sorbonne Paris Cité
179423568	26-09-11	05-12-17	Economie gestion	Paris 9
182118703	26-09-11	07-07-20	Matériaux, Milieux et chimie	Paris 11
81323557	24-05-13	08-07-20	Sciences de l'ingénieur	Compiègne
81323557	24-05-13	07-07-20	Psychologie	Clermont-Ferrand 2
81323557	24-05-13	07-07-20	Biologie	Bordeaux 2
81323557	08-07-17	10-12-19	Biologie	Institut National agronomique Paris-Grignon

TABLE 3 – Synthèses de données relatives à une sélection de variables permettant de statuer sur un cas d'homonymie entre auteurs de thèses (Cécile Martin)

— date de soutenance,
— identifiant de la thèse,
— établissement soutenance,
— identifiant établissement de soutenance.

Les variables dont les valeurs sont toujours les mêmes ne permettent pas non plus d'ajouter une plus-value à l'analyse du cas de Cécile Martin :

— langue de la thèse,
— genre.

3.2.2 Données utiles à l'analyse

Les variables qui comptent des données redondantes :

— date de mise à jour dans theses.fr,
— publication dans theses.fr.,
— identifiant auteur,
— discipline - prédi.

L'établissement de soutenance peut toutefois, en fonction des cas, apporter des éléments de localisation géographique et le titre de la thèse, des pistes de discrimination supplémentaires. Ils seront consultés en fonction du besoin.

A la lecture de la Table 3, il ressort qu'il existe 4 identifiants différents pour les 7 cas de soutenances enregistrés pour l'auteur "Cécile Martin". *A priori*, il y aurait

donc *a minima* 4 auteurs différents.

3.2.3 Cas d'homonymie ou publications multiples ?

Intéressons-nous aux cas d'auteurs dont l'identifiant est le même (81323557). On remarque que trois d'entre eux ont été publiés le même jour dans *theses.fr*, il pourrait ainsi s'agir d'un problème de recopie de données dans la base, d'autant plus que les disciplines concernées sont très différentes (Sciences de l'ingénieur, Psychologie et Biologie).

Concernant la dernière thèse, enregistrée avec ce même "identifiant auteur" mais à une date de publication différente, on peut noter qu'elle concerne elle aussi la biologie. Toutefois, les thèses ont été soutenues dans des établissements différents, éloignés géographiquement :

— Bordeaux 2,
— Institut National agronomique Paris-Grignon.

Le fait que cette dernière thèse soit ultérieure aux précédentes a pu permettre de réutiliser l'identifiant (81323557) déjà créé dans la base.

Les titres de thèses sont les suivants :

— "Système laitier et filière lait au Mexique : contraintes de développement, stratégies d'acteurs, enjeux de leur coevolution.",
— "Caractérisation electrophysiologique et pharmacologique des canaux ioniques : sodium, calcium, actives par l'ATP, des cellules myometriales, effets de la gestation et de l'ocytocine."

En outre, les domaines de biologie concernent plutôt des spécialités différentes :

— les sciences biologiques fondamentales et appliquées, d'une part,
— les neurosciences, d'autre part.

Si l'on soustrait les valeurs aberrantes rencontrées sur l'"identifiant auteur", l'ensemble des éléments d'analyse nous invite à poser l'hypothèse suivante : les 7 publications de thèses sous l'auteur "Cécile Martin" concernent uniquement des cas d'homonymies et non des cas de publications multiples.

4 Outliers et résultats anormaux

4.1 Confrontation des variables

Après avoir examiné les homonymes parmi les noms d'auteurs, il serait pertinent de réaliser une analyse similaire pour les directeurs de thèse, certains d'entre eux affichant un nombre exceptionnellement élevé de thèses supervisées.

Il s'agit ici de détecter des *outliers* ou individus atypiques, c'est-à-dire des individus très différents des autres. Il est très important de les repérer car ils peuvent fortement biaiser les analyses sur un échantillon statistique, notamment s'ils sont erronés.

Ainsi, afin de ne pas fausser les données, les valeurs "directeur de thèse inconnu" sont retirées de la base.

D'après la Table 4, on observe que jusqu'à près de 200 thèses peuvent être encadrées par certains directeurs.

