Projet de Statistique Appliquée

– Mémoire –

Machine Learning à l'Assemblée Nationale

- Projet encadré par Jules Depersin et Nicolas Schreuder -

Auteures :
Clémentine Abed Meraim
Romane Gajdos
Clara Le Gallic-Ach
Poppée Mongruel

Novembre 2020 - Mai 2021

Sommaire

Introduction

Si de nombreux domaines d'application de la science des données sont aujourd'hui bien connus du grand public et entrent petit à petit dans nos vies quotidiennes (voiture autonome, reconnaissance faciale, ...), certaines thématiques le sont moins, en-dehors des grands scandales qui leur sont liés. C'est le cas du champ politique. Le développement de la science des données, allant de pair avec la collecte et l'ouverture des données, a permis le déploiement de nouveaux outils d'apprentissage automatique dont l'exploitation change d'une part fondamentalement le fonctionnement politique (campagnes électorales ciblées, prédiction de l'impact d'une politique, ...) et ouvre d'autre part de nouveaux champs dans le domaine de la recherche en politique. En 2008, Barack Obama et son équipe de campagne ont ainsi croisé des techniques avancées d'analyse de données et de publicité ciblée au sein de la sphère politique. Au total, au cours des campagnes américaines de 2020, 200 millions de dollars ont été dépensés pour la création de publicités ciblées sur Facebook par des algorithmes d'intelligence artificielle, contre 643 000 dollars en 2008 pour la campagne d'Obama [1]. Plus retentissantes sont peut-être les affaires telles que celle autour de l'entreprise Cambridge-Analytica dans le cadre du Brexit, accusée d'avoir collecté les données personnelles de 50 millions d'utilisateurs de Facebook, dans le but de cibler des messages favorables au camp du Leave [2].

Si ces exemples montrent davantage des cas d'utilisation de certaines données des citoyen.ne.s par les responsables politiques pour comprendre et orienter les votes, il est intéressant de se saisir de la contraposée de ces applications, c'est-à-dire d'utiliser les données disponibles pour tenter d'analyser et de comprendre les votes de ces décideur.se.s politiques. Cette démarche semble d'autant plus pertinente qu'elle fait écho aux politiques d'*open data* mises en place par de plus en plus d'institutions gouvernementales. C'est le cas de l'Assemblée nationale en France, qui a lancé en 2015 sa plateforme d'*open data* [3].

L'Assemblée nationale est la Chambre basse du Parlement français, le Sénat en étant la Chambre haute. Elle est composée de 577 député.e.s élu.e.s au suffrage universel direct pour une durée de cinq ans. Chaque député.e représente une circonscription et peut être affilié.e à un parti politique (seuls 4.2% des député.e.s ne sont affilié.e.s à aucun parti). L'Assemblée nationale a deux rôles principaux dans la vie politique française : le vote de la loi et le contrôle de l'action du gouvernement. Le vote d'une loi est initié par la proposition d'un texte de loi par le Premier ministre (il s'agit alors d'un projet de loi) ou par un.e député.e (on parle dans ce cas de proposition de loi). Le texte est ensuite transmis à la commission compétente sur le sujet, où il sera discuté et amendé par les député.e.s membres de cette commission. Une fois adopté en commission, le texte est présenté dans l'hémicycle à l'ensemble des député.e.s. Après débat, chaque article et chaque amendement doivent être votés par les député.e.s. Il y a donc plusieurs votes pour un même texte. Une fois adopté par l'Assemblée nationale, le texte est soumis au Sénat. La Chambre haute peut adopter le texte tel quel ou peut le modifier, auquel cas les modifications devront être votées à nouveau par l'Assemblée nationale. Le second rôle de l'Assemblée nationale est le contrôle du gouvernement. Elle dispose à cet effet de trois outils principaux : le vote de confiance, la motion de censure et l'engagement de la responsabilité du gouvernement sur un texte.

Au-delà de ses fonctions centrales au sein de l'appareil législatif, l'Assemblée nationale permet d'observer les rapports entre les différents groupes politiques français. Elle permet également de comprendre l'évolution de ces rapports à travers les positionnements relatifs des différents groupes en fonction des propositions de lois mais aussi en fonction du contexte politique global. Cette entité centrale du système institutionnel français offre de plus la possibilité d'analyser les rapports des partis politiques à certains sujets, de même que le comportement et la cohésion interne de ces partis.

En nous penchant sur les données de l'*open data* de l'Assemblée Nationale dans le cadre de notre projet de Statistique appliquée, dont le sujet initial était "Machine Learning à l'Assemblée nationale", nous étions à la fois libres dans le choix de l'angle d'approche et dans celui des méthodes mises en oeuvre. Après nous être renseignées

sur les spécificités de l'Assemblée et nous être appropriées les données, nous nous sommes demandé dans quelle mesure les données de vote des député.e.s permettaient de rendre compte du fonctionnement de cette institution.

Dans un premier temps, nous avons produit une synthèse sur l'*open data* afin de mieux saisir le sens de cette notion floue et trop souvent utilisée dans les médias sans sa définition précise. Dans un second temps, nous nous sommes penchées sur les données elles-mêmes et avons étudié les thèmes nous semblant les plus caractéristiques du fonctionnement de l'Assemblée : la cohésion au sein des partis, les demandeurs des scrutins (*ie* les identités des instances soumettant un projet de loi au vote des député.e.s), et l'abstention. Ces trois entrées nous ont également permis de passer en revue les principales entités agissant au sein de l'Assemblée : les partis, les député.e.s et les demandeurs. Nous avons ensuite cherché à retrouver ces caractéristiques à partir des données disponibles. Nous avons d'abord réfléchi à la possibilité de prédire l'appartenance d'un.e député.e à un parti à partir de ces données. Plus précisément, nous avons cherché à déterminer si l'historique de vote permettait d'identifier à lui seul l'appartenance à un parti. Ensuite, nous avons tenté de prédire l'abstention ou le vote d'un.e député.e en fonction du thème du scrutin à partir de ses caractéristiques politiques et socio-économiques.

1 L'open data de l'Assemblée Nationale : une mise à disposition récente de données au cœur de la vie démocratique...

Nous avons travaillé sur les données provenant du site d'*open data* de l'Assemblée nationale. Cet *open data* met à disposition les archives et les données actualisées des XIVème et XVème législatures. La création de ce site en 2015 s'inscrit dans le contexte plus général d'ouverture des données des administrations publiques en réponse à une demande accrue de transparence politique.

La notion d'*open data* est apparue pour la première fois en 1995 dans un rapport de l'Académie des sciences américaines [4] et désignait alors le partage des données dans le cadre de l'échange de connaissances et de savoirs, compris comme des biens publics (non rivaux et non exclusifs). L'*open data* désigne aujourd'hui la politique de diffusion volontaire des données, principalement issues des administrations publiques, mais aussi les données elles-mêmes : ouvertes et libres d'utilisation.

Les principes de Sebastopol, énoncés par une trentaine d'activistes du numérique lors d'une réunion en 2007, délivrent une définition commune de la notion d'*open data* [5]. Pour être qualifiées d'ouvertes, les données doivent être :

— complètes,
— primaires (telles que collectées à la source),
— opportunes (mises à disposition rapidement afin de préserver leur valeur),
— accessibles,
— exploitables,
— non discriminatoires (accessibles à tou.te.s sans aucune obligation préalable),
— non propriétaires (aucune entité n'exerce de contrôle exclusif sur les données),
— libres de droit.

En France, le droit à disposer de telles données relatives au fonctionnement de l'Etat remonte à la Déclaration des droits de l'homme et du citoyen de 1789, dont l'article 15 fait état du "droit de demander compte à tout agent public de son administration". Plus récemment, la loi Commission d'accès aux documents administratifs (CADA) de 1978 garantit l'accès de chacun.e aux données produites ou détenues par les administrations. L'*open data* inverse le processus en rendant les données ouvertes par défaut.

Cette ouverture des données publiques fait écho à la demande croissante de transparence et de responsabilité

publique, dans le contexte de ce que Pierre Cahuc et Yann Algan appellent la "société de défiance" [6]. L'*open data* permettrait une plus grande efficacité de l'action publique et une plus grande intégrité des responsables publics, tout en renforçant la participation des citoyen.ne.s. L'*open data* a donc un réel rôle à jouer dans le développement d'une démocratie plus représentative. Enfin, l'ouverture des données sous un format réutilisable permettrait de proposer de nouvelles ressources pour l'innovation économique et sociale. Dans le cas de la France, la plateforme *data.gouv* permet aux citoyen.ne.s d'enrichir, de modifier et d'interpréter les données mises à disposition par les services publics, en vue de co-produire des informations d'intérêt général.

En 2016, la France a présidé le Partenariat pour un gouvernement ouvert qui rassemble 75 Etats et des centaines d'organisations de la société civile qui "œuvrent pour la transparence de l'action publique, la participation citoyenne et l'innovation démocratique" [7]. En 2019 la France a consolidé son rôle dans la politique mondiale d'*open data* en se plaçant deuxième selon l'OURdata Index de l'OCDE et troisième selon l'Open Data Maturity Report de la Commission européenne.

L'ouverture des données de l'Assemblée nationale depuis juin 2015 est ainsi au coeur de ce processus. Les député.e.s sont en effet directement élu.e.s par les citoyen.ne.s et les représentent à l'Assemblée. Si les données dont nous nous sommes saisies dans ce rapport ne concernent que peu des sujets relatifs à l'intégrité et la corruption, elles donnent la possibilité aux citoyen.ne.s d'exercer leur droit de regard sur les comportements de vote de leurs représentant.e.s. Dans le contexte actuel de défiance des citoyen.ne.s vis-à-vis du monde politique, ce droit semble fondamental. Le président de l'Assemblée nationale a ainsi décrit, au moment de l'ouverture de ces données, "cette évolution technologique [comme] une avancée démocratique majeure pour notre pays. L'ouverture des données améliore la transparence de l'action parlementaire et favorise la participation des citoyens à la vie publique." [8].

2 ...permettant de réaliser un état des lieux de la XVème législature...

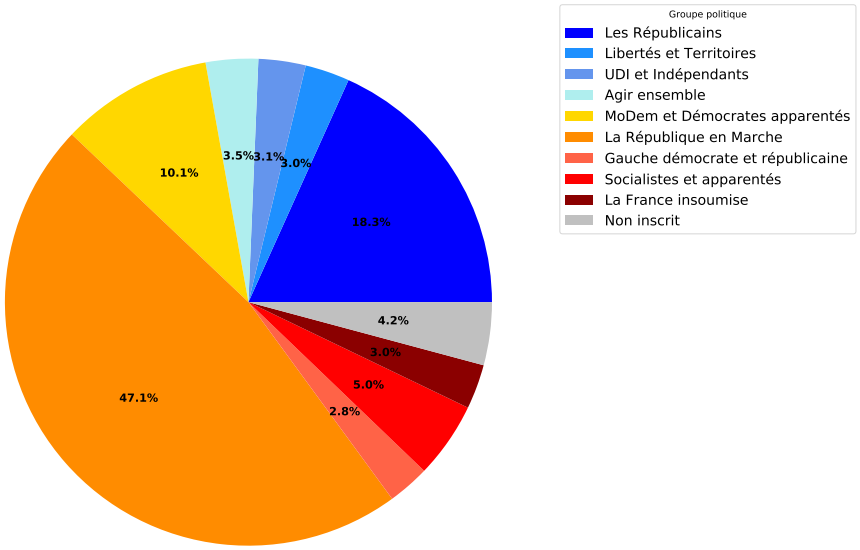
2.1 Présentation de la base et statistiques descriptives générales : un premier aperçu des données mobilisées dans ce rapport

Nous avons construit notre base de données à partir des données de votes des député.e.s français.e.s lors des scrutins publics de la XVème législature en cours (entre le 4 juin 2017, début de la législature, et le 6 novembre 2020, date à laquelle nous avons récupéré les données).

Cette base de données est constituée de 290 496 observations sur les votes des député.e.s lors des 3117 scrutins publics de la XVème législature s'étant tenus entre le 4 juin 2017 et le 6 novembre 2020. Chaque observation correspond à un scrutin, déterminé par le nom du projet de loi, le ou les groupes politiques à l'origine du scrutin, et la date du vote. Nous avons accès à la position de l'ensemble des député.e.s pour chaque vote. En effet, pour chaque scrutin, nous disposons des informations suivantes : le résultat du vote, le nombre total de votants, le nombre de votants s'étant exprimés "pour", le nombre de "contre", le nombre de non-votants (qui sont exclus du vote pour raison constitutionnelle), de non-votants volontaires (qui s'excluent d'eux-mêmes, par exemple pour éviter un conflit d'intérêt) et d'abstentionnistes. Aux individus présents au moment du scrutin mais se déclarant non-votants ou abstentionnistes, il convient de rajouter l'ensemble des député.e.s absent.e.s au moment du vote (qui ne sont pas visibles dans les données fournies par l'Assemblée) pour obtenir le nombre d'abstentionnistes réel. Dans le présent rapport, nous considérons les votant.e.s comme les député.e.s ayant voté pour ou contre, et les abstentionnistes comme regroupant les non-votant.e.s, les abstentionnistes présent.e.s au moment du scrutin et l'ensemble des député.e.s absent.e.s au moment du scrutin. Enfin, nous disposons de plusieurs caractéristiques concernant les député.e.s : leur nom, genre, âge, origine (circonscription, département et région) et leur catégorie

socio-professionnelle. Nous avons récupéré ces données nominatives sur le site de l'*open data*, sur le site officiel de l'Assemblée nationale et sur Wikipédia.

FIGURE 1 – Composition politique de l'Assemblée Nationale.



La **Figure 1** représente les 577 député.e.s de l'Assemblée Nationale. Certain.e.s député.e.s ne sont apparenté.e.s à aucun parti (on compte 24 non inscrits), et on pourra distinguer par la suite des sous-groupes parmi ces principaux partis. De plus, seul.e.s 575 député.e.s ont participé au moins une fois au vote. C'est donc ce chiffre que l'on retiendra comme étant le nombre total de votants, plutôt que les 577 député.e.s théoriques.

Une première analyse nous indique qu'en moyenne, un.e député.e vote pour 505 projets de lois différents, avec un écart-type de 289 : certains individus participent à beaucoup de scrutins différents, et d'autres à très peu. La plus faible participation est ainsi de 5 scrutins, contre 2165 pour la participation la plus élevée. Il y a en moyenne 482 abstentionnistes par scrutin, avec un écart-type de 83 : on compte donc un fort phénomène d'abstentionnisme des élu.e.s. Par ailleurs, sur l'ensemble des scrutins proposés, 27.7% ont été adoptés.

2.2 Analyse du fonctionnement de l'Assemblée nationale et de ses spécificités : les partis, les demandeurs, l'abstention

Après avoir réalisé quelques statistiques descriptives générales sur notre base, nous nous sommes concentrées sur trois axes caractéristiques de l'Assemblée nationale, afin de mettre en lumière son fonctionnement interne.

