

Mesurer l'activité de l'Assemblée nationale à partir de ses données ouvertes

Définitions, dénominateurs et incertitudes des indicateurs du *Radar parlementaire*

Sylvain Desroziers

Citoyen. Sans affiliation institutionnelle, sans financement, sans parti.

sylvain.desroziers@radar-parlementaire.fr

Rédaction assistée par Claude (Anthropic) ; voir la déclaration en section 15.

14 août 2026

Version de référence : radar-parlementaire.fr/methode.html

Résumé

L'Assemblée nationale publie en données ouvertes le détail nominatif de ses scrutins publics, ses amendements et la composition de ses groupes. Ces fichiers sont accessibles à tous et lus par presque personne : leur format suppose des compétences que la citoyenneté ne devrait pas exiger. Le *Radar parlementaire* les transforme en indicateurs consultables, et ce document spécifie ces indicateurs.

Nous décrivons chaque mesure publiée par le site : sa définition formelle, son numérateur, son dénominateur, la population sur laquelle elle porte, les observations qu'elle exclut et l'incertitude qui l'entoure. Sont couverts l'accord de vote entre députés et entre groupes, la participation, la cohésion et la dissidence, deux réductions descriptives (analyse en composantes principales et positionnement multidimensionnel), un modèle de points idéaux estimé avec ses intervalles par bootstrap de blocs, le traitement de l'abstention, le réseau de cosignatures d'amendements et la détection de sujets sous contrôle du taux de fausses découvertes.

Deux principes gouvernent l'ensemble. Aucun taux n'est publié sans le dénombrement qui le produit : un pourcentage privé de son dénominateur n'est pas une information, c'est une opinion présentée comme un fait. Et aucun indicateur ne délivre de verdict : ces mesures décrivent une part observable de l'activité parlementaire, elles n'évaluent ni la qualité d'un élu ni la justesse d'une position. Le document est écrit pour être vérifié, y compris — et surtout — par qui souhaite démontrer qu'il se trompe.

Mots-clés — données ouvertes parlementaires ; scrutins publics ; points idéaux ; cohésion de groupe ; cosignatures ; transparence algorithmique ; reproductibilité.

Table des matières

1	Introduction	3
2	Avant toute formule : quelle est la question ?	4
3	Des fichiers bruts au cube de votes	6
4	Proximité de vote : compter les accords, puis seulement les diviser	9
5	Présence, ligne de groupe et dissidence	11

6 Deux cartes descriptives : ACP et positionnement multidimensionnel	14
7 Le modèle de points idéaux : expliquer et prédire les votes binaires	16
8 L’abstention est une position à analyser	19
9 Amendements et réseau de cosignatures	20
10 Détecter les sujets qui montent	21
11 Lire les sorties sans leur faire dire plus qu’elles ne disent	22
12 Reproduire et auditer	23
13 Limites connues	23
14 Disponibilité des données, du code et de ce document	24
15 Déclaration sur l’usage d’outils d’assistance à la rédaction	24
Licence de ce document	25
Références	25

1 Introduction

1.1 Une transparence qui existe sans être accessible

L'*open data* de l'Assemblée nationale est une réussite institutionnelle réelle et une impasse pratique. Chaque scrutin public y figure avec le vote nominatif de chaque député ; chaque amendement avec son auteur, ses cosignataires et son sort. Rien n'est caché. Mais la donnée est distribuée en archives compressées de plusieurs centaines de mégaoctets, en JSON issus d'une conversion XML dont les conventions ne sont documentées nulle part, et sans aucun indicateur dérivé. La transparence est juridiquement acquise et matériellement hors de portée.

L'espace laissé vide se remplit tout seul. Il se remplit de chiffres cités sans source, de classements de « députés les plus absents » dont le dénominateur n'est jamais donné, de pourcentages d'accord calculés sur des échantillons choisis après coup. Ces chiffres circulent parce qu'ils sont les seuls disponibles sous une forme lisible, et non parce qu'ils sont justes.

Le *Radar parlementaire* occupe cet espace autrement : il recalcule chaque indicateur depuis les fichiers officiels et publie, avec chaque résultat, ce qui permet de le contester.

1.2 Ce que ce document ajoute au site

Un site qui affiche « 42 % de présence » demande à être cru. Un site qui affiche « 42 %, soit 178 scrutins sur 424 où l'élu pouvait voter, après retrait des scrutins antérieurs à son entrée en fonction et des non-votes imposés par une fonction incompatible » demande à être vérifié. La seconde formulation ne vaut que si la règle qu'elle invoque est écrite quelque part, précisément, une fois pour toutes. C'est ce document.

Il remplit trois fonctions distinctes.

Une spécification. Chaque mesure affichée par le site est définie ici formellement. Le code implémente cette spécification ; en cas de divergence, c'est le code qui a tort. Cette contrainte a déjà servi : plusieurs incohérences entre deux chemins de calcul d'une même grandeur ont été découvertes en écrivant les définitions plutôt qu'en lisant le code.

Un support pédagogique. Les notions mobilisées — dénominateur conditionnel, indice de cohésion, variable latente, bootstrap de blocs, correction pour tests multiples — ne sont pas hors de portée d'un lecteur attentif, mais elles s'expliquent mal en légende d'un graphique. Elles sont introduites ici avec des exemples de petite taille, calculables à la main.

Une prise pour la contradiction. C'est la fonction la plus importante. Un indicateur sur la vie politique sera contesté, et il doit l'être : un député dont le taux de dissidence est publié a le droit de savoir comment il a été obtenu, sur quels scrutins, avec quelle définition de la ligne de son groupe, et à partir de quel seuil le chiffre cesse d'être significatif. Sans ce document, cette contestation ne peut porter que sur la conclusion. Avec lui, elle porte sur la méthode — c'est-à-dire qu'elle devient un débat qu'on peut trancher.

1.3 Trois exigences

Reproductibilité. Les données sources sont publiques, le code est publié sous licence libre, et les tables intermédiaires sont archivées à chaque mise à jour. Un lecteur disposant d'une machine ordinaire peut reconstruire l'intégralité des chiffres du site.

Réfutabilité. Chaque indicateur est accompagné de sa population de référence, de ses exclusions et, lorsque cela a un sens, de son incertitude. Une mesure dont l'échantillon est trop mince n'est

pas affichée avec une précaution en petits caractères : elle n'est pas affichée.

Absence de verdict. Le projet ne classe pas les députés en bons et en mauvais. Il ne pondère pas les scrutins par leur importance supposée, il ne construit pas de note de synthèse, il n'infère aucune intention. La section 11 récapitule, indicateur par indicateur, ce que chacun ne démontre pas. Cette retenue n'est pas de la tiédeur : c'est la condition pour qu'un chiffre reste utilisable par des lecteurs qui ne partagent pas les mêmes convictions.

1.4 Ce dont ce travail hérite

Aucune des méthodes employées ici n'est nouvelle, et c'est délibéré : sur un objet politique, l'originalité méthodologique est un risque, pas un atout. L'analyse quantitative des scrutins nominatifs est un domaine constitué depuis près d'un siècle. L'indice de cohésion d'un groupe remonte à Rice [8]. L'estimation d'une position latente à partir de votes binaires suit la lignée de Poole et Rosenthal [6, 7], dans la variante probabiliste formalisée par Clinton, Jackman et Rivers [2] ; l'application à un parlement européen multipartisan, plus proche du cas français que le bipartisme américain, doit beaucoup à Hix, Noury et Roland [3]. Les outils statistiques auxiliaires sont classiques : indice de similarité de Jaccard [4], positionnement multidimensionnel de Torgerson [9], bootstrap par blocs de Künsch [5] pour tenir compte de la dépendance entre scrutins d'une même séance, et contrôle du taux de fausses découvertes de Benjamini et Hochberg [1] pour la détection de sujets.

L'apport de ce projet n'est donc pas méthodologique. Il est dans l'application systématique de ces méthodes à la législature en cours, dans leur mise à jour automatique, et dans la publication intégrale des dénominateurs.

1.5 Plan

La section 2 pose les unités d'observation et le périmètre. La section 3 décrit le passage des fichiers bruts aux matrices de vote. Les sections 4 à 10 définissent les indicateurs, dans l'ordre où le site les présente. La section 11 récapitule les lectures abusives à éviter, et la section 12 indique comment tout reconstruire.

2 Avant toute formule : quelle est la question ?

Un calcul politique n'est jamais seulement une formule. Il suppose aussi une question, une unité d'observation et un périmètre. Ici, les unités élémentaires sont :

- un **scrutin public** j ;
- un **député** i ;
- un **vote nominatif** de i au scrutin j ;
- lorsque les amendements sont disponibles, un **amendement** et l'ensemble de ses signataires.