La base de données répertorie environ 129 091 directeurs de thèse différents, avec une proportion de 2,8 % ayant des identifiants distincts. Ce classement des identifiants de directeurs suggère que certains "identifiants directeurs" pourraient être erronés ou manquants, comme l'indique la Table 2.

D'après ces premiers résultats, la question est de savoir si le nombre anormalement élevé de thèses encadrées par des directeurs s'explique par :

— des cas d'homonymies (*cf.* cas de Cécile Martin),
— des cas fidèles à la réalité, qui montrent que certains directeurs encadrent un nombre anormalement élevé des thèses.

En effet, selon la Table 5, 24,3 % des identifiants directeurs sont manquants ou aberrants.

Afin de savoir si l'encadrement d'un grand nombre de thèses par un directeur est possible, il est nécessaire de se focaliser sur un cas spécifique, le cas de Jean-Michel Scherrmann qui est le directeur ayant encadré le plus grand nombre de thèses au

Directeur de these	Directeur de these (nom, prénom)	Nombre de thèses	Nombre d'identifiants directeurs
Jean-Michel Scherrmann	Scherrmann Jean-Michel	208	1
Francois-Paul Blanc	Blanc Francois-Paul	201	1
Pierre Brunel	Brunel Pierre	195	3
Michel Bertucat	Bertucat Michel	173	1
Guy Pujolle	Pujolle Guy	172	1

TABLE 4 – Extrait du classement du nombre de thèses encadrées par directeurs par ordre décroissant

Identifiants	Nombre
manquant	18432
1	875
3	658
7	622
8	521
6	486
2	423
9	239
59375140	208
26730774	205

TABLE 5 – Extrait des "identifiants directeurs" les plus représentés dans la base et leur nombre respectif

Variables	Données uniques associées
Discipline prédi	Biologie, Psychologie, Medecine
Établissement de soutenance	Paris 5, Paris 6
Identifiants établissements	26404788, 27787087
Établissements - rec	Université de Paris, Sorbonne Université

TABLE 6 – Extrait des valeurs uniques de variables associées aux soutenances de thèses encadrées par Jean-Michel Scherrmann

sein de la base de données.

4.2 Le cas Jean-Michel Scherrmann

Comme ce directeur ne possède qu'un seul numéro identifiant (59375140), la confrontation de plusieurs variables permettra de répondre à la question suivante : "le professeur Scherrmann a-t-il réellement encadré 208 thèses ?".

Les variables conservées pour l'analyse sont les suivantes :

— discipline prédi,
— établissement de soutenance,
— identifiant établissement,
— établissements - rec.

Elles sont recensées dans la Table 6 .

En effet, les disciplines sont relatives au domaine de la santé :

— Biologie,

— Psychologie,
— Médecine.

Les établissements de soutenance sont localisés à Paris avec deux types d'identifiants :

— Paris 5 (26404788),
— Paris 6 (27787087).

L'hypothèse selon laquelle ce professeur a encadré un très grand nombre de thèses est confirmée par l'analyse des données présentes dans la base.

En synthèse, des données aberrantes ont été relevées, notamment en ce qui concerne les numéros d'identifiants (24,3 % de données erronées ou absentes environ) ou encore les noms de directeurs de thèses ("directeur de thèse inconnu"), mais cela ne remet pas en question la qualité de la base de données.

Des cas de directeurs de thèses avec plusieurs numéros identifiants ont été relevés mais ils ne concernent qu'un faible pourcentage de la proportion totale des directeurs (2,8 %) et ce qui réduit la possibilité d'homonymie à un faible taux.

Enfin la confrontation de variables pour le cas d'un directeur ayant encadré un grand nombre de thèses ne montre pas de données contradictoires. Il semble donc que certains directeurs fassent le choix d'encadrer un grand nombre de thèses au cours de leur carrière (jusqu'à environ 200 contre une moyenne de 3 par directeurs).

5 Résultats préliminaires

5.1 Évolution des choix de langues d'écriture de thèses

La figure 11 représente les choix de langues d'écriture de thèses, entre 1984 et 2018. Les langues d'écriture sont catégorisées de la manière suivante :

— français,
— anglais,
— bilingue (français et anglais),
— autre (toutes les autres langues de rédaction en dehors des catégories ci-dessus).