2.2.1 L'Assemblée nationale, une instance parlementaire : une vue d'ensemble des positions politiques intra et inter-groupe

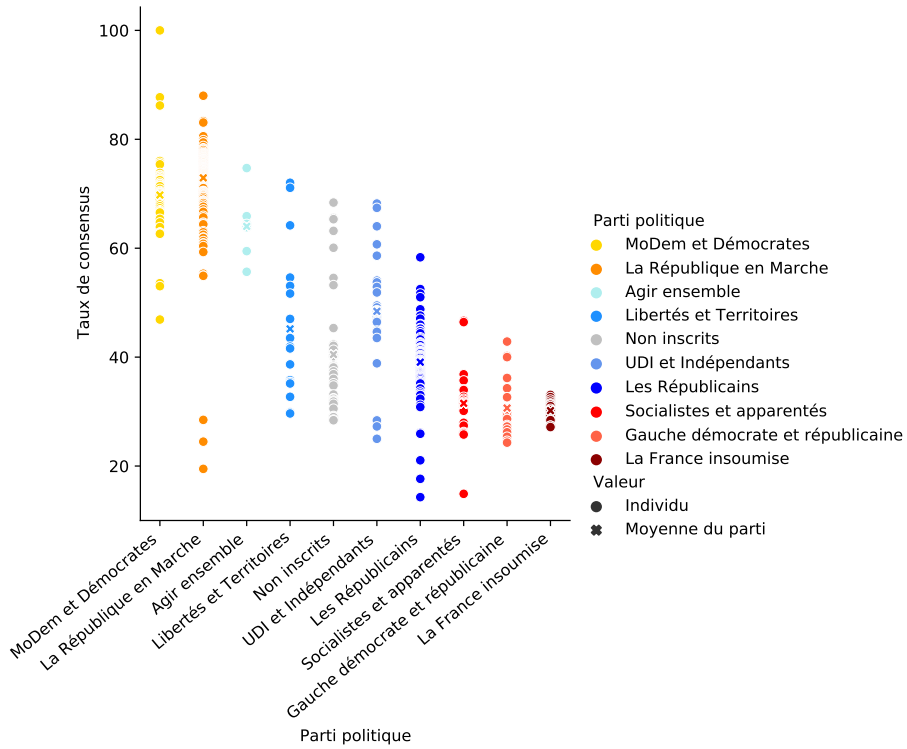
Dans un premier temps, nous nous sommes intéressées au lien entre un.e député.e et son parti. En effet, 96% des député.e.s sont élu.e.s au nom d'un parti et revendiquent au moment de leur élection leur appartenance à ce parti. Nous avons donc cherché, d'une part, à étudier les positions de ces partis les uns par rapport aux autres à l'Assemblée nationale. D'autre part, nous avons tenté d'évaluer la cohésion au sein de chaque parti.

L'accord de chaque député.e à la décision finale de l'Assemblée : le poids du vote

Afin d'établir des tendances générales entre et/ou au sein des partis politiques, nous avons décidé d'étudier la position des député.e.s vis-à-vis des décisions finales de l'Assemblée nationale. Nous avons pour cela créé un indice d'accord : en observant pour chaque votant s'il était en accord ou en opposition avec le résultat adopté pour chaque scrutin, nous avons pu établir une moyenne d'accord comprise entre 0 et 100. Plus cet indice est proche de 100, plus l'individu est en accord avec les décisions de l'Assemblée (c'est-à-dire qu'il a souvent voté pour une loi qui a ensuite été adoptée, ou voté contre une loi qui a ensuite été refusée). L'écart-type de cet indice au sein de chaque parti donne une mesure de la cohésion du parti. La cohésion des partis est une donnée importante dans les systèmes parlementaires, où les député.e.s suivent souvent la ligne officielle de leur groupe politique. Pour certain.e.s auteur.ice.s, " la discipline de vote est [même] devenue, comme le démontre sa constance, une caractéristique de la Cinquième République" [9]. Autrement dit, quand il s'agit de voter, la cohésion des groupes politiques est relativement élevée. Cela est dû à la proximité idéologique des membres d'un même groupe politique, mais également à des pressions exercées sur les député.e.s dans le cadre de stratégies de vote.

Nous avons représenté, dans la **Figure 2**, l'indice d'accord de chaque député.e selon son parti politique. Cela nous permet de visualiser l'accord individuel de chaque député.e face aux décisions finales de l'Assemblée, mais également d'avoir une vue d'ensemble sur la cohésion au sein des partis politiques et entre les partis.

FIGURE 2 – Position et dispersion des député.e.s face aux décisions de l'Assemblée Nationale.



On observe des résultats cohérents avec la composition politique de l'Assemblée. Le MoDem et La République en Marche (LREM), majoritaires dans l'hémicycle, apparaissent comme les partis les plus en accord avec les décisions adoptées par l'Assemblée, avec des moyennes respectives de consensus de 73 et 70. Les partis les plus à gauche dans le spectre politique sont ceux les plus en désaccord avec les résultats finaux, mais aussi les moins dispersés, comme faisant bloc face aux décisions majoritaires. En effet on observe respectivement pour les Socialistes, la Gauche

Démocrate et Républicaine et la France Insoumise des moyennes d'accord de 31, 30 et 30, ainsi que des écarts types respectifs de 6, 5 et 1. Les Républicains jouent également un rôle important dans l'opposition. Ils constituent le deuxième groupe majoritaire de l'Assemblée (117 député.e.s) et affichent un accord moyen de 39 et un écart-type de 6. On remarque aussi que l'individu le plus en désaccord avec les décisions de l'Assemblée Nationale fait parti des Républicains. Seuls 14% de ses votes étaient en accord avec la décision finale.

Concernant la dispersion intra-parti, on observe quelques député.e.s LREM dont les votes s'écartent beaucoup de la majorité du groupe. Cela peut notamment s'expliquer par la taille du groupe (322 député.e.s), un effectif largement supérieur aux autres groupes, mais aussi par l'histoire de ce parti récent qui rassemble des député.e.s de diverses origines politiques. Par ailleurs, en s'attardant sur le groupe des Non-inscrits, on remarque une moindre dispersion de ceux-ci, alors qu'ils n'ont *a priori* aucune raison de s'accorder sur leurs votes. Il est donc relativement étonnant de remarquer que ce groupe n'est pas le plus disparate, alors qu'il est un groupe non homogène de député.e.s non membres ni apparenté.e.s aux groupes parlementaires. Les "Non-inscrits" regroupent en effet aussi bien Marine Le Pen (Rassemblement National ou RN, parti non inscrit à l'Assemblée) que Cédric Villani (ancien LREM). En observant plus précisément leurs origines politiques, on peut diviser ce parti entre une moitié de député.e.s de droite et d'extrême droite (RN, Ligue du Sud, Debout la France), un quart de centriste (étiquette "Divers Centre") et un quart de député.e.s de gauche (Les Nouveaux Démocrates et Génération Ecologie).

Cette première analyse nous permet de mettre en évidence une tripartition de l'Assemblée nationale, visible sur la **Figure 2** :

— une majorité dispersée et en accord avec les décisions finales de l'Assemblée,
— une opposition de droite moyennement dispersée et moyennement en accord avec les décisions finales de l'Assemblée,
— une opposition de gauche peu dispersée et peu en accord avec les décisions finales de l'Assemblée.

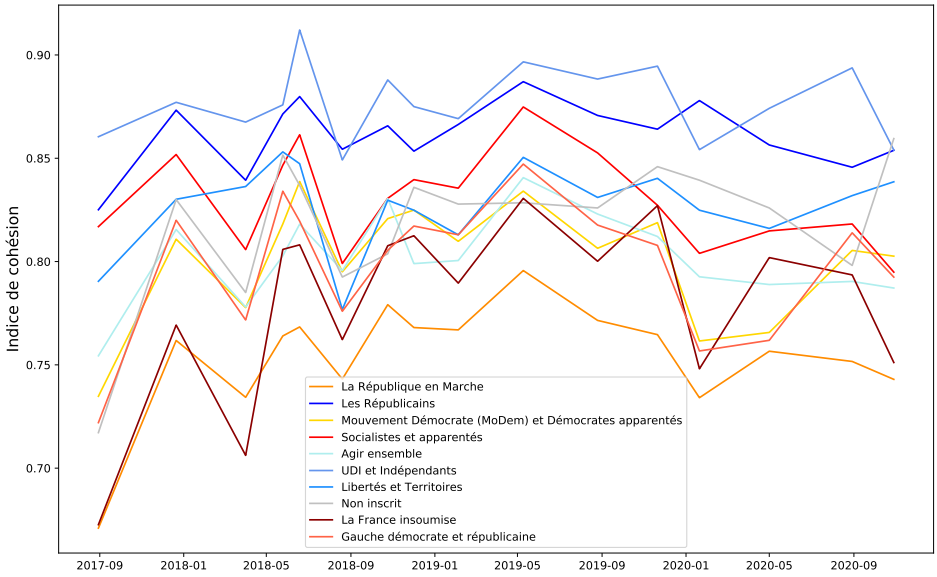
Cette étude permet ainsi de représenter globalement la situation du spectre politique vis-à-vis des décisions adoptées à l'Assemblée nationale. Pour autant, si l'indice que nous avons construit nous donne des résultats satisfaisants et très informatifs, deux nuances sont à lui apporter. Tout d'abord, il ne prend pas en compte l'abstention. Or l'abstention est une donnée cruciale dans l'analyse des votes, en particulier à l'Assemblée où elle est très élevée. De plus, l'indice d'accord que nous avons construit est une représentation statique de la position des député.e.s et des partis par rapport aux décisions finales de l'Assemblée nationale. Cette représentation ne rend donc pas compte de la recomposition de certains partis ou de l'évolution de certain.e.s député.e.s d'un bord politique à un autre, que ce soit à l'occasion d'un scrutin particulier ou de mutations plus globales du contexte politique. Il est donc important d'une part d'intégrer l'abstention dans notre mesure de la cohésion des partis, et d'autre part d'apporter une dimension temporelle à notre analyse.

L'indice de cohésion de chaque parti et son évolution dans le temps : le poids de l'abstention

Afin de mieux représenter et analyser les positions politiques intra et inter-partis, nous avons donc décidé de rajouter un poids à l'abstention et avons privilégié une analyse dynamique. Nous nous sommes éloignées de l'approche individuelle de la partie précédente et avons opté pour une approche par parti. Nous avons cherché à représenter la cohésion interne de chaque parti et son évolution au cours du temps.

Pour ce faire, nous avons cette fois utilisé un indice référencé dans la littérature académique. Nous avons choisi l'*Agreement Index* développé par Simon Hix, Abdul Noury et Gérard Roland en 2007. Ces chercheurs ont souhaité remédier à l'absence de prise en compte de l'abstention dans de précédentes analyses, dans le but d'étudier la pertinence du schéma parlementaire comme réponse au déficit démocratique de l'Union européenne [10]. Cet indice prend des valeurs comprises entre 0 et 1 ; 0 correspondant à un groupe complètement divisé et 1 à un groupe dans

FIGURE 3 – Évolution de la cohésion des partis au cours de la XVe législature.



lequel tous les individus votent la même chose. Il est calculé pour chaque scrutin à partir des données de vote de tou.te.s les député.e.s d'un parti selon la formule suivante [11] :

$$AI = \frac{\max(P, C, A) - \frac{1}{2}(P + C + A - \max(P, C, A))}{P + C + A} \tag{1}$$

où P = nombre de votes "pour", C = nombre de votes "contre", et A = nombre d'abstentions.

Nous avons calculé l'indice de cohésion de chaque parti en faisant des moyennes tous les 200 scrutins. Nous avons représenté l'évolution de cet indice dans la **Figure 3**.

On observe sur ce graphique des indices de cohésion globalement élevés, compris entre 0.70 et 0.90, ce qui traduit bien une forte "discipline de vote" [9]. On observe également une répartition droite/gauche de ces indices : la droite, représentée par les partis en bleu, a une plus grande cohésion que la gauche, représentée par les partis en rouge. On note de plus que le parti de gouvernement, La République en Marche, a l'indice de cohésion le plus faible sur l'ensemble de la période. Ce dernier point est cohérent avec la forte dispersion de l'indice d'accord pour ce même parti (**Figure 2**). A l'inverse, Les Républicains apparaissent très dispersés dans leurs votes lorsque l'on s'intéresse à l'indice d'accord, mais la cohésion du parti apparaît beaucoup plus importante lorsque l'on regarde l'indice de cohésion. De même, la France insoumise apparaît très peu dispersée sans prise en compte de l'abstention (**Figure 2**) alors qu'elle l'est beaucoup plus dès que l'on considère l'abstention : c'est le parti ayant l'indice de cohésion le plus faible après La République en Marche (**Figure 3**). La tripartition observée ici diffère donc de celle obtenue sans prise en compte de l'abstention : l'opposition de droite est plus dispersée que l'opposition de gauche, qui elle-même est plus dispersée que la majorité. Cette comparaison illustre l'importance de la prise en compte de l'abstention dans l'analyse des votes et la nécessité de choisir un indice adapté à ce que l'on cherche à analyser. En effet, l'indice d'accord est pertinent pour étudier le choix du vote ("pour" ou "contre"), puisqu'il permet étudier

l'accord moyen de chaque parti à la décision finale de l'Assemblée nationale, laquelle reste valide même sans prise en compte de l'abstention. A l'inverse, l'indice de cohésion rend compte de l'importance de l'abstention dans le fonctionnement des partis et de l'Assemblée nationale. Cette double entrée permet de mettre en évidence les stratégies différenciées des partis d'opposition : l'opposition de gauche a davantage tendance à faire bloc contre la majorité en votant à l'inverse de celle-ci, alors que l'opposition de droite tient sa cohésion d'un recours concerté à l'abstention.

L'évolution dans le temps de la cohésion au sein des partis est à mettre en lien autant avec le contexte politique global et certains événements propres à chaque parti (dissensions, changement de tête de file,...) qu'avec le vote de scrutins particuliers. On remarque tout d'abord que les moments de forte cohésion sont les mêmes pour tous les partis : mai-juin 2018, novembre 2018 et mai-juin 2019 par exemple. Le pic de cohésion autour de mai-juin 2018 peut être en partie expliqué par la proposition de révision constitutionnelle contenue dans les projets de lois "pour une démocratie plus représentative, responsable et efficace", présentés pour la première fois à l'Assemblée le 9 mai 2018. La forte hausse de la cohésion en mai 2019 est quant à elle liée à la tenue des élections européennes qui ont resserré les député.e.s autour de la ligne de leur parti. Si l'on s'intéresse aux partis pris indépendamment les uns des autres, on remarque que certains partis ont un indice de cohésion stable dans le temps, comme les Républicains. Ce parti, fondé en 2002 en tant qu'Union pour un Mouvement Populaire (UMP) avant d'être renommé en 2015, est bien ancré politiquement et peu sujet aux dissensions. A l'inverse, La France Insoumise a une cohésion très variable dans le temps, qui illustre les conflits internes autour de la ligne stratégique du parti et du manque de démocratie interne [12]. On remarque également que la cohésion au sein de La République en Marche, bien que la plus faible de tous les partis, augmente de manière régulière entre septembre 2017 et mai 2019. Cela traduit la consolidation de ce parti créé par Emmanuel Macron en avril 2016 et encore relativement nouveau au début de la législature. La rupture à la mi-2019 correspond aux dissensions grandissantes au sein du parti LREM, qui se traduisent par une augmentation du nombre de démissions et d'exclusions : on en compte 8 entre septembre 2017 et février 2019, contre 32 entre l'été 2019 et décembre 2020 [13].

Les deux indices étudiés nous donnent donc une première image des rapports de force intra et inter-partis à l'Assemblée. L'historique de votes ("pour" ou "contre") des député.e.s permet d'identifier les député.e.s et les partis les plus en accord avec les décisions finales de l'Assemblée et celles et ceux qui sont majoritairement dans l'opposition. La prise en compte de l'abstention permet quant à elle de dégager des rapports de force légèrement différents, mais peut-être plus représentatifs du fonctionnement réel de l'Assemblée nationale. En effet, le vote des lois n'est qu'une partie de l'activité des député.e.s, qui ne sont que peu présent.e.s dans l'hémicycle. Si le vote "pour" ou "contre" est l'enjeu principal d'un processus démocratique, "voter" ou "ne pas voter" est, dans le cas de l'Assemblée, un enjeu loin d'être secondaire, et qui peut en partie déterminer l'issue finale du vote.

2.2.2 Les demandeurs des scrutins : un proxy permettant de visualiser le clivage entre la majorité et l'opposition à l'Assemblée

Une autre particularité caractérisant un scrutin est l'identité du demandeur, c'est-à-dire l'entité qui soumet la proposition de loi au vote des député.e.s. Cette information a retenu notre attention puisqu'elle est susceptible d'influencer le vote des député.e.s : ce dernier dépendant de leur parti d'appartenance, il dépend aussi *a fortiori* du groupe proposant le scrutin. Afin d'affiner notre étude des votes en fonction des partis, nous avons ainsi étudié les résultats des scrutins en fonction de leur demandeur, c'est-à-dire en fonction de l'instance qui en est à l'origine. Une telle analyse pourrait notamment donner un aperçu de l'articulation des rapports entre partis et de leur influence relative.