Le projet répond à des questions descriptives : « qui vote de la même façon ? », « un groupe a-t-il une ligne nette ? », « quels termes deviennent plus présents ? », « qui cosigne avec qui ? ». Il ne permet pas, à lui seul, de répondre à « pourquoi ce député a-t-il voté ainsi ? », « ce député travaille-t-il bien ? » ou « deux groupes sont-ils alliés ? ». Ces dernières questions nécessitent du contexte politique, qualitatif et institutionnel.

À retenir. Un pourcentage n'est interprétable qu'avec son dénominateur et sa population de référence. C'est pourquoi les fonctions du projet renvoient les comptes bruts avec les taux.

2.1 Qui fait quoi : les fonctions qui déplacent un chiffre

Plusieurs calculs de ce document dépendent de la fonction exercée par un député. Ces fonctions sont familières au monde parlementaire et opaques pour tout le monde sinon ; les voici, avec l'endroit exact où chacune intervient. Ce n'est pas une présentation de l'Assemblée : c'est la liste des rôles dont l'oubli fausse un résultat.

Fonction	Ce que c'est	Où elle change un chiffre
Président de l'Assemblée	Élu par ses pairs pour la durée de la législature, il dirige l'institution. Son siège surélevé a donné le mot <i>perchoir</i> , employé comme synonyme de la fonction. Il ne prend pas part aux votes qu'il préside.	Non-vote structurel PAN : retiré du dénominateur de présence (§5.1).
Président de séance	La présidence tourne au cours de la journée : des vice-présidents se relaient au perchoir. Celui qui préside à l'instant du scrutin ne vote pas.	Non-vote structurel PSE , même retrait.
Membre du Gouvernement	Un député nommé ministre cesse de siéger et son suppléant le remplace ; la source continue néanmoins d'enregistrer ses non-votes.	Non-vote structurel MG , même retrait.
Rapporteur	Député chargé d'examiner un texte au nom d'une commission et d'en proposer les corrections. Deux co-rapporteurs cosignent par dizaines des amendements purement rédactionnels, quels que soient leurs groupes.	Amendements de rapporteur exclus du réseau de cosignatures (§9).
Commission	Formation restreinte où les textes sont examinés avant la séance publique. L'essentiel du travail législatif s'y fait, le plus souvent sans vote nominatif publié .	Hors périmètre : invisible dans tous les indicateurs.
Groupe politique	Réunion d'au moins quinze députés. Il structure le temps de parole, les postes et les positions de vote.	Population de référence de la cohésion, de la ligne et de la dissidence.

Les trois premières lignes ont un point commun qui justifie leur traitement identique : dans les trois cas, **c'est l'institution qui empêche le vote, pas l'élu qui s'y dérobe**, et la source le déclare. La cinquième ligne dit l'inverse : le travail en commission est réel, massif, et ce document n'en mesure rien — non par choix, mais parce qu'il ne laisse pas de trace nominative exploitable.

Prudence d'interprétation. La conséquence pratique est qu'un député très investi en commission, ou occupé à présider, peut apparaître peu actif ici. Le taux de présence répond à « a-t-il voté quand il pouvait voter ? », jamais à « travaille-t-il ? ».

3 Des fichiers bruts au cube de votes

3.1 Nettoyer sans inventer de données

Les fichiers de l'Assemblée sont des JSON issus d'un format XML. Ils peuvent encoder une liste d'un seul élément comme un dictionnaire, ou une valeur absente comme un objet particulier. Le module `src/radar/parse.py` homogénéise ces formes puis produit des tables Parquet : `deputes`, `scrutins`, `votes`, `positions_groupe` et, si demandé, `amendements`.

La table `votes` a une ligne par couple observé (i, j) et quatre états possibles : *pour*, *contre*, *abstention* et *non-votant*. Une absence de ligne dans une table n'est jamais convertie silencieusement en un vote contre ou en une absence : ce serait fabriquer une observation.

3.2 Une matrice par position

Pour n députés et m scrutins sélectionnés, le projet construit quatre matrices booléennes de taille $n \times m$:

$$\begin{aligned} P_{ij} &= \mathbf{1}(i \text{ vote pour au scrutin } j), & C_{ij} &= \mathbf{1}(i \text{ vote contre au scrutin } j), \\ A_{ij} &= \mathbf{1}(i \text{ s'abstient au scrutin } j), & N_{ij} &= \mathbf{1}(i \text{ est non-votant au scrutin } j). \end{aligned}$$

Le vote est dit *exprimé* lorsqu'il vaut pour, contre ou abstention :

$$V_{ij} = P_{ij} + C_{ij} + A_{ij}.$$

Ces quatre matrices sont exclusives : un couple (i, j) relève d'au plus une d'entre elles. V_{ij} vaut donc 1 si le député a exprimé une position, et 0 sinon.

3.3 Le non-votant : la quatrième matrice, et pourquoi elle compte

N_{ij} demande plus d'explications que les trois autres, car elle recouvre une situation que le vocabulaire courant confond avec l'absence.

Cinq états, et non quatre. Un couple (i, j) peut se trouver dans cinq situations, dont la cinquième n'a pas de matrice parce qu'elle se déduit des autres :

État	Ce que contient la source	Codage
Pour / contre / abstention	une ligne avec la position	P , C ou A vaut 1
Non-votant	une ligne, position <code>nonVotant</code> , plus un code de cause	$N_{ij} = 1$
Absent	aucune ligne	les quatre matrices valent 0

La différence entre les deux dernières lignes est la clé. Un *absent* est un député dont la source ne dit rien : il pouvait voter, il ne l'a pas fait, et on ignore pourquoi. Un *non-votant* est un député que l'Assemblée déclare comme n'ayant pas pris part au vote, **en indiquant le motif**. Ce motif n'est pas une explication a posteriori : c'est une donnée du fichier source.

Les motifs observés. Sur la 17^e législature, la source n'emploie que trois codes, et tous trois désignent une fonction incompatible avec le fait de voter :

Code	Fonction	Portée
MG	membre du Gouvernement	un ministre ne vote pas
PSE	président de séance	celui qui préside au moment du scrutin
PAN	président de l'Assemblée nationale	le titulaire du perchoir

On note $\mathcal{S}$ l'ensemble de ces codes et cause_{ij} le motif déclaré. Un non-vote est dit **structurel** lorsque $\text{cause}_{ij} \in \mathcal{S}$.

Pourquoi N_{ij} n'entre pas dans V_{ij} . Un non-votant n'a exprimé aucune position : il ne peut donc être ni d'accord ni en désaccord avec quiconque, et l'inclure dans un taux d'accord fabriquerait une concordance là où il n'y a pas de vote. N n'apparaît pour cette raison ni dans V , ni dans le numérateur d'accord K de la section 4.

Mais N_{ij} n'est pas pour autant sans usage : c'est elle qui permet de distinguer, au dénominateur de présence, ce qu'un député n'a pas fait de ce qu'on l'a empêché de faire. Elle réapparaît sous la forme du terme R_i de la section 5.1, et c'est son seul emploi.

À retenir. Un non-vote structurel n'est pas une absence : c'est l'institution qui interdit le vote, pas l' élu qui s'y soustrait. Le confondre avec de l'absentéisme revient à reprocher à un député d'avoir présidé la séance ou d'être entré au Gouvernement.

Prudence d'interprétation. La réciproque est tout aussi importante : le retrait ne vaut que pour ce que la source justifie explicitement. Élargir $\mathcal{S}$ au-delà des motifs déclarés permettrait de faire disparaître n'importe quelle absence en la requalifiant, et le dénominateur cesserait d'être opposable.

3.4 La matrice signée, et ce qu'elle écrase

Pour les analyses factorielles exploratoires, le projet utilise enfin une matrice signée :

$$S_{ij} = P_{ij} - C_{ij} \in \{-1, 0, +1\}.$$

Cette matrice a une propriété qu'il faut énoncer avant de s'en servir : elle **n'a que trois valeurs pour cinq états**. Le 0 y recouvre à lui seul l'abstention, le non-vote structurel, l'absence, et le cas d'un député non encore élu. Ces quatre situations sont politiquement très différentes et deviennent indiscernables.

Ce n'est pas une remarque de principe. Sur la 17^e législature, un scrutin public réunit en moyenne moins du tiers de l'Assemblée : **environ trois quarts des cellules de S valent 0 sans que personne se soit abstenu**. Autrement dit, S décrit d'abord qui était là, et seulement ensuite qui a voté quoi. Les conséquences sur l'ACP sont détaillées en section 6.1 ; elles sont assez sérieuses pour que la matrice signée soit réservée aux cartes descriptives et jamais employée à situer un député en particulier.

