



Analyse de la Distribution des Dates de soutenance des  
Thèses en France entre 1985 et 2020  
et  
Analyse de la Répartition de leurs langues d'écriture

Florian GROLLEAU

Novembre 2024

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	L'intérêt d'une étude sur les conditions de déroulement des doctorats . . . . .	4
1.2	Le choix de l'étude des langues d'écriture . . . . .	4
1.3	Les dates de soutenance : un exemple de mutation des processus doctorants . . . . .	4
<b>2</b>	<b>Méthodes et données</b>	<b>4</b>
2.1	Présentation des données . . . . .	4
2.2	Fiabilité des données . . . . .	5
2.3	Gestion des <i>outliers</i> . . . . .	7
2.4	Outils utilisés . . . . .	8
<b>3</b>	<b>Résultats</b>	<b>8</b>
3.1	Introduction . . . . .	8
3.2	La répartition des dates de soutenances des thèses . . . . .	9
3.3	Analyse de la distribution des langues d'écriture des thèses . . . . .	12
<b>4</b>	<b>Discussion</b>	<b>13</b>
4.1	Plan de la discussion . . . . .	13
4.2	Analyse critique des résultats . . . . .	13
4.3	Interprétation des résultats concernant les dates de soutenance . . . . .	14
4.4	Interprétation des résultats concernant les langues d'écriture . . . . .	14
4.5	Conclusion . . . . .	14
<b>5</b>	<b>Références</b>	<b>14</b>

## Table des figures

1	Pourcentage de données manquantes par variable . . . . .	5
2	Pourcentage de données manquantes par variable en fonction du statut . . . . .	6
3	Corrélations de données manquantes les plus importantes par variable . . . . .	6
4	Répartition du nombre de thèses encadrées au delà de 50 thèses par tranche de 50. . . . .	7
5	Répartition des thèses encadrées par discipline des encadrants nommés Pierre Martin. . . . .	8
6	La surreprésentation des thèses soutenues en Janvier . . . . .	9
7	Nombre de thèses soutenues par mois sans la date du 1er Janvier . . . . .	10
8	Nombre de thèses soutenues par mois en fonction de l'année . . . . .	11
9	Thèses soutenues par année et par mois . . . . .	11
10	Évolution de l'usage des langues d'écriture . . . . .	12
11	Évolution des usages différenciés des langues d'écriture par discipline . . . . .	13

# **1 Introduction**

## **1.1 L'intérêt d'une étude sur les conditions de déroulement des doctorats**

De la même manière que dans la plupart des autres pays, l'obtention d'un diplôme de doctorat en France passe par plusieurs années de recherche sur un sujet de thèse bien précis. Celle-ci est validée ou non à l'occasion d'une soutenance de ladite thèse qui conditionne l'obtention du doctorat. Si au premier abord il semblerait que les processus d'obtention d'un doctorat sont uniformes à l'échelle nationale, il s'avère qu'ils diffèrent assez nettement dans leur encadrement et leur réalisation à l'échelle nationale.

De fait, avec l'arrivée de nombreux outils numériques et la constitution de base de données dématérialisées, les données sur les thèses effectuées ou en cours de traitement acquiescent de certains changements qui affectent à différents échelons les bases culturelles de la recherche académique. Ces informations sur les thèses, regroupées en bases de données témoignent de différences notables de traitement des thèses en fonction des disciplines, de l'évolution au cours du temps, de l'influence du ou des directeurs de thèse, de la proximité naturelle des chercheurs avec la recherche académique anglophone, ou encore de la sensibilité inégale à l'emploi des nouvelles technologies pour la publication et la classification des thèses.

C'est pourquoi nous nous sommes basés ici sur les bases de données produites par les outils numériques à l'initiative de la recherche française entre 1985 et 2020. Nous nous sommes particulièrement appuyé sur le site de référencement des thèses produites ou en cours en France : thèses.fr et de son jeu de données portant principalement sur les métadonnées de ces thèses. La question suivante est naturellement apparue : Comment cette lecture des données pourrait-elle nous permettre de mettre en évidence des changements et des patterns de comportements significatifs qui modifient les processus d'obtention d'un doctorat ?

## **1.2 Le choix de l'étude des langues d'écriture**

Si le français fut par le passé la langue de nombreux domaines scientifiques, sa place a été depuis très largement remise en question à l'international. Toutefois il importe de se demander si ce changement s'applique aussi à l'échelle nationale, et pour quelles raisons. La langue d'écriture revêt d'une importance toute particulière quand il s'agit de mesurer les degrés d'ouvertures académiques d'une discipline à l'échelle internationale. En effet, l'étude du choix personnel des doctorants quant à la langue d'écriture est un potentiel reflet de certaines pratiques académiques françaises, et peut également permettre de mesurer l'apport de la recherche universitaire française au sein de la recherche mondiale. Ainsi, nous nous sommes interrogés sur en quoi le choix des langues d'écritures est à la fois un vecteur et un témoin d'implication de la recherche française dans la recherche internationale ?

## **1.3 Les dates de soutenance : un exemple de mutation des processus doctorants**

D'autre part nous avons pu au cours de nos recherches, mettre en évidence via l'exploitation des métadonnées des marqueurs significatifs de rupture de certaines pratiques académiques. Les dates de soutenances en sont un témoin et observent des changements particulièrement sensibles sur une très courte période. Elles sont en partie un exemple d'une régularisation et d'une normalisation de la constitution des bases de données nationales des thèses réalisées ou en cours. Subsiste alors la question suivante, comment ces ruptures rendent compte de changement plus généraux dans la constitution des bases de données des thèses françaises ?

# **2 Méthodes et données**

## **2.1 Présentation des données**

Les données que nous avons exploité proviennent du site thèses.fr. La démarche du site s'inscrit plus globalement dans une volonté de créer des bases de données publiques pour la recherche

française, à l'image de la recherche chinoise. Le site [thèses.fr](#) s'appuie le modèle d'archive TEL, qui cherche à rendre plus indépendantes des financements privés la publication des ouvrages de recherche (insufflé en partie par le "Plan S" européen). Initialement au format .csv, nous avons exploité ces données sous l'environnement de traitement statistiques RStudio (utilisant le langage R). Principalement constitué de métadonnées sur la réalisation des thèses françaises entre 1985 et 2020, on retrouve des données de tout type, catégorielles, continues, de différents formats : Date, chaînes de caractères, entiers naturels, etc. Cependant les données brutes présentaient parfois des incohérences qui faussaient certains résultats. Nous en déclinerons ici certains cas particuliers.