Il convient de préciser que nous avons considéré les différents demandeurs en respectant les mêmes appellations

Table 1 – Taux moyen d'adoption des scrutins en fonction du demandeur.

Demandeur	Abstention	Demandeur	Abstention
Gouvernement	100%	Socialistes et apparentés	24,9%
Conférence des Présidents	92,6%	Gauche démocrate et républicaine	23,9%
LREM	86,6%	Agir Ensemble	20%
Commission	75%	Les Républicains	19,9%
MoDem	48,8%	Nouvelle Gauche	17,6%
Libertés et Territoires	32,3%	France Insoumise	17,3%
UDI, Agir Ensemble, Indépendants	26,8%	Ecologie, Démocratie, Solidarité	11,1%

que dans la base de données : ces derniers ne coïncident pas tout à fait avec les différents groupes d'affiliation politique étudiés précédemment, et cela sur plusieurs plans. D'une part, certaines instances de l'Assemblée autres que des partis politiques peuvent proposer des scrutins, comme la Commission, le Gouvernement ou encore la Conférence des Présidents. De plus, les propositions de scrutins n'émanent souvent pas d'un seul groupe, mais résultent de la collaboration entre plusieurs groupes différents. Pour simplifier les innombrables associations possibles qui en résultent, nous avons considéré séparément chaque groupe comme demandeur de tels scrutins (il y a ainsi des scrutins redondants dans les statistiques effectuées). D'autre part, certains groupes ont connu des évolutions au cours du temps (scission, fusion, ...) et l'approche statique faite ici nous a amené à les considérer comme distincts : nous avons distingué par exemple le groupe Agir Ensemble, qui est issu d'une scission du groupe UDI, Agir Ensemble et Indépendants, ou encore le groupe Écologie, Démocratie, Solidarité, qui est issu d'une scission de LREM et qui disparaît par ailleurs en octobre 2020. Il nous paraissait en effet pertinent de ne pas regrouper ces partis, dans le sens où de telles évolutions étaient souvent le reflet d'un changement de ligne directrice d'un parti et impliquaient des comportements différents de la part des député.e.s : les considérer de façon indépendante permet d'offrir un perspective plus fine sur leur influence relative et leur articulation par rapport aux autres partis.

Il ressort tout d'abord de cette analyse selon les demandeurs des scrutins que les groupes politiques de l'Assemblée proposent en moyenne 234 scrutins chacun (sur la période étudiée où sont proposés au total 3117 scrutins), mais avec un écart-type de 245, ce qui traduit une très forte dispersion entre les groupes concernant la fréquence de leurs propositions de lois. Parmi les groupes qui proposent le plus de scrutins, on observe dans l'ordre : Les Républicains (737 scrutins proposés), La France Insoumise (613 scrutins), La Gauche Démocrate et Républicaine (539 scrutins), Les Socialistes et apparentés (437 scrutins), et La République En Marche (321 scrutins). Du côté des groupes qui en proposent le moins - en dehors des instances qui sortent des groupes d'affiliation politique comme la Commission (4 scrutins proposés) ou le Gouvernement (7 scrutins) - on retrouve Agir Ensemble (5 scrutins), MoDem (43 scrutins) et Écologie Démocratie Solidarité (45 scrutins). Ceci paraît cohérent avec la composition politique de l'Assemblée, puisque les groupes qui occupent le moins la position de demandeurs correspondent (à l'exception de MoDem) à d'assez petits groupes exerçant donc éventuellement moins d'influence. A l'inverse, les groupes qui proposent de nombreux scrutins sont d'une part les plus grand groupes politiques de l'Assemblée (LREM et LR), et d'autre part certains groupes importants de l'opposition, comme La France Insoumise.

A partir de ces premières informations nous avons fait figurer dans la **Table 1** le taux d'adoption moyen des scrutins proposés selon les demandeurs, et ceci par ordre décroissant. Cela permet de déterminer quels sont les groupes dont les propositions de loi sont le plus souvent acceptées à l'Assemblée, c'est-à-dire quels sont les groupes ayant le plus de soutien de la part des autres partis.

On remarque ainsi que certains groupes politiques voient leurs propositions adoptées avec beaucoup de succès : c'est le cas de LREM dont les scrutins sont adoptés en moyenne à 86.6%. Les groupes dont les propositions de

TABLE 2 – Taux moyen d'abstention (réelle) en fonction du demandeur du scrutin.

Demandeur	Abstention	Demandeur	Abstention
Commission	88,8%	UDI, Agir Ensemble, Indépendants	85,4%
LREM	87,5%	France Insoumise	85%
Socialistes et apparentés	87,3%	Gauche démocrate et républicaine	84,3%
MoDem	87%	Nouvelle Gauche	84,3%
Agir Ensemble	86,9%	Les Républicains	83,9%
Ecologie, Démocratie, Solidarité	86,9%	Conférence des Présidents	14,9%
Libertés et Territoires	85,9%	Gouvernement	11,1%

loi sont les plus adoptées sont les groupes de la majorité (LREM et MoDem). De même, tous les projets de loi proposés par le gouvernement sont adoptés. A l'inverse, les partis minoritaires et d'opposition proposent des lois qui sont en majorité refusées : on compte seulement 11.1% de succès pour Écologie, Démocratie, Solidarité, et respectivement 17.3% et 17.6% pour La France insoumise et la Nouvelle Gauche. Ces observations semblent donc confirmer l'hypothèse de petits groupes hétérogènes faisant bloc face à une importante majorité. Il est intéressant de remarquer que le groupe La France Insoumise fait à la fois partie des demandeurs les plus fréquents mais également de ceux dont les scrutins sont le plus souvent rejetés. La proposition de scrutins semble donc constituer un moyen privilégié pour ce parti de l'opposition d'affirmer sa présence et de tenter d'atténuer la domination du groupe majoritaire, avec cependant peu de succès et beaucoup de difficultés à faire consensus. Au-delà de ces résultats polarisés, coexistent beaucoup de groupes au succès intermédiaire. Il s'agit de groupes se plaçant plutôt au centre du spectre politique (Libertés et Territoires ; UDI, Agir Ensemble, Indépendants ; Socialistes et apparentés), ce qu'illustrent bien les résultats moyens de leurs propositions de lois. Ce sont en effet des partis au positionnement indépendant qui ne se revendiquent ni de la majorité ni de l'opposition et sont en ce sens moins clivants que ces derniers. Il est également intéressant de remarquer que Les Républicains, avec leur position de deuxième groupe le plus important de l'Assemblée et leurs nombreux scrutins proposés, se placent parmi les demandeurs au taux d'adoption moyen de scrutin les plus bas, se situant donc bien comme la forte opposition de droite.

On retrouve donc bien la partition majorité/opposition déjà mise en évidence, avec un fort taux d'adoption pour la majorité. Cela confirme l'importance de la discipline de vote : les député.e.s appartenant à la majorité votent pour les propositions de loi de celle-ci et votent contre celles de l'opposition, tandis que les député.e.s de l'opposition (minoritaires) votent pour leurs propositions de lois mais contre celles de la majorité. Entre ces deux extrêmes, les partis intermédiaires votent aux côtés de la majorité ou de l'opposition en fonction des scrutins. Cette analyse permet de plus d'affiner nos observations en mettant en avant les groupes des Républicains et de La France Insoumise comme les partis les plus actifs, affirmant la place de l'opposition à l'Assemblée nationale.

Par ailleurs, le constat, fait précédemment, du rôle majeur que pouvait jouer l'abstention à l'Assemblée nous a amenées à analyser l'abstention moyenne des député.e.s en fonction des demandeurs des scrutins dans la **Table 2**, afin de voir si celle-ci variait en fonction du groupe politique soumettant la proposition de loi aux votes.

On remarque que l'abstention ne semble pas dépendre du demandeur de façon significative : le taux d'abstention est d'environ 85% pour tous les groupes politiques, à l'exception du Gouvernement et de la Conférences des Présidents. Il s'agit de deux instances qui ne proposent que peu de scrutins mais qui mobilisent très fortement les député.e.s, avec un taux d'abstention très faible (de l'ordre de 10 à 15%). Cela laisse suggérer que la mobilisation d'un.e député.e quant à un scrutin donné dépend moins du demandeur que de l'objet de ce scrutin. En effet, on n'observe pas de phénomène de boycott de certains demandeurs par un ou plusieurs partis. Malgré le rôle clef de l'abstention à l'Assemblée mis en évidence précédemment, le vote "pour" ou "contre" reste au coeur du

jeu démocratique. Ainsi, l'opposition de beaucoup de député.e.s à La France Insoumise passe par une très faible adoption des scrutins proposés par ce parti, et non pas par une abstention record à toutes ses propositions de loi.

Ces résultats, tout comme la construction de l'indice d'accord précédemment, sont à nuancer au sens où il s'agit d'une approche statique qui n'explique donc que partiellement les différences de comportement des votant.e.s en fonction du demandeur du scrutin. En particulier, l'évolution de ce comportement n'est pas prise en compte.

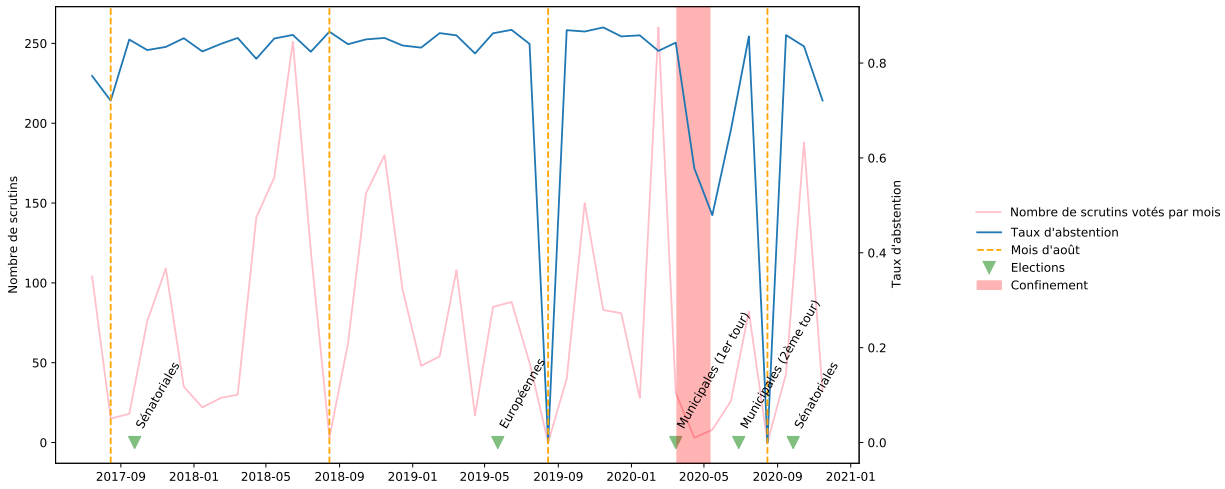
2.2.3 L'abstention : un paramètre crucial du vote à l'Assemblée Nationale

Alors qu'initialement la sociologie politique s'intéressait principalement au vote, la montée de l'abstention à partir du milieu des années 80 a placé l'étude des formes de démobilisation électorale au coeur de la discipline, avec des travaux fondateurs comme ceux de Daniel Gaxie en France [14] et ceux de Frances Fox Piven et Richard A. Cloward aux Etats-Unis [15]. La sociologie de l'abstention est aujourd'hui une thématique centrale de la sociologie politique qui cherche à mettre en évidence les déterminants de l'abstention. Une forte abstention est communément considérée comme étant le signe d'une démocratie malade. Les travaux sur l'abstention sont souvent centrés sur l'abstention des citoyen.ne.s, et sur les raisons qui poussent certains groupes sociaux à la démobilisation électorale. On peut mentionner à cet effet les travaux de Cécile Braconnier et Jean-Yves Dormagen sur l'abstention des milieux populaires [16] ou ceux de Vincent Tiberj sur l'abstention des jeunes [17]. Cependant, l'abstention au sein du monde politique n'est que peu étudiée. Pourtant, l'activité de vote des député.e.s est au coeur de la démocratie représentative, et l'abstention à l'Assemblée peut être, au même niveau que l'abstention des citoyen.ne.s, symptomatique d'un système démocratique en crise. Il est intéressant de noter que les caractéristiques socio-économiques et professionnelles des député.e.s les placent en-dehors des traditionnels bastions de l'abstention : la moyenne d'âge au sein de l'Assemblée est de 53 ans (avec un écart-type de 11), et plus de la moitié des député.e.s appartiennent aux "cadres de la fonction publique, professions intellectuelles et artistiques" ou aux "cadres d'entreprise" (306 député.e.s sur 575). On est donc loin des populations étudiées dans les travaux cités précédemment. Pourtant, l'abstention à l'Assemblée Nationale est très élevée : en moyenne, sur les 3117 scrutins que nous étudions, l'abstention est de 84%, avec un écart-type de 14. Il est important de noter que voter à l'Assemblée ne représente qu'une partie de l'activité professionnelle des député.e.s, là où les citoyen.ne.s ne sont appelé.e.s à voter que quelques fois dans l'année, ce qui ne rentre que peu en conflit avec leurs activités professionnelles. Cela conduit nécessairement à une forte abstention à l'Assemblée. De plus, les député.e.s peuvent être géographiquement éloigné.e.s du Palais Bourbon, et la sociologie traditionnelle de l'abstention a pu mettre en évidence le fait que l'éloignement du bureau de vote était un des facteurs expliquant la démobilisation électorale des citoyen.ne.s. Nous avons, dans cette partie, cherché à étudier l'abstention dans le cadre particulier de l'Assemblée nationale, afin d'en analyser les évolutions et les déterminants.

Dans un premier temps, nous nous sommes penchées sur les données dont nous disposions sur l'abstention des député.e.s pour étudier les variations du taux d'abstention. Nous avons construit une frise chronologique allant du début de la XVème législature jusqu'à la date où nous avons récupéré les données (novembre 2020), que nous présentons dans la **Figure 4**. Celle-ci comporte le taux d'abstention calculé tous les mois, ainsi que le nombre de scrutins votés chaque mois. Nous y avons également fait figurer des événements importants permettant de donner quelques pistes d'explication aux variations du taux d'abstention (élections, mois d'août, confinement).

On remarque tout d'abord la faible activité de l'Assemblée Nationale à chaque mois d'août. C'est le mois de vacances des député.e.s. En août 2019 et août 2020 aucun scrutin n'a été soumis au vote : cela explique les deux pics du taux d'abstention qui tombe à zéro. En août 2017, peu de scrutins ont été soumis au vote. Si les scrutins ont été proposés de manière rapprochée, cela a pu permettre aux député.e.s de se mobiliser davantage, ce qui expliquerait

FIGURE 4 – Evolution du taux d'abstention de la XVème législature, nombre de scrutins par mois et principaux événements de la vie politique et sociale française.