3.5 L'éligibilité : ne pas pénaliser un mandat qui n'avait pas commencé

Le cube contient enfin E_{ij} , l'indicatrice que le mandat de i couvre la date du scrutin j . Si d_j est la date du scrutin, $d_i^{\text{début}}$ et d_i^{fin} les bornes du mandat, alors :

$$E_{ij} = \mathbf{1} \left(d_i^{\text{début}} \leq d_j \leq d_i^{\text{fin}} \right).$$

Pour un mandat en cours, la borne de fin est considérée comme future. Cette matrice est essentielle pour les mesures de présence : un élu arrivé en cours de législature ne pouvait pas participer aux scrutins antérieurs.

3.6 Le parcours d'un texte, en quatre mots

La suite parle d'amendements, d'articles et de « votes sur l'ensemble ». Ces termes désignent des moments distincts de la fabrication d'une loi, et c'est leur ordre qui explique pourquoi ils ne pèsent pas pareil.

1. Un **texte** est déposé : *projet de loi* s'il vient du Gouvernement, *proposition de loi* s'il vient d'un parlementaire. Il est découpé en **articles**, numérotés.
2. Chaque article peut être modifié par des **amendements** — des propositions de retouche, parfois d'un seul mot. N'importe quel député peut en déposer, et un texte disputé en attire des milliers.
3. L'Assemblée vote, dans l'ordre : d'abord les amendements un par un, puis chaque article, puis enfin **l'ensemble du texte**. Ce dernier vote est le seul qui engage sur la loi entière.
4. Le texte part au Sénat, revient éventuellement — c'est la *navette* —, et en cas de désaccord persistant une *commission mixte paritaire* réunit les deux chambres pour chercher un compromis.

Cette mécanique produit une asymétrie massive et purement arithmétique : **une seule loi très disputée génère des milliers de votes d'amendement et un unique vote sur l'ensemble**. Sur la 17^e législature, environ six scrutins publics sur sept portent sur un amendement.

Une conséquence suit immédiatement. Un vote d'amendement est souvent tactique : on peut approuver un aménagement technique tout en combattant le texte, ou rejeter une retouche que l'on juge insuffisante sans s'opposer au principe. Agréger tous les scrutins « toutes portées confondues » revient donc à mesurer surtout de la tactique parlementaire, et à laisser une loi très amendée peser plus lourd que deux ans de votes solennels.

3.7 Portée politique d'un scrutin

C'est pourquoi le projet classe les scrutins par **portée**, selon ce sur quoi ils portent. Le titre est classé par motifs réguliers en une catégorie, puis en une portée :

Catégorie	Portée	Lecture
Amendement, sous-amendement	detail	Travail technique et tactique sur le texte.
Article, motion de procédure, déclaration	intermediaire	Enjeu intermédiaire.
Ensemble d'un texte, motion de censure	texte	Vote qui engage le plus nettement une position sur le texte.

Le classement est déterministe : il dépend de la syntaxe du titre, pas d'un codage politique fait à la main. Les analyses peuvent être recalculées séparément pour les trois portées. Cette séparation évite qu'une longue séquence d'amendements écrase, par son seul nombre, les votes sur l'ensemble d'un texte.

3.8 Ordinaire, solennel, motion de censure : le type est autre chose

La portée décrit *sur quoi* l'Assemblée vote. Le **type** décrit *comment* elle vote, et la source le publie séparément. Les deux ne se recouvrent pas, et il faut les tenir distincts pour ne pas compter deux fois la même précaution.

Code	Type	Ce qui le caractérise
SP0	scrutin public ordinaire	Le cas courant. Le vote a lieu depuis le pupitre, au fil de la séance, souvent devant un hémicycle clairsemé.
SPS	scrutin public solennel	Annoncé à l'avance, inscrit à l'ordre du jour, généralement le mardi après les questions au Gouvernement. La présence y est bien plus forte.
MOC	motion de censure	Procédure distincte : seuls les soutiens de la motion votent, et l'adoption exige la majorité absolue des membres de l'Assemblée, non des votants.

Le scrutin solennel est rare — quelques dizaines sur une législature — et presque toujours de portée **texte** : on solennise le vote sur l'ensemble d'une loi, pas sur un sous-amendement. L'inverse n'est pas vrai : la majorité des votes sur l'ensemble d'un texte reste des scrutins ordinaires.

Ce type est disponible comme filtre (**types_vote**), mais les indicateurs publiés ne l'utilisent pas : la portée est le découpage retenu, parce qu'elle sépare des populations bien plus grandes et repose sur le contenu du vote plutôt que sur son protocole.

Prudence d'interprétation. La motion de censure ne se lit pas comme les autres scrutins. Ne pas la voter n'est pas s'y opposer : c'est le mode normal d'expression de ceux qui ne la soutiennent pas. La source n'y publie d'ailleurs ni « contre » ni abstention, et un taux de participation calculé dessus n'aurait pas de sens.

Le projet calcule aussi la *contestation* du scrutin :

$$q_j = \frac{\min(n_{\text{pour},j}, n_{\text{contre},j})}{\max(n_{\text{pour},j} + n_{\text{contre},j}, 1)}.$$

Elle est proche de 0 pour un scrutin quasi unanime et de 0,5 pour un scrutin très partagé. On peut l'utiliser pour retirer les votes joués d'avance, qui gonflent mécaniquement l'accord général.

4 Proximité de vote : compter les accords, puis seulement les diviser

4.1 La définition entre deux députés

Pour une paire de députés (i, k) , le nombre de scrutins où les deux ont exprimé un suffrage est

$$M_{ik} = \sum_{j=1}^m V_{ij} V_{kj}.$$

Le nombre d'accords est

$$K_{ik} = \sum_{j=1}^m (P_{ij} P_{kj} + C_{ij} C_{kj} + A_{ij} A_{kj}).$$

Le taux d'accord est donc

$$T_{ik} = \frac{K_{ik}}{M_{ik}}.$$

Une absence ou un non-vote ne compte ni comme accord ni comme désaccord : il sort du dénominateur. L'abstention, en revanche, est une position et deux abstentions concordent.

Dans le code, les matrices de tous les K_{ik} et M_{ik} sont obtenues d'un coup par produits matriciels :

$$K = PP^T + CC^T + AA^T, \quad M = VV^T.$$

C'est exactement la formule précédente, calculée pour toutes les paires à la fois.

4.2 Exemple jouet

Considérons six scrutins. A et B votent toujours pareil ; C est opposé à A, sauf qu'ils s'abstiennent tous les deux au cinquième scrutin ; D ne vote qu'une fois, comme A.

	s_1	s_2	s_3	s_4	s_5	s_6
A	pour	pour	contre	contre	abst.	pour
B	pour	pour	contre	contre	abst.	pour
C	contre	contre	pour	pour	abst.	contre
D	pour	—	—	—	—	—

Alors $T_{AB} = 6/6 = 100\%$, $T_{AC} = 1/6 \approx 16,7\%$ et $T_{AD} = 1/1 = 100\%$. Le dernier résultat est exact mais peu informatif. Le projet masque un taux quand M_{ik} est inférieur à un seuil, fixé par défaut à 30 scrutins communs (et à 20 pour les seuls votes de portée `texte`). La diagonale est également masquée : un député est toujours d'accord avec lui-même.

À retenir. Un « 100% d'accord » sans le nombre de scrutins communs est souvent un faux signal. Le seuil ne rend pas une paire peu observée moins proche : il reconnaît que sa proximité n'est pas mesurée de façon suffisamment stable.

4.3 De deux députés à deux groupes : deux questions légitimes

Soient G et H deux groupes. Après exclusion des paires sous le seuil, le projet publie deux conventions :

$$\bar{T}_{GH} = \frac{1}{|\mathcal{P}_{GH}|} \sum_{(i,k) \in \mathcal{P}_{GH}} T_{ik},$$

$$T_{GH}^{\text{agrégé}} = \frac{\sum_{(i,k) \in \mathcal{P}_{GH}} K_{ik}}{\sum_{(i,k) \in \mathcal{P}_{GH}} M_{ik}}.$$

Ici $\mathcal{P}_{GH}$ est l'ensemble des paires retenues, avec une paire interne à G comptée une seule fois.

La première formule est une **moyenne non pondérée des paires**. Elle répond à : « si je tire un député de G et un de H , quelle est leur proximité moyenne ? ». La seconde est un **quotient de sommes** : une paire observée sur 400 scrutins y pèse quatre fois une paire observée sur 100. Elle répond à : « sur tous les votes communs de toutes les paires, quelle part est concordante ? ».

Ce ne sont pas deux estimations concurrentes d'une vérité unique ; elles ont des pondérations et donc des questions différentes. Le nombre de paires et le nombre de scrutins communs accompagnent les deux mesures.