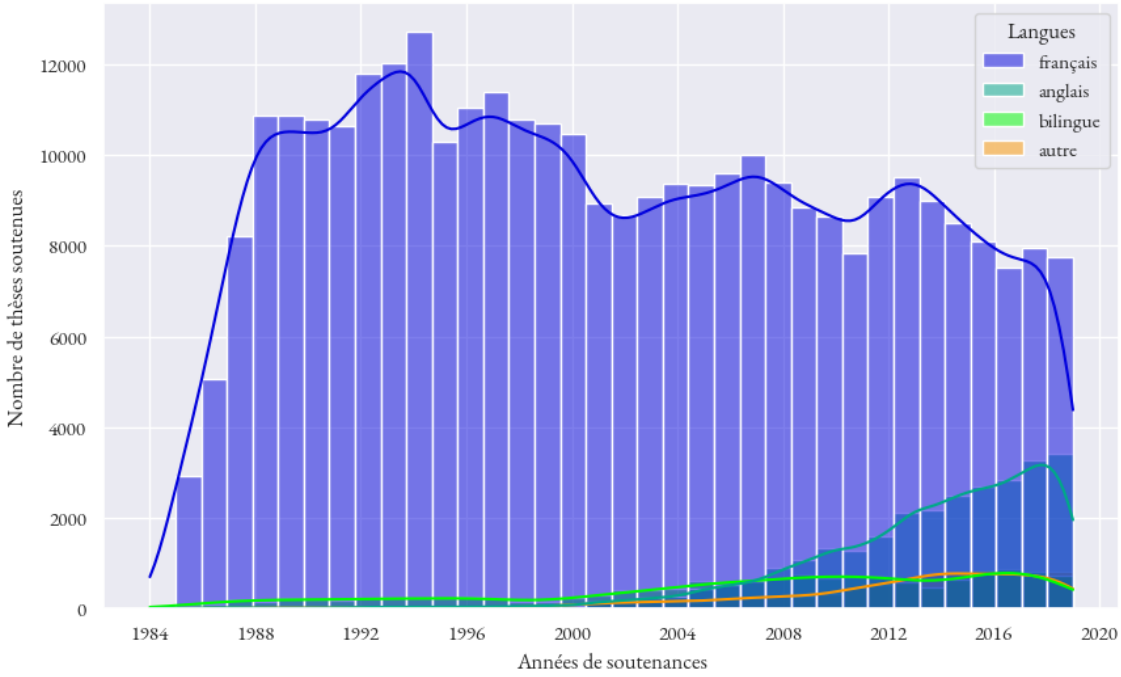

FIGURE 11 – Évolution du choix de la langue d'écriture de thèses par années, entre 1984 et 2018

La Figure 12 représente la proportion de choix de langues d'écriture de thèses au cours des deux dernières décennies.

5.2 Analyse de cette évolution

5.2.1 Thèses soutenues en français

Entre 1984 et 1994, le nombre annuel de thèses soutenues en français a augmenté, passant de 3000 à plus de 12000. Ensuite, de 1994 à 2018, ce nombre a diminué pour atteindre environ 8000 thèses par an.