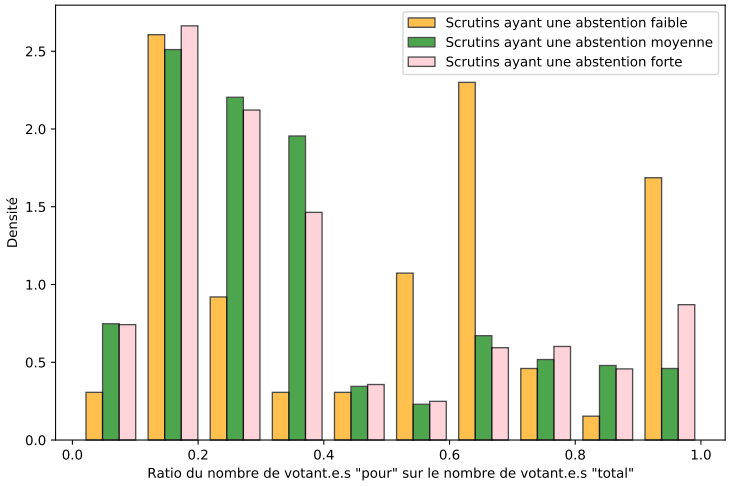
la baisse de l'abstention. A l'inverse, en août 2018, l'abstention a augmenté alors que le nombre de scrutins soumis au vote était très faible. On peut supposer que l'investissement des député.e.s était particulièrement fort au début de la législature en 2017 mais a diminué par la suite. On remarque ensuite que les mois où des élections ont lieu sont des mois où l'abstention est élevée ou en hausse. Ce schéma semble être peu lié au nombre de scrutins soumis sur la période concernée. On peut alors soumettre l'hypothèse que les périodes d'élections sont des périodes où les député.e.s sont moins actif.ve.s à l'Assemblée nationale car plus investi.e.s dans leurs campagnes électorales (dans la mesure où certain.e.s député.e.s exercent ou sont candidats à un autre mandat). Enfin, on note une importante baisse de l'abstention à partir du milieu du premier confinement de mars 2020. On peut considérer que les député.e.s ont été plus sollicité.e.s et se sont davantage investi.e.s au Palais Bourbon dans le cadre de la gestion du premier confinement et de sa levée en mai 2020.

Dans un second temps, nous avons tenté d'expliquer ces variations du taux d'abstention au cours de la législature : qu'est ce qui fait que les député.e.s se mobilisent pour un scrutin plutôt qu'un autre ? On pourrait par exemple s'attendre à ce que les député.e.s ne se déplacent pas pour un vote très consensuel, c'est-à-dire "joué d'avance". Nous n'avons donc pas ici cherché à établir les déterminants sociaux de l'abstention des député.e.s, mais à étudier une possible corrélation entre l'abstention et la polarisation des votes.

Pour ce faire, nous avons créé un ratio d'abstention, correspondant au rapport du nombre d'abstentionnistes sur le nombre de votants total, pour chaque scrutin. La distribution de cette variable présente une forte concentration pour les valeurs hautes : 80% des scrutins ont une abstention supérieure à 80%. A l'inverse, seulement 4% des scrutins ont une abstention inférieure à 60%. Il faut également noter que la queue de la distribution reste épaisse et prend des valeurs très faibles : 2% des scrutins présentent une abstention comprise entre 4% et 20%. De manière à rendre compte de cette distribution du taux d'abstention, nous avons distingué trois groupes de scrutins : les scrutins ayant une abstention faible (inférieure à 20%), les scrutins ayant une abstention moyenne (entre 20 et 80%) et les scrutins ayant une abstention très forte (supérieure à 80%).

Nous avons calculé ensuite, pour chacun de ces groupes de scrutins, le ratio de votes "pour", correspondant au rapport du nombre de votants "pour" sur le nombre de votants total (pour, contre et abstention). Ce rapport rend compte de la polarisation du vote : si le rapport est très élevé ou très faible, alors le vote est très polarisé ;

FIGURE 5 – Ratio du nombre de votants “pour” sur le nombre de votants total en fonction du niveau d’abstention par scrutin.



à l’inverse, si le rapport est moyen, alors le scrutin sera plus mitigé. On représente les densités du ratio de votes “pour” dans les trois groupes de scrutins sur la **Figure 5**. Les distributions rose et verte (abstention moyenne et forte) se ressemblent fortement. A l’inverse, on voit que la distribution orange (abstention faible) est très différente, et semble presque être le symétrique des deux autres distributions par rapport à la droite $x = 0.5$. Les scrutins où l’abstention est moyenne à forte sont les scrutins qui sont largement rejetés : les 3/4 des scrutins ont un ratio votants pour/votants total inférieur à 40%. Ces scrutins semblent donc très consensuels. A l’inverse, lorsque l’abstention est faible, les scrutins sont plus souvent adoptés, mais de manière moins consensuelle : les votes “pour” dépassent les 50% pour seulement 56% des scrutins. On observe donc une corrélation négative entre abstention et polarisation des votes.

Il est possible d’interpréter cette corrélation négative comme une relation causale entre les deux variables. Les scrutins où l’abstention est la plus forte sont des scrutins qui ne font que peu débat et, de ce fait, mobilisent moins les député.e.s. Les député.e.s pourraient avoir l’impression que tout est joué d’avance, et ne prendraient donc pas la peine de voter. A l’inverse, des scrutins ayant une abstention faible sont des scrutins moins consensuels : les député.e.s se mobilisent plus pour ces scrutins, pour lesquels leurs opinions divergent.

Ces premières statistiques descriptives rendent ainsi compte du fonctionnement particulier de l’Assemblée nationale. Tout d’abord, l’étude de la cohésion au sein des partis nous a permis de mettre en évidence l’importance de la discipline de vote au Palais Bourbon, ainsi que son hétérogénéité entre les partis. Nous avons de plus remarqué que la cohésion au sein des partis n’était pas corrélée à la position de ceux-ci dans le spectre majorité/opposition. L’étude des demandeurs a quant à elle mis en évidence le poids de chaque parti à l’Assemblée nationale, ainsi que sa plus ou moins grande proximité avec la majorité. Enfin, l’ensemble de cette partie a permis d’inscrire l’étude des données de vote de l’Assemblée nationale dans la lignée des travaux sur l’abstention, puisque cette dernière est une donnée fondamentale du vote des député.e.s, au même titre que le vote “pour” ou le vote “contre”.

3 ...et une analyse de l'évolution des positions des député.e.s et des groupes politiques au cours de la législature ainsi que la prédiction de l'abstention des député.e.s.

L'analyse de la cohésion au sein de chaque parti ainsi que celle de l'adoption moyenne des scrutins en fonction des demandeurs a mis en évidence l'importance du rôle joué par la discipline de vote à l'Assemblée nationale. Il semblait donc intéressant d'étudier d'une part l'évolution de la position des député.e.s au sein de leur parti, mais également l'évolution des positions des partis les uns par rapport aux autres. D'autre part, nous avons constaté l'importance de l'abstention dans le fonctionnement de l'Assemblée nationale et nous avons donc décidé de prédire si un.e député.e s'abstiendrait ou non à un scrutin particulier.

3.1 Clustering et analyse en composantes principales : modéliser les positions relatives des député.e.s et des groupes politiques

Après avoir analysé la cohésion des partis politiques, nous nous sommes intéressées plus précisément à la position des député.e.s. Nous cherchons ainsi à savoir si, à partir du comportement de vote des député.e.s, nous pouvons retrouver les groupes politiques qui composent l'Assemblée Nationale : cette question recouvre la partie clustering de notre modélisation. Nous nous demandons également s'il est possible de mettre en évidence une certaine structure du vote des député.e.s, à travers une analyse en composantes principales.

Afin d'analyser la position des député.e.s selon leur comportement en terme de votes, nous avons considéré d'une part leur historique de vote (qui correspond à la position de l'individu pour chaque scrutin : "pour", "contre" ou "non votant"), et d'autre part des indices synthétiques que nous avons construits. Ces indices sont, pour chaque individu :

- le taux d'abstention (ratio entre le nombre de scrutins pour lesquels l'individu était absent et le nombre de scrutins total),
- l'indice d'accord (présenté ci-dessus), qui mesure le taux d'accord de l'individu à la décision finale de l'Assemblée,
- le taux de "pour" (ratio entre le nombre de fois où l'individu s'est prononcé "pour" et le nombre de fois où il a voté),
- le taux de "contre".

On peut modéliser les député.e.s par un ensemble de points $(X_1, ..., X_n)$, le vecteur X_i de l'individu i étant dans un espace à 3117 dimensions si l'on considère son historique de votes (il y a 3117 scrutins), et à 4 dimensions si l'on considère les indices que l'on a construits. Face à ce positionnement des député.e.s en grande dimension, nous cherchons alors à résoudre deux problèmes :

- peut-on retrouver la composition politique de l'Assemblée dans l'espace de votes, que ce soit la traditionnelle opposition droite/gauche, ou un partitionnement qui se rapproche des différents partis politiques ?
- peut-on trouver une structure du positionnement des député.e.s dans un espace à deux dimensions qui soit interprétable ?

Ces questions nous amènent à réaliser en premier lieu un clustering des député.e.s, puis une analyse en composantes principales.

3.1.1 Clustering : retrouver les groupes politiques de l'Assemblée à partir du vote des député.e.s

Nous cherchons tout d'abord à partitionner les 575 député.e.s en un nombre de groupes (clusters) qu'il faudra déterminer. Notre jeu de données étant particulièrement important, nous avons choisi d'utiliser la méthode des K-moyennes. Cet algorithme détermine la partition de données $S = S_1, ..., S_k$ (où k est le nombre de clusters) qui minimise l'inertie, *i.e* la distance entre les points à l'intérieur de chaque cluster :

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2}$$

où μ_i est la moyenne (le centroïde) des points x dans l'ensemble S_i.

La particularité de cette méthode est qu'il n'existe pas un unique partitionnement possible : la difficulté est de choisir le nombre de clusters k qui mette le mieux en évidence la structure des données. Nous chercherons donc par la suite le nombre optimal de clusters pour partitionner les données. Dans un premier temps, nous choisissons arbitrairement $k = 2$ clusters, afin de voir si cette séparation duale des député.e.s selon leur historique de vote permet de retrouver le clivage gauche/droite traditionnel. Nous comparons la composition politique des deux groupes obtenus dans la **Figure 6**.

Si cette première partition ne ressemble pas à première vue à un regroupement selon l'axe gauche/droite, nous pouvons l'analyser comme une partition à partir d'un axe majorité/opposition. En effet, le premier cluster peut s'apparenter à un groupe d'opposition, regroupant principalement des Républicains (42,3%) et des Socialistes (11,7%), tandis que le deuxième cluster correspond plutôt au groupe de la majorité parlementaire, étant essentiellement composé de député.e.s appartenant à LREM (78,5%) et au MoDem (13,5%). Cette séparation fait sens dans la mesure où les élections présidentielles puis législatives de 2017 ont quelque peu remis en cause la traditionnelle opposition gauche/droite avec l'émergence du parti LREM.

Après cette première vue d'ensemble à travers une partition duale des député.e.s, nous nous demandons s'il est possible de retrouver les différents groupes politiques de l'Assemblée à partir des seuls votes des député.e.s. Nous cherchons donc à savoir quel nombre k de clusters partitionne au mieux les données. Pour déterminer ce nombre, nous utilisons la règle du coude, qui indique un nombre optimal de $k = 7$ clusters. Le détail de cette méthode se trouve en **Annexe 4.1**. Bien que ce nombre de clusters ne recouvre pas parfaitement les dix groupes politique présents à l'Assemblée, nous obtenons dans la **Table 3** des regroupements politiques cohérents. En effet, certains clusters sont composés presque essentiellement de député.e.s issu.e.s du même parti. On peut ainsi identifier un cluster de droite, comprenant 24 membres, dont 91,67% de Républicains. Le cluster le plus important, comprenant 160 membres, correspond lui aussi à un groupe de droite, composé à 51,88% de Républicains, ainsi que de député.e.s provenant de LREM, de Libertés et Territoires, et de l'UDI. A l'opposé du spectre politique, un cluster de gauche est formé de 47 député.e.s, se répartissant de manière homogène entre Gauche Démocrate, Socialistes et France insoumise. Les député.e.s LREM, les plus présent.e.s à l'Assemblée (269 membres), se répartissent en quatre groupes différents dans lesquels ils sont chaque fois majoritaires. Ces groupes de respectivement 167, 84, 80 et 33 député.e.s sont complétés par des député.e.s positionné.e.s au centre-droit (MoDem, UDI, Libertés et Territoires, Agir ensemble, et Non inscrit). Cet éclatement du groupe majoritaire de l'Assemblée en différents clusters reflète bien l'importante dispersion de ce groupe politique, mise en évidence précédemment (rappelons en effet que LREM a l'indice de cohésion le plus faible sur l'ensemble de la période).

Ces résultats permettent donc de confirmer nos premières observations concernant la cohésion des partis politiques. Le fait qu'un clustering effectué sur l'historique de vote des député.e.s ne recoupe pas parfaitement les partis

FIGURE 6 – Partitionnement de l'Assemblée en deux groupes selon l'historique de vote des député.e.s.

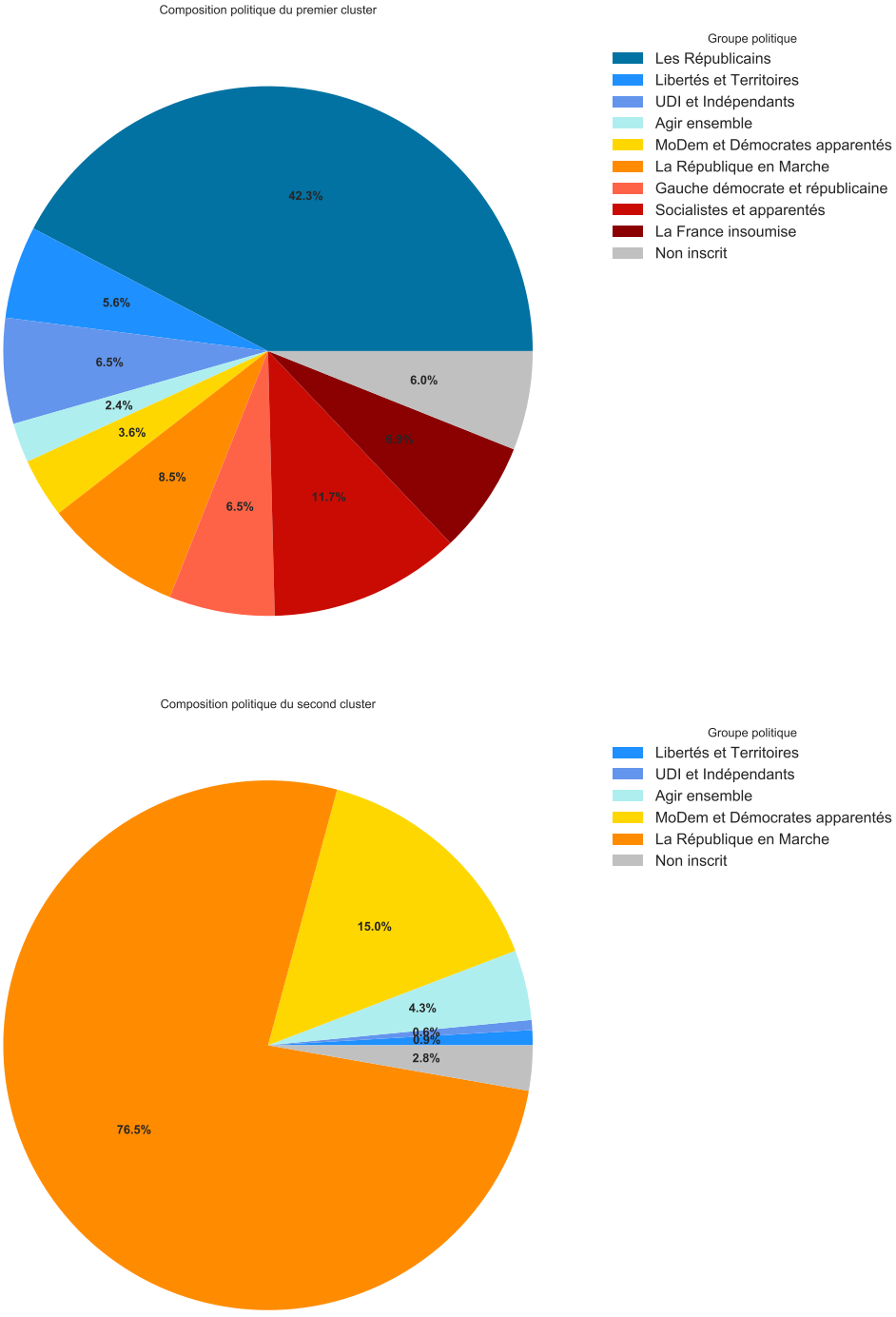


TABLE 3 – Partitionnement des député.e.s en sept clusters selon leur historique de vote.