4.4 Cohésion interne

La cohésion d'un groupe est la convention $\bar{T}_{GG}$: l'accord moyen entre les paires distinctes de ses membres. Elle mesure une homogénéité de votes, pas l'existence d'une consigne ni l'intensité des débats internes. Un effectif doit toujours être affiché à côté : une moyenne calculée sur trois binômes est beaucoup plus instable qu'une moyenne calculée sur plusieurs milliers.

4.5 Comparer les portées et vérifier la taille d'échantillon

Pour une paire de groupes, le projet recalcule les taux sur les trois ensembles disjoints de scrutins : détail, intermédiaire et texte. Un écart entre la portée **détail** et la portée **texte** peut cependant venir de deux causes : un vrai changement de comportement ou le fait que le second ensemble est plus petit.

La contre-épreuve est un test de rééchantillonnage. Si la portée **texte** contient m_T scrutins, on tire sans remise m_T scrutins parmi les scrutins de détail, B fois (par défaut $B = 40$), et l'on recalcule l'accord. Si le taux observé sur les textes est extérieur au minimum–maximum des B tirages, il est peu plausible que la seule taille d'échantillon explique l'écart. Cette procédure est un contrôle intuitif ; elle ne remplace pas un modèle causal.

5 Présence, ligne de groupe et dissidence

5.1 Participation : un numérateur, deux retraits au dénominateur

Pour le député i , le nombre de votes exprimés est

$$X_i = \sum_j V_{ij} E_{ij}.$$

Le nombre de scrutins éligibles est $L_i = \sum_j E_{ij}$. Le nombre de non-votes structurels reprend la matrice N et l'ensemble de motifs $\mathcal{S}$ introduits en section 3.3 :

$$R_i = \sum_j N_{ij} E_{ij} \mathbf{1}(\text{cause}_{ij} \in \mathcal{S}).$$

Le dénominateur et le taux sont alors

$$D_i = \max(L_i - R_i, 0), \quad \text{participation}_i = \frac{X_i}{D_i},$$

avec une valeur manquante si $D_i = 0$.

Les deux retraits répondent à deux objections distinctes. Le premier (E_{ij}) écarte les scrutins antérieurs — ou postérieurs — au mandat : on ne reproche pas à un élu arrivé en cours de législature les votes tenus avant son arrivée. Le second (R_i) écarte les non-votes que l'institution impose. Le plancher à zéro n'a pas de rôle statistique : un non-vote structurel suppose un mandat en cours, donc $R_i \leq L_i$ par construction, et le **max** ne protège que d'une source incohérente.

Un cas limite qui n'en est pas un. Le retrait R_i peut être du même ordre que L_i , et c'est attendu. Le président de l'Assemblée ne vote pas tant qu'il préside : sur les votes qui engagent de la 17^e législature, la présidente est déclarée non-votante pour cause **PAN** sur la quasi-totalité des scrutins, et a exprimé un suffrage à chacun des scrutins restants. Son taux est donc de 100 % sur un dénominateur de quelques dizaines de votes, et non de 8 % sur l'ensemble. Les deux nombres sont issus des mêmes données ; un seul répond à la question « à quelle fraction des scrutins où cet élu pouvait voter a-t-il voté ? ».

À retenir. Ce cas illustre pourquoi le dénominateur doit être publié à côté du taux, et non seulement calculé. « 100 % sur 19 votes » et « 8 % sur 221 » décrivent la même personne ; sans les effectifs, le lecteur ne peut pas savoir lequel des deux répond à sa question.

Prudence d'interprétation. Un D_i petit rend le taux instable : à 19 scrutins, un vote de plus ou de moins déplace le taux de cinq points. Le dénominateur sert à lire le taux, pas seulement à le justifier.

Publier un taux et le comparer sont deux choses

De cette instabilité découle une règle qui n'est pas une exclusion. Un taux dont le dénominateur est inférieur à un seuil $D_{\min}$ — fixé à 20, le même nombre qu'exige déjà un taux d'accord sur les votes qui engagent — reste calculé, reste publié, et reste accompagné de son effectif. Ce qu'il ne fait plus, c'est entrer dans la *distribution de comparaison* : la médiane, le maximum, les rangs et la représentation graphique de la population.

La raison est arithmétique. Ces quantités servent de repère à tous les autres députés, et un taux mesuré sur quelques scrutins les fixe sans les mériter. Le cas s'est produit : la présidente de l'Assemblée, qui ne vote pas tant qu'elle préside, a exprimé un suffrage à chacun des dix-neuf scrutins où elle n'était pas empêchée. Son taux de 100 % est exact, et il devenait « le député le plus assidu de la législature », au-dessus de qui a voté 190 fois sur 245. Le maximum de l'Assemblée était fixé par son plus petit dénominateur.

À retenir. Écarter un taux d'une comparaison n'est pas le juger faible. C'est reconnaître qu'il ne répond pas à la même question que les autres, faute de porter sur la même assiette. La fiche concernée l'écrit explicitement plutôt que d'afficher un rang que rien ne fonde.

Le taux mesure la participation aux scrutins publics, et rien d'autre : ni la présence en commission, ni le travail en circonscription, ni les interventions en séance sans vote.

5.2 Le vote par délégation : ce que « présence » recouvre

Un député empêché peut confier son vote à un collègue, qui l'exprime en son nom ; le règlement en autorise une par député. Le vote est alors juridiquement celui du délégant, et il entre au numérateur X_i de la section 5.1 exactement comme un vote émis en personne.

Cette convention est la seule défendable — la délégation transfère l'exercice du vote, pas son imputation — mais elle a une conséquence que le mot « présence » masque. On publie donc le compte séparément. En notant Δ_{ij} l'indicatrice d'un vote délégué :

$$\delta_i = \sum_j V_{ij} E_{ij} \Delta_{ij}, \quad \text{part déléguée}_i = \frac{\delta_i}{X_i}.$$

Le dénominateur est ici X_i , le nombre de suffrages exprimés, et non D_i : la question posée est « parmi les votes émis en son nom, combien l'ont été par un tiers ? », qui porte sur le numérateur de la présence et non sur son assiette.

Les ordres de grandeur important pour lire ce chiffre. La délégation n'est pas un cas limite : elle représente environ 15 % des suffrages exprimés sur l'ensemble des scrutins, et près de 23 % sur les seuls votes qui engagent — c'est-à-dire précisément l'assiette des mesures principales de ce projet. La médiane par député y avoisine 20 %, et les valeurs extrêmes dépassent 90 %.

À retenir. Le taux de la section 5.1 mesure les suffrages *émis au nom* du député, non sa présence physique dans l'hémicycle. Pour un député dont la moitié des votes sont délégués, ces deux lectures divergent franchement, et c'est la première qui est exacte.

Ce que la source ne permet pas. Le fichier de scrutin nomme le délégant et signale que son vote a été délégué ; il ne nomme jamais le collègue qui l’a porté. On ne peut donc ni classer les récipiendaires, ni reconstituer les binômes de délégation. Le site affiche pour cette raison un compte et une part, jamais un porteur : la donnée n’existe pas, et la déduire serait l’inventer.

Prudence d’interprétation. Une part déléguée élevée ne se lit pas comme un manquement. Elle peut résulter d’une mission, d’un congé, d’une maladie ou d’une fonction — autant de motifs que la source ne donne pas. Le chiffre dit comment le vote a été émis, jamais pourquoi.

5.3 Reconstruire la ligne à partir du dépouillement

Pour un groupe g et un scrutin j , on recompte les trois effectifs :

$$n_{gj}^+, \quad n_{gj}^-, \quad n_{gj}^0, \quad n_{gj} = n_{gj}^+ + n_{gj}^- + n_{gj}^0.$$

La part de la position dominante est

$$h_{gj} = \frac{\max(n_{gj}^+, n_{gj}^-, n_{gj}^0)}{n_{gj}}.$$

La source publie une position majoritaire de groupe, mais le projet ne l’utilise pas pour classer les députés : il recalcule les comptes à partir des votes nominatifs. La ligne observée est la position dominante uniquement si :

1. il y a au moins cinq votants dans le groupe ;
2. il n’y a pas d’égalité au sommet ;
3. $h_{gj} > 0,5$.

Pourquoi une majorité stricte ? Avec 7 pour, 5 contre et 5 abstentions, la position la plus fréquente ne rassemble que $7/17 \simeq 41\%$. Les dix autres votes ne sont pas dix infractions à une ligne : ils révèlent qu’aucune position ne domine. Le scrutin est classé *groupe partagé*, pas *fracture*.