## 2.2 Fiabilité des données

Pour ce qui est de la fiabilité des données, certaines catégories des méta-données des thèses présentaient des absences éparses voire très nombreuses.

Durant la phase de *data wrangling* nous avons produits plusieurs graphiques afin de mettre en lumière ces disparités des données manquantes, par exemple :

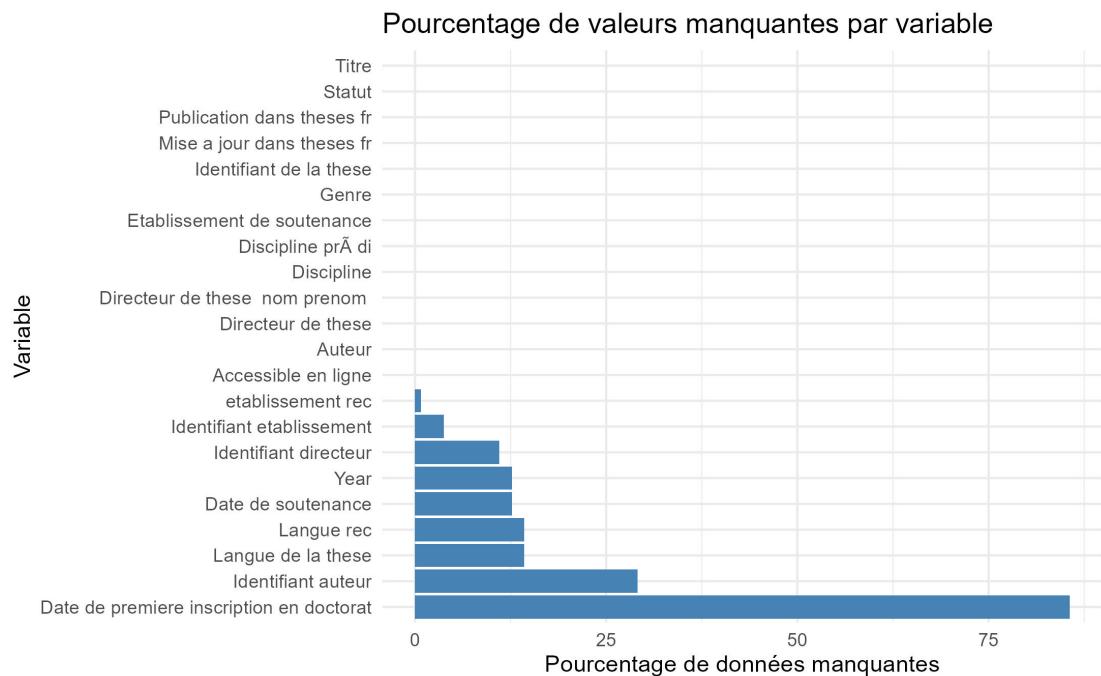


FIGURE 1 – Pourcentage de données manquantes par variable

Principalement involontaires, ces données manquantes sont parfois corrélées entre elles. Par exemple dans la figure ci-dessous on remarque qu'il existe une très forte corrélation entre le statut de la thèse (en cours ou soutenue).

En effet, l'identifiant auteur est quasiment systématiquement absent quand la thèse est en cours, ce qui semble logique puisque l'on se voit attribué un identifiant auteur qu'au moment de la publication de la thèse et non de l'entier du processus doctorant. De la même manière une date de soutenance ne peut être renseignée lorsque le doctorant est en cours de thèse, celle-ci étant définie au dernier jour de sa thèse.

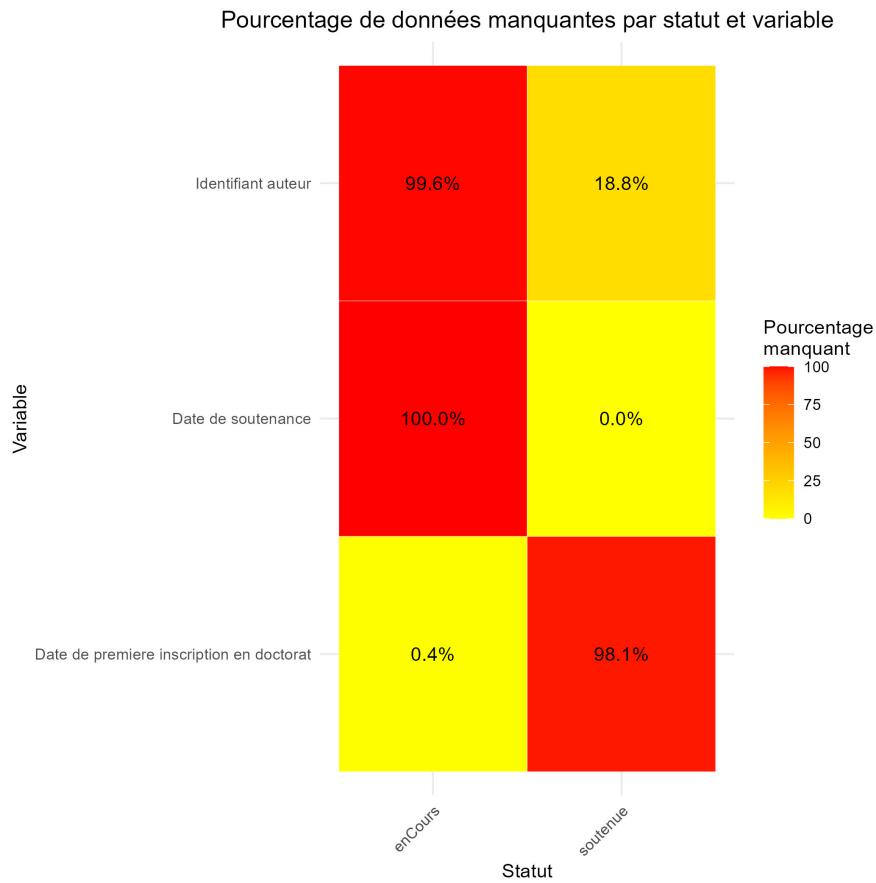


FIGURE 2 – Pourcentage de données manquantes par variable en fonction du statut

Plus généralement, on remarquera que la plupart des données manquantes sont issues de corrélations logiques, liées tantôt à une période particulière (les thèses antérieures à 1990 ne sont pas rigoureusement renseignées numériquement), ou encore l'identifiant directeur très souvent lié à l'identifiant auteur (dans le cas où l'encadrant encadre pour la première fois).