Cluster / Parti	1	2	3	4	5	6	7
Les Républicains	0%	0%	0%	51.88%	0%	0%	**91.67**%
Libertés et Territoires	0%	**6.38**%	3.03%	6.25%	2.04%	0%	0%
UDI et Indépendants	1.19%	0%	0%	**8.75**%	0.68%	2.5%	0%
Agir ensemble	3.57%	0%	6.06%	0.62%	**8.16**%	2.5%	0%
MoDem et apparentés	11.9%	0%	6.06%	3.75%	**19.73**%	13.75%	0%
La République en Marche	**83.33**%	0%	78.79%	10.62%	63.27%	80.0%	4.17%
Gauche démocrate	0%	**21.28**%	0%	3.75%	0%	0%	0%
Socialistes et apparentés	0%	**36.17**%	0%	7.5%	0%	0%	0%
La France insoumise	0%	**36.17**%	0%	0%	0%	0%	0%
Non inscrit	0%	0%	6.06%	**6.88**%	6.12%	1.25%	4.17%
Effectif total (membres)	84	47	33	160	147	80	24

Les pourcentages les plus importants pour chaque parti sont spécifiés en gras.

politiques de l'Assemblée, autant en termes de nombre que de composition des groupes, est révélateur du comportement de vote des député.e.s. Celui-ci ne s'apparente pas à une pure discipline de parti, puisque des votes contraires peuvent s'exprimer au sein d'un même groupe, tandis que des membres de différents partis peuvent se retrouver sur une même ligne de vote. Afin de mieux rendre compte de cette complexité, nous représentons graphiquement ce positionnement des député.e.s en termes de vote dans la partie suivante.

3.1.2 Analyse en composantes principales : visualiser les positions relatives des député.e.s

Nous cherchons à présent à représenter plus précisément la structure du vote des député.e.s, à travers une analyse en composantes principales. Il s'agit de visualiser le nuage de points $X_1, ..., X_n$ des député.e.s dans un plan (le vecteur X_i de l'individu i étant dans l'espace des votes de dimension 3117). On cherche le sous-espace linéaire H de dimension 2 qui minimise l'inertie du nuage de points, *i.e* qui minimise la déformation du nuage une fois projeté en deux dimensions :

$$\underset{H}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \|X_i - P_H X_i\|^2 \tag{3}$$

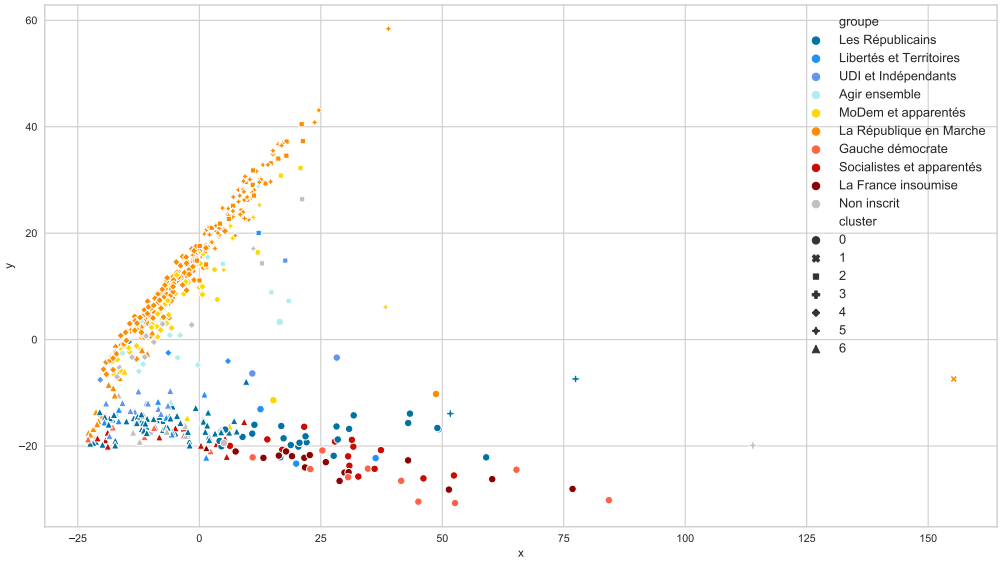
où $P_H X_i$ est le projeté orthogonal de X_i sur H.

Cette méthode permet à la fois de visualiser la forme du nuage des député.e.s et de synthétiser les relations entre les variables d'intérêt.

ACP sur l'historique de vote des député.e.s

Nous commençons par analyser le positionnement dans député.e.s dans l'espace des votes en 3117 dimensions. Comme pour chacun des 3117 scrutins, les député.e.s ont le choix entre voter pour, voter contre, ou s'abstenir, les individus peuvent être représentés dans un espace à 6046 dimensions, où chaque dimension correspond à un choix de vote exprimé par scrutin. En effet, les résultats du scrutin n°1 peuvent se décomposer en deux indicatrices : voter "pour" et s'abstenir. Nous omettons l'indicatrice voter "contre" afin d'éviter la co-linéarité des variables. Par ailleurs, il se peut qu'à certains scrutins il n'y ait aucun vote "pour" (ou respectivement aucun "contre", ou aucune abstention), si bien que nous n'obtenons pas systématiquement deux indicatrices par scrutin. C'est pour cette raison que nous passons d'un espace de 3117 dimensions à un espace de 6046 dimensions.

FIGURE 7 – Représentation des député.e.s en ACP selon leur historique de vote.



La projection du nuage de points des député.e.s depuis l'espace des votes dans un espace en deux dimensions est obtenue à la **Figure 7** : chaque point représente un.e député.e, et l'on précise son parti politique et le cluster auquel il ou elle appartient. On remarque une nette séparation des individus de part et d'autre de la bissectrice de la figure, correspondant à une séparation majorité/opposition. En effet, la partie supérieure du graphique regroupe essentiellement des député.e.s LREM et MoDem, tandis que les député.e.s dans l'opposition (Les Républicains, Socialistes et apparentées, France insoumise, Gauche démocrate et républicaine, ...) sont dans la partie inférieure. Les différents clusters recoupent ce partitionnement, et l'on remarque que plus les député.e.s s'éloignent de l'origine le long de l'axe des abscisses, plus ils appartiennent à un cluster de gauche. On peut ainsi interpréter cette ACP comme une projection sur un axe majorité/opposition d'une part, et sur un axe gauche/droite d'autre part. Les axes étant des combinaisons linéaires de l'historique des votes, nous ne pouvons donner qu'une approximation de leur signification, puisque les composantes principales de cette ACP, détaillées dans la **Table 4**, ne sont pas interprétables telles quelles. En effet, cette table indique les corrélations entre les variables d'intérêt (le choix de vote pour chacun des 3117 scrutins) et les composantes principales. Or, d'une part, en considérant que les corrélations sont significatives à partir d'un certain seuil (par exemple 0.5 en valeur absolue), on remarque qu'aucune corrélation n'est suffisamment élevée pour qu'une variable soit considérée comme explicative d'un axe. D'autre part, même si l'une de ces corrélations était significative, on ne pourrait pas interpréter la contribution d'une variable telle que "voter pour au scrutin n°1" dans la structure d'un axe. C'est pour cette raison que nous avons réalisé une seconde ACP, en considérant non pas l'historique de vote des député.e.s, mais leurs caractéristiques de vote.

ACP sur les caractéristiques de vote des député.e.s

Nous avons donc ensuite choisi de représenter le positionnement des député.e.s selon leurs caractéristiques de vote, en utilisant les indices définis précédemment : leur taux d'abstention, leur taux de votes "contre", leur taux de vote "pour", et leur indice d'accord. Les résultats de cette ACP en quatre dimensions sont présentés dans la

TABLE 4 – Corrélation entre les variables d'intérêt et les composantes principales (ACP sur l'historique de vote).

Variables	composante 1	composante 2	composante 3	composante 4
Voter "pour" au scrutin n°0	0.008	0.009	0.01	-0.001
S'abstenir au scrutin n°1	-0.01	-0.003	0.009	-0.008
Voter "pour" au scrutin n°1	0.02	-0.01	-0.03	0.004
S'abstenir au scrutin n°2	-0.01	-0.006	-0.01	-0.004

Cette table contient les quatre premières variables indicatrices sur les 6046 qui ont permis de réaliser l'ACP.

FIGURE 8 – Représentation des député.e.s en ACP selon leurs caractéristiques de vote.

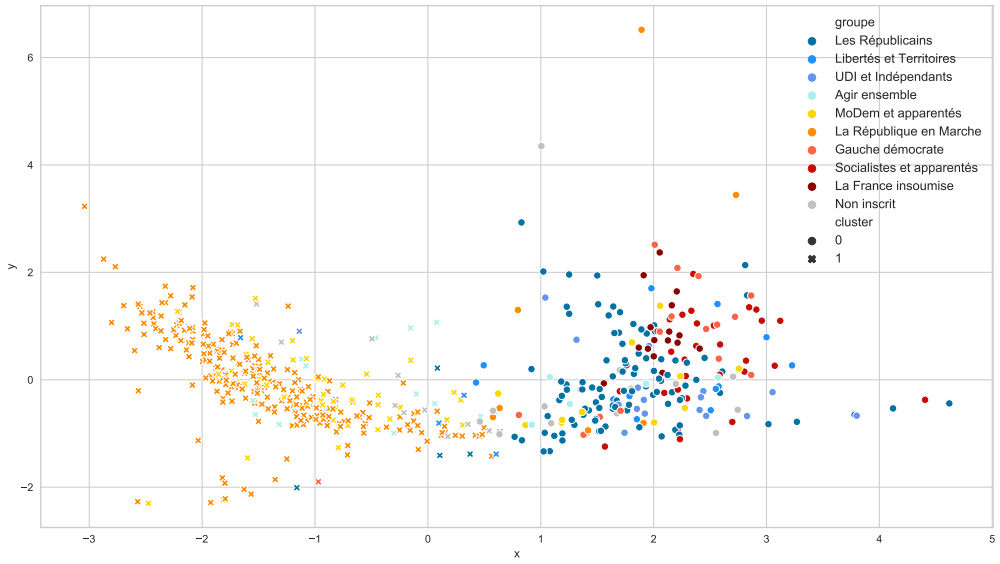


Figure 8 : à nouveau, chaque point correspond à un.e député.e, pour qui nous précisons son parti et son cluster d'appartenance. Nous avons choisi cette fois-ci le partitionnement en deux clusters, qui rendait compte d'une séparation majorité/opposition ; et c'est bien cette même séparation que l'on observe sur la projection du nuage des député.e.s dans le plan. En effet, l'ensemble de gauche sur le graphique correspond aux député.e.s LREM et MoDem, appartenant principalement au deuxième cluster ; tandis que l'ensemble de droite correspond aux individus appartenant aux partis d'opposition, principalement dans le premier cluster.

L'avantage de partir de l'ensemble en quatre dimensions correspondant aux indices de comportement de vote est que l'on peut cette fois-ci interpréter les axes dans lesquels les député.e.s sont projeté.e.s. On peut en effet voir dans la **Table 5** que les corrélations entre les variables d'intérêt et les composantes principales sont significatives. L'axe des abscisses est une combinaison linéaire de l'indice d'accord et des taux de votes "pour" et "contre", tandis que l'axe des ordonnées correspond essentiellement au taux d'abstention. Ainsi, plus l'on monte le long de l'axe des ordonnées, moins l'individu s'abstient. Plus l'on s'éloigne de l'origine sur l'axe des abscisses, plus l'individu émet de votes "pour", moins il émet de votes "contre" et moins son vote exprimé est en accord avec la décision finale de l'Assemblée.

Ces deux premières ACP mettent chacune en lumière une certaine structure des député.e.s au sein de l'Assemblée nationale. D'un côté, leur positionnement par rapport à l'historique des votes révèle ce qu'avait déjà suggéré le

TABLE 5 – Corrélation entre les variables d'intérêt et les composantes principales (ACP sur les caractéristiques de vote des député.e.s).

Variables	composante 1	composante 2	composante 3	composante 4
Taux de votes "contre"	**-0.56**	-0.15	0.36	**-0.72**
Taux de votes "pour"	**0.56**	0.10	-0.44	**-0.68**
Indice d'accord	**-0.53**	-0.21	**-0.81**	0.05
Taux d'abstention	0.27	**-0.95**	0.07	0.03

Les corrélations les plus importantes (en valeur absolue) sont spécifiées en gras.

travail de clustering, à savoir une structure en termes de majorité/opposition. Leur positionnement selon leurs caractéristiques de vote met quant à lui en évidence l'importance de l'abstention et de l'accord des député.e.s vis-à-vis des décisions finales de l'Assemblée dans la structure politique de l'hémicycle.

Si cette analyse en ACP nous éclaire sur le positionnement des député.e.s les un.e.s par rapport aux autres, le travail préalable de clustering a également permis d'appréhender leur positionnement par rapport aux groupes politiques. Il a ainsi mis en avant la plus ou moins grande cohésion des partis politiques, qui, nous l'avons vu dans la partie précédente, fluctue au cours de la législature. C'est pour cela que nous avons décidé de compléter cette analyse statique de la structure politique de l'Assemblée par une analyse dynamique, en cherchant à mettre en évidence les évolutions des positions des député.e.s tout au long de la législature.

3.1.3 Evolution de la représentation en ACP

Afin d'analyser l'évolution du positionnement des député.e.s les un.e.s par rapport aux autres dans le temps, nous avons réalisé une ACP par année depuis le début de la législature. Nous avons pour cela considéré séparément l'historique de votes de 2017, 2018, 2019 et 2020 (à noter qu'il y a eu 357 scrutins en 2017, 1255 en 2018, 804 en 2019 et 701 en 2020). De plus, afin de comparer l'évolution du positionnement des député.e.s entre eux mais également par rapport à leur parti, nous avons fait figurer sur le graphique un député fictif pour chaque parti. Ce député fictif correspond à la moyenne des points des membres du parti une fois projetés dans le plan, et est calculé chaque année. Cela nous permet d'estimer l'évolution du positionnement moyen d'un parti par rapport aux autres, tout en rendant compte de la plus ou moins grande cohésion d'un parti.

Les représentations de ces ACP par année sont disponibles dans les **Figures 9a, 9b, 9c et 9d**. La première observation que nous pouvons faire est que la structure générale précédemment mise en évidence par l'ACP globale sur l'historique de votes (**Figure 7**) semble stable dans le temps : on retrouve chaque année le même positionnement des député.e.s, avec d'une part la majorité et de l'autre l'opposition. Si le réseau de points semble pivoter après 2017 (la branche composée de la majorité se rapprochant de la verticale et la branche de l'opposition se rapprochant de l'horizontale), l'écartement entre les branches reste sensiblement le même. Comme vu précédemment, les axes n'étant pas interprétables en eux-mêmes, ce sont les déformations du réseau de points que l'on cherche à analyser. La déformation observée étant seulement une rotation, elle ne traduit aucune évolution interprétable. Concernant la structure globale, non plus de tou.te.s les député.e.s mais des députés moyens (fictifs), on remarque la formation progressive de 3 pôles : la majorité, l'opposition de droite et l'opposition de gauche. Si ces pôles sont déjà visibles en 2018, les députés fictifs appartenant à chacun de ces pôles se rapprochent fortement les uns des autres en 2019 et 2020. En 2020, on observe ainsi trois groupes de députés moyens :

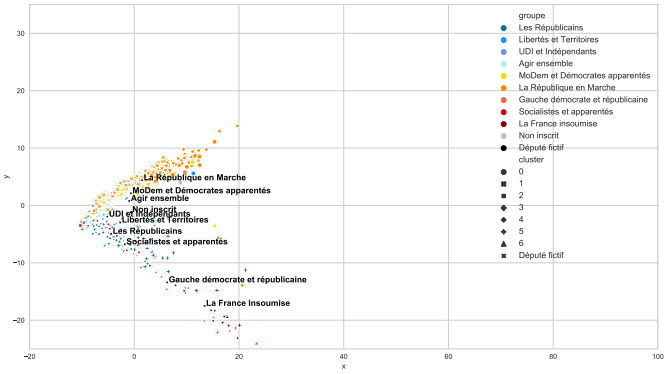
— la majorité : La République en Marche, Agir ensemble, MoDem ;
— l'opposition de droite : UDI et Indépendants, Les Républicains, Non inscrits ;
— l'opposition de gauche : Socialistes et apparentés, Gauche démocrate et républicaine, La France insoumise.

