CERGY PARIS UNIVERSITÉ

Rapport de projet des Soutenances de Thèse en France de 1984 à 2018

Auteur : Bi Kristian Gohi

Enseignant : Matthieu Cisel

Juillet 2024

Table des Matières

I. Introduction

L'analyse des soutenances de thèse en France entre 1984 et 2018 permet d'obtenir des informations précieuses sur les tendances et les évolutions dans le domaine de la recherche doctorale. Cette étude vise à examiner les données concernant ces soutenances, en mettant l'accent sur la distribution des mois de soutenance, l'évolution des langues d'écriture des thèses, et la gestion des données manquantes et des anomalies.

II. Présentation des Données

1. Charger et Prétraiter les Données

Après le chargement de notre donnée PhD_v4, nous avons constaté qu'elle contenait 448578 lignes et 24 colonnes. Les colonnes 'Unnamed: 0' et 'index' étaient des index générés lors du précédent nettoyage. Nous avons supprimé la colonne 'index', utilisé la colonne 'Unnamed: 0' pour l'index de notre base de données et renommé la colonne 'Discipline_prÃ©di' en 'Discipline_prédi'.

Les types de données actuels de toutes les colonnes étaient de type "object", signifiant que pandas considérait chaque valeur comme une chaîne de caractères ou un objet générique. Cependant, certaines colonnes devraient probablement être d'autres types de données, tels que des dates ou des nombres, pour permettre des analyses plus précises.

Les colonnes 'Date de première inscription en doctorat' et 'Date de soutenance' ont été transformées en datetime. Les colonnes 'Year', 'Identifiant auteur', 'Identifiant directeur', 'Identifiant établissement' ont été converties en int sans traiter les valeurs manquantes. Les autres colonnes ont été laissées dans leurs types initiaux.

Le tableau 1 montre le type de données après transformation.

	Column	Non-Null Count	Dtype
0	Auteur	448578	object
1	Identifiant auteur	288202	Int64
2	Titre	448571	object
3	Directeur de these	448542	object
4	Directeur de these (nom prenom)	448542	object
5	Identifiant directeur	304564	Int64
6	Etablissement de soutenance	448554	object
7	Identifiant etablissement	396965	Int64
8	Discipline	448555	object
9	Statut	448553	object
10	Date de premiere inscription en doctorat	64322	datetime64[ns]
11	Date de soutenance	390942	datetime64[ns]
12	Year	390942	Int64
13	Langue de la these	448555	object
14	Identifiant de la these	448555	object
15	Accessible en ligne	448555	object
16	Publication dans theses.fr	448555	object
17	Mise a jour dans theses.fr	447847	object
18	Discipline_predi	448024	object
19	Genre	448019	object
20	etablissement_rec	444945	object
21	Langue_rec	383908	object

Tableau 1. Présentation des colonnes et leurs types après modification

III. Analyse des Données Manquantes

1 Évaluation des Valeurs Manquantes

L'analyse du jeu de données révèle des proportions préoccupantes de valeurs manquantes dans plusieurs variables clés, comme le montre la figure 1. Ces manques peuvent avoir des répercussions importantes sur la fiabilité et la validité des analyses ultérieures.

Date de première inscription en doctorat (86%) : Ce taux extrêmement élevé, indiquant que plus des trois quarts des données sont manquantes, soulève des interrogations majeures quant à la collecte ou à l'enregistrement de cette information. Il est crucial d'identifier l'origine de ces manques et de mettre en place des solutions adéquates pour les imputer ou les exclure du jeu de données.

Identifiant auteur (36%) et Identifiant directeur (32%) : Un tiers des données manquantes pour ces identifiants critiques peut entraver la capacité à associer correctement les thèses à leurs auteurs ou directeurs respectifs. Cela pourrait biaiser les analyses de collaboration ou de productivité individuelle.

Langue_rec (14%), Date de soutenance (13%) et Year (13%) : Bien que moins importantes que les variables précédentes, ces proportions de données manquantes (environ 13%) peuvent néanmoins affecter les analyses temporelles ou linguistiques. Il est essentiel d'évaluer l'impact potentiel de ces manques et de déterminer les stratégies appropriées pour les traiter.