5.4 Taux de dissidence

Sur les seuls votes $\mathcal{L}_i$ où une ligne nette existe pour le groupe de i , la dissidence est

$$\text{dissidence}_i = \frac{\sum_{j \in \mathcal{L}_i} \mathbf{1}(\text{vote de } i \neq \text{ligne de son groupe au scrutin } j)}{|\mathcal{L}_i|}.$$

Une personne avec trop peu de votes comparables est écartée du classement (minimum de 50 par défaut). Cette mesure décrit un écart à la majorité *observée* du groupe : elle ne prouve ni une opposition idéologique ni une désobéissance à une consigne privée.

L’incertitude du taux

Ce taux est une proportion estimée sur un nombre fini de votes : il a une marge, et elle se calcule. Le projet publie l’intervalle de Jeffreys au niveau $1 - \gamma$, c’est-à-dire les quantiles de la loi $\text{Beta}(k + \frac{1}{2}, |\mathcal{L}_i| - k + \frac{1}{2})$ où k est le nombre de votes dissidents.

Le choix de Jeffreys plutôt que de l’intervalle de Wald usuel $\hat{p} \pm z\sqrt{\hat{p}(1 - \hat{p})/n}$ n’est pas indifférent ici. Wald se dégrade précisément là où se trouve la majorité des députés : proche de zéro, il produit des bornes négatives et son taux de couverture réel s’effondre. Or la moitié de l’Assemblée est sous 3 % de dissidence. L’intervalle de Jeffreys reste dans $[0, 1]$ par construction et se comporte correctement aux extrêmes.

À retenir. L'ordre de grandeur importe plus que la formule : l'intervalle fait environ un point pour un député médian et jusqu'à quatre points pour les taux élevés. Un rang de dissidence recouvre donc, en pratique, plusieurs dizaines de positions que les données ne départagent pas. Le site publie le rang *et* le nombre de députés dont l'intervalle recouvre celui du député consulté.

5.5 Scrutins serrés et scrutins clivants

Un scrutin est dit serré lorsque

$$e_j = |n_{\text{pour},j} - n_{\text{contre},j}|$$

est faible (au plus 10 voix par défaut). À l'inverse, la part de votes dissidents parmi les votes auxquels une ligne de groupe s'applique permet de construire un taux de fracture du scrutin. Les deux notions sont différentes : un vote peut être très serré mais parfaitement discipliné, ou largement adopté tout en divisant fortement certains groupes.

6 Deux cartes descriptives : ACP et positionnement multidimensionnel

6.1 Analyse en composantes principales (ACP)

L'idée, avant la formule

Chaque député est décrit par une liste de m nombres : sa colonne de votes, codée +1, -1 ou 0. C'est un point dans un espace à m dimensions, où m vaut plusieurs milliers — donc invisible. L'ACP cherche le *plan* sur lequel projeter ce nuage en perdant le moins possible de sa forme, un peu comme on cherche l'angle sous lequel photographier une sculpture pour qu'elle reste reconnaissable.

Elle ne suppose rien sur la politique. Elle constate seulement que les colonnes de votes ne se répartissent pas au hasard : si deux députés votent presque toujours pareil, leurs points sont voisins, et le nuage s'étire dans certaines directions plus que d'autres. L'ACP nomme ces directions.

Le calcul

Premier temps, centrer. On retire à chaque colonne de S sa moyenne :

$$(S_c)_{ij} = S_{ij} - \bar{S}_{.j}, \quad \bar{S}_{.j} = \frac{1}{n} \sum_i S_{ij}.$$

Après cette opération, un coefficient ne dit plus « ce député a voté pour » mais « ce député s'est écarté de la position moyenne du scrutin, et dans quel sens ». C'est ce qu'on veut : un vote acquis à l'unanimité ne distingue personne et devient une colonne de zéros, tandis qu'un vote serré garde toute son amplitude.

Second temps, décomposer. On écrit la décomposition en valeurs singulières

$$S_c = U \Sigma V^T,$$

où Σ est diagonale, de coefficients $\sigma_1 \geq \sigma_2 \geq \dots \geq 0$ appelés *valeurs singulières*. Chaque colonne de V est une direction de l'espace des scrutins — une combinaison de votes ; la valeur singulière

associée mesure l'amplitude du nuage dans cette direction. Les coordonnées des députés sur les deux premiers axes sont les deux premières colonnes de $U\Sigma$.

L'inertie est la dispersion totale du nuage, $\sum_q \sigma_q^2$. La part portée par l'axe r est

$$\frac{\sigma_r^2}{\sum_q \sigma_q^2}.$$

Une part élevée sur le premier axe signifie qu'une seule direction résume l'essentiel des écarts entre députés — c'est-à-dire qu'un clivage domine. Elle ne dit rien de la *nature* de ce clivage.

Les axes peuvent être retournés sans que le résultat change : $U\Sigma$ et $-U\Sigma$ décrivent la même configuration. Leurs signes ne sont donc pas des étiquettes politiques, et « à gauche du graphique » ne veut pas dire « à gauche de l'hémicycle » autrement que par convention d'affichage.

La limite principale : ce que le zéro cache

L'ACP travaille sur S , et S code par 0 quatre situations distinctes (section 3.4). Elle traite donc une absence **exactement comme un point milieu** entre le pour et le contre. C'est une imputation, et c'est la plus forte hypothèse de tout ce document.

Elle est d'autant plus lourde que les zéros sont majoritaires : sur la 17^e législature, environ trois quarts des cellules du cube complet ne portent aucun suffrage exprimé. Les conséquences sont mesurables et vont dans un sens prévisible :

- un député souvent absent a une colonne proche de zéro ; après centrage, son point se retrouve près du centre du nuage, quelles que soient ses convictions ;
- la *distance au centre* de la carte est donc corrélée à l'assiduité. Sur les données de référence, la corrélation entre la valeur absolue de la coordonnée sur l'axe 1 et le nombre de suffrages exprimés dépasse 0,7 ;
- la zone centrale de la carte mêle en conséquence deux populations qu'il ne faut pas confondre : des députés réellement intermédiaires, et des députés peu présents.

Cette faiblesse est propre aux méthodes factorielles appliquées à des données lacunaires ; elle n'affecte pas le taux d'accord de la section 4, qui écarte explicitement les scrutins non partagés, ni le modèle de points idéaux de la section 7, qui ne calcule la vraisemblance que sur les votes observés.

Prudence d'interprétation. L'ACP répond correctement à « quelle est la structure d'ensemble des votes ? » et incorrectement à « où se situe ce député ? ». Pour situer une personne, c'est le modèle de points idéaux qu'il faut lire, avec son intervalle. Le site n'emploie l'ACP que pour des vues d'ensemble, jamais sur une fiche individuelle.

À retenir. Une part d'inertie n'est pas une part de vérité. Elle est calculée sur la matrice telle qu'elle a été codée, imputations comprises : elle mesure la concentration du nuage, pas la qualité de la description.

6.2 Positionnement multidimensionnel (MDS)

Le MDS répond à la même question par l'autre bout. Plutôt que de projeter des colonnes de votes, il part directement des proximités deux à deux et cherche une carte plane où les distances graphiques reproduisent au mieux les distances mesurées. On pose $d_{ik} = 1 - T_{ik}$: voter pareil, c'est être près.

L'intérêt est qu'il s'appuie sur T_{ik} , dont le dénominateur ne retient que les scrutins réellement partagés — l'imputation massive qui affecte l'ACP ne s'y produit donc pas. Une imputation

subsiste néanmoins, mais bien plus limitée : les paires sous le seuil de scrutins communs n'ont pas de taux, et leur distance est provisoirement remplacée par la moyenne des distances disponibles, faute de quoi la double centralisation serait impossible. Ces paires sont peu nombreuses, et cette valeur de remplacement les place au centre du nuage : la même prudence que pour l'ACP s'applique à elles, à cette échelle réduite.

La matrice des distances est ensuite doublement centrée :

$$B = -\frac{1}{2}JD^{\circ 2}J, \quad J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top.$$

Les deux vecteurs propres associés aux plus grandes valeurs propres de B donnent les coordonnées. Le MDS cherche une carte qui préserve autant que possible les distances d'accord ; il est plus directement lié aux proximités de paires, mais il est plus coûteux.

Prudence d'interprétation. Une carte rapproche des comportements observés. Elle ne démontre pas une idéologie, une stratégie commune ou une cause du vote.

7 Le modèle de points idéaux : expliquer et prédire les votes binaires

7.1 Le modèle statistique

L'ACP décrit une configuration ; le modèle de points idéaux pose un mécanisme probabiliste. Il ne conserve que les votes pour et contre. Pour un député i et un scrutin j :

$$Y_{ij} = \mathbf{1}(i \text{ vote pour}), \quad \mathbb{P}(Y_{ij} = 1) = \text{logit}^{-1}(\beta_j^\top x_i - \alpha_j),$$

où

$$\text{logit}^{-1}(z) = \frac{1}{1 + e^{-z}}.$$

Le vocabulaire, qui vient d'ailleurs

Ce modèle est celui de la *réponse à l'item*, construit à l'origine pour analyser des examens : les élèves y ont une aptitude, les questions une difficulté et un pouvoir discriminant. Le vocabulaire a suivi jusqu'ici, et deux termes sur trois induisent en erreur si on les prend au pied de la lettre.