On retrouve la même tripartition que celle mise en évidence par l'analyse de l'accord et de la cohésion des partis. Cette polarisation en trois groupes rend bien compte des évolutions récentes de la vie politique française, avec l'apparition du parti LREM qui a sappé le traditionnel bipartisme lors des élections présidentielles puis législatives de 2017. Les trois groupes mis en évidence sont ainsi à la fois distincts par leur cohésion et par leurs positions politiques. On remarque également que la dispersion des député.e.s les un.e.s par rapport aux autres est beaucoup plus faible en 2017 que durant les années suivantes ; c'est en particulier le groupe LREM qui se disperse dans le temps. On voit en effet que ses membres tendent à s'écarter de leur député fictif moyen, qui paraît stable dans le temps. Ceci confirme nos premières analyses sur la cohésion de ce nouveau parti, qui tend à connaître plus d'hétérogénéité au cours de la législature. A l'inverse, si La France insoumise se disperse davantage au cours du temps, son député fictif se rapproche progressivement de l'origine du V de l'ACP. On retrouve donc la faible voire décroissante cohésion de ce parti, à laquelle on peut ajouter une légère tendance au rapprochement vers les autres partis de gauche, comme nous l'avons vu avec la tripartition de l'Assemblée. Cependant, et cela semble logique, la branche de la majorité est moins épaisse que la banche de l'opposition. D'un certain point de vue, la majorité est donc moins dispersée que l'opposition, qui elle rassemble des partis aussi différents politiquement que La France insoumise et Les Républicains. Enfin, l'année 2018 présente une ACP avec une dispersion beaucoup plus forte que les autres. C'est aussi l'année où le plus de scrutins ont été votés à l'Assemblée ; on peut supposer que cette année a connu un nombre plus important de propositions de loi non consensuelles. La plus grande dispersion observée est ici aussi en accord avec la **Figure 3**, où les plus fortes fluctuations de l'indice de cohésion des partis était observée pour 2018.

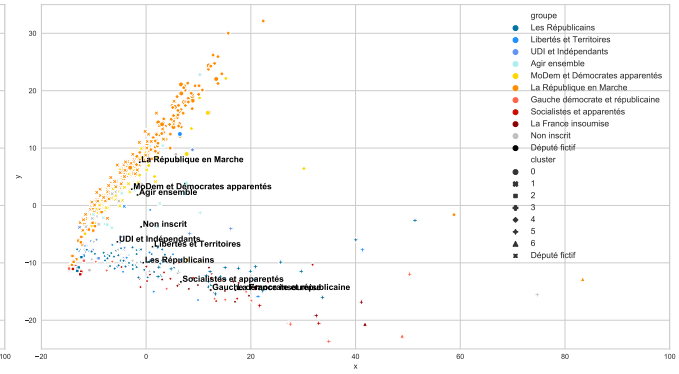
Cette analyse des positions relatives des député.e.s et de leur évolution au cours de la législature confirme donc nos premières observations sur la cohésion des partis. Le fait que l'on ne puisse partitionner parfaitement les député.e.s en des clusters correspondant aux partis politiques, malgré la forte discipline de vote, s'explique d'une part par la polarisation majorité/opposition au sein de l'Assemblée et d'autre part par la diversité des déterminants du vote que nous n'étudions pas ici. Apparaissent ainsi trois pôles majeurs autour desquels gravitent les élu.e.s : la majorité parlementaire, l'opposition de droite et l'opposition de gauche. Cette tripartition est également mise en évidence par l'analyse en composantes principales, qui souligne le rôle significatif de l'abstention dans les positions relatives des député.e.s. Tout comme la cohésion des groupes politiques, ces relations intra- et inter-partis fluctuent au fil de la législature, mettant en lumière une scission progressive entre deux pôles de l'opposition.

FIGURE 9 – Evolution de la représentation en ACP des député.e.s.

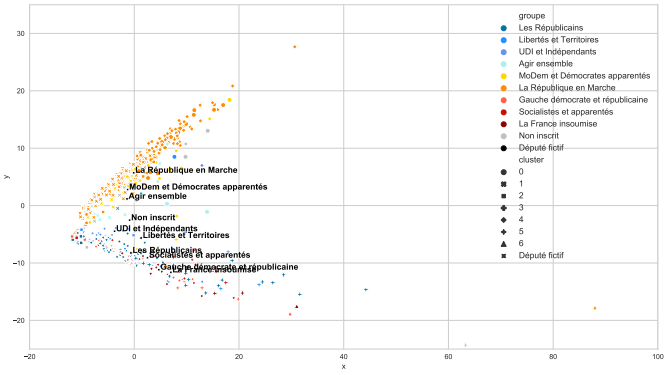
(a) ACP des député.e.s selon leurs votes de 2017



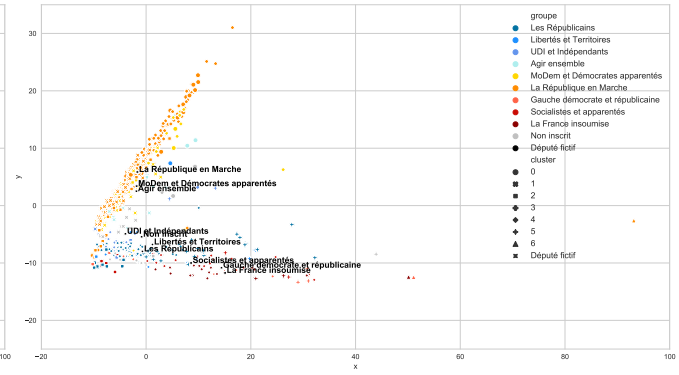
(b) ACP des député.e.s selon leurs votes de 2018



(c) ACP des député.e.s selon leurs votes de 2019



(d) ACP des député.e.s selon leurs votes de 2020



3.2 Voter ou ne pas voter ? Prédire la mobilisation des député.e.s pour un thème donné selon leur profil politique et socio-économique

3.2.1 Un détour par la sociologie du vote

Après avoir étudié les caractéristiques de vote de l'Assemblée Nationale et leurs évolutions sous le prisme de la partition traditionnelle de l'Assemblée Nationale par parti, nous avons cherché à ajouter une dimension socio-économique aux profils des député.e.s. Nous avons pour cela considéré la catégorie socio-professionnelle, le genre, l'âge et la région d'origine de chaque député.e afin d'enrichir leurs profils. Comme précisé dans notre première partie, nous avons récupéré ces données sur Wikipédia et sur le site de l'Assemblée Nationale. Alors que notre étude de l'abstention présentée précédemment s'intéressait seulement aux corrélations entre abstention et polarisation des votes, nous nous inscrivons ici de plein-pied dans la sociologie politique et plus précisément dans la sociologie du vote, qui étudie le vote et ses déterminants sociologiques. Si cette approche porte généralement sur le vote des citoyen.ne.s, nous nous sommes ici penchées sur les profils des élu.e.s et leur mobilisation sur certains sujets.

En outre, cette réflexion répond à la pensée de *La politique des deux axes* développée par Vincent Tiberj [18]. Celle-ci s'intéresse au processus de vote, *i.e* à la décision d'un votant lors d'un certain scrutin dans une certaine conjoncture, plus qu'à la structure du champ politique. En effet, si la mobilisation des citoyen.ne.s ou des élu.e.s lors de scrutins a longtemps été considérée comme une réponse logique à leurs déterminants sociaux, ce postulat est aujourd'hui remis en question. Par exemple, il a longtemps été admis que le clivage gauche/droite chez les citoyen.ne.s reposaient entièrement sur leurs caractéristiques sociales et que quelle que soit la conjoncture, leur vote répondrait à un clivage de classe. Si le pouvoir explicatif des déterminants sociaux reste significatif, on observe aujourd'hui une montée des déterminants culturels (tels que le niveau de diplôme) et conjoncturels [19]. Ainsi le vote ne reposerait pas seulement sur le déterminisme social du votant, mais également sur sa position culturelle face à des enjeux de plus courte durée. On parle de courte durée dans le sens où les débats publics le sont, et non pas les enjeux en eux-mêmes. On considère en effet l'écologie, l'immigration, la parité ou encore l'utilisation des nouvelles technologies (haine sur internet, diffusion d'images, ...) et leurs enjeux actuels comme des facteurs structurant des systèmes partisans. Il s'agit donc non pas d'une réelle mutation d'un système de "vote sur clivage" vers un "vote sur enjeu", mais d'une multiplicité des axes à considérer dans la sociologie du vote.

Nous avons donc voulu explorer cette multiplicité dans notre étude du processus de vote. Nous considérons les caractéristiques socio-économiques des député.e.s en plus de leur profil politique, dans la logique de la sociologie politique traditionnelle. Mais surtout, nous étudions le vote selon des thèmes précis (par exemple le climat ou la haine sur internet), afin d'ajouter la dimension culturelle et conjoncturelle dont l'importance a été mise en valeur plus récemment.

3.2.2 Résumé de notre démarche

Nous avons ajouté la catégorie socio-professionnelle, le genre, l'âge et la région d'origine de chaque député.e à nos précédentes variables afin de prédire la mobilisation des député.e.s sur un scrutin donné. Si nous avons d'abord pensé à prédire le vote d'un.e député.e (vote pour ou vote contre) pour un scrutin donné, ce choix ne semblait que peu pertinent. En effet, la position "pour" ou "contre" dépend fortement de l'usage ou non de la négation dans l'intitulé d'une proposition. Par exemple, pour deux scrutins tels que "Projet de loi en faveur de la PMA pour tou.te.s" et "Projet de loi contre la PMA pour tou.te.s", les votes "pour" et "contre" seraient inversés. Ce constat rend la prédiction des votes compliquée à partir des méthodes classiques de Traitement Naturel du Langage (Natural Language Processing, NLP) et des données dont nous disposons. Nous avons alors choisi de prédire la mobilisation d'un.e député.e sur un scrutin, *i.e* prédire s'il ou elle va s'abstenir ou aller voter. En effet nous avons

précédemment observé un fort taux d'abstention moyen qui diminue pour les scrutins les moins consensuels (**Figure 5**). Les député.e.s sembleraient donc se mobiliser principalement sur des thèmes pour lesquels les avis sont à la fois les plus tranchés et les plus divisés. Il paraît donc intéressant de se pencher sur la propension des député.e.s à voter ou non pour certains scrutins.

Dans notre modèle de prédiction, chaque député.e est décrit.e par ses caractéristiques de vote et les informations socio-économiques précédemment citées. Concernant les thèmes des scrutins, nous les avons établis nous-mêmes à partir des intitulés des propositions de notre base de données. Pour ce faire, nous avons assigné à chaque proposition le thème qui lui est associé à l'aide de méthodes de NLP appliquées aux intitulés des propositions et d'un travail de clustering appliqué aux poids de chaque mot dans chaque proposition dans une matrice de *Term Frequency - Inverse Document Frequency* (TF-IDF).

Chaque intitulé se limitant à une seule phrase, nous avons quelques imprécisions dues au manque de richesse sémantique mais nous avons pu les contrôler. Les imprécisions résident dans le fait qu'avec un nombre élevé de clusters nous obtenons un nombre satisfaisant de thèmes très divers mais certains clusters regroupent en réalité toutes sortes de propositions mal partitionnées. A l'inverse, quand nous choisissons un faible nombre de clusters dans notre modèle, celui-ci ne retrouve pas suffisamment de thèmes exploitables. Notre but étant d'étudier seulement certains thèmes que nous trouvons intéressants, nous avons choisi de paramétrer 35 clusters afin d'avoir le choix parmi tous ces principaux thèmes. Le détail des méthodes de Preprocessing, d'Embedding et de Clustering utilisées est disponible en **Annexe 4.3**.

3.2.3 Présentation de la Random forest

Afin de prédire l'abstention ou la mobilisation d'un.e député.e pour un thème donné, nous avons choisi d'implémenter une Random forest.

Notre modèle prend en entrée un vecteur X_i représentant un profil i de député.e. Ce profil se décompose en une partie politique, composée des caractéristiques suivantes : l'indice d'accord, le taux d'abstention moyen, le taux de vote "pour" moyen, le taux de vote "contre" moyen ; et une partie socio-économique, composée des caractéristiques suivantes : la région d'origine, la catégorie socio-professionnelle, le genre, l'âge, le parti politique. Les données quantitatives dites politiques sont calculées sur la période 2017-2018. En effet, dans un souci de cohérence chronologique, il nous a fallu séparer les données de référence (dites du passé) des données que nous voulions prédire (dites du futur). Nous cherchons donc à prédire les labels sur la période 2019-2020. Les labels du modèle sont définis tels que $Y_i \in \{0, 1\}$ est la variable prenant la valeur 0 si l'individu i décrit s'abstient et 1 si celui-ci se mobilise sur un scrutin pour le thème T_j donné. Les Random forest ne sont appliquées que sur les données d'un seul thème T_j ($j \in [\![1, 35]\!]$) à la fois.

Les thèmes sur lesquels nous avons choisi de travailler sont à la fois ceux que le clustering a correctement prédits et ceux qui étaient encore débattus en 2019 et 2020 (*i.e* présents sur la période de référence et la période de prédiction). En effet certains thèmes précis comme le Pacte ferroviaire ("*Loi n° 2018-515 du 27 juin 2018 pour un nouveau pacte ferroviaire*"), Avenir professionnel ("*Loi n° 2018-771 «Avenir professionnel» du 5 septembre 2018*") ou Relations commerciales agricoles ("*Loi n° 2018-938 du 30 octobre 2018 pour l'équilibre des relations commerciales dans le secteur agricole et alimentaire et une alimentation saine, durable et accessible à tous*") ne présentent aucune donnée sur la période 2019-2020, sur laquelle nous cherchons à effectuer des prédictions. Si ces sujets ont suffisamment été débattus (par de nouvelles propositions et de nouveaux amendements) en 2017 et 2018 pour avoir été reconnus comme *thèmes* par notre méthode (cf. **Annexe 4.3**), ils faisaient référence à une loi précise qui a été votée avant 2019. Ces thèmes ne sont donc plus présents sur la période étudiée.

Pour les thèmes restants, nous avons pu prédire pour chaque individu s'il allait voter ou non. Il est important

de préciser que nous avons équilibré nos données : nous travaillons sur une répartition égale du taux d'abstention et de participation. En effet, le taux d'abstention moyen étant de 80% (*i.e* le taux de 1 dans nos labels), nos résultats auraient été biaisés par un tel déséquilibre.

Afin de mesurer les performances de notre prédiction nous avons regroupé dans la **Table 9** pour chaque thème donné, la précision de la base de train, celle de la base de test ainsi que son rappel et son F1-score. En effet, les performances de telles prédictions s'évaluent de plusieurs manières. Ici, si l'on considère que prédire qu'un.e député.e votera à un scrutin donné correspond à lui attribuer la valeur "Vrai", on distingue les "Vrais positifs" - prédiction du vote d'un.e député.e, tandis que dans les faits il ou elle vote réellement - des "Faux positifs" - prédiction du vote d'un.e député.e, alors que dans les faits il ou elle s'abstient. On peut alors chercher d'une part à calculer la qualité de notre prédiction, c'est-à-dire sa précision, qui correspond au nombre de "Vrai" correctement prédits ("Vrai positifs") sur le nombre total de prédiction de "Vrai" (sommes des "Vrais positifs" et des "Faux positifs"). D'autre part, on peut chercher à calculer son exhaustivité, aussi appelée rappel, c'est-à-dire le nombre de "Vrai positifs" sur le nombre de "Vrai" réels, autrement dit la capacité du modèle à repérer les "Vrais" réels. On a ainsi :

$$\text{Précision} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux positifs}} \tag{4}$$

$$\text{Rappel} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}} \tag{5}$$

On peut donc par exemple avoir parfois une très bonne précision mais un très mauvais rappel. Une mesure permettant de combiner ces deux variables est le F1-Score, qui est donné par :

$$\text{F1-Score} = \frac{2 * \text{Précision} * \text{Rappel}}{(\text{Précision} + \text{Rappel})} \tag{6}$$

Cette mesure pondère de façon égale précision et rappel, leur donnant la même importance. Le calcul de ce score permet d'avoir une vision des performances de notre algorithme en tenant compte de ces deux dimensions.