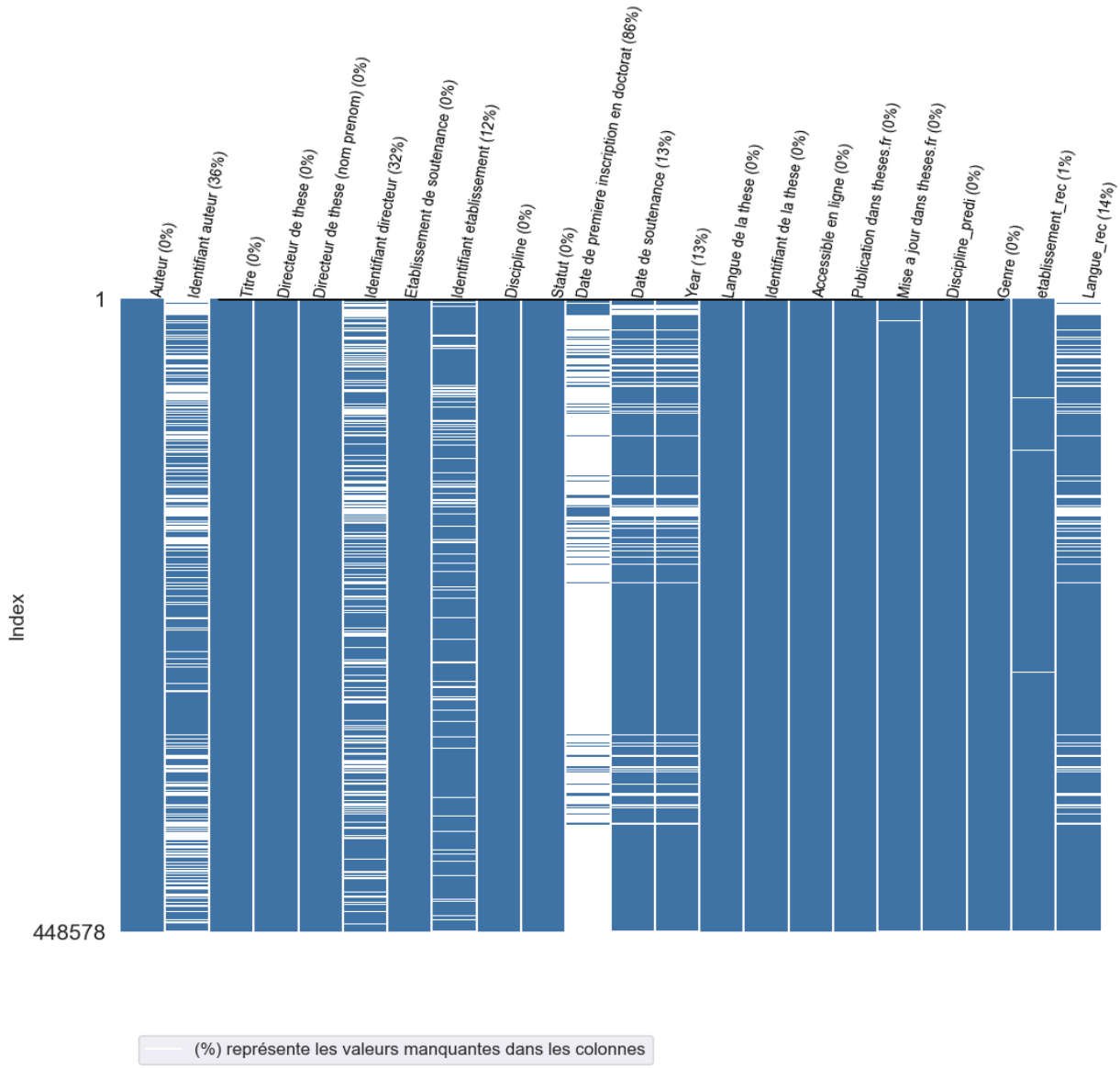

Figure 1. Répartition des données manquantes par pourcentage

2 Variables avec Peu ou Pas de Données Manquantes

Heureusement, la majorité des autres variables, telles que Auteur, Titre, Directeur de thèse, Discipline, etc., ne présentent pas de données manquantes ou un taux très faible. Cela indique une bonne qualité des données pour ces attributs, ce qui est favorable pour les analyses basées sur ces informations.

3 Remarque concernant le Traitement des Données Manquantes

Il est important de souligner que les données manquantes ne seront pas traitées dans le cadre de l'analyse actuelle. Cette analyse vise uniquement à identifier et à quantifier les données manquantes afin de mieux comprendre leurs implications potentielles pour les analyses ultérieures. Le choix de la stratégie de traitement des données manquantes dépendra des objectifs spécifiques de l'analyse et des caractéristiques des données manquantes elles-mêmes.

4 Analyse par Heatmap

La Figure 2 présente une analyse par heatmap des données manquantes, comparant les thèses "en cours" et "soutenues". Cette visualisation permet de comparer rapidement les modèles de données manquantes entre les deux groupes de thèses.

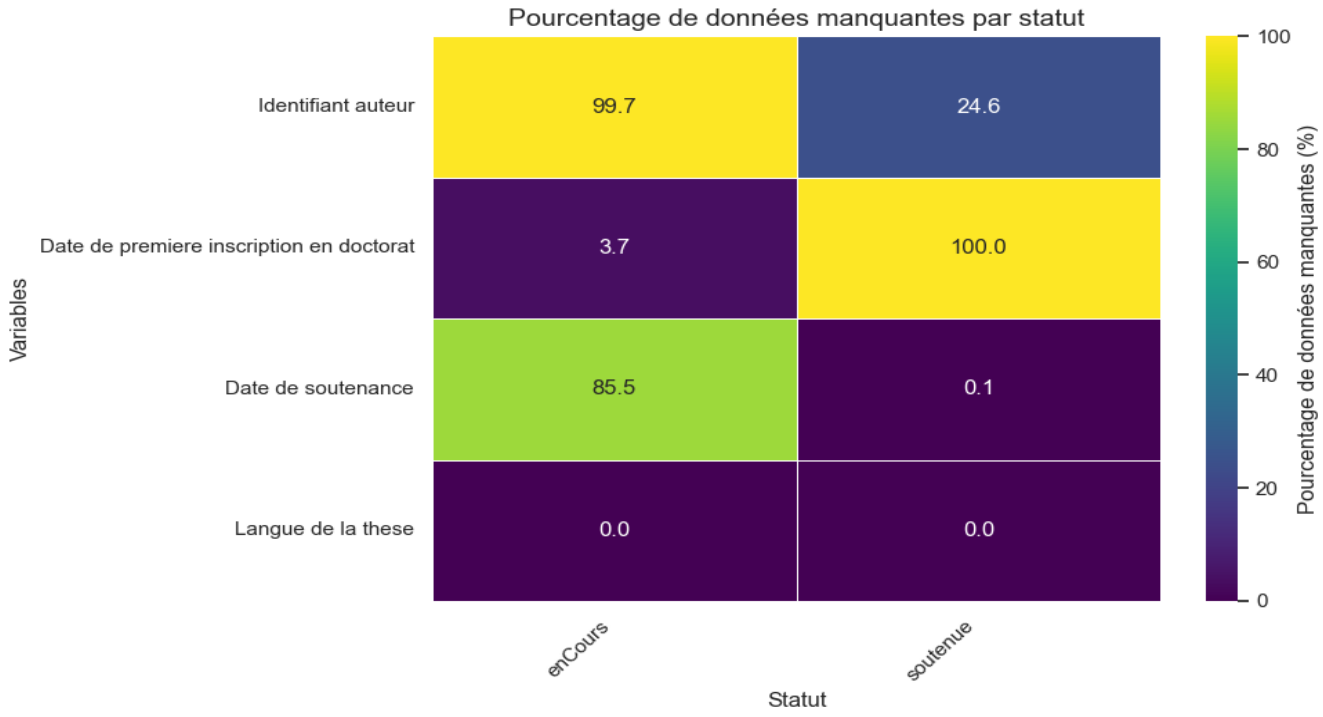

Figure 2. Disparités des données manquantes dans les thèses : Une analyse par statut (en cours vs soutenue)

4.1 Analyse par Variable

Identifiant auteur :

Thèses en cours : 99,7% des données manquantes.
Thèses soutenues : 24,6% des données manquantes.