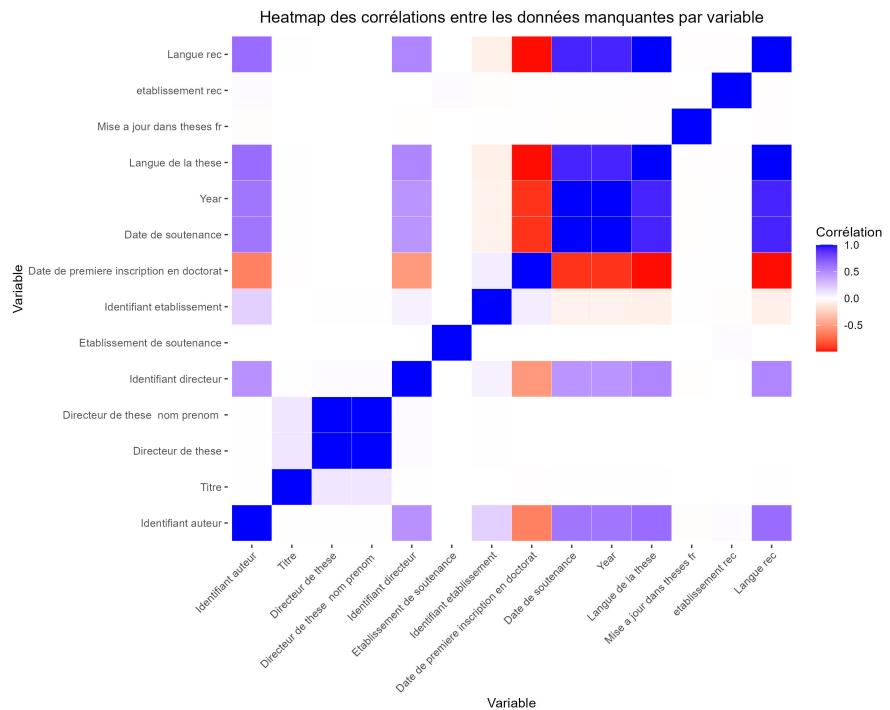


FIGURE 3 – Corrélations de données manquantes les plus importantes par variable

### 2.3 Gestion des *outliers*

L'une des première étapes du traitement des données a été l'identification et l'explication des *outliers*. Ceux-ci sont particulièrement probants au niveau du nombre de thèses encadrées par certains superviseurs.

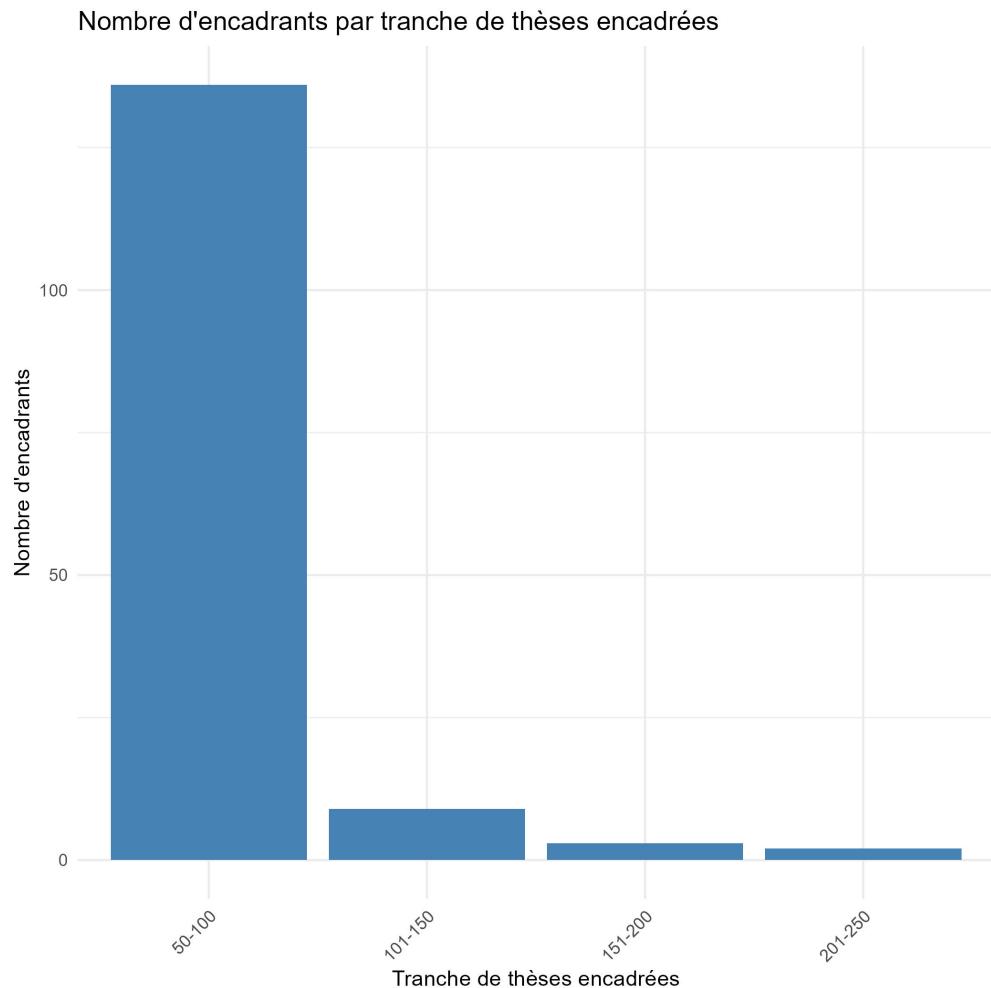


FIGURE 4 – Répartition du nombre de thèses encadrées au delà de 50 thèses par tranche de 50.