Deux approches différentes

Comme énoncé précédemment, nous avons choisi de calculer les indices (taux de vote "contre", taux de vote "pour", indice d'accord et taux d'abstention) sur la période 2017-2018 afin de travailler sur la période 2019-2020. A partir de ces données, deux choix s'offraient à nous : calculer ces indices par thème ou bien sur tous les scrutins confondus. Calculer les indices sur tous les scrutins nous permet de déterminer le profil d'un.e député.e indépendemment du thème, pour ensuite entraîner nos données par thème. L'autre méthode, pour laquelle on calcule les indices par thème, nous permet de préciser le comportement de l'individu conditionnellement au thème choisi, restreignant ainsi nos données aux thèmes avant même le travail d'entraînement et de prédiction.

Si la seconde méthode semble plus pertinente, nous avons choisi d'implémenter les deux. Il s'avère finalement que les résultats des deux méthodes ne diffèrent que peu et semblent répondre aux mêmes logiques, malgré une légère baisse de performance pour la seconde méthode (calcul des indices par thème). En dépit de ce léger défaut de performance, nous choisissons finalement la méthode où l'on calcule les indices par thème, qui nous apparaît la plus pertinente. Les résultats de la première méthode (calcul des indices sur tous les scrutins confondus) sont cependant à retrouver en **Annexe 4.4**.

TABLE 6 – Performances de la Random Forest selon le thème avec calcul des indices au sein du thème donné.

Résultats / **Thèmes**	Précision train	Précision test	Recall	F1 Score
Sécurité sociale	0.651	0.654	0.654	0.653
Finances	0.630	0.627	0.627	0.626
Immigration	0.621	0.608	0.608	0.605
Réforme justice	0.645	0.583	0.583	0.581
Haine sur internet	0.654	0.572	0.572	0.571
Violences sexuelles et sexistes	0.628	0.562	0.562	0.562

Conclusions de l'approche choisie

Dans la **Table 6**, on peut tout d'abord remarquer que les performances de nos prédictions sont assez homogènes d'un thème à l'autre (autour de 0.6). S'il est impossible de prédire parfaitement le vote ou l'abstention à partir des simples données dont nous disposons, il apparaît tout de même que notre prédiction est meilleure que le hasard. Cette homogénéité d'un thème à l'autre laisse supposer que les variables sont aussi informatives pour chaque thème : l'arbre construit à partir de mêmes variables donne des performances similaires pour des thèmes aux données pourtant très hétérogènes (cf. **Table 8**). Il est aussi intéressant de remarquer que le rappel et la précision sont à chaque fois très proches : le modèle est très équilibré entre ces deux mesures de performance. Ainsi, le modèle ne prédit pas mieux en proportion le fait de voter ou le fait de ne pas voter (on prédit environ autant de faux négatifs que de faux positifs). Ceci est en partie dû à l'équilibrage des données (le fait d'avoir 50% d'abstention dans les données d'entraînement permet d'éviter de biaiser la prédiction) mais aussi à une assez bonne séparation des données par la Random forest. Cependant, le principe de la Random forest étant d'entraîner de nombreux arbres aléatoires sur des sous-ensembles de données différents, cela rend nos prédictions plus difficilement interprétables. La Random forest est une "boîte noire" : il est complexe de déterminer ce qui fait qu'un.e député.e se mobilise pour l'un ou l'autre thème. Cette difficulté est accrue par la nature qualitative de la plupart des variables considérées. Nous avons néanmoins tenté de déterminer l'importance relative des différentes variables, en ôtant à chaque fois l'une d'entre elles de notre modèle et en comparant les performances du modèle ainsi tronqué avec celles de la **Table 6**. Il en ressort alors une importance très similaire de chaque variable (la région d'origine, la catégorie socio-professionnelle, le genre, l'âge, le parti politique et les indices politiques suivants : l'indice d'accord, le taux d'abstention moyen, le taux de vote "pour" moyen, le taux de vote "contre" moyen) .

Certaines limites de notre modèle sont cependant à préciser. Tout d'abord, il faut noter que les thèmes qui présentent le plus de sur-entraînement sont en réalité ceux pour lesquels nous avions très peu de données. Il est donc naturellement difficile d'avoir de bonnes prédictions sur un sujet que l'on ne connaît que très peu (le modèle s'adapte aux données d'entraînement mais a une assez mauvaise capacité de généralisation). C'est par exemple le cas des scrutins concernant la haine sur internet. De plus, comme présenté précédemment, certains thèmes très débattus en 2017-2018 n'ont pas donné lieu à de nouveaux débats en 2019-2020, du fait du vote des lois concernées avant 2019. Ces thèmes sont donc exclus de notre modèle, ce qui réduit sa portée.

Deuxièmement, comme évoqué à travers les évolutions de la sociologie du vote, la conjoncture est un point primordial à prendre en compte. En effet, l'évolution du contexte culturel influence fortement la propension à voter d'un.e député.e. Cet angle d'approche semble particulièrement pertinent dans le cas du thème du climat. Dans la **Table 8** nous remarquons que le thème *Energie et climat* présente le plus fort taux d'abstention (88% en moyenne).

Pour autant, le contexte culturel actuel autour de la nouvelle Loi Climat a conduit beaucoup de député.e.s à voter le projet de loi Climat et résilience le 4 mai 2021 (qui ne figure pas dans notre base de données que nous avons arrêtée en novembre 2020) : le taux d'abstention était en effet de 29% [20]. Si l'on avait tenté de prédire l'abstention d'un.e député.e à ce scrutin, notre prédiction aurait été très éloignée de la réalité car notre modèle n'aurait pas pris en compte le contexte culturel, politique et militant actuel.

Enfin, si nous avons en partie axé ce rapport et plus particulièrement cette dernière section sur l'abstention, il est important de mentionner que les chiffres de l'abstention sont à interpréter avec prudence. Nous illustrons cette remarque en reprenant l'exemple précédent. Un taux d'abstention moyen de 88% sur le thème *Energie et climat*, mis en regard de l'urgence climatique et des mobilisations massives pour le climat, pose question dans le cadre d'une démocratie représentative. On pourrait alors se réjouir des 29% d'abstention observés lors du vote de la Loi Climat le 4 mai 2021. Cependant, l'abstention dit peu du contenu de la loi votée, comme l'illustrent les débats actuels quant à ce projet de loi et son accord avec les propositions de la Convention citoyenne pour le climat. C'est d'ailleurs pour cela que nous avions choisi de prédire l'abstention plutôt que le vote "pour" ou "contre" : prédire le vote "pour" ou "contre" suppose de tenir compte de la présence d'une négation dans l'intitulé de la proposition de loi, ce que les méthodes de NLP implémentées ici ne permettaient pas de faire. De plus, l'abstention ne dit rien des modifications des textes entre ceux produits par les commissions et ceux votés *in fine* à l'Assemblée. Précisons également qu'un même projet de loi est débattu plusieurs fois à l'Assemblée Nationale à travers des réécritures et des amendements, chaque scrutin ne représentant donc pas toujours le même enjeu. Ainsi une analyse statique et ponctuelle de la position de l'Assemblée nationale par rapport à un scrutin précis demande un certain recul quant au contexte culturel et délibératif. Finalement, si la participation à l'Assemblée, comprise comme l'inverse de l'abstention, peut s'interpréter comme une mobilisation, elle n'est pas toujours le gage d'un intérêt ni d'un engagement des député.e.s en accord avec la volonté des citoyen.ne.s.

Conclusion

Ce sujet sur le Machine Learning à l'Assemblée Nationale nous invitait à une grande liberté dans le choix des pistes de réflexion, ce qui nous a permis de découvrir différents aspects du vote des député.e.s. Les diverses approches explorées cherchaient toutes à expliquer dans quelle mesure les données de vote des député.e.s permettent de rendre compte du fonctionnement particulier de l'Assemblée nationale. En ayant recours à une grande partie des données concernant la XVème législature disponibles sur le site de l'open data de l'Assemblée nationale, nous avons pu mettre en évidence quelques grandes conclusions.

Un premier tour d'horizon nous a permis d'identifier plusieurs facteurs déterminants du fonctionnement de l'Assemblée nationale. Le premier facteur que nous avons pris en compte est l'appartenance des député.e.s à un parti politique. Nous avons ainsi mis en évidence une tripartition de l'Assemblée, qui regroupe des partis politiquement proches et qui confirme l'importance de la discipline de vote au sein des partis : la majorité est très dispersée mais en accord avec les décisions finales de l'Assemblée ; l'opposition de droite est moyennement dispersée et moyennement en accord avec les décisions finales de l'Assemblée ; enfin, l'opposition de gauche est peu dispersée et peu en accord avec les décisions finales de l'Assemblée. Par ailleurs, la prise en compte de l'abstention dans l'analyse de la cohésion des partis montre une nouvelle partition, dans laquelle l'opposition de droite est moins dispersée que l'opposition de gauche. Cela souligne l'importance de la prise en compte de l'abstention dans l'analyse des votes des député.e.s, ainsi que les différentes stratégies mises en place par chaque parti pour défendre ses opinions à l'Assemblée.

Après avoir montré l'importance de la discipline de vote au sein des partis, nous nous sommes demandé dans quelle mesure le vote d'un.e député.e dépendait du demandeur du scrutin. Comme l'on pouvait s'y attendre, les partis dont les scrutins sont le plus souvent adoptés sont les partis de la majorité (LREM et le MoDem). Les partis de l'opposition, qu'ils soient de droite ou de gauche, ont des taux d'adoption plus faibles. Le vote "pour" ou "contre" d'un.e député.e dépend donc en partie du demandeur du scrutin. A l'inverse, l'abstention ne dépend pas du demandeur.

Finalement, nous avons choisi d'inscrire cette première partie dans une analyse plus globale de l'abstention, élément clef tant dans la compréhension du comportement de vote des député.e.s que dans celle de l'issue finale d'un vote. Cette analyse a notamment démontré la corrélation négative entre abstention et polarisation des votes : les scrutins moins consensuels mobilisent plus les député.e.s.

Après avoir montré l'importance des partis et de l'abstention dans le fonctionnement de l'Assemblée nationale, nous avons cherché à déterminer dans quelle mesure les données mises à disposition sur le site de l'*open data* de l'Assemblée permettaient de rendre compte de ces particularités.

Nous avons d'abord cherché à retrouver le parti d'appartenance des député.e.s à partir de leurs caractéristiques de vote. Nous avons utilisé des méthodes de clustering sur l'historique de vote des député.e.s, et avons représenté ce clustering (et son évolution par année) grâce à une analyse en composantes principales. Si les clusters construits ne recoupent pas les partis politiques, ils se rapprochent d'un partitionnement majorité/opposition. Les ACP construites par année mettent en évidence une tendance à la formation de trois grands ensembles : la majorité parlementaire, l'opposition de droite et l'opposition de gauche. Ce résultat est à mettre en relation avec les conclusions tirées de l'étude de la cohésion des partis. Les trois groupes mis en évidence sont ainsi à la fois distincts par leur cohésion et par leurs positions politiques.

Dans un deuxième temps, nous avons cherché à prédire l'abstention ou le vote d'un.e député.e à un scrutin sachant son thème, à partir de ses caractéristiques de votes mais également de ses caractéristiques socio-professionnelles. Nous avons implémenté une Random forest dont les performances de prédictions sont homogènes selon les thèmes

et plutôt satisfaisantes au regard de la multiplicité des facteurs entrant en jeu dans le phénomène de l'abstention.

Ainsi, les simples données de votes des député.e.s permettent de rendre compte du fonctionnement spécifique de l'Assemblée nationale, et en particulier de l'importance de l'abstention et de l'appartenance à un parti. En ajoutant des éléments relatifs au contexte politique ainsi que certaines données socio-professionnelles des député.e.s, nous avons pu donner du sens aux données de vote et mieux appréhender les enjeux de la XVème législature, comme le fractionnement du parti de gouvernement La République en Marche ou encore la tripartition de l'Assemblée. L'ensemble de cette démarche, à laquelle s'ajoute la prédiction, même imparfaite, de l'abstention des député.e.s en fonction du thème des scrutins, s'inscrit au coeur de la logique d'*open data* et lui donne tout son sens.

Nous avons mentionné précédemment les précautions à prendre dans l'analyse et l'interprétation des statistiques de l'abstention. Or un des facteurs déterminants de l'abstention des député.e.s n'a pas été étudié dans le présent rapport et pourrait apporter un degré de compréhension supplémentaire de ce phénomène : l'appartenance des député.e.s à des commissions, indépendemment de leur parti d'appartenance. En effet, les député.e.s participent davantage aux scrutins dépendant de leurs commissions et sur lesquels ils et elles sont compétent.e.s. A l'inverse, ils et elles ont tendance à ne pas participer aux scrutins portant sur des sujets éloignés de leurs domaines d'expertise. Les données de l'Assemblée nationale peuvent ainsi encore être utilisées pour approfondir les analyses menées dans le cadre de ce projet, ou pour se saisir de nombreux sujets non abordés ici.

4 Annexe

4.1 Règle du coude utilisée dans le clustering

Nous apportons dans cette section des précisions sur la façon dont nous avons obtenu un nombre optimal de $k = 7$ clusters dans la partie 4.1.1 en utilisant l'algorithme des K-moyennes. En effet, un nombre trop important de clusters peut conduire à un partitionnement trop fragmenté des données, tandis qu'un nombre trop faible peut conduire à des clusters trop généralistes ; dans les deux cas, cela ne permet pas de mettre en évidence une structure intéressante des données. Afin de déterminer le nombre optimal de clusters pour partitionner les données, nous avons donc eu recours à la règle du coude. Il s'agit de représenter graphiquement la distorsion du modèle en fonction du nombre de clusters choisi, et de déterminer le point d'infléchissement de cette courbe, c'est-à-dire déterminer quand la distorsion ne se réduit plus significativement. Cela correspond au moment à partir duquel ajouter un cluster supplémentaire n'apporte que peu de gains en terme de modélisation des données. L'indice de distorsion que nous utilisons permet de mesurer la variance du cluster : plus celle-ci est faible, plus les données regroupées au sein d'un cluster sont homogènes. Cet indice est défini comme suit :

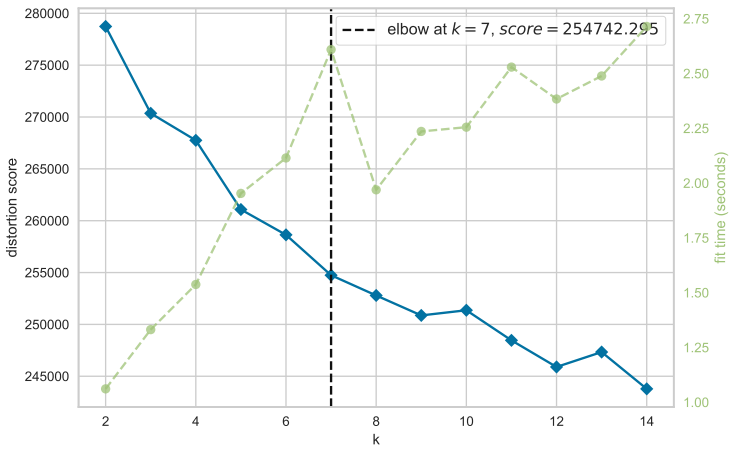
$$\frac{1}{n}\sum_{i=1}^{n}\|x_i - \mu_i\|^2 \tag{7}$$

où μ_i est le centroïde le plus proche de l'observation x_i.