Date de première inscription en doctorat :

Thèses en cours : 3,7% des données manquantes.
Thèses soutenues : 100% des données manquantes.

Date de soutenance :

Thèses en cours : 85,5% des données manquantes.
Thèses soutenues : Données quasi-complètes.

Langue de la thèse : Données complètes pour toutes les thèses.

4.2 Interprétation des Tendances et des Modèles

Thèses encours :

Forte proportion de données manquantes pour les dates de soutenance (85,5%), indiquant que ces informations ne sont pas encore disponibles pour les thèses en cours.

Proportion moindre de données manquantes pour les dates de première inscription (3,7%), suggérant que ces informations sont généralement saisies lors de l'inscription.

Thèses soutenues :

Données de date de soutenance quasi-complètes, indiquant que cette information est généralement enregistrée à la fin du processus doctoral.

Proportion notable de données manquantes pour les identifiants d'auteur (24,6%), suggérant des erreurs ou des omissions dans la collecte de ces données.

3.4.3 Explication du Pattern entre la Date de Soutenance et la Date de Première Inscription

Le pattern observé entre les dates de soutenance et de première inscription peut s'expliquer par plusieurs facteurs :

- **Processus administratif** : Les dates de première inscription et de soutenance sont souvent enregistrées simultanément ou dans le cadre d'une procédure administrative structurée. Les thèses en cours, n'ayant pas encore de date de soutenance définie, peuvent manquer d'enregistrement complet de la date de première inscription.

- **Mise à jour des bases de données** : Les thèses plus anciennes, issues de cycles de collecte de données antérieurs, peuvent ne pas avoir toutes leurs informations mises à jour dans la base de données actuelle, ce qui explique les données manquantes pour certaines années.

- **Statut de la thèse** : Les thèses en cours, en raison de leur nature évolutive et de leur statut actuel, sont plus susceptibles d'avoir des informations incomplètes, en particulier pour les dates de soutenance qui ne sont pas encore définies.

IV. Détection d'un Problème dans les Données

Cette section vise à analyser les données sur les soutenances de thèse afin d'identifier d'éventuels problèmes de qualité ou de cohérence. Les analyses portent sur la distribution des mois de soutenance, l'évolution des soutenances au 1er janvier et la proportion moyenne de soutenances par mois.

1. Distribution des mois de soutenance

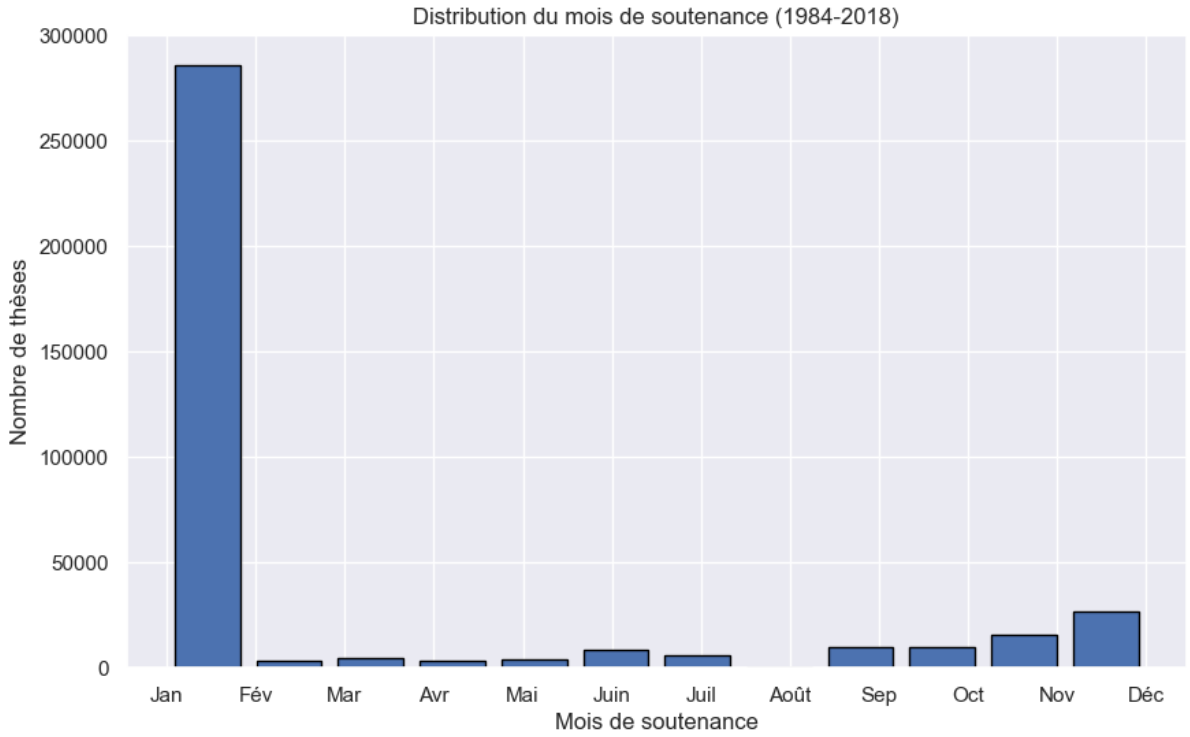

Figure 3. Distribution du mois de soutenance de 1984 à 2018

Observations :

Période de soutenance privilégiée : Janvier est le mois le plus fréquent pour les soutenances, suivi de décembre, novembre, octobre et septembre.

Influence des vacances et des périodes académiques : Les soutenances sont rares pendant les vacances d'été (juillet et août) et augmentent considérablement avant et après les vacances d'hiver et la rentrée académique.

Tendances et anomalies : Le pic élevé en janvier et le creux en août sont les principales anomalies, probablement dues à des influences institutionnelles et académiques.

Interprétation :

La distribution des mois de soutenance met en évidence une forte concentration des soutenances en début et fin d'année académique, avec une influence notable des vacances scolaires.