Comme on peut l'observer certains "encadrants" de thèses auraient encadré jusqu'à plus de 250 thèses. Si rapporté au nombre d'encadrants total ces chiffres semblent insignifiants, ils témoignent pour autant d'une forme de tradition de la filiation dans le monde de la recherche. C'est-à-dire que certains encadrants (très souvent motivés par un désir de capitaliser un nombre considérable de citations, mentions) n'hésite pas à s'affilier à l'encadrement d'une thèse, quand bien même ils n'y sont pas lié, et n'observe qu'un lien de discipline, ou encore une relation particulière (ancien tuteur de l'encadrant par exemple), avec l'encadrant principale.

Pour autant il est aussi possible que cette présence des *outliers*, notamment dans les encadrants et leurs nombre de thèses encadrées soient purement involontaires. C'est le cas par exemple des homonymies qui peuvent involontairement créer des *outliers*. Un nom-prénom récurrent peut être susceptible d'engendrer ces *outliers* en sommant les thèses encadrées de plusieurs personnes physiques et les regroupant naturellement sous un même nom-prénom. En témoigne l'exemple d'un nom-prénom de thèse récurrent : "Pierre Martin" :

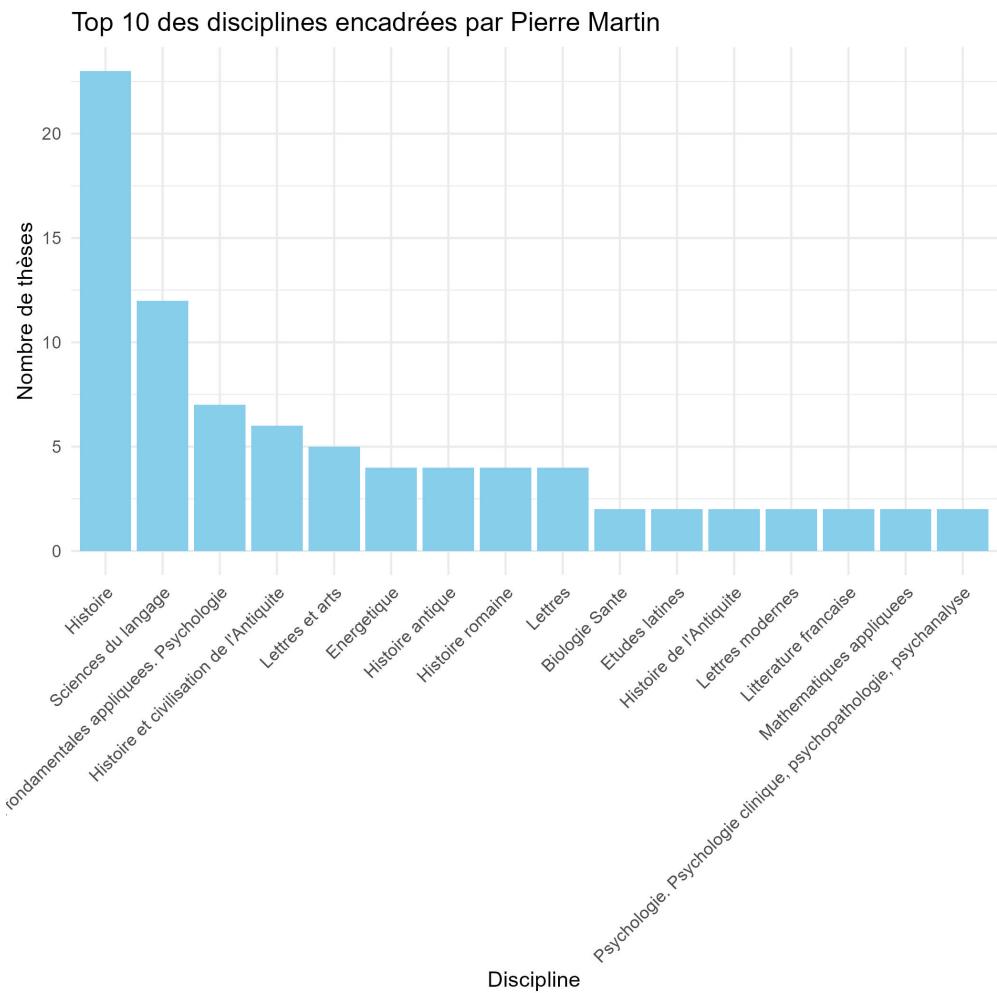


FIGURE 5 – Répartition des thèses encadrées par discipline des encadrants nommés Pierre Martin.

## 2.4 Outils utilisés

L'extraction et la visualisation des données ont été conduites à l'aide du langage de développement statistique R. Par ailleurs nous nous sommes servis de l'environnement de développement Rstudio. Les principales librairies (*packages*) utilisés ont été :

- dplyr : pour la manipulation des data frames créés
- lubridate : pour une lecture standardisée des dates
- ggplot2 : pour la visualisation
- readr : afin de lire les données d'origine
- naniar : dédiée à la visualisation et à l'identification de données manquantes
- visdat : dédiée à la visualisation et à l'intégrité des données
- tydverse : utilisée afin d'ordonner certains types de données

# 3 Résultats

## 3.1 Introduction

De l'exploitation des données ressort des résultats avec peu d'ambiguïté. Des ruptures très nettes sont observables sur les deux variables sur lesquelles nous nous sommes penchés. On remarque principalement que la dimension du temps (non seulement pour les dates de soutenances, mais aussi pour les langues d'écriture) est capitale pour comprendre la plupart des changements

(changement induits par une évolution temporelle). Les résultats convergent pour les deux variables, vers,

- D'une part : une rigueur accrue dans le renseignement des métadonnées d'une thèse dans les BDD (ici : Base de données) françaises.
- D'autre part : une diversification, non pas du type de données, mais de la valeur de cette donnée, qu'elle soit catégorielle ou continue.