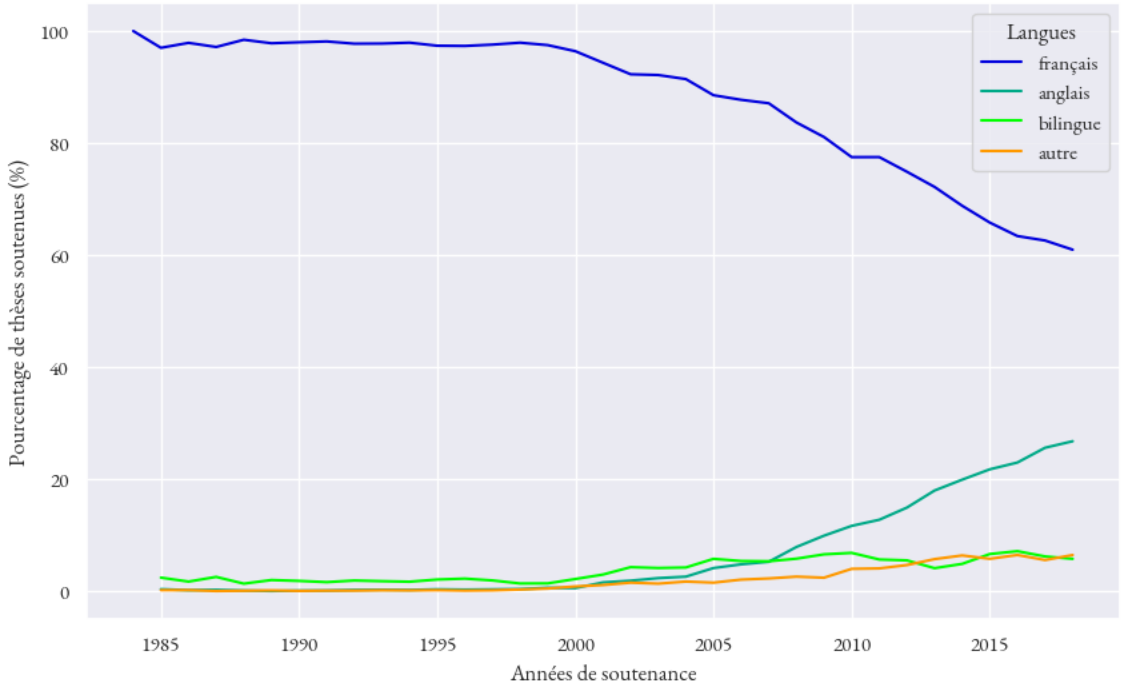

FIGURE 12 – Évolution de la proportion des langues d'écriture de thèses, par années, entre 1984 et 2018

La proportion des thèses soutenues en français est demeurée stable (environ 97 %) de 1984 à 2000, puis a progressivement diminué jusqu'en 2018, année où elle a atteint 60 %.

5.2.2 Thèses soutenues en anglais

Entre 1984 et 2000, le nombre de thèses soutenues en anglais est resté pratiquement nul. Il a ensuite augmenté progressivement pour dépasser 3000 thèses par an en 2018.

La proportion des thèses soutenues en anglais est restée négligeable jusqu'aux années 2000, puis a connu une croissance continue pour représenter entre 25 % et 30 % de toutes les thèses soutenues en 2018.

5.2.3 Thèses bilingues et soutenues dans d'autres langues

Entre 1984 et 2000, le nombre de thèses soutenues en bilingue (anglais-français) ou dans d'autres langues était quasiment nul. À partir de 2000, ce nombre a augmenté progressivement pour atteindre près de 1000 thèses par an en bilingue et autant dans d'autres langues que le français ou l'anglais en 2018.

La proportion des thèses soutenues en bilingue anglais-français ou dans d'autres langues est restée insignifiante jusqu'aux années 2000, puis a régulièrement augmenté

pour représenter environ 10 % de toutes les thèses soutenues dans chacune de ces catégories en 2018.

5.2.4 Conclusion

Entre 2000 et 2018, on observe une augmentation de l'utilisation de l'anglais comme langue d'écriture des thèses, particulièrement après 2007. Le nombre de thèses écrites en français a légèrement diminué, tandis que la catégorie bilingue (anglais-français) est restée stable mais minoritaire. Enfin, les thèses dans des langues autres que le français et l'anglais sont restées marginales. Ces tendances peuvent s'expliquer par la globalisation de la recherche académique et l'importance croissante de l'anglais comme langue scientifique internationale. Comme le souligne Martin (2015), "*la visibilité donnée sur le Web aux thèses a un impact sur les doctorants, les docteurs, les directeurs de thèses et établissements de soutenance qui réagissent spontanément aux erreurs, manquements constatés et visibles publiquement*"[1]. La publication de thèses en ligne favorise l'utilisation de l'anglais pour atteindre un public plus large et améliorer la visibilité des travaux de recherche.

Références

[1] Martin, I. (2015). *Le signalement des thèses de doctorat*. I2D - Information, données documents, Volume 52(1), 46-47.