2. Représentation de la distribution du mois de soutenance pour chaque année

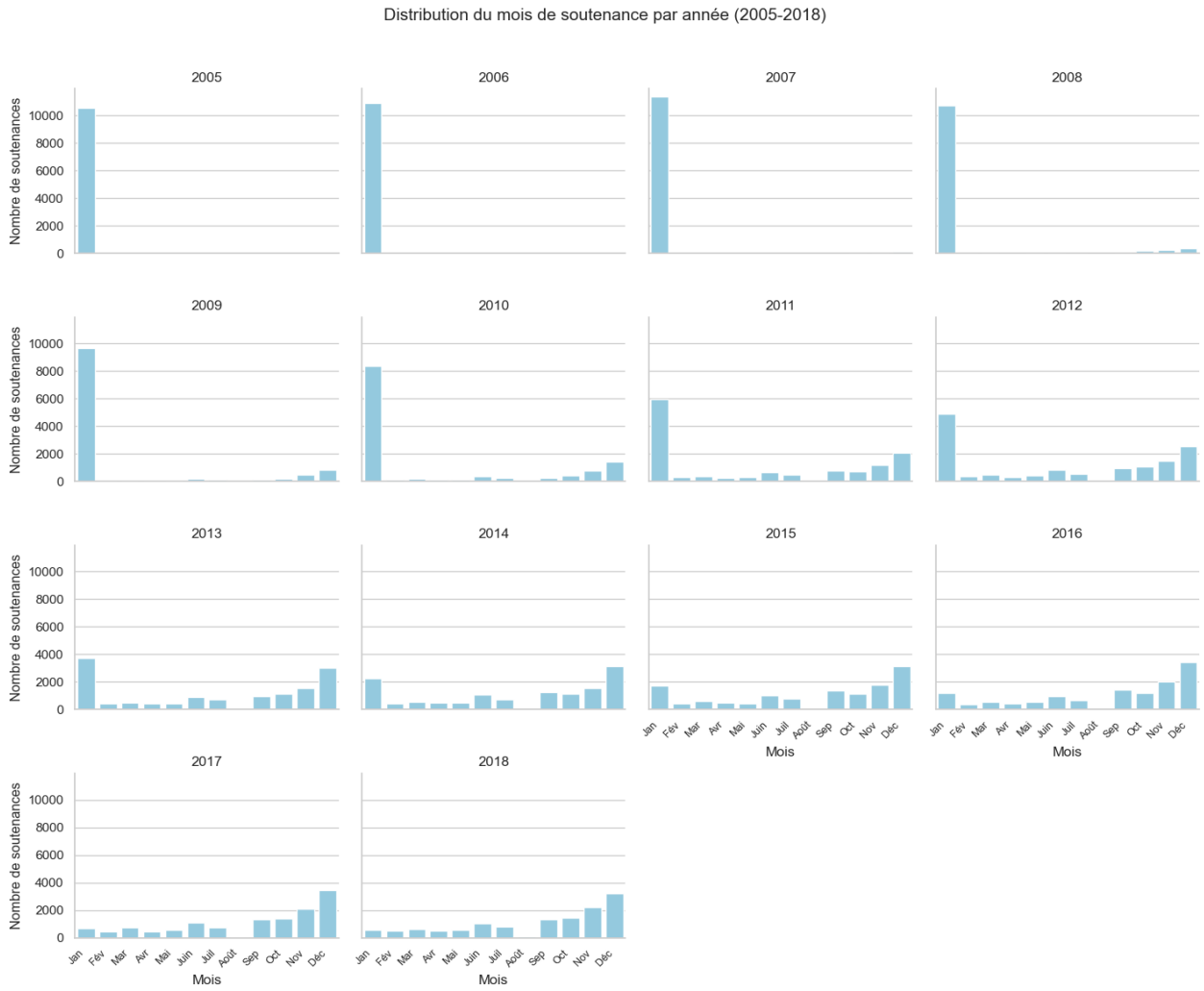

Figure 4. **Représentation de la distribution du mois de soutenance pour chaque année de 2005 à 2018**

Observations :

Variabilité intra-annuelle : On observe une variabilité significative du nombre de soutenances au cours de l'année, avec des pics en janvier et décembre, et des creux en été (août).

Tendance annuelle : Le nombre total de soutenances par an augmente globalement sur la période, passant d'environ 10 000 en 2005 à plus de 32 000 en 2018.

Changement de pic de soutenance : À partir de 2014, le pic des soutenances change de janvier à décembre.

Interprétation :

L'analyse par année confirme la concentration des soutenances en début et fin d'année, avec une évolution notable du pic de janvier vers décembre à partir de 2014. Cette évolution peut s'expliquer par des changements dans les calendriers académiques, les politiques institutionnelles ou les préférences des doctorants.