### 3.2 La répartition des dates de soutenances des thèses

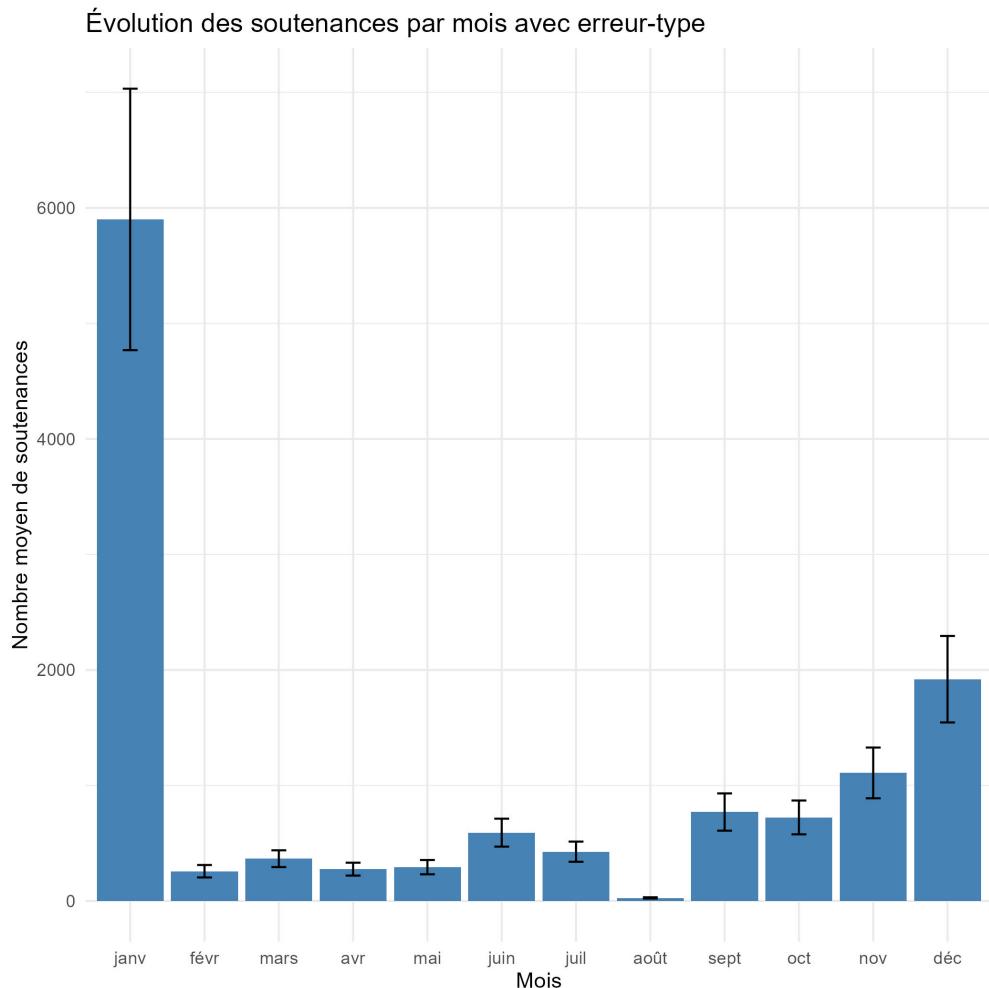


FIGURE 6 – La surreprésentation des thèses soutenues en Janvier

On distingue très nettement une surreprésentation des thèses soutenues en Janvier pour la période 1985-2020. Tandis que les autres valeurs semblent à peu près identiques et cohérentes, particulièrement pour le mois d'Août et pour le mois de Décembre.

Pour autant d'autres visualisations des valeurs de cette variable nous conduisent à une lecture des résultats plus appropriée :