On cherche donc le nombre de clusters k tel que tous les clusters retenus minimisent la distance entre leur centroïde et les observations associées : il s'agit de minimiser la distance intra-classe.

Nous obtenons, dans la **Figure 10**, un infléchissement de la décroissance de la distorsion à partir de $k = 7$ clusters. C'est donc ce nombre que nous conservons.

FIGURE 10 – Règle du coude associée à l'algorithme des K-moyennes.



4.2 Sélection des variables pour le clustering

Notre démarche entamée en première partie a permis de mettre en valeur différents facteurs clés du fonctionnement de l'Assemblée : à la fois la forte discipline de vote, mais aussi la cohésion relative des partis, l'important phénomène de l'abstention ainsi que l'influence du demandeur dans le succès d'un scrutin. Il nous a ainsi paru naturel de tenir compte de ces éléments pour les méthodes de clustering et d'ACP que nous avons mises en oeuvre. Nous avons donc utilisé à la fois l'historique de vote des député.e.s mais aussi d'autres indices synthétiques que nous avons construits, permettant d'inclure dans une certaine mesure ces éléments : taux d'abstention, indice d'accord, taux de votes "pour" et taux de votes "contre". La construction de ces variables n'a évidemment pas été faite au hasard et nous tenions à préciser la démarche qui a mené à la sélection des variables finales.

Nous avions dans la première partie souligné l'importance de la discipline de vote, de l'abstention mais aussi du demandeur, influençant chacun dans une certaine mesure le comportement des député.e.s. Tout d'abord, représenter les positions des député.e.s à partir de leur historique de vote nous apparaissait crucial pour voir si ce dernier suffisait à décrire l'appartenance d'un.e député.e à son parti.

Nous avons cependant choisi de construire d'autres indices susceptibles d'enrichir la base : le taux d'accord, le taux d'abstention, de "pour" et de "contre", mais également le "comportement moyen" de l'élu.e en fonction du demandeur du scrutin. Ce dernier ajoutait cependant de nombreuses variables, à savoir deux fois le nombre de demandeurs, pour obtenir la moyenne des votes "pour" et "contre" de chaque député.e en fonction du demandeur du scrutin, soit 28 variables supplémentaires. Ces indices, qui tentaient d'inclure l'identité du demandeur comme susceptible d'influencer le comportement de vote d'un.e député.e, augmentaient ainsi sensiblement la complexité du modèle, en passant de 4 indices à 32. Il nous a ainsi paru important de chercher à mesurer l'impact de telles variables, à savoir à quel point elles étaient susceptibles d'affiner et d'améliorer le clustering, afin d'arbitrer entre l'augmentation de la précision liée à leur ajout et l'augmentation de la complexité du modèle.

Nous avons donc comparé à chaque fois les clusterings obtenus avec et sans l'ajout de ces variables liées aux demandeurs et avons constaté que cela n'améliorait pas sensiblement ces derniers (notamment que cela ne regroupait pas mieux les député.e.s en clusters bien distincts selon les partis). On peut notamment l'observer dans l'exemple suivant où l'on compare le clustering selon les caractéristiques des député.e.s (en utilisant les indices synthétiques) avec et sans les variables relatives aux demandeurs : nous avons regroupé les proportions obtenues pour chaque méthode dans la **Table 7**. Tout d'abord, en séparant les député.e.s en deux clusters, on retrouve à chaque fois dans une certaine mesure la polarisation majorité/opposition, le cluster n°1 regroupant dans chaque cas plutôt la majorité, avec une forte proportion de député.e.s LREM et Modem et apparentés, et le cluster n°2 s'apparentant à l'opposition, avec notamment une proportion significative de député.e.s LR ou encore de la France insoumise. En comparant les deux clusterings, on remarque notamment qu'entre leurs clusters n°1 respectifs, aucune proportion des groupes politiques représentés ne varie de façon significative lorsqu'on ne considère plus les indices relatifs aux demandeurs (on n'a notamment jamais de variation de plus de 1 point de pourcentage). Les clusters n°2 sont également très semblables dans les deux cas. On compte néanmoins légèrement plus (à hauteur de 1% en proportion) de députés LREM dans le cluster 1 et moins dans le cluster 2 (on passe de 3% à 1%) lorsqu'on retire les indices liés aux demandeurs. On pourrait éventuellement supposer que l'ajout des variables liées aux demandeurs sépare légèrement moins bien les données selon l'axe majorité/opposition que l'on cherchait à retrouver dans notre partitionnement en deux clusters. En effet, les autres répartitions des groupes politique étant assez similaires d'un cluster à l'autre (avec et sans indices demandeurs), mieux regrouper les député.e.s LREM revient dans une certaine mesure à mieux regrouper la majorité. On peut cependant opposer à cela le fait que dans le cas du partitionnement où l'on ajoute les indices relatifs aux demandeurs, le cluster n°2 regroupe en proportion plus de député.e.s LR (48%

contre 44%), qui est un groupe de l'opposition. On ne peut donc pas réellement affirmer que la séparation est moins bonne. Cependant les variations en proportion sont si faibles que l'on peut négliger l'apport lié aux variables des demandeurs.

Ainsi globalement ces deux clusterings (avec et sans les variables relatives aux demandeurs) sont extrêmement similaires. Cela peut s'expliquer en partie car l'information apportée par ces statistiques moyennes en fonction du demandeur étaient déjà partiellement contenues par les indices synthétiques que nous avions créés (indice d'accord par exemple). Les variables relatives aux demandeurs ne sont donc pas décisives dans le partitionnement réalisé. Nous avons ainsi choisi de ne pas prendre en compte ces variables dans la démarche de clustering et d'ACP.

TABLE 7 – Comparaison de deux partitionnements des député.e.s.

Clustering	Selon les indices		Selon les indices et les demandeurs	
Numéro de cluster	1	2	1	2
Les Républicains	1%	44%	0%	48%
Libertés et Territoires	1%	6%	1%	6%
UDI et Indépendants	1%	7%	1%	7%
Agir ensemble	4%	3%	5%	1%
MoDem et apparentés	14%	5%	15%	2%
La République en Marche	76%	3%	75%	1%
Gauche démocrate	0%	7%	0%	7%
Socialistes et apparentés	0%	13%	0%	13%
La France insoumise	0%	7%	0%	8%
Non inscrit	3%	6%	4%	5%
Effectif total (membres)	230	345	217	358

4.3 Précisions concernant les méthodes de NLP utilisées dans le partitionnement par thème des propositions

Les intitulés des scrutins proposés sont représentés par un ensemble $(X_1, ..., X_{3117})$ où X_i est l'intitulé i. Les mots formant ces intitulés constituent notre vocabulaire défini comme l'ensemble $(Y_1, ..., Y_{296})$ où Y_j représente un mot. Un thème sera défini par un sous-ensemble de vecteurs du vocabulaire.

Preproccesing :

Nous allons prendre un exemple pour illustrer le déroulé du traitement. Considérons pour cela le vecteur X_{2793}, donné sous la forme suivante : "*l'amendement n° 630 de M. Aubert à l'article premier du projet de loi relatif à l'énergie et au climat (première lecture).*"

Le travail de preprocessing consiste à nettoyer les données. Ainsi, pour chaque vecteur : nous enlevons les caractères spéciaux, les nombres et les accents ; nous transformons toutes les lettres en lettres minuscules ; nous enlevons les *Stopwords* français ainsi que certains mots propres aux propositions n'apportant pas de sens supplémentaire aux vecteurs (ex : 'amendement', 'article', 'lecture', 'motion', 'projet', 'loi', ...). Tous les vecteurs sont également *tokenisés*, *i.e* les intitulés sont transformés en liste de tokens. Ici les tokens sont les mots non supprimés et transformés. Dans notre exemple, le vecteur X_{2793} devient ainsi : *['aubert', 'énergie', 'climat']*.

Nous avons finalement décidé de supprimer les mots peu fréquents afin d'éliminer les noms de famille ou autre précisions ne pouvant constituer un thème. Le vecteur X_{2793} est alors tel que : *['energie', 'climat']*.

Construction de la matrice de pondération :

Nous avons choisi d'appliquer la méthode de ***Term Frequency - Inverse Document Frequency*** (TF-IDF). Il s'agit d'une méthode de pondération de l'importance des mots prenant en compte la fréquence des mots par vecteur et dans l'ensemble des vecteurs. On le formalise de la façon suivante :

On définit la fonction *Term Frequency* d'un mot Y_j dans un intitulé X_i telle que :

$$TF_{j,i}(Y_j, X_i) = \frac{\text{Nombre d'occurence du mot } Y_j \text{ dans l'intitulé } X_i}{\text{Nombre de mots total dans l'intitulé } X_i} \tag{8}$$

Avec les mêmes notations et $df(Y_j)$ = *le nombre de documents où le terme* Y_j *apparaît*, on définit la fonction *Inverse Document Frequency* d'un mot Y_j telle que :

$$IDF(Y_j) = \log(\frac{n+1}{df(Y_j)+1}) + 1 \tag{9}$$

où n=3117 est le nombre total d'intitulés.

Finalement nous calculons donc la pondération de chaque mot dans chaque intitulé, définie telle que :

$$TFIDF_{j,i}(Y_j, X_i) = TF_{j,i}(Y_j, X_i).IDF(Y_j) \tag{10}$$

Nous normalisons les données en norme $L2$. Nous obtenons finalement une matrice $TFIDF$ de dimension (3117x296), de terme général $TFIDF_{j,i}(Y_j, X_i)$. Pour le clustering, nous utiliserons les vecteurs lignes $TFIDF_i$ dont les 296 composantes représentent la pondération de chaque composante du vocabulaire $(Y_1, ..., Y_{296})$ dans l'intitulé X_i.

Clustering :

Afin de partitionner les propositions des scrutins selon leurs intitulés, nous choisissons d'appliquer la méthode de clustering des K-moyennes. Les vecteurs $TFIDF_1, ..., TFIDF_{3117}$ synthétisant le sens de nos intitulés, le clustering peut être considéré comme une partition par thème.

En effet, comme précisé dans l'équation (2), l'algorithme des K-moyennes minimise la distance entre les vecteurs de chaque cluster, *i.e* rapproche les intitulés de thème proche. Nous avons choisi $k = 35$, de sorte que nous obtenons finalement 35 thèmes partitionnant les 3117 propositions. Pour reprendre notre exemple, l'intitulé X_{2793} est assigné au cluster n°33 regroupant les propositions autour du thème du climat. La **Table 8** énonce les thèmes des clusters les plus pertinents. Ceux que nous ne gardons pas regroupaient des intitulés trop variés pour les désigner par une thématique commune. Nous indiquons face à chaque thème le taux d'abstention moyen pour ces thèmes. La moyenne est calculée sur tous les député.e.s et tous les scrutins du thème.

4.4 Résultats de l'approche non retenue pour la Random forest

Les résultats sont proches de la **Figure 6** et sont même légèrement meilleurs, mais cette approche nous apparaît moins pertinente. L'interprétation des résultats reste tout de même similaire et l'implémentation de cette méthode nous a également permis de nous rendre compte de certaines limites temporelles par exemple. En outre, cette approche nous a aussi permis de confirmer que la précision et le recall de notre méthode sont toujours du même

TABLE 8 – Abstention moyenne selon le thème des scrutins.

Thème	Abstention moyenne
Sécurité sociale	77%
Finances	70%
Retraite universelle n°1	76%
Relations commerciales agricoles	75%
Immigration	72%
Démocratie représentative	74%
Avenir professionnel	61%
Réforme justice	73%
Bioéthique	85%
Retraite universelle n°2	70%
Notre-Dame	78%
Gaspillage et économie circulaire	79%
Pacte Ferroviaire	57%
Haine sur internet	72%
Violences sexuelles et sexistes	70%
Energie et climat	88%

ordre et que le sur-apprentissage correspond aux thèmes les moins représentés. Ainsi, malgré la logique moins pertinente de cette approche, les interprétations peuvent être en relation avec celles de la **Figure 6**.

TABLE 9 – Performances de la Random Forest selon le thème avec calcul des indices indépendamment du thème donné

Résultats / **Thème**	Précision train	Précision test	Recall	F1 Score
Sécurité sociale	0.668	0.653	0.653	0.653
Finances	0.648	0.631	0.631	0.629
Retraite universelle n°1	0.697	0.675	0.675	0.673
Relations commerciales agricoles	x	x	x	x
Immigration	0.969	0.635	0.635	0.634
Démocratie représentative	0.702	0.673	0.673	0.672
Avenir professionnel	x	x	x	x
Réforme justice	0.636	0.584	0.584	0.584
Bioéthique	0.700	0.680	0.678	0.678
Retraite universelle n°2	0.672	0.652	0.651	0.652
Notre-Dame	0.725	0.646	0.646	0.646
Gaspillage et économie circulaire	0.695	0.662	0.662	0.662
Pacte Ferroviaire	x	x	x	x
Haine sur internet	0.670	0.606	0.606	0.605
Violences sexuelles et sexistes	0.655	0.591	0.591	0.591
Energie et climat	0.731	0.649	0.649	0.648

Les “x” correspondent au thèmes dont les données sont absentes pour la période 2019-2020.

Liste des figures

Liste des tableaux

Références

[1] www.csis.org/blogs/technology-policy-blog/evolving-role-artificial-intelligence-and-machine-learning-us-politics

[2] www.liberation.fr/planete/2018/03/26/sans-cambridge-analytica-il-n-y-aurait-pas-eu-de-brexit_1638940/

[3] www.francetvinfo.fr/politique/l-assemblee-nationale-offre-acces-a-ses-donnees-en-open-data_964437.html

[4] National Research Council, *On the Full and Open Exchange of Scientific Data*, Washington, DC : The National Academies Press, 1995.

[5] https ://opengovdata.org/

[6] Algan Y., Cahuc P, *La Société de défiance. Comment le modèle social français s'autodétruit*, CEPREMAP, 2016.

[7] www.etalab.gouv.fr/retour-sur-le-4e-sommet-du-partenariat-pour-un-gouvernement-ouvert-pgo-a-paris-en-decembre-2016.

[8] www.franceculture.fr/politique/lassemblee-nationale-va-ouvrir-ses-donnees

[9] Reignier D., *La discipline de vote dans les assemblées parlementaires sous la cinquième République*, Droit. Université du Droit et de la Santé - Lille II, 2011.

[10] Hix S., Noury A. et Roland G., *Democratic Politics in the European Parliament*, Cambridge, Cambridge University Press, 2007.

[11] https ://datan.fr/statistiques/aidecohesion

[12] www.mediapart.fr/journal/france/061017/la-france-insoumise-la-democratie-interne-fait-debat ?onglet=full

[13] www.lefigaro.fr/politique/le-scan/depuis-2017-le-groupe-lrem-perd-un-depute-tous-les-deux-mois-et-demi-20191113

[14] Gaxie D., *Le Cens caché. Inégalités culturelles et ségrégation politique*, Le Seuil, « Hors collection », 1978.

[15] Piven F. et Cloward R., *Why Americans Don't Vote*, Pantheon, 1988.

[16] Braconnier C., Dormagen J-Y., *La démocratie de l'abstention : aux origines de la démobilisation électorale en milieu populaire*, Gallimard, 2007.

[17] Tiberj V, «La démocratie à l'épreuve de la jeunesse : une (ré)génération politique ? », *INJEP, Analyses Synthèses*, no 46, mars 2021.

[18] https ://www.cairn.info/revue-francaise-de-science-politique-2012-1-page-71.htm

[19] Gethin A., Martinez-Toledano C., Piketty T., *Clivages politiques et inégalités sociales*, Hautes Etudes, EHESS/Gallimard/Seuil, 2021, pp 96-112.

[20] www.francetvinfo.fr/meteo/climat/loi-climat-l-assemblee-nationale-adopte-le-projet-en-premiere-lecture_4610961.html