3 Nombre Moyen de Soutenances par Mois avec Erreur-Type

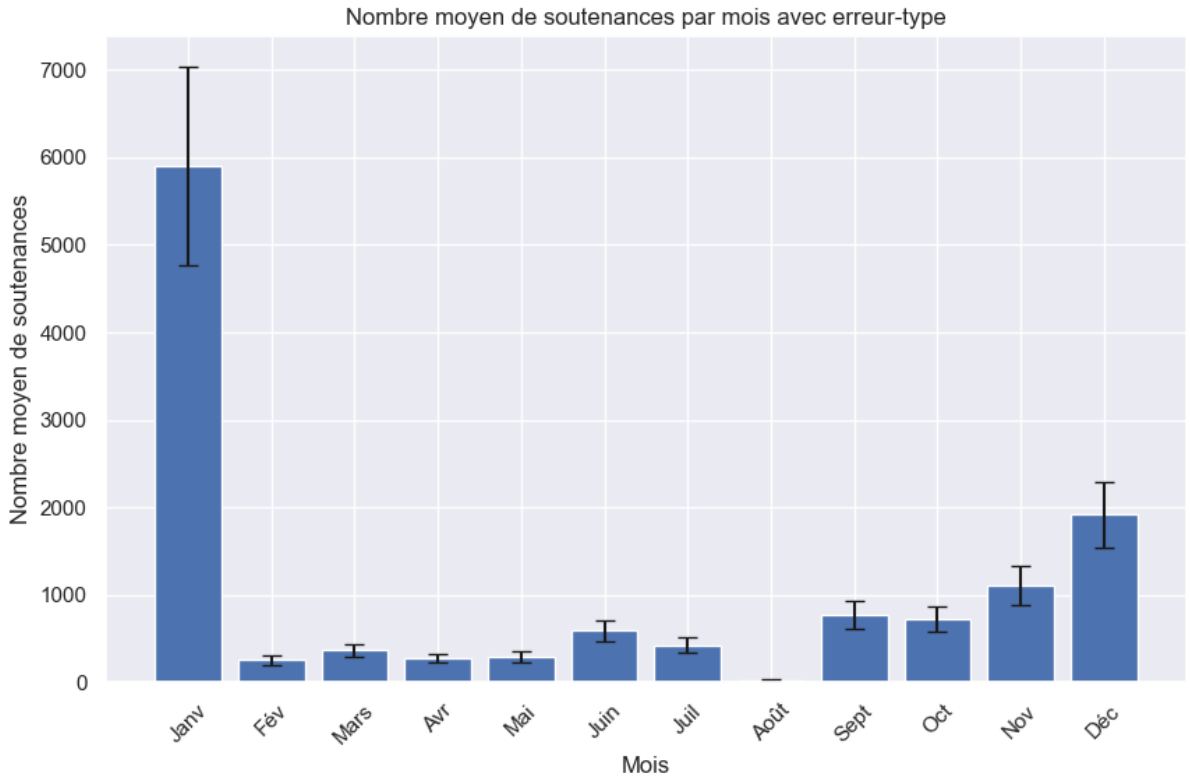

Figure 5. Nombre moyen de soutenances par mois avec erreur-type de 2005 à 2018

Observations :

Pics de soutenances : Janvier et décembre se distinguent comme les mois avec le plus grand nombre moyen de soutenances, avec janvier à 5 900 soutenances et décembre à 1 920 soutenances.

Creux de soutenances : Août a le nombre le plus bas de soutenances, avec une moyenne de seulement 27 soutenances.

Variabilité des soutenances : Les erreurs-types montrent une variabilité notable dans les moyennes mensuelles, en particulier pour janvier et décembre.

Mois intermédiaires : Les mois de février à novembre présentent des nombres de soutenances plus modérés et relativement stables.

4.4 Évolution des Soutenances du 1er Janvier

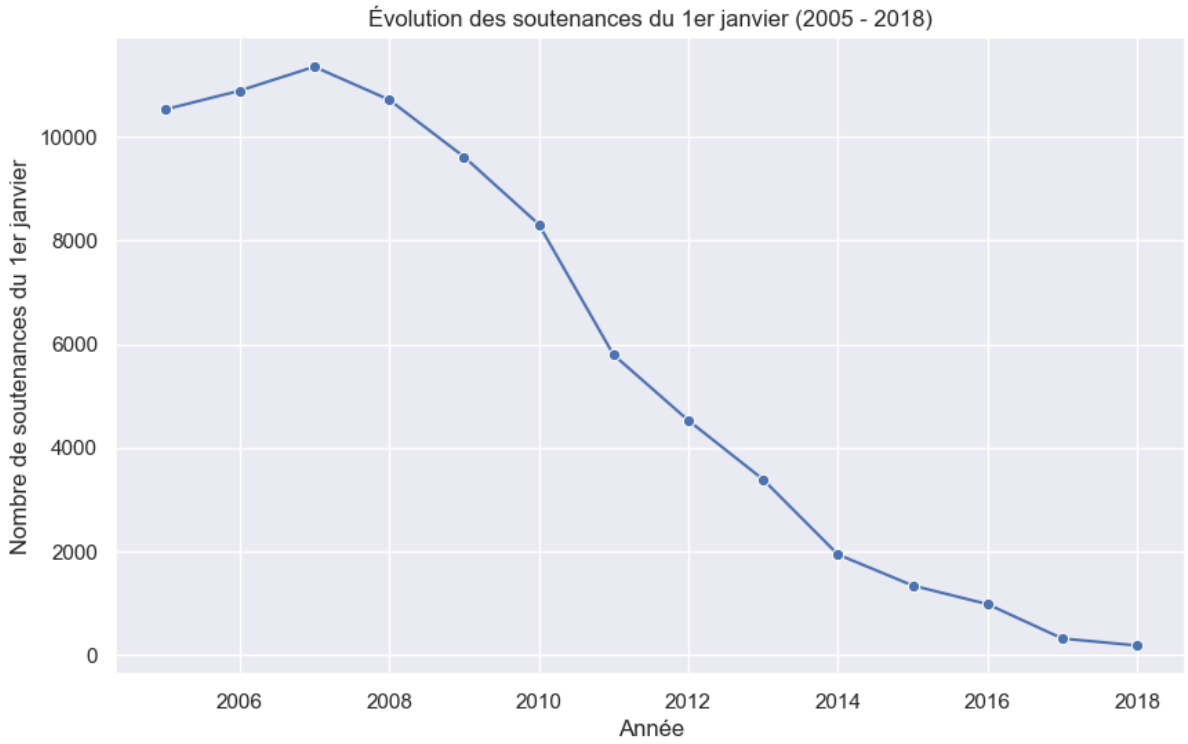

Figure 6. Évolution des soutenances du 1er janvier 2005 au 2018

Observations :

Augmentation initiale : Entre 2005 et 2007, le nombre de soutenances au 1er janvier augmente légèrement.

Diminution continue : À partir de 2008, le nombre de soutenances au 1er janvier diminue constamment, avec un déclin particulièrement marqué à partir de 2011.

Tendance à long terme : La proportion des soutenances au 1er janvier diminue fortement de 2005 à 2018.

Interprétation :

Cette analyse révèle une diminution importante du nombre de soutenances au 1er janvier sur la période étudiée. Cette tendance peut s'expliquer par divers facteurs, tels que des changements dans les pratiques de recherche, les calendriers académiques ou les politiques de financement.