### Répartition des dates de soutenance par mois (sans le 1er janvier)

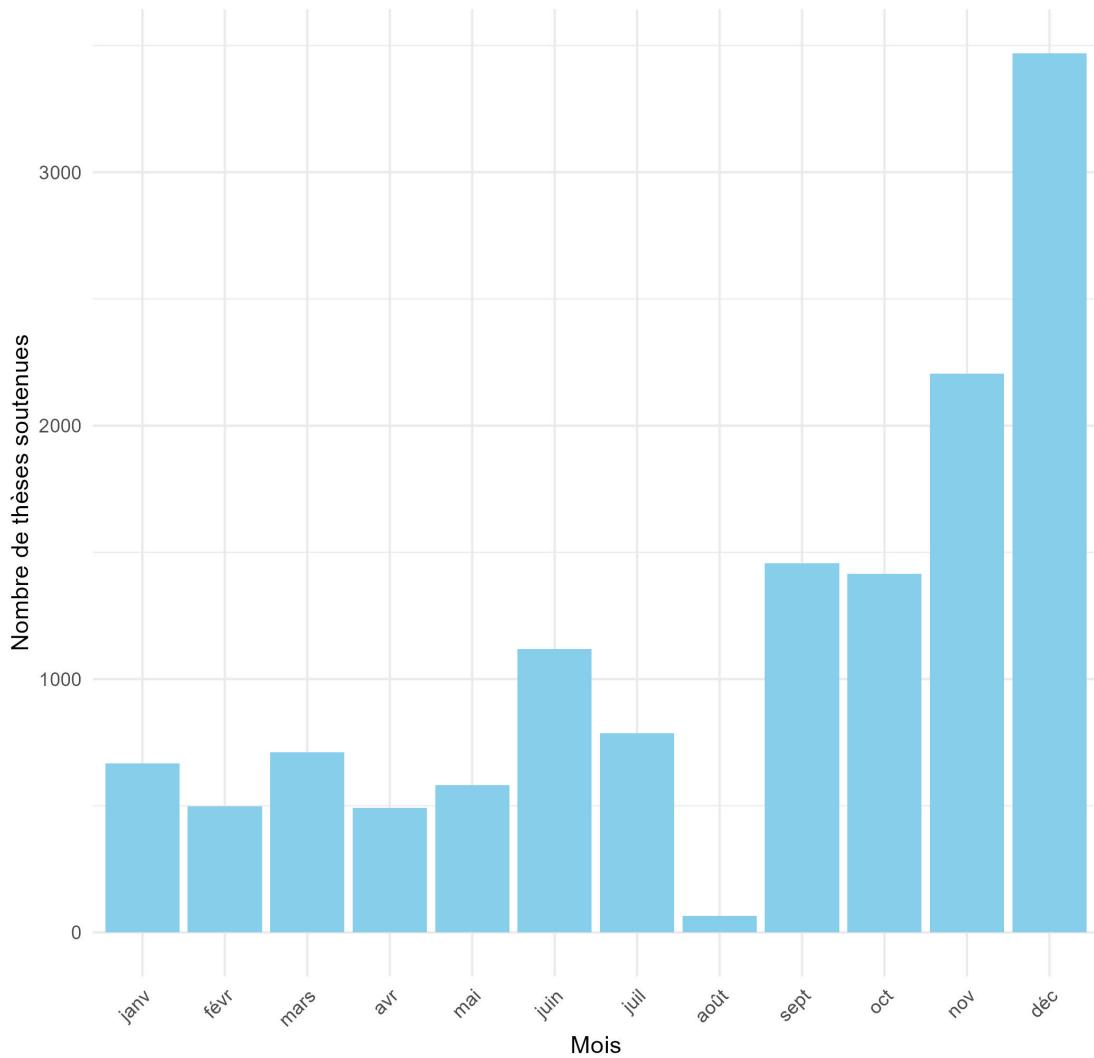


FIGURE 7 – Nombre de thèses soutenues par mois sans la date du 1er Janvier

Dès lors que la date du 1er Janvier est enlevé, la répartition mensuelle des thèses soutenues devient plus cohérente, avec un d'août relativement creux et des mois de Novembre et de Décembre assez importants. Ainsi la seule date du 1er Janvier (et non le mois tout entier) est responsable de ce déséquilibre (volontairement renseignée).

Cependant, avec une autre mise en perspective nous pouvons avoir une lecture divergente des résultats. Il apparaît que les thèses soutenues un 1er Janvier sont uniquement surreprésentées pour les années antérieures à 2016. Pour cette année et les suivantes, le mois de décembre devient le mois avec le plus de thèses soutenues (environ 3500 par an). Le graphique ci-dessous nous permet une lecture encore plus visuelle de ces disparités en fonction de l'année.

Répartition des dates de soutenance par mois pour chaque année (2005-2018)

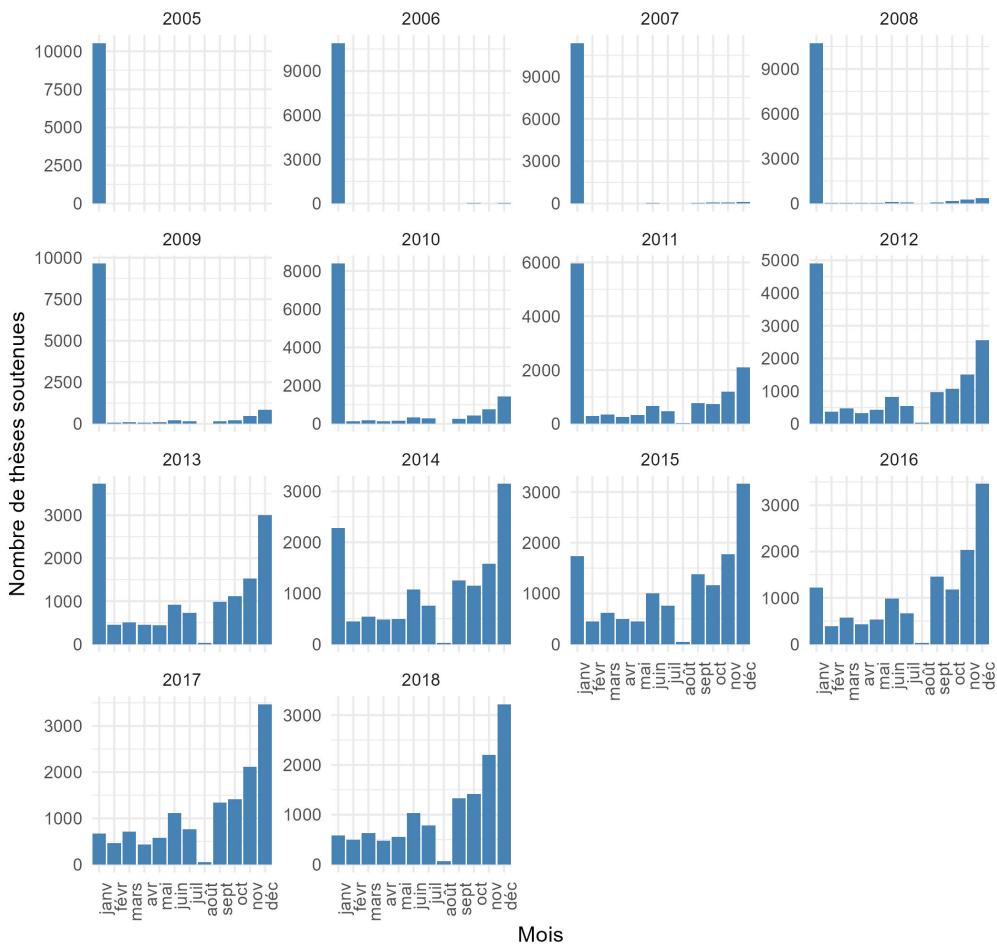


FIGURE 8 – Nombre de thèses soutenues par mois en fonction de l’année

Le graphique ci-dessous nous permet une lecture encore plus visuelle de ces disparités en fonction de l’année.

Nombre de soutenances par mois et année (incluant le 1er janvier)

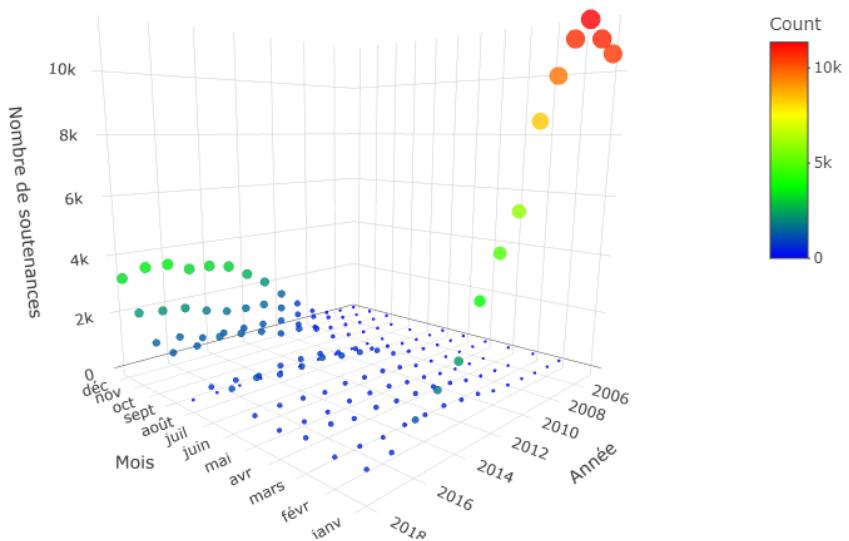


FIGURE 9 – Thèses soutenues par année et par mois

### 3.3 Analyse de la distribution des langues d'écriture des thèses

Quant à la distribution des langues d'écriture des thèses, les résultats de plusieurs visualisations mettent en évidence de probantes évolutions :