Symbole	Nom consacré	Ce que c'est ici
$x_i \in \mathbb{R}^d$	position latente	la position que l'on cherche. « Latente » signifie non observée : elle n'est écrite nulle part dans les données et n'est connue que par ses effets sur les votes.
$\beta_j \in \mathbb{R}^d$	discrimination	la direction et la netteté avec lesquelles le scrutin j sépare l'Assemblée. Rien à voir avec le sens ordinaire du mot.
$\alpha_j \in \mathbb{R}$	difficulté	le terme le plus mal nommé : il ne mesure aucune difficulté, il situe le <i>seuil</i> du vote — à partir de quelle position on bascule vers le « pour ».

Comment lire l'équation

Une image suffit à fixer les idées. Rangeons les députés sur une ligne. Chaque scrutin y trace une frontière : à gauche on vote contre, à droite on vote pour. Alors, en dimension 1 :

- le **point de bascule** α_j/β_j (défini si $\beta_j \neq 0$) dit où passe la frontière — c'est la position à laquelle un député a exactement une chance sur deux de voter pour ;
- $|\beta_j|$ dit à quel point elle est nette. Grand, la frontière est une coupure franche et la position prédit presque parfaitement le vote ; proche de zéro, le passage est un dégradé et le scrutin traverse les camps au lieu de les opposer.

Estimer le modèle revient donc à chercher le rangement des députés qui rend le plus grand nombre de scrutins explicables par une telle frontière.

À retenir. Une position latente n'est pas une opinion : c'est la coordonnée qui prédit le mieux les votes observés, sur les textes effectivement soumis pendant la période. Changez les textes, vous changez l'axe.

Les abstentions ne sont pas transformées en demi-votes. Elles sont exclues de ce modèle parce qu'il est binaire, non parce qu'elles seraient dépourvues de sens politique. Les analyses dédiées de la section 8 les traitent explicitement.

7.2 Les filtres avant estimation

Le modèle retire les scrutins avec trop peu de pour/contre, les députés avec trop peu de votes binaires et, par défaut, les scrutins très peu contestés :

$$\frac{\min(n_{\text{pour},j}, n_{\text{contre},j})}{n_{\text{pour},j} + n_{\text{contre},j}} \geq 0,025.$$

Un vote quasi unanime peut être politiquement important mais apporte très peu d'information pour ordonner les députés sur un axe. Ce filtre sert à stabiliser l'estimation, pas à déclarer ces scrutins sans intérêt.

7.3 Estimation alternée et pénalisation

Le projet maximise une vraisemblance logistique pénalisée. Schématiquement, il cherche à maximiser :

$$\ell = \sum_{i,j \text{ observés}} [Y_{ij} \log(p_{ij}) + (1 - Y_{ij}) \log(1 - p_{ij})] - \frac{\lambda_x}{2} \sum_i \|x_i\|^2 - \frac{\lambda_\beta}{2} \sum_j \|\beta_j\|^2.$$

Les difficultés α_j ne sont pas pénalisées : un scrutin peut légitimement être très largement acquis. Les paramètres sont ajustés par régressions logistiques pénalisées alternées :

1. positions x_i fixées, estimer les paramètres de chaque scrutin ;
2. paramètres des scrutins fixés, estimer les positions des députés ;
3. recommencer jusqu'à stabilisation de la log-vraisemblance.

Chaque régression logistique est mise à jour par IRLS (*iteratively reweighted least squares*), c'est-à-dire un pas de Newton utilisant les poids $p_{ij}(1 - p_{ij})$. Le code traite les nombreux scrutins en parallèle dans des produits matriciels, ce qui rend possible le bootstrap.

La pénalisation de β_j est importante : avec une séparation parfaite, le maximum de vraisemblance non pénalisé tend vers $|\beta_j| = \infty$. On obtiendrait des frontières verticales et des positions artificiellement tassées. La valeur de λ_β n'est pas choisie à l'œil : elle est comparée par validation croisée.

7.4 Le problème d'identification

Le produit $\beta_j^\top x_i$ reste le même si l'on change simultanément l'échelle et l'orientation de l'espace. Après estimation, le projet :

1. centre les positions ;
2. les réduit à un écart-type unitaire ;
3. les tourne sur leurs axes principaux ;
4. oriente le premier axe pour qu'un groupe d'ancrage soit du côté négatif.

L'ancrage rend deux exécutions lisibles dans le même sens. Il ne donne pas une signification intrinsèque (« gauche », « droite ») à l'axe : il fixe seulement le miroir.

7.5 Évaluer l'ajustement : classification et APRE

La classification compare $\mathbf{1}(p_{ij} \geq 0,5)$ au vote observé. Mais elle favorise les scrutins très déséquilibrés : prédire toujours la majorité est déjà souvent correct. L'APRE (*aggregate proportional reduction in error*) compare donc les erreurs du modèle aux erreurs de ce prédicteur naïf :

$$\text{APRE} = \frac{E_0 - E_{\text{modèle}}}{E_0},$$

où E_0 est la somme des effectifs minoritaires de chaque scrutin, et $E_{\text{modèle}}$ le nombre d'erreurs de classification. $\text{APRE} = 0$ signifie « pas mieux que toujours choisir la majorité » ; $\text{APRE} = 1$ signifie aucune erreur.

Pour choisir le nombre de dimensions et la pénalisation, une part des cellules de vote est mise de côté au hasard. Le modèle est estimé sur le reste puis évalué sur ces votes non vus. Une dimension supplémentaire qui n'améliore que l'ajustement d'apprentissage est du *surajustement* — elle a mémorisé du bruit propre à l'échantillon — et ne doit pas être lue comme un nouveau clivage politique.

Le retrait porte sur des **cellules** et non sur des scrutins entiers, et ce choix est contraint : retirer un scrutin entier priverait le modèle de ses paramètres α_j et β_j , si bien qu'il n'y aurait plus rien à prédire dessus.

Prudence d'interprétation. Il faut reconnaître une tension entre cette procédure et le bootstrap de la section 7.6. Ce dernier tire des blocs entiers au motif — exact — que les scrutins d'une même séance ne sont pas indépendants. La validation croisée, elle, retire des cellules isolées : le modèle voit alors la plupart des votes du même scrutin et peut en estimer précisément les paramètres, ce qui rend la prédiction des cellules retirées plus facile qu'une véritable prévision hors échantillon. Le protocole est celui d'usage pour ce type de modèle, mais il est plus indulgent que le bootstrap sur la même donnée. La conclusion sur le nombre de dimensions doit être lue comme une borne supérieure de ce qui est justifié, non comme une démonstration.

7.6 Incertitude par bootstrap de blocs

Une position estimée est une estimation, non une position mesurée sans erreur. Le projet tire avec remise des *blocs* de scrutins, réestime le modèle à chaque tirage, puis prend les quantiles empiriques. Pour un niveau $1 - \gamma$, l'intervalle est :

$$[\hat{x}_i^{(\gamma/2)}, \hat{x}_i^{(1-\gamma/2)}].$$

Les scrutins d’une même séance sont corrélés : ils peuvent porter sur le même texte et le même rapport de force. Les tirer indépendamment reviendrait à surestimer la quantité d’information. Le bootstrap regroupe donc, si possible, par dossier législatif, sinon par séance, sinon par date. Les intervalles sont plus larges que sous un bootstrap scrutin par scrutin, mais plus honnêtes.

À retenir. Deux intervalles qui se recouvrent ne démontrent pas l’égalité des positions, mais ils doivent empêcher de lire un petit écart de rang comme une différence certaine.

8 L’abstention est une position à analyser

8.1 Taux d’abstention

Pour un groupe, un député, une catégorie ou une portée, le taux est

$$\frac{\#\{\text{abstentions}\}}{\#\{\text{pour, contre ou abstention}\}}.$$

Les absences et non-votes ne sont pas dans le dénominateur : sinon on mesurerait un mélange d’abstention volontaire et d’absentéisme.

8.2 Consigne, retrait ou indétermination

Chaque abstention est jointe à la ligne recalculée de son groupe. Elle est classée :

- **consigne**, si le groupe a une ligne nette d’abstention ;
- **retrait**, si le groupe a une ligne nette pour ou contre ;
- **indéterminée**, si aucune ligne nette n’existe ou si le groupe a trop peu voté.