5 Proportion Moyenne de Soutenances (sans le Mois de Janvier)

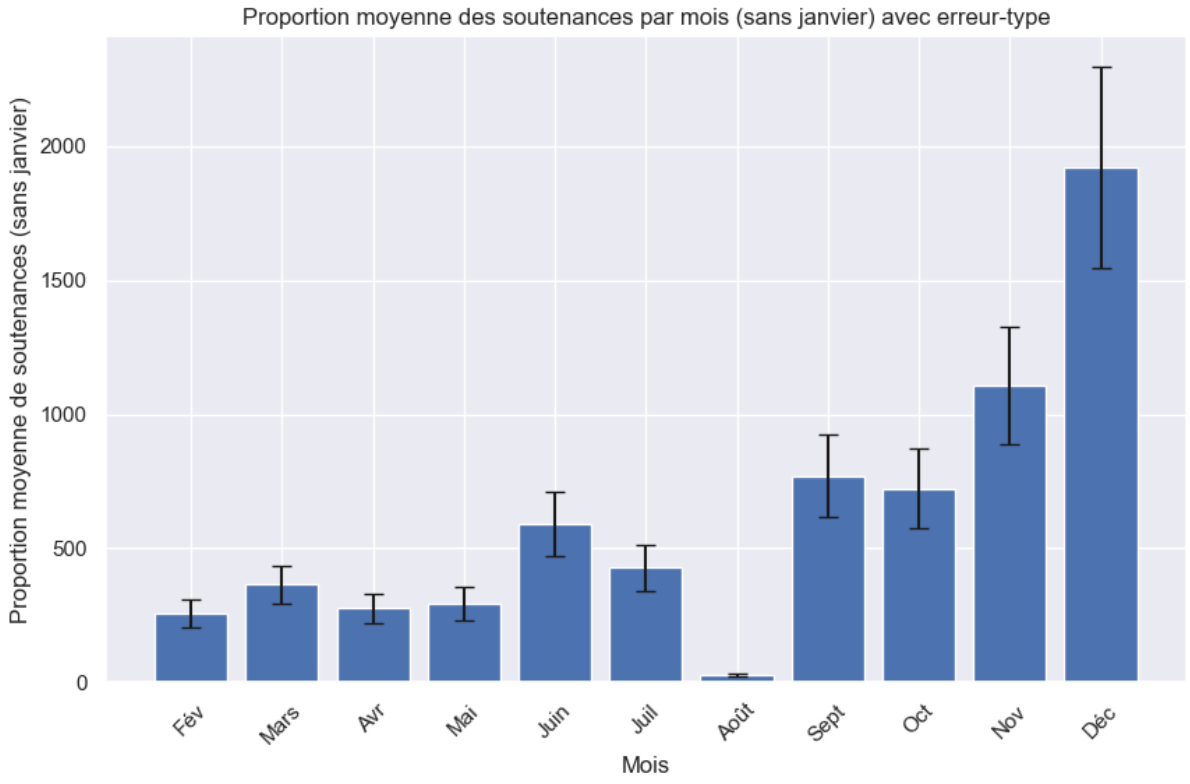

Figure 7. Proportion moyenne des soutenances par mois (sans janvier) avec erreur-type

Observations :
Décembre est le mois avec la proportion moyenne de soutenances la plus élevée.

Le mois d'été (août) est la proportion la plus faible.

Les autres mois présentent des proportions relativement stables.

Interprétation :

Cette analyse confirme que, hors janvier, décembre est le mois le plus privilégié pour les soutenances, suivi de novembre et octobre. Les mois d'été restent les moins actifs.

V. Outliers et Résultats Anormaux

1. Enquête sur les Homonymes : Le Cas de Cécile Martin

L'analyse des données concernant les thèses de Cécile Martin révèle plusieurs travaux attribués à des auteurs portant ce même nom.

Objectif : Comprendre les raisons de ces résultats et proposer des interprétations possibles.

Méthodologie :

Analyse des informations disponibles pour chaque thèse de Cécile Martin :

Titre de la thèse

Discipline

Établissement de soutenance

Date de soutenance

Identifiant auteur

Identification des éléments distinctifs entre les différentes thèses.

Formulation d'hypothèses explicatives.

Résultats :

Thèse	Titre	Discipline	Établissement	Date	Identifiant Auteur	Observations
1	L'invention de l'ecran. De l'ecran de cheminee...	Études cinématographiques et audiovisuelles	Sorbonne Paris Cité	2017	203208145	
2	Système laitier et filière lait au Mexique	Sciences biologiques fondamentales et appliquées	Institut national agronomique Paris-Grignon	2000	81323557	
3	Concurrence, prix et qualité de la prise en charge	Sciences économiques	Paris 9	2014	179423568	
4	Modélisation et critères de combustibilité	Génie des procédés industriels	Compiègne	2001	81323557	Même identifiant auteur que thèse 2
5	Caractérisation électrophysiologique et pharmaco...	Neurosciences	Bordeaux 2	1991	81323557	Même identifiant auteur que thèses 2 et 4
6	Influence du pH ruminal sur la digestion des prot	Sciences biologiques fondamentales et appliquées	Clermont-Ferrand 2	1994	81323557	Même identifiant auteur que thèses 2, 4 et 5
7	Déposition d'énergIV. Outliers et résultats anormauxie par production de paires...	Physique	Paris 11	1989	182118703	

Tableau 2. Récapitulatif des problèmes du Cas de Cécile Martin

Interprétations possibles :

Homonymie réelle : Plusieurs personnes portant le même nom ont soutenu des thèses. Les informations disponibles ne permettent pas de confirmer ou d'infirmer cette hypothèse.

Erreur de saisie : Des erreurs de saisie lors de l'enregistrement des données pourraient expliquer la présence de plusieurs thèses sous le même nom.

Identifiant auteur non unique : L'identifiant auteur ne semble pas être unique, comme le montrent les thèses 2, 4, 5 et 6. Cela pourrait indiquer un problème de gestion des identifiants ou une absence de standardisation.

Changement de nom : Cécile Martin pourrait avoir changé de nom après la soutenance de certaines thèses.

Propositions d'interprétations des résultats :

Coïncidence : Il est possible que ces personnes partagent simplement le même nom sans aucun lien entre elles, à l'exception de l'univers académique.

Lien familial : Certaines Cécile Martin pourraient être apparentées ou liées à d'autres par le sang ou le mariage, ce qui expliquerait la similarité de leur nom.

Nom de famille commun : Le nom Martin est relativement courant en France et il est possible que ces personnes n'aient aucun lien familial ou personnel, mais partagent simplement un nom de famille commun.

Noms composés : Il est également possible que certaines Cécile Martin aient un deuxième prénom ou un nom de famille composé, ce qui pourrait expliquer pourquoi elles apparaissent dans la recherche initiale.