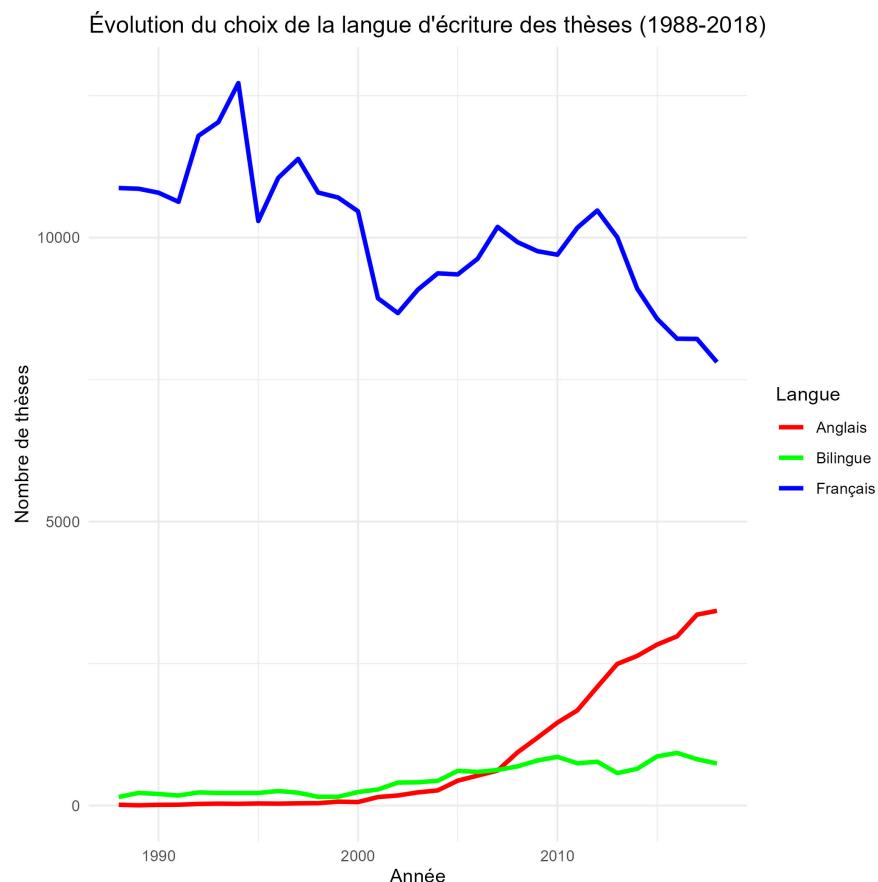


FIGURE 10 – Évolution de l'usage des langues d'écriture

On peut lire dans ce premier résultat que le nombre (et donc la proportion) de thèses écrites en français décroît, principalement au profit de thèses uniquement écrites en anglais. En 2005 moins de 1000 thèses étaient écrites uniquement en anglais, en 2018 elles ne sont pas loin de 4000. Il est à noter que si la langue d'écriture française est de moins en moins usitée au profit de la langue anglaise seule, les thèses dites bilingues (rédigées en français et anglais) connaissent une augmentation relative, leur nombre ayant doublé entre 1985 et 2018.

Par ailleurs au vue des formes respectives des 3 courbes, ils nous semble important de mentionner que d'autres facteurs semblent intervenir dans la récession de l'usage du française comme langue d'écriture des thèses.

Dans ce deuxième graphique nous décomposons plus en détails l'usage des langues d'écriture que l'on a couplé à certaines disciplines.

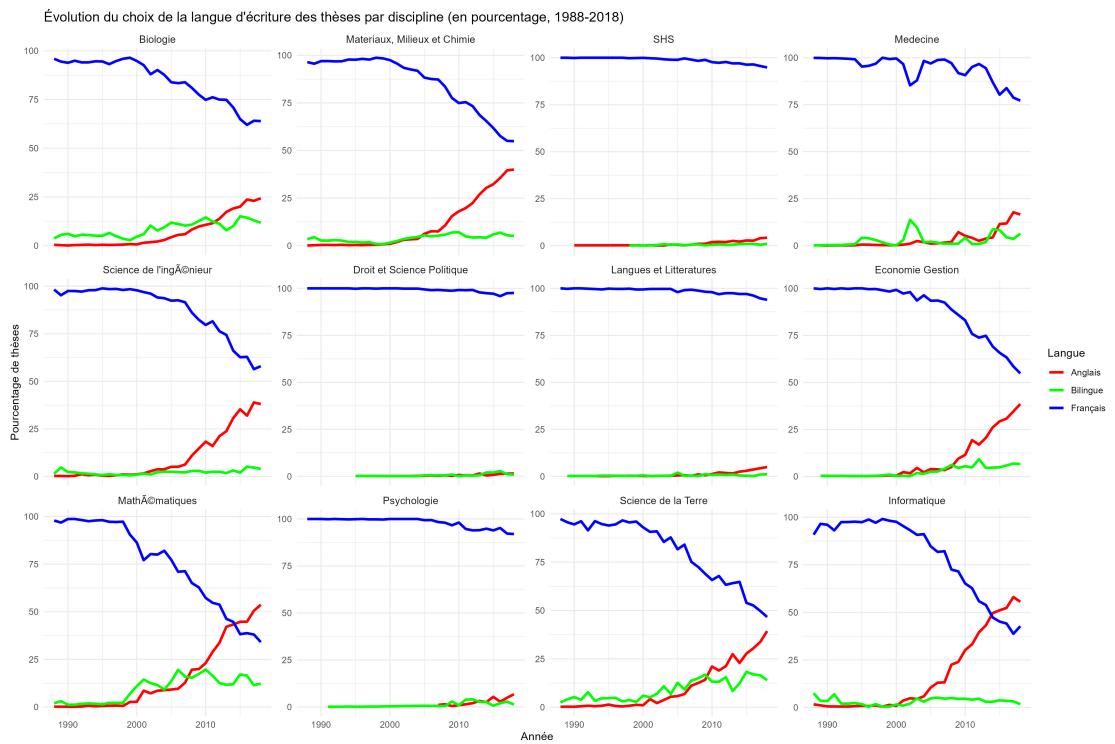


FIGURE 11 – Évolution des usages différenciés des langues d’écriture par discipline

On remarque que la distribution des langues d’usage ne suit pas la même évolution en fonction de la discipline concernée. Pour exemple on observe que le droit et la science politique n’ont laissé que très peu de place aux catégories bilingue et anglaise. Elle n’enregistre qu’une chute de 4 à 5 pourcent du français au profit des deux autres types.