Les parts sont divisées par le *total de toutes les abstentions*, y compris les indéterminées. Cette troisième catégorie mesure une limite de classement ; la retirer du dénominateur ferait artificiellement gonfler les deux premières.

8.3 Tester l’idée d’une position intermédiaire

Le modèle de points idéaux est estimé sans abstentions. Pour chaque scrutin où les trois populations sont assez nombreuses, on compare les médianes des positions : m_+ pour les pour, m_- pour les contre et m_0 pour les abstentionnistes. Après avoir posé $b = \min(m_+, m_-)$ et $h = \max(m_+, m_-)$, la position relative est

$$r = \frac{m_0 - b}{h - b}.$$

Une valeur entre 0 et 1 place la médiane des abstentionnistes entre les deux camps ; $r = 0,5$ la place à mi-chemin. C’est un test *hors modèle* : l’information d’abstention n’a pas servi à construire l’axe. Il ne prouve pas que chaque abstention est un compromis ; il évalue cette proposition scrutin par scrutin.

8.4 Quand l’abstention aurait pu faire basculer l’issue

Si un texte est adopté, $e = n_{\text{pour}} - n_{\text{contre}} > 0$. Il suffit de déplacer e abstentionnistes vers le contre pour atteindre l’égalité, qui fait échouer l’adoption : le seuil est donc $a \geq e$.

Si le texte est rejeté, $e = n_{\text{contre}} - n_{\text{pour}} \geq 0$. Le pour doit dépasser le contre, pas seulement l’égaliser : le seuil est $a > e$, soit au moins $e + 1$ abstentionnistes. La fonction ne prétend pas

connaître leurs intentions ; elle établit seulement qu'ils étaient assez nombreux pour modifier l'issue sous ce scénario arithmétique.

9 Amendements et réseau de cosignatures

9.1 Volumes et adoption

Pour un député ou un groupe, le projet compte les amendements dont l'auteur est identifié. La source publie pour chacun un *sort* — adopté, rejeté, tombé, retiré, non soutenu — ou rien du tout lorsque le texte n'a pas encore été discuté. En notant A l'ensemble des amendements déposés par l'auteur et $A_{\text{ex}} \subseteq A$ ceux dont la source publie un sort :

$$\text{taux d'adoption} = \frac{\#\{a \in A_{\text{ex}} : \text{sort}(a) = \text{adopté}\}}{\#A_{\text{ex}}}, \quad \text{indéfini si } A_{\text{ex}} = \emptyset.$$

Le dénominateur est l'examen, pas le dépôt, et l'écart n'est pas marginal : sur la 17^e législature, 54 485 des 123 224 amendements enregistrés n'ont pas de sort publié, soit 44 %. Diviser par les dépôts revient à compter comme des échecs des amendements dont l'Assemblée n'a rien dit ; le taux d'un député baisserait alors à chaque nouveau dépôt sur un texte en cours de navette, sans qu'aucun vote ait eu lieu. Le nombre d'amendements en attente est publié à côté du taux, pour que la différence entre $\#A$ et $\#A_{\text{ex}}$ reste lisible.

Ce taux est descriptif : l'adoption dépend notamment de la recevabilité, de la procédure, de la majorité, du sujet et du statut de l'auteur. Comparer seulement des taux sans le volume de dépôts serait trompeur.

9.2 Construire les liens

Un amendement signé par A, B et C crée les trois liens non orientés A–B, A–C et B–C. Si D est la matrice d'incidence députés $\times$ amendements, avec $D_{ia} = 1$ si le député i a signé l'amendement a , alors

$$W = DD^T.$$

Pour $i \neq k$, W_{ik} est le nombre d'amendements cosignés ensemble. Sur la diagonale, W_{ii} est le nombre total d'amendements signés par i : ce n'est pas un lien avec soi-même.

Le réseau exclut par défaut les amendements de rapporteur et les amendements ayant plus de dix signataires. Le premier filtre évite de confondre une tâche institutionnelle avec une alliance. Le second évite qu'un dépôt collectif de groupe crée mécaniquement tous les liens internes possibles.

9.3 Affinité de Jaccard

Le compte W_{ik} favorise les députés les plus actifs. L'affinité utilisée pour classer les paires est donc l'indice de Jaccard :

$$J_{ik} = \frac{W_{ik}}{s_i + s_k - W_{ik}},$$

où $s_i = W_{ii}$ est le nombre d'amendements signés par i . Il s'agit de la taille de l'intersection des deux répertoires de signatures divisée par leur union. Les députés avec moins de 20 signatures sont masqués : une petite base peut produire un Jaccard de 1 sans signal robuste.

9.4 Part des liens entre groupes

Pour deux groupes distincts G et H , le volume de liens est

$$L_{GH} = \sum_{i \in G} \sum_{k \in H} W_{ik}.$$

Pour un lien interne, la matrice symétrique le compte deux fois et la diagonale doit être retirée :

$$L_{GG} = \frac{1}{2} \left(\sum_{i \in G} \sum_{k \in G} W_{ik} - \sum_{i \in G} W_{ii} \right).$$

La part de la production de liens de G orientée vers H est

$$\pi_{G \rightarrow H} = \frac{L_{GH}}{\sum_Q L_{GQ}}.$$

La même convention – chaque paire une fois, jamais la diagonale – est employée au numérateur et au dénominateur. Ainsi, les parts d’une ligne totalisent 1 et les groupes deviennent comparables.

9.5 Les « courtiers » : sortir de son groupe plus que ne l’impose sa taille

La part brute de liens hors groupe est mécaniquement plus grande dans un petit groupe. Pour le député i du groupe G , le projet la compare à l’attendu sous mélange aléatoire :

$$\text{part observée}_i = \frac{\text{liens de } i \text{ hors de } G}{\text{tous les liens de } i}, \quad \text{part attendue}_i = \frac{N - |G|}{N - 1}.$$

Le ratio observé/attendu vaut 1 sous cette référence. Un ratio supérieur à 1 signale une ouverture hors groupe plus forte que ce que la seule taille du groupe produirait. Les non-inscrits sont exclus de cette mesure : « hors de son groupe » n’a alors pas de référence bien définie.

10 Détecter les sujets qui montent

10.1 Corpus et unités de comptage

Le corpus hebdomadaire contient les titres de scrutins et, lorsque disponible, les titres et les premiers caractères des exposés sommaires d’amendements. Les textes sont passés en minuscules sans accents, tokenisés en mots et bigrammes, puis filtrés : mots vides, jargon légistique, termes de procédure et noms de députés ne doivent pas devenir des « sujets ».

L’unité pertinente est le **document qui mentionne au moins une fois un terme**, non le nombre d’occurrences. Si un mot est répété dix fois dans un long document, ce document contribue une fois à $n_{t,w}^{\text{docs}}$. Avec N_t documents durant la semaine t , la fréquence comparable est

$$f_{t,w} = \frac{n_{t,w}^{\text{docs}}}{N_t}.$$

Cela empêche qu’une semaine comportant des textes longs ou beaucoup de dépôts soit confondue avec un changement de thème.

10.2 Attendu binomial et score de poussée

Sur les semaines de référence R , la fréquence de référence du terme w est

$$p_w = \frac{\sum_{t \in R} n_{t,w}^{\text{docs}}}{\sum_{t \in R} N_t}.$$

Dans la semaine courante de taille N , l'effectif attendu est $\mu_w = Np_w$. Sous l'hypothèse de stabilité, le nombre de documents mentionnant le terme suit approximativement

$$X_w \sim \text{Binomiale}(N, p_w).$$

Le projet calcule la probabilité de queue $\mathbb{P}(X_w \geq x_w)$ et un score de classement

$$z_w = \frac{x_w - \mu_w}{\sqrt{\mu_w + 1}}.$$

Le score classe l'ampleur de la hausse ; la p -valeur vérifie si elle est compatible avec le hasard sous le modèle binomial. Pour un terme absent de la référence, une petite fréquence plancher évite une p -valeur artificiellement nulle sur une seule mention.

10.3 Tests multiples : Benjamini–Hochberg

Des milliers de mots et d'expressions peuvent être testés chaque semaine. Au seuil de 5% sans correction, on attendrait beaucoup de faux signaux. Les p -valeurs ordonnées $p_{(1)} \leq \dots \leq p_{(m)}$ sont converties en q -valeurs de Benjamini–Hochberg :

$$q_{(r)} = \min_{s \geq r} \left(\frac{m}{s} p_{(s)} \right),$$

bornée par 1. Filtrer à $q \leq 0,05$ contrôle le taux de fausses découvertes attendu à 5% parmi les signaux retenus, sous les hypothèses usuelles de la procédure. Enfin, des n -grammes parfaitement redondants sont fusionnés pour ne pas faire passer plusieurs fragments d'un même titre pour plusieurs sujets.