Conclusion :

Le cas de Cécile Martin met en lumière les difficultés liées à la gestion des homonymes dans les bases de données bibliographiques. Une analyse approfondie des données et la mise en place de procédures de vérification et de standardisation sont nécessaires pour garantir la fiabilité des informations.

2. Détection des Outliers

Nous remarquons à la figure 8 que la distribution du nombre de thèses par directeur de thèse n'est pas normale. Pour la détection des outliers, nous allons utiliser l'écart interquartile (IQR).

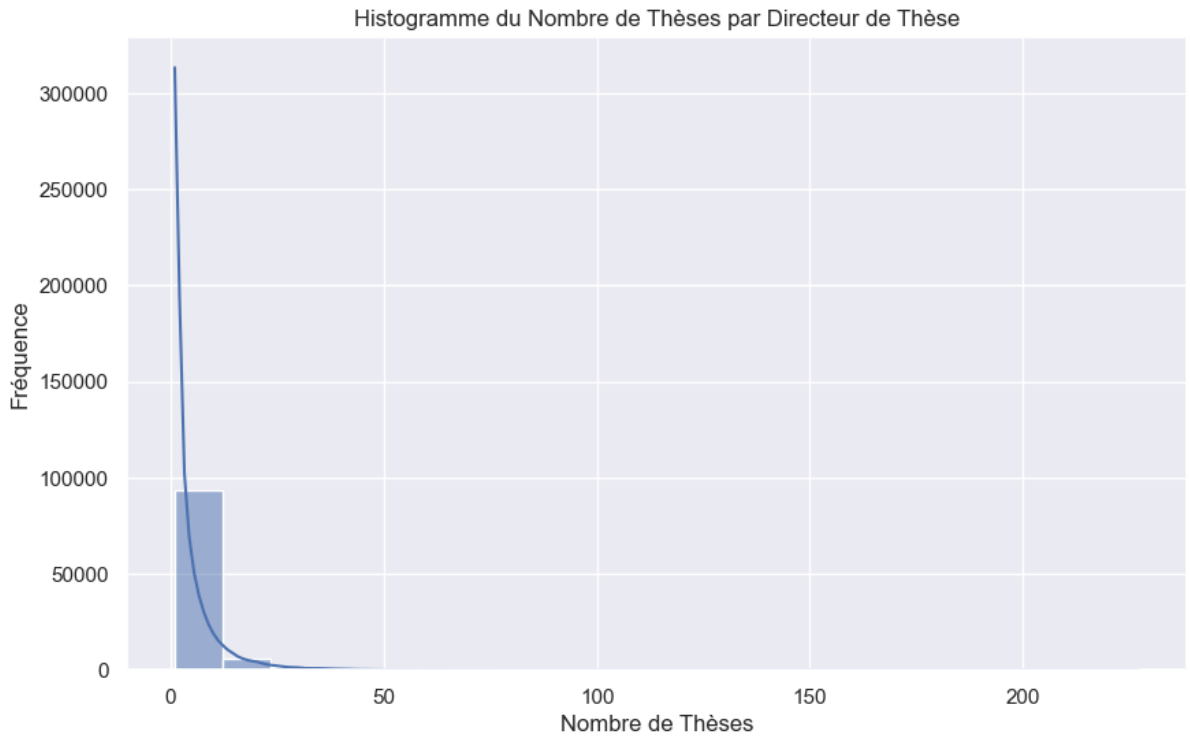

Figure 8. Distribution des valeurs aberrantes

Méthodologie:

Création d'un nouveau jeu de données : J'ai créé un nouveau jeu de données à partir du jeu de données initial, en me concentrant sur les directeurs de thèse. J'ai conservé les colonnes "Directeur de thèse (nom prénom)" et "Nombre de thèses" pour chaque directeur/directrice.

Calcul des Outliers : J'ai identifié les individus ayant encadré un nombre anormalement élevé de thèses en calculant le troisième quartile (Q3) et l'écart interquartile (IQR) du nombre de thèses encadrées. J'ai considéré comme outliers les directeurs ayant encadré plus de Q3 + 1.5 * IQR thèses.

Enquête sur les Outliers : Pour enquêter sur ces outliers, j'ai effectué une jointure entre le dataframe des outliers et le dataframe initial sur la colonne "Directeur de thèse (nom prénom)". Cela m'a permis d'obtenir les identifiants des directeurs outliers et de vérifier s'il y avait des erreurs dans les données.

Résultats :

Voici un exemple de la jointure pour les cinq premiers outliers :

Directeur de thèse (nom prénom)	Nombre de thèses_outliers	Identifiant directeur	Nombre de thèses_directeur
Blanc Francois-Paul	227	26730774	201
Scherrmann Jean-Michel	209	59375140	208
Brunel Pierre	206	26756625	194
Brunel Pierre	206	30616123	1
Brunel Pierre	206	116052384	1

Tableau 3. Présentation du problème des valeurs identifier comme valeurs aberrantes

En examinant les données, j'ai remarqué que certains directeurs outliers avaient des identifiants différents pour le même nom et prénom, ce qui pourrait indiquer des erreurs de saisie ou des homonymes. Par exemple, "Brunel Pierre" a trois identifiants différents et des nombres de thèses encadrées différents. Pour vérifier s'il s'agit d'erreurs ou d'homonymes, j'ai effectué des recherches supplémentaires sur ces directeurs en utilisant des sources externes, telles que les sites web des universités ou les bases de données de publications.

Dans le cas de "Brunel Pierre", j'ai trouvé qu'il y avait effectivement plusieurs directeurs de thèse portant ce nom dans différentes universités. Par conséquent, il est probable que les données pour "Brunel Pierre" soient correctes et qu'il s'agisse d'homonymes plutôt que d'erreurs. Cependant, pour confirmer cela, il serait nécessaire de vérifier chaque thèse individuellement pour s'assurer que le directeur est bien le même "Brunel Pierre" pour chaque thèse.

Conclusion :

Pour déterminer si les outliers sont des erreurs ou des homonymes, il est nécessaire d'effectuer des recherches supplémentaires en utilisant des sources externes et de vérifier chaque thèse individuellement pour s'assurer que le directeur est bien le même

pour chaque thèse. Une analyse approfondie et la mise en place de procédures de vérification et de standardisation sont cruciales pour garantir la fiabilité des données académiques.