A l’inverse certaines matières comme les Sciences de la Terre ou l’Economie et Gestion laissent progressivement plus de places aux langues d’écriture bilingue et surtout anglaise. On observe que c’est surtout à partir des années 2010 qu’une hausse significative de l’usage d’autres langues d’écriture que le français est enregistré.

Enfin on remarquera que pour l’informatique la langue d’écriture des thèses est depuis 2014 l’anglais et non plus le français. C’est d’ailleurs l’une des disciplines qui enregistrent la croissance la plus rapide de l’anglais en tant que langue d’écriture.

## 4 Discussion

### 4.1 Plan de la discussion

Dans cette partie du rapport nous nous attarderons sur l’interprétation des résultats traités plus haut. Il conviendra en premier lieu de porter un regard critique sur ces travaux, dont les potentielles limites subsistantes. Puis nous nous pencherons sur deux interprétations de ces résultats, l’une portant sur les Dates de soutenance, l’autre sur les langues d’écriture des thèses. Enfin, nous adopterons un regard plus systémique sur la manière dont les processus d’obtention d’un doctorat ont pu évolué, au regard de ces deux interprétations.

### 4.2 Analyse critique des résultats

La première limite de ces résultats est probablement leur fiabilité, qui dépend directement de la complétude des données introduites dans les outils statistiques et de visualisation. Or, comme notre méthode l’a décliné plus haut, les absences de données systématiques pour certaines variables (par exemple : la date de soutenance ou l’identifiant auteur) sont susceptibles de biaiser

les tendances observées. Par ailleurs, le changement de normes académiques a pu affecté les différentes données éprouvées et autorisé des erreurs d'appréciation qui conduisent à des atrophies ou hypertrophies dans la représentation de certaines disciplines (quasiment inexistantes il y a 30 ans pour certaines, quand d'autres comptent chaque année plusieurs milliers de thèses).

Ainsi, la représentativité des disciplines en est affectée, variant fortement en fonction des périodes et des sources d'enregistrement, et entraînant une perception biaisée des tendances de fond.

Nous aurions pu également nous appuyer sur d'autres jeux de données, car certaines limitations techniques ont restreint notre capacité à observer certaines anomalies subtiles. Par exemple, des visualisations plus avancées de la corrélation entre les langues d'écriture et des types de collaboration internationale (des degré d'intégrations dans la recherche internationale) auraient permis de mettre en évidence des sensibilités d'évolution différentes.

#### **4.3 Interprétation des résultats concernant les dates de soutenance**

Nous avons mis en évidence le fait que certaines dates de soutenances (dont le 1er Janvier) étaient surreprésentées avant 2016. Les causes de cette surreprésentation sont multiples. D'une part le référencement numérique tardif de certaines branches de la recherche française a pu conduire à un manque de rigueur dans la saisie des-dites dates de soutenances. Willison, J. W. (2011). D'autre part, il est possible que les processus de standardisation post 2016 soient bien plus performants que les précédents, qui dataient par défaut les thèses soutenues le 1er Janvier. Beaucoup plus rare mais possible, l'automatisation du remplissage de la date de soutenance autorise une certaine rigueur qui n'était probablement pas présente.

#### **4.4 Interprétation des résultats concernant les langues d'écriture**

La première interprétation que l'on peut faire à partir des résultats concernant le choix des langues d'écriture est que toutes les disciplines ne reconnaissent pas l'importance de l'anglais comme moyen pour plus de considération. Cela peut s'expliquer par exemple pour le cas des Sciences Politiques, dont les travaux de recherche sont généralement propres à chaque pays et donc beaucoup plus hermétique à une transmission du savoir extra-nationale. A l'inverse, les disciplines comme l'informatique ou les mathématiques relèvent plus aisément de savoirs universelles qui, s'ils sont rédigés dans la langue de la recherche académique internationale (l'anglais), peuvent impliquer un certains rayonnement. Gunnarsson, B. L. (2009). Enfin la question des doctorants étrangers en France a pu potentiellement affecter la langue d'écriture, à nouveau les mathématiques et l'informatique sont des milieux plus perméables à une population étudiante (voire encadrante) hétéroclite. Burgess, J. (2016)

#### **4.5 Conclusion**

Ainsi, l'analyse des données concernant le référencement des thèses nous a permis de conclure sur l'existence d'un "capital de proximité" entre certains pans de la recherche scientifique française et la recherche scientifique mondiale, qui provoque des mutations dans la réalisation d'un doctorat et de la soutenance de thèses. Par ailleurs nous avons pu mettre en évidence la présence de culture académique, notamment dans les méthodes, qui reflètent voire accentuent certaines rivalités académiques déjà très présentes dans certaines disciplines en France. En outre, ces travaux nous ont en partie permis de bâtir un lien entre métadonnées de thèse et "climat" général de la recherche académique française.

### **5 Références**

- Burgess, J. (2016). PhD Examination and the Institutional Context : A Study of Practices and Outcomes. *Studies in Higher Education*, 41(3), 342-358.
- Gunnarsson, B. L. (2009). Language Choice in Academic Writing : A Comparative Study of Linguistic Practices in Swedish and English Research Articles. *Journal of English for Academic Purposes*, 8(3), 190-206.
- Willison, J. W. (2011). Planning for the PhD : Administrative Aspects and Issues of Doctoral Supervision. *Higher Education Quarterly*, 65(2), 160-174.