11 Lire les sorties sans leur faire dire plus qu'elles ne disent

Indicateur	Ce qu'il décrit	Ce qu'il ne démontre pas
Accord de vote	Concordance de positions exprimées	Motif commun, alliance durable, identité idéologique.
Participation	Vote aux scrutins où l'élu pouvait voter	Volume total de travail parlementaire.
Dissidence	Écart à une majorité observée de groupe	Violation d'une consigne privée ou désaccord idéologique.
Point idéal	Axe latent qui prédit des votes binaires	Étiquette intrinsèque ou position politique totale.
Cosignature	Travail d'initiative rendu visible par les signatures	Alliance générale ou causalité entre députés.
Sujet montant	Hausse relative de mentions dans le corpus	Importance normative ou attention de toute la société.

Une bonne lecture procède dans cet ordre :

1. vérifier le périmètre (dates, portée, élus inclus) ;
2. lire le numérateur et le dénominateur ;
3. comparer, si nécessaire, les scrutins de détail et les votes qui engagent ;
4. regarder les seuils d'exclusion et l'incertitude ;
5. revenir au titre, au texte et au contexte politique du scrutin.

12 Reproduire et auditer

Les calculs décrits ici sont implémentés dans :

- `src/radar/parse.py` : normalisation, catégories, portées et positions de groupe ;
- `src/radar/analyze.py` : cube, accord, cohésion, participation, dissidence et cartes ;
- `src/radar/ideal.py` : modèle de points idéaux, validation et bootstrap ;
- `src/radar/abstention.py` : analyses spécifiques aux abstentions ;
- `src/radar/cosign.py` : réseau de cosignatures ;
- `src/radar/topics.py` : corpus et détection de poussées.

Pour reconstruire les tables depuis les données ouvertes de l'Assemblée :

```
uv run radar update
```

Les tests unitaires du projet contiennent également des cubes synthétiques : ils vérifient notamment les faux 100% d'accord, la différence entre moyenne de taux et quotient de sommes, l'orientation du modèle et l'APRE.

À retenir. La reproductibilité ne consiste pas seulement à pouvoir relancer le code. Elle suppose de publier les définitions, les seuils, les exclusions et les incertitudes qui rendent un chiffre discutable.

13 Limites connues

Les limites qui suivent ne sont pas des réserves de style. Chacune borne une conclusion que le lecteur pourrait être tenté de tirer.

Le scrutin public n'est pas le vote. L'immense majorité des décisions de l'Assemblée se prend à main levée, sans trace nominative. Les scrutins publics sont un sous-ensemble non aléatoire : ils sont demandés, souvent précisément parce qu'un groupe veut rendre des positions visibles. Tout ce qui est mesuré ici porte sur cette fraction, et sur elle seule.

La participation mesurée n'est pas la présence. Un député présent en commission, en circonscription ou en mission n'apparaît pas dans nos dénominateurs. Le taux publié dit à quelle fraction des scrutins publics auxquels il pouvait prendre part un élu a exprimé une position ; il ne dit rien de son travail parlementaire pris dans son ensemble.

Le point idéal est un artefact utile, pas une opinion. L'axe estimé est la direction qui prédit le mieux les votes observés. Son orientation est fixée par convention, son échelle est arbitraire, et son interprétation substantielle dépend entièrement des textes soumis au vote pendant la période. Deux législatures ne sont pas comparables sur cette échelle.

L'appartenance de groupe est datée. Les changements de groupe en cours de mandat sont traités à la date d'effet publiée par l'Assemblée. Les mesures agrégées par groupe portent donc sur une composition qui a varié à l'intérieur de la période d'observation.

La délégation n'est pas neutralisée dans la dissidence. La part déléguée est désormais définie et publiée (section 5.2), mais un seul indicateur en tire les conséquences : la présence, qui l'affiche à côté de son numérateur. Le taux de dissidence, lui, traite un vote délégué comme tous les autres, alors qu'un écart à la ligne du groupe a pu être matériellement produit par le collègue porteur. Recalculé sur les seuls votes émis en personne, le classement déplace le député médian de quelques rangs, et bien davantage ceux qui délèguent souvent. Publier les deux versions, ou au moins cette sensibilité, reste à faire.

La participation reste publiée sans intervalle. La position latente est accompagnée de son intervalle de bootstrap (section 7.6) et la dissidence du sien (section 5.4) ; le taux de présence, lui, est encore un point nu. C'est aussi une proportion, son intervalle se calcule de la même façon, et rien ne justifie l'asymétrie sinon qu'elle n'a pas encore été levée. La lacune est plus étroite qu'elle n'y paraît — le dénominateur vaut 245 pour les trois quarts des députés, ce qui donne des intervalles serrés — mais elle est réelle pour les élus arrivés en cours de législature.

La cosignature n'est pas la collaboration. Signer un amendement est un acte peu coûteux, parfois collectif par défaut au sein d'un groupe. Le réseau décrit une pratique de signature, dont le rapport au travail réel est une question ouverte que ces données ne tranchent pas.

14 Disponibilité des données, du code et de ce document

Données sources. Jeux de données ouverts de l'Assemblée nationale, <https://data.assemblee-nationale.fr>, sous Licence Ouverte ETALAB. Le projet enregistre pour chaque archive l'en-tête Last-Modified renvoyé par le serveur ; la date de la source est publiée sur le site à côté de la date de calcul.

Code. Le paquet `radar` est publié sous licence AGPL-3.0-or-later. Le choix de cette licence est délibéré : un outil dont l'objet est la transparence ne peut pas être réutilisé pour produire des chiffres dont la méthode resterait fermée.

Tables intermédiaires. Elles ne sont pas encore archivées : chaque mise à jour reconstruit les tables et remplace les précédentes. Reconstruire l'état du site à une date passée suppose donc aujourd'hui que les archives amont soient encore disponibles dans la même version — ce que rien ne garantit. L'archivage d'un instantané par mise à jour est prévu ; tant qu'il n'est pas en place, cette limite est réelle et il faut la connaître.

Ce document. Il est publié au format PDF avec le site, à l'adresse <https://radar-parlementaire.fr/methode/note-methodologique.pdf>, et sa source L^AT_EX est versionnée avec le code.

15 Déclaration sur l'usage d'outils d'assistance à la rédaction

Ce document, ainsi qu'une part substantielle du code du projet, ont été rédigés avec l'assistance de **Claude**, modèle de langage développé par **Anthropic**, utilisé de manière interactive.

Le périmètre de cette assistance a été large : formulation et structuration du texte, mise en forme L^AT_EX, formalisation mathématique des définitions, rédaction et révision du code Python, et travail critique sur les choix de méthode — plusieurs incohérences entre indicateurs, dont deux définitions divergentes d'une même grandeur, ont été identifiées au cours de ces échanges.

La responsabilité du contenu n'en est pas partagée pour autant. Les choix de méthode, les seuils,

les décisions de ne pas publier une mesure jugée trop fragile, la vérification des résultats et l'ensemble du texte publié relèvent de l'auteur, qui en répond seul. Une erreur dans ce document est une erreur de l'auteur.

Cette déclaration figure ici pour la même raison que tout le reste du document. Un projet qui exige des autres qu'ils exposent leur méthode ne peut pas dissimuler la sienne, y compris lorsqu'elle porte sur ses propres outils de travail. Mentionner un tel usage devient une pratique attendue dans l'édition scientifique ; nous considérons qu'elle l'est tout autant dès lors qu'on publie des chiffres sur la vie politique.

Licence de ce document

Ce texte est mis à disposition sous licence [Creative Commons Attribution 4.0 International](#) (CC BY 4.0). Il peut être cité, reproduit, traduit et adapté, y compris pour le contredire, à condition d'en attribuer la paternité.

Références

- [1] Y. Benjamini et Y. Hochberg. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, 57(1) :289–300, 1995.
- [2] J. Clinton, S. Jackman et D. Rivers. The Statistical Analysis of Roll Call Data. *American Political Science Review*, 98(2) :355–370, 2004.
- [3] S. Hix, A. Noury et G. Roland. Dimensions of Politics in the European Parliament. *American Journal of Political Science*, 50(2) :494–511, 2006.
- [4] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37 :547–579, 1901.
- [5] H. R. Künsch. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, 17(3) :1217–1241, 1989.
- [6] K. T. Poole et H. Rosenthal. A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*, 29(2) :357–384, 1985.
- [7] K. T. Poole et H. Rosenthal. *Congress : A Political-Economic History of Roll Call Voting*. Oxford University Press, 1997.
- [8] S. A. Rice. *Quantitative Methods in Politics*. Alfred A. Knopf, New York, 1928.
- [9] W. S. Torgerson. Multidimensional Scaling : I. Theory and Method. *Psychometrika*, 17(4) :401–419, 1952.