VI. Résultats Préliminaires

1. Évolution du Choix de la Langue d'Écriture au Cours des Deux Dernières Décennies

L'analyse de l'évolution du choix de la langue d'écriture des thèses de 2000 à 2020 met en lumière plusieurs tendances intéressantes (Figure 9).

Tout d'abord, l'utilisation de l'anglais a connu une augmentation significative. Cette progression peut être attribuée à la reconnaissance internationale de l'anglais comme langue de communication scientifique et académique. L'anglais est souvent perçu comme un moyen d'atteindre un public mondial plus large, ce qui augmente la visibilité et l'impact potentiel des travaux de recherche.

Parallèlement, le français reste la langue dominante pour la rédaction des thèses tout au long de la période étudiée, bien que son pourcentage relatif ait légèrement diminué au fil des années. Cette diminution pourrait être le reflet de l'internationalisation croissante des publications scientifiques, où les chercheurs privilégient les revues internationales souvent anglophones pour maximiser la diffusion et la reconnaissance de leurs travaux.

De plus, le nombre de publications bilingues (français et anglais) a augmenté de manière régulière. Cette tendance suggère une volonté de toucher un public plus large tout en préservant l'identité francophone. Les publications bilingues permettent aux chercheurs de communiquer leurs résultats à la fois à un public local, francophone, et à la communauté scientifique internationale.

Enfin, bien que minoritaire, l'utilisation d'autres langues que le français et l'anglais a également connu une hausse. Cette augmentation peut s'expliquer par une diversification des collaborations internationales, permettant aux chercheurs de publier dans des langues correspondant aux régions de leurs partenaires ou aux publics cibles spécifiques.

En résumé, ce graphique montre non seulement les préférences linguistiques pour la rédaction des thèses en France au cours des deux dernières décennies, mais aussi des tendances plus larges dans le monde de la recherche académique en réponse à la mondialisation et aux dynamiques locales.

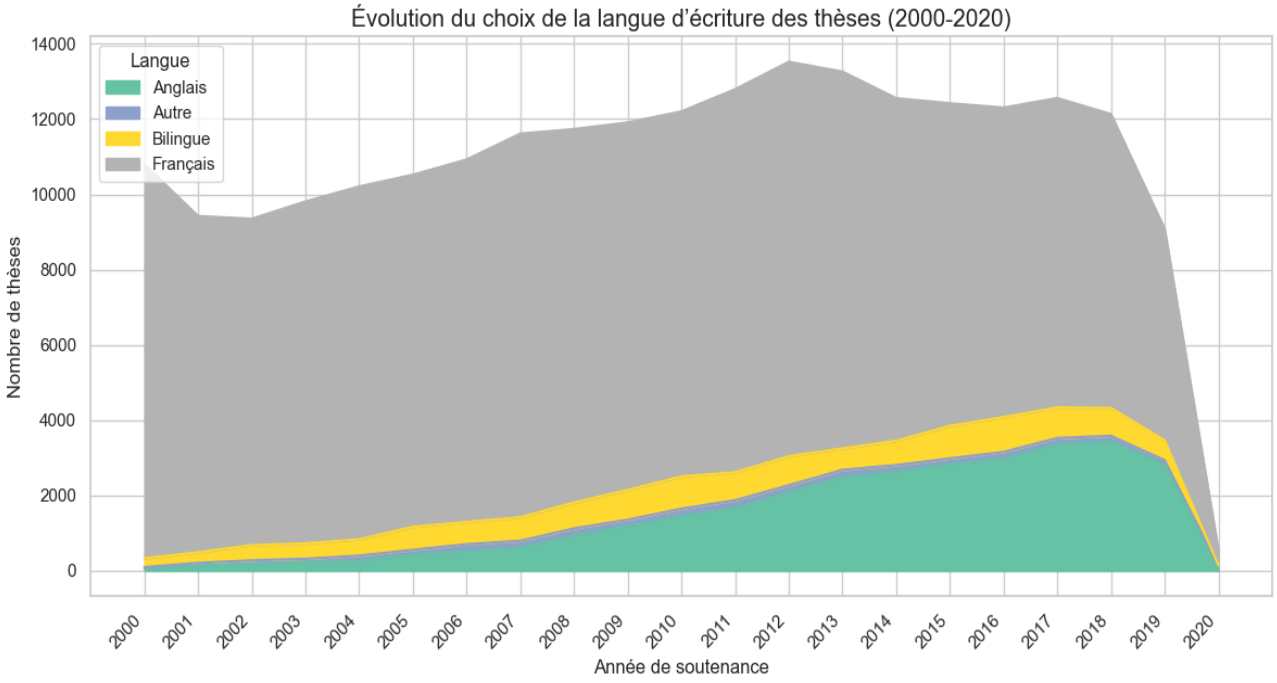

Figure 9. Évolution du choix de la langue d'écriture des thèses de 2000 à 2020

VII. Conclusion

L'analyse des soutenances de thèse en France entre 1984 et 2018 a permis de mettre en évidence plusieurs tendances et variations dans le temps, les différentes disciplines et les établissements de recherche. Cette analyse a également révélé une augmentation régulière du nombre de soutenances de thèse au fil des années, ainsi que des différences significatives entre les domaines et les établissements de recherche.

Ces résultats soulignent l'importance de continuer à surveiller et à analyser les soutenances de thèse pour mieux comprendre les évolutions et les tendances dans le domaine de la recherche doctorale en France. Ils peuvent également servir de base pour des études futures sur les facteurs qui influencent les soutenances de thèse, tels que les politiques de financement, les opportunités de carrière et les normes de publication.

En outre, la gestion des données manquantes et des anomalies est essentielle pour garantir la qualité et la fiabilité des analyses. Des efforts doivent être déployés pour améliorer la collecte et la gestion des données, notamment en standardisant les identifiants et en mettant en place des procédures de vérification rigoureuses.

Enfin, la diversité linguistique et l'évolution du choix des langues d'écriture des thèses soulignent la nécessité de développer des compétences linguistiques et de promouvoir l'ouverture internationale dans le domaine de la recherche. Les chercheurs devraient être encouragés à publier dans différentes langues pour toucher un public plus large et maximiser l'impact de leurs travaux.