BRÉF – Base de données Révisée des Élu×es de France

Vincent Labatut, Noémie Février, Guillaume Marrel

AVIGNON UNIVERSITÉ

BRÉF – Base de données Révisée des Élu·es de France

Rapport technique

Vincent Labatut[1,3]
Noémie Février[2,3]
Guillaume Marrel[2,3]

21 septembre 2020

[1] Laboratoire Informatique d'Avignon – LIA EA 4128
[2] Laboratoire Biens, Normes, et Contrats - LBNC EA 3788
[3] Agorantic - FR 3621

{prenom.nom}@univ-avignon.fr

LBNC

LABORATOIRE INFORMATIQUE D'AVIGNON

agorantic
Culture, Patrimoines, Sociétés numériques
Fédération de Recherche

Sommaire

Ce document détaille l'élaboration de la *Base de données Révisée des Élu·es de France* (BRÉF) à partir d'une source principale, le *Répertoire National des Élus* (RNE) et de plusieurs sources secondaires, les bases de données de l'Assemblée Nationale, du Sénat et du Parlement Européen. Cette base a vocation à être étendue ultérieurement, en exploitant plus complètement ces sources secondaires, et à plus long terme en intégrant de nouvelles bases de données et des apports ponctuels.

Contexte. Ce rapport, BRÉF et le code source associé [1] ont été élaborés dans le cadre du travail doctoral de Noémie Février au LBNC [2], sur un financement de la FR Agorantic [3]. Les données traitées sont destinées à être rendues publiques à terme, mais sont à la date de ce rapport uniquement accessibles à certains membres d'Agorantic.

Version. Il s'agit de la version 1 de ce rapport, qui décrit la version 1.0.2 de la BRÉF, datant du 18 juillet 2020 (cf. l'Annexe F pour la liste complète des versions). Celle-ci contient les données du RNE correspondant à l'extraction historique de juillet 2018, c'est à dire en principe tous les élus de France entre 2001 et 2018. Elle comporte également les mandats de sénateur·ices (1959–2020), député·es national·es (2002–2020) et député·es européen·nes (1978–2020), respectivement extraits des bases de données du Sénat, de l'Assemblée Nationale, et du Parlement Européen (publiquement accessibles en ligne), en Janvier 2020.

Des mises à jour ultérieures sont possibles (et probables). La dernière version en date de ce rapport est disponible sur HAL [4].

Organisation. Nous nous intéressons d'abord à notre source primaire. Dans la Section 1, nous décrivons sommairement l'état des données constituant le RNE, dans la version qui nous a été fournie, et présentons dans la Section 2 l'analyse que nous en avons faite pour y détecter un certain nombre de problèmes.

Nous nous tournons ensuite, dans la Section 3, vers les méthodes que nous avons appliquées pour résoudre ces problèmes. Celles-ci impliquent notamment l'exploitation des sources secondaires, que nous décrivons plus succinctement.

Enfin, nous effectuons dans la Section 4 une description de la base obtenue, ainsi que ce ses limites et des problèmes restant à traiter.

Mise en forme. Nous adoptons la convention de mise en forme suivante : les valeurs extraites des bases de données ainsi que les noms des tables de données utilisent une `police d'écriture à chasse fixe`.

1. https://github.com/CompNet/BrefInit – Note : pas encore public.
2. https://lbnc.univ-avignon.fr/
3. https://agorantic.univ-avignon.fr/
4. https://hal.archives-ouvertes.fr/hal-02886580 – Note : pas encore public.

1 Répertoire national des élus

1.1 Présentation du RNE

Le RNE est une base de données gérée par le *Bureau des élections* (BdE), un service dépendant du du *Ministère de l'Intérieur*, et autorisée par décret [5]. Il a pour principal objectif d'informer le Parlement, le gouvernement et les citoyen·nes sur les élu·es et les mandats qu'ils occupent. Il vise aussi à permettre de vérifier que ces mandats sont conformes à la législation, notamment pour ce qui touche au cumul des mandats [6]) et à la parité [7]. Le ministère centralise les données, mais leur saisie est effectuée au niveau des préfectures, principalement au moment des élections.

Le RNE n'est pas accessible directement au grand public, probablement parce qu'il contient des informations sensibles (les adresses des élu·es, par exemple). Pour cette raison, nous ne savons pas précisément ce qu'il contient, outre les données qui ont été rendues publiques par le ministère. En effet, comme le prévoit la *règle de communication des informations* [8], le ministère est tenu de fournir à toute personne qui en ferait la demande une extraction *partielle* du RNE, de laquelle les données personnelles jugées sensibles sont supprimées.

Jusqu'en janvier 2019, cette demande était réalisée auprès d'une préfecture, qui la relayait ensuite au ministère. Il était possible de demander deux types d'extractions : *historique* vs. *photographique*. Une extraction **historique** effectuée à une date donnée contient l'évolution des mandats depuis la création du RNE, en 2001, jusqu'à la date concernée. En revanche, une extraction **photographique** se limite à la liste des élu·es en poste *à la date concernée*. Après janvier 2019, la loi sur les données ouvertes [9] a modifié la politique de communication des documents administratifs. Le ministère a mis en place un nouveau mode d'accès, en publiant régulièrement des extractions photographiques du RNE sur `data.gouv.fr`, le site de données ouvertes du gouvernement français [10], avec une fréquence de 3 mois. Pour le ministère, une telle publication systématique et régulière des données a vocation à remplacer la communication discrétionnaire mise en œuvre jusque là. Si ce nouveau mode de communication présente l'avantage d'offrir un accès grandement facilité au RNE, en revanche cette ouverture s'accompagne d'un appauvrissement des données, qui peut constituer un problème majeur pour le chercheur. D'une part, les extractions mises à disposition sur `data.gouv.fr` sont exclusivement photographiques (et non pas historiques), et d'autre part plusieurs attributs décrivant les mandats sont désormais omis des données fournies : identifiant unique des élu·es, motifs de fin de mandat et de fonction, et nuance politique. Nous revenons en Section 4.4 sur les problèmes occasionnés par ce changement.

5. *Décret n°2001-777 du 30 août 2001* – https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000224644

6. *Loi organique n°2014-125 du 14 février 2014 interdisant le cumul de fonctions exécutives locales avec le mandat de député ou de sénateur* – https://www.legifrance.gouv.fr/eli/loi_organique/2014/2/14/INTX1302979L/jo/texte

7. *Loi n°2000-493 du 6 juin 2000 tendant à favoriser l'égal accès des femmes et des hommes aux mandats électoraux et fonctions électives* – https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000400185

8. *Circulaire n°NOR/INT/A/06/00090C* du ministère de l'intérieur, § 3.1–3.2 – https://www.conseil-constitutionnel.fr/sites/default/files/as/root/bank_mm/dossiers_thematiques/presidentielle_2007/INTA0600090C.pdf

9. Loi n°2016-1321 du 7 octobre 2016 pour une République numérique – https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/

10. https://www.data.gouv.fr/en/datasets/repertoire-national-des-elus-1/

1.2 Structure des données brutes

Au titre de la règle de communication des informations mentionnée en Section 1.1, nous avons sollicité en tant que chercheurs le BdE du ministère, par l'intermédiaire de la préfecture du Vaucluse, le 9 février 2017, afin d'obtenir une extraction du RNE. Brigitte Hazart, alors correspondante RNE du BdE, nous a fourni une extraction *historique* réalisée le 17/07/2018, qui constitue notre source principale dans la constitution de la BRÉF. Cette date sera appelée **date d'extraction** dans le reste de ce document.

1.2.1 Tables de données

L'extraction prend la forme de 11 fichiers texte, qui sont listés dans la Table 1. Nous supposons que cette décomposition est due à des contraintes techniques. En effet, le processus d'extraction a été réalisé via *BusinessObjects* sur une simple machine de bureau, probablement incapable de traiter l'ensemble des données (environ 500 000 élu·es) en une seule fois. Le découpage en plusieurs fichiers permet à la fois de contourner ce problème de mémoire et de distribuer le processus d'extraction dans le temps, en le réalisant en plusieurs itérations.

Les 11 fichiers mentionnés décrivent 8 tables de données distinctes, chacune consacrée à un type de mandat spécifique. La Table 1 indique le sigle que nous associons à chaque table de données (et donc à chaque type de mandat), et que nous utiliserons dans le reste de ce document pour la désigner.

Table	Sigle	Fichier(s)	Lignes
Conseiller·es municipal·es	`CM`	`A Tous CM 01 30.txt`	366 361
		`B Tous CM 31 60.txt`	425 478
		`C Tous CM 61 95.txt`	432 123
		`D Tous CM OM.txt`	15 451
Conseiller·es de communautés de communes	`EPCI`	`E Tous Membres EPCI.txt`	127 940
Conseiller·es général·es et départemental·es	`CD`	`F Tous CD.csv`	12 728
Conseiller·es régional·es	`CR`	`G Tous CR.txt`	5 818
Député·es	`D`	`H Tous Deputes.txt`	2 488
Sénateur·rices	`S`	`I Tous Senateurs.txt`	1 076
Député·es européen·nes	`DE`	`J Tous RPE.txt`	253
Maires	`M`	`K Tous Maires.txt`	113 522
Total			1 503 238

Table 1. Fichiers constituant l'extraction historique du RNE datant du 17/07/2018, et sigles associés aux tables de données correspondantes.

Au cours d'une réunion qui a eu lieu le 08/10/2019 avec Carla Banens, la correspondante RNE actuelle du BdE, et Olivier Le Mer, le responsable technique du RNE, nous avons pu déterminer que la base de données maintenue par le ministère (et qui n'est pas accessible directement au public) ne possède *pas* la structure décrite dans la Table 1. La structure réelle se rapprocherait plus d'une division en *trois* tables principales, `Elu`, `Mandat` et `Fonction`, plus des tables annexes (notamment pour les aspects territoriaux). Nous qualifierons cette base d'**interne** dans le reste du document, pour la distinguer des données auxquelles nous avons effectivement accès.

L'extraction qui nous est fournie, et à plus forte raison les extractions disponibles sur `data.gouv.fr`, ne sont donc que des *vues partielles* de la base interne. Outre le fait que le processus d'extraction est associé à un filtrage des données, il faut aussi souligner qu'il n'est, selon toute vraisemblance, pas exempt d'erreurs. En effet, comme nous le verrons plus loin, nous avons détecté un certain nombre de valeurs incohérentes qui semblent incompatibles avec la structure supposée de la base interne. Par exemple, dans `CM`, la date de naissance de `Robert BLANCHET` est parfois le `22/11/1939` et parfois le `21/11/1939`. Ceci n'est pas cohérent

avec une centralisation de cette information (auquel cas la date serait toujours exactement la même), et nous supposons donc que l'erreur est introduite au moment de l'extraction des données.

1.2.2 Champs

La Table 2 répertorie les champs présents dans les différentes tables de données. Elle est divisée en trois parties arrangées verticalement et séparées par des doubles-lignes horizontales. Ces parties correspondent respectivement aux aspects géographiques, aux informations individuelles des élu·es, et à la dimension temporelle des mandats.

Dimension spatiale. La partie *supérieure* recense les champs relatifs au découpage administratif. Chaque table diffère des autres de par le territoire concerné par le mandat qu'elle représente. La notion de circonscription européenne (`DE`) correspond à la plus grande aire géographique. Elle a évolué au cours du temps, variant du pays entier à une région ou ensemble de régions. Dans le cas de l'extraction que nous traitons, il s'agit uniquement des régions existant avant le redécoupage de 2016 [11]. Les conseiller·es régional·es (`CR`) sont rattachés à une région, et l'extraction contient à la fois des régions antérieures et postérieures à 2016.

Le département apparaît dans toutes les tables, exceptée `DE`. Ceci est dû au fait que les circonscriptions législatives (`D`), les cantons (`CD`) et les communes (`CM`, `EPCI`, `M`) ne sont pas numérotés de façon unique au niveau national : le même numéro est susceptible d'apparaître dans chaque département. Par exemple, il y a de nombreuses circonscriptions portant le numéro `1` : tout dépend du département considéré. Le numéro du département doit donc être précisé afin de les différencier. Dans le cas d'`EPCI`, on trouve à la fois le département de l'EPCI, mais aussi celui de la commune de rattachement de l'élu (i.e. la commune dans laquelle il a été élu). Pour les sénateur·rices (`S`), l'utilisation du terme *département* est un abus de langage, car il s'agit en réalité d'une circonscription qui recouvre, outre les départements, d'autres types de territoires (collectivités d'outre-mer, français de l'étranger, etc.).

Dimension individuelle. La partie *centrale* de la Table 2 indique les informations relatives à l'élu·e : numéro dans le RNE, date de naissance, nom et prénom, sexe, nuance politique, et profession. Ces champs-là sont quasiment tous présents dans chaque table de données, à l'exception du sexe dans `EPCI`. Nous supposons que la base réelle du RNE contient cette information (et bien d'autres), mais qu'elle a été omise par erreur lors de l'extraction de cette table en particulier.

Dimension temporelle. La partie *inférieure* de la Table 2 contient les informations relatives aux mandats et fonctions : dates de début et de fin, nature et motif de fin. Tous ces champs sont également présents dans toutes les tables, là aussi à quelques exceptions près (probablement explicables de la même façon).

Dans la terminologie du RNE, le **mandat** correspond à la position élective (député, sénateur, etc.) tandis que le terme **fonction** recouvre une fonction exécutive dans le cadre d'un mandat. Ainsi, une personne peut être élue à la position de conseiller·e municipal·e, puis à celle de maire au sein de ce conseil municipal. Cet exemple soulève le problème de la redondance potentielle entre les tables `CM` et `M` : en principe, la seconde devrait être incluse dans la première, puisque *Maire* est une fonction possible du mandat *Conseiller municipal*, parmi d'autres comme *Adjoint au maire* (nous décrivons en Section 3.5.1 comment nous avons testé cette hypothèse).

Sur la base de notre étude préliminaire du RNE, nous faisons l'hypothèse qu'une ligne correspond *au plus* à un mandat. Autrement dit, la période couverte ne doit pas corres-

11. *Loi n°2015-29 du 16 janvier 2015 relative à la délimitation des régions, aux élections régionales et départementales et modifiant le calendrier électoral* – https://www.legifrance.gouv.fr/eli/loi/2015/1/16/INTX1412841L/jo/texte

Champ	Nature	CD	CM	CR	D	DE	EPCI	M	S
Code de la circonscription européenne	ID	–	–	–	–	✓	–	–	–
Libellé de la circonscription européenne	Nom	–	–	–	–	✓	–	–	–
Code de la région	ID	–	–	✓	–	–	–	–	–
Libellé de la région	Nom	–	–	✓	–	–	–	–	–
Code du département	ID	✓	✓	✓	✓	–	✓	✓	✓
Libellé du département	Nom	✓	✓	✓	✓	–	–	✓	✓
Code de la circonscription législative	ID	–	–	–	✓	–	–	–	–
Libellé de la circonscription législative	Nom	–	–	–	✓	–	–	–	–
Code du canton	ID	✓	–	–	–	–	–	–	–
Libellé du canton	Nom	✓	–	–	–	–	–	–	–
Numéro SIREN de l'EPCI	ID	–	–	–	–	–	✓	–	–
Libellé de l'EPCI	Nom	–	–	–	–	–	✓	–	–
Code département de l'EPCI	ID	–	–	–	–	–	✓	–	–
Code INSEE de la commune	ID	–	✓	–	–	–	✓	✓	–
Libellé de la commune	Nom	–	✓	–	–	–	✓	✓	–
Population de la commune	Entier	–	✓	–	–	–	–	✓	–
Numéro de l'élu·e	ID	✓	✓	✓	✓	✓	✓	✓	✓
Date de naissance	Date	✓	✓	✓	✓	✓	✓	✓	✓
Nom de l'élu·e	Nom	✓	✓	✓	✓	✓	✓	✓	✓
Prénom de l'élu·e	Nom	✓	✓	✓	✓	✓	✓	✓	✓
Code de sexe	Catégorie	✓	✓	✓	✓	✓	–	✓	✓
Nuance politique	Catégorie	✓	✓	✓	✓	✓	✓	✓	✓
Code de profession	ID	✓	✓	✓	✓	✓	✓	✓	✓
Libellé de profession	Catégorie	✓	✓	✓	✓	✓	✓	✓	✓
Libellé de mandat	Catégorie	✓	–	✓	✓	✓	–	✓	✓
Date de début de mandat	Date	✓	✓	✓	✓	✓	✓	✓	✓
Date de fin de mandat	Date	✓	✓	✓	✓	✓	✓	✓	✓
Motif de fin de mandat	Catégorie	✓	✓	✓	✓	✓	✓	✓	✓
Libellé de fonction	Catégorie	✓	✓	✓	✓	–	✓	✓	✓
Date de début de fonction	Date	✓	✓	✓	✓	–	✓	✓	✓
Date de fin de fonction	Date	✓	✓	✓	✓	–	✓	✓	✓
Motif de fin de fonction	Catégorie	✓	✓	✓	✓	–	✓	✓	✓

Table 2. Champs présents dans les différentes tables de données de l'extraction historique du RNE considérée.

pondre à plusieurs mandats consécutifs : dans ce cas-là, les données devraient plutôt être décomposées en plusieurs lignes (au moins une par mandat). Par contre, la période peut être plus courte qu'un mandat complet, si l'élu·e ne l'a pas mené à son terme, et/ou l'a commencé postérieurement aux élections régulières. En règle générale, on suppose donc qu'une ligne correspond à un mandat, sauf si l'élu a occupé plusieurs fonctions au cours de son mandat : chacune fait alors l'objet d'une ligne spécifique (les informations relatives à leur mandat commun étant strictement identiques).

Remarques générales. Certains de ces champs sont parfois non-renseignés : concrètement, cela correspond dans les fichiers à une cellule *vide*. D'après notre interprétation, cette absence de valeur peut s'interpréter de deux façon bien distinctes. D'une part, et cela semble être le cas le plus courant, le champ peut être laissé vide car il n'est **pas pertinent** pour le cas considéré. Par exemple, si l'élu·e n'occupe pas de fonction durant le mandat considéré, il n'y a pas de raison de renseigner de libellé de fonction. Ou bien si le mandat considéré est en cours à la date d'extraction, la notion de date de fin ne s'applique pas. D'autre part, le champ peut être pertinent mais sa valeur non-renseignée car **inconnue**.

Il s'agit donc là, à proprement parler, d'un manque du RNE : l'information devrait être présente mais ce n'est pas le cas. On observe notamment cette situation pour un bon nombre

de fonctions terminées pour lesquelles le motif de fin n'est pas indiqué (cf. Table 3).

2 Problèmes détectés

Nous avons produit un ensemble de scripts en *Langage R*[12] dans le but d'effectuer de façon automatique une batterie de tests sur les données brutes du RNE. Ce code source est disponible librement en ligne[1]. Il faut toutefois souligner que, avant d'être exécuté, il nécessite d'obtenir par ailleurs les données du RNE. De plus, les tests ne constituent qu'une partie du traitement implémenté dans ce programme. Les aspects techniques sont décrits dans la documentation fournie avec les scripts, et ne sont pas repris en détail ici : nous nous intéressons plutôt aux aspects fonctionnels et aux résultats des tests.

Nous présentons ces résultats au travers de statistiques agrégées et d'exemples sélectionnés, mais il faut souligner que le détail des cas problématiques détectés par nos scripts est automatiquement stocké dans des fichiers dédiés. Bien sûr, rien ne permet d'affirmer que nos tests sont exhaustifs, et il est possible que d'autres problèmes que ceux présentés ci-après nous aient échappés. Inversement, malgré le soin apporté à leur élaboration, il est possible que ces tests détectent des faux positifs, i.e. considèrent comme erronés des cas en réalité corrects.

Nous distinguons deux types de tests bien distincts. Un test de **validité** vise à détecter les valeurs qui sont erronées en soi. C'est par exemple le cas d'une date qui serait ultérieure à la date d'extraction. Un test de **cohérence** a pour but d'identifier des contradictions entre plusieurs valeurs de la table. Ce type de test détecte par exemple une ligne dans laquelle la date de début de mandat est ultérieure à la date de fin correspondante.

Dans la suite de cette section, nous présentons les problèmes détectés en considérant trois catégories de champs : les noms propres et labels (Section 2.1), les dates (Section 2.2), et les identifiants (Section 2.3). Nous terminons avec les problèmes identifiés de façon plus globale, en recoupant plusieurs de ces catégories (Section 2.4). Chaque problème est identifié de façon unique par un code de la forme **Px** (où **x** est un numéro), et l'ensemble des problèmes que nous avons détectés est listé dans la Table 9. Ces codes sont également utilisés en Section 3, au moment de décrire les opérations permettant de traiter les problèmes correspondants.

2.1 Noms propres et labels

Notre évaluation des problèmes se porte d'abord sur les anthroponymes (noms de famille et prénoms, Section 2.1.1) et les toponymes (noms de cantons, communes, régions, etc., Section 2.1.2), puis sur les labels (nuance politique, nature du mandat ou de la fonction, motif de fin de mandat, profession, Section 2.1.3).

2.1.1 Anthroponymes

Les noms de personnes présentent différents types de problèmes, relatif à leur forme, à leur complétude (le fait qu'ils soient renseignés ou pas) et à leur unicité.

Variabilité. Un examen des noms de personnes présents dans le RNE révèle une certaine hétérogénéité, notamment typographique.

L'usage des **signes diacritiques** (accents, cédilles, etc.) n'est pas systématique (**P1**). Ils ont tendance à être présents dans les prénoms, qui sont écrits en minuscules, mais absents des noms de familles, qui sont quant à eux en capitales. Citons par exemple la conseillère municipale `Stéphanie LEFEVRE` (au lieu de `LEFÈVRE`). Mais cela n'est pas toujours vrai, comme l'illustrent les maires `Claude HOUZÉ` et `Madeleine HOUZE` (un même nom rendu de deux façons différentes).

12. https://www.r-project.org/

On observe la présence d'**espaces** consécutives (**P2**), i.e. plusieurs caractères espaces qui se suivent. Cela se produit à la fois dans les noms de famille, par exemple la maire `Martine KSZAK  PEYROUZERE`, et les prénoms, par exemple le maire `Pierre  André BIDAULT`.

L'utilisation de la **ponctuation** (**P3**), et en particulier des **tirets** hauts (-), se fait de façon arbitraire. Dans les prénoms composés, on observe parfois l'utilisation d'un tiret séparateur, comme pour la maire `Marie-Claire BONNAND`, et parfois simplement une espace, comme pour le maire déjà donné en exemple `Pierre  André BIDAULT`. Par ailleurs, le prénom de la maire `Marie - Thérèse BONNEAU` rajoute à cela des espaces séparant le tiret des deux parties du prénom composé, ce qui constitue une erreur typographique. La même remarque s'applique à des cas comme le conseiller municipal `Jean -Claude CHAUDE`, avec cette fois une seule espace incorrecte.

La représentation des **noms d'usage** (**P4**) est visiblement inconsistante. Pour les femmes mariées en particulier, il arrive que le nom de naissance et le nom d'usage apparaissent simultanément. Mais cela peut prendre plusieurs formes distinctes, comme par exemple l'ajout du mot *épouse* (ex. `Céline BAUWENS ÉPOUSE LONGUEPEZ` dans `CM`), voire du diminutif *ép.* (ex. `Christelle DELANNOY ÉP. LEMOINE`), ou même *ép* sans le point (ex. `Josette CARLIER ÉP DEPOORTER`. On rencontre également la forme parenthésée `Marlène LACHAUD ( EPOUSE VEYRET )`. À noter que ces derniers exemples illustrent aussi l'utilisation inconsistante de la ponctuation, en l'occurrence du point et des parenthèses (les espaces sont aussi utilisées de façon incorrecte dans le dernier exemple). Enfin, le mot `née` est également employé, par exemple `Nicole FERFOURI NÉE MARZOLF` dans `CM`. On décompte 2 665 noms de ce type dans tout le RNE. Il arrive également qu'un nom composé soit utilisé pour séparer le nom de naissance du nom d'usage, avec là-aussi des incohérences, notamment dans l'ordre utilisé pour positionner ces deux noms. Par exemple, la même maire est désignée sous les noms d'`Anne-Marie BRUZEAUD-SOUCAZE` et d'`Anne-Marie SOUCAZE-BRUZEAUD`, son nom de naissance étant `BRUZEAUD`.

Complétude. La Table 3 indique le nombre de valeurs manquantes pour chaque table de données, en incluant seulement les champs pour lesquels il manque *au moins* une valeur (les autres ne sont pas représentés). De plus, il s'agit uniquement des manques *inexplicables*, c'est à dire qui ne peuvent pas être justifiés par la valeur d'un autre champ (une notion sur laquelle nous revenons plus loin). On observe une grande hétérogénéité entre les tables de données et les champs considérés.

Pour ce qui concerne les personnes, on observe que les prénoms ne sont pas renseignés du tout dans certaines tables bien spécifiques. Pour `CM`, le prénom est manquant (**P5**) dans 39 lignes, dont 4 se retrouvent également dans `M` et 2 dans `EPCI`. Une vérification manuelle révèle que, pour des raisons extrêmement mystérieuses, il s'agit quasi-exclusivement des prénoms `Marie-Danielle` et `Marie-Danièle`.

Unicité. Il n'y a en général pas de relation biunivoque entre les noms et les entités dans le RNE. D'une part, il arrive qu'un même nom soit utilisé pour désigner des entités différentes, et d'autre part, la même entité peut apparaître sous plusieurs noms différents. Pour les personnes, on rencontre dans le RNE d'une part, évidemment, des homonymes complets, i.e. des personnes possédant exactement les mêmes noms et prénoms, et d'autre part il arrive que la même personne soit désignée par différents noms.

Pour les **noms de famille**, il arrive que la même personne soit désignée par plusieurs noms différents en raison des différents types de variabilité listés auparavant (P1–P4). C'est par exemple le cas de la plupart des personnes désignées par un nom d'usage. Ainsi, `Céline BAUWENS ÉPOUSE LONGUEPEZ` est aussi désignée par `BAUWENS LONGUEPEZ`, un nom intégrant le marqueur `EPOUSE` (ou `EP`) mentionné précédemment. Ou encore `Anne-Marie BRUZEAUD`, une maire (déjà mentionnée) dont le nom prend trois formes distinctes dans le RNE : `BRUZEAUD`, `BRUZEAUD-SOUCAZE`, et `SOUCAZE-BRUZEAUD`. On rencontre également ce qui semble être des coquilles (**P6**), comme par exemple `PEUGOT` pour `PEUGEOT`, `BILLOD` pour `BILLOT`, `LEGENTIL` pour

Pb	Champ	CD	CM	CR	D	DE	EPCI	M	S
P39	Code du département	0	0	43	0	–	4 481	0	0
P45	Code INSEE de la commune	–	0	–	–	–	3 293	0	–
P12	Libellé de la commune	–	0	–	–	–	5 527	0	–
P23	Date de naissance	0	8	0	0	0	0	2	0
P5	Prénom de l'élu	0	39	0	0	0	2	4	0
P16	Nuance politique	0	245 345	0	0	0	8 955	2 885	0
P17	Code de profession	2	3 206	0	0	0	7	150	0
P17	Libellé de profession	2	3 206	0	0	0	7	150	0
P25	Date de début de mandat	0	0	0	0	0	1 269	0	0
P18	Motif de fin de mandat	0	588	1 807	1 131	151	2 834	20	0
P31	Date de début de fonction	0	0	0	0	–	599	0	0
P32	Date de fin de fonction	82	11 489	24	0	–	9 141	1 284	0
P19	Motif de fin de fonction	144	8 789	42	2	–	9 425	2 402	0
Total		230	272 670	1 916	1 133	151	45 540	6 897	0

Table 3. Nombre de valeurs manquantes *inexplicables* par table de donnée et par champ. Les valeurs dont l'absence peut être justifiée ne sont pas décomptées. Les champs ne comportant, pour l'ensemble des tables, aucune valeur manquante inexplicable, sont omis. Les tirets (–) signalent les cas où un champ donné est complètement absent d'une table. La colonne *Pb* contient le code du problème correspondant décrit dans le texte.

`LE GENTIL`, ou encore `DOLMAZON` pour `DOLAMAZON`.

La même remarque peut être faite relativement aux **prénoms** (P6), pour lesquels on rencontre des coquilles comme `Rémi` pour `Rémy SENNEVILLE` ou `Frédérique` pour `Frédéric BUISSON` ; des prénoms incomplets comme `Jean` pour `Jean-Clément CASSAN` ; ou de complètes erreurs comme `Jean` pour `Louis CRUSOL`. Certaines des coquilles peuvent être détectées en comparant les attributs des différentes occurrences de la même personne. Il faut souligner que pour certaines de ces personnes apparaissant sous plusieurs noms distincts, on rencontre dans le RNE autant d'identifiants que de noms, ce qui pose d'importants problèmes discutés plus loin (cf. Section 2.3.2).

2.1.2 Toponymes

Comme pour les noms de personnes, les problèmes relatifs aux noms de lieux touchent à la variabilité de leur forme, à leur complétude, et à leur unicité.

Variabilité. On observe pour les noms de lieux une certaine variabilité, en particulier typographique. Certains des problèmes détectés sont similaires à ceux déjà mentionnés pour les noms de personnes, d'autres sont spécifiques aux noms de lieux, voire à certains types de lieux.

La **casse** varie (**P7**) entre l'utilisation d'une capitale initiale et uniquement des capitales. Par exemple, le canton `Pléneuf-Val-André` apparaît également sous la forme `PLENEUF-VAL-ANDRE`).

À l'instar des anthroponymes, les **signes diacritiques** ne sont pas utilisés de façon systématique pour les toponymes (P1). Comme illustré avec le canton donné en exemple ci-dessus (accents), ils ont tendance à être absents des noms typographiés en capitales. Mais cela n'est pas toujours vrai, par exemple la communauté de communes `CA DU GRAND BESANÇON` a droit à une cédille.

Les noms de lieux présentent les mêmes problèmes d'utilisation irrégulière de la **ponctuation** que les anthroponymes (P3). Par exemple, des espaces sont utilisés comme séparateurs dans certains noms de canton comme `AMIENS 6 SUD`, alors que pour `Amiens-2` ou `CARMAUX-NORD` il s'agit de tirets. On peut aussi citer l'utilisation sporadique de guillemets comme dans l'EPCI `CC "TERRES TOULOISES"`.

Dans les noms de cantons, les **numéros** sont exprimés aussi bien sous forme de nombres arabo-indiens (ex. `Bastia-2`) que romains (ex. `LYON III`), ce qui constitue une autre irrégularité (**P8**).

Les noms de communes et d'EPCI correspondant à des entités disparues sont respectivement suffixés par les mentions `archivée` et `archivé` (**P9**), par exemple : `CHALARONNE CENTRE (archivé)` (`EPCI`) et `Le Poizat (archivée)` (`CM`, `M`). D'une part, un champ séparé semblerait plus adapté au stockage de cette information. D'autre part, la forme exacte de cette mention varie, on la trouve même parfois présente plusieurs fois dans le même nom (ex. `CA SAINT LO AGGLO - archivé (archivé)`). On peut également constater l'utilisation de plusieurs espaces consécutives, comme déjà relevé pour les anthroponymes (P2).

Certains noms de lieux apparaissent avec ou sans **article** (**P10**), comme par exemple le canton de `La côte vermeille`, qui est aussi désigné par `Côte vermeille`) (tout court). Il s'agit du *même* lieu désigné par deux noms *distincts*. Mais par ailleurs, certains lieux *distincts* sont désignés par le même nom avec ou sans l'article, en particulier les communes. Par exemple, `LE MAGNY` (code INSEE `36109`), `LES MAGNY` (`70317`) et `MAGNY` (`02433`) sont trois communes différentes.

L'adjectif **Saint** (**P11**) apparaît parfois écrit en entier et d'autres fois dans sa forme abrégée *St*. Par exemple, les communes de `Saint-Médard` et `Auge St Medard` (qui illustrent en outre une nouvelle fois l'utilisation inconsistante du tiret).

Complétude. Comme indiqué dans la Table 3, le nom de la commune n'est pas indiqué (**P12**) dans 5 527 lignes de `EPCI`, ce qui représente 4% des lignes de cette table de données. Par exemple, le nom de la commune de rattachement de *Yves DELARUELLE* pour son mandat de *15/04/2014–11/03/2016* n'est pas indiqué (ni son code INSEE, ni son département). Les conseiller·es d'EPCI étant par ailleurs conseiller·es municipal·es, l'information manquante peut théoriquement être retrouvée dans *CM*. Cependant, il faut préciser que la correspondance n'est pas parfaite (cf. Section 2.4.5).

Unicité. L'existence d'une relation bi-univoque entre les lieux et leurs noms dépend du type de lieu considéré. L'absence d'une telle relation n'est pas un problème en soi, mais nous la discutons ici car c'est une caractéristique importante des données.

Ainsi, c'est le cas pour les **circonscriptions européennes**, qui n'ont pour l'instant jamais été modifiées. Certains scrutins ont adopté une circonscription unique au niveau national, d'autres ont utilisé une division supra-régionale donnée, sans qu'il y ait de conflit de nom entre les deux.

C'est également le cas des **régions** en théorie, et ce malgré leur réforme de 2016, car les noms des nouvelles régions et de celles qui ont été modifiées diffèrent des noms préexistants. Cependant, en pratique, dans le RNE on trouve deux cas de régions portant deux noms distincts : `Centre` est aussi désigné par `Centre-Val de Loire`, et `Pays de la Loire` par `Pays de Loire`. Nous rangeons ce problème dans la catégorie des coquilles (P6).

Les noms des **départements** ne posent pas de problème pour la période décrite dans le RNE (2001–2018), car le dernier changement de nom de département a eu lieu en 1990 (les *Côtes-du-Nord* sont devenues les *Côtes-d'Armor*). Mais le but de ce projet est d'intégrer des informations plus anciennes, et concrètement cela va se produire dès le recoupement avec les sources secondaires (cf. Section 3). La relation biunivoque entre les départements et leurs noms ne sera plus garantie à ce moment-là. Cela ne serait pas un problème si ce n'était pas également le cas pour les numéros de département (cf. 2.3).

Les noms de **cantons** prennent généralement quatre formes possibles. S'il s'agit d'un canton recouvrant exactement une commune, il porte le même nom qu'elle (ex. `Villenave-d'Ornon`). S'il regroupe plusieurs communes, son nom correspond souvent à son chef-lieu (ex. `Varennes-sur-Allier`). S'il recouvre une partie d'une commune, il peut être numéroté (ex. `Dijon-1`), associé à un point cardinal (ex. `Besançon-Est`), ou à un quartier (ex. `Marseille-Belsunce`). Enfin, Saint-Pierre-et-Miquelon constitue un cas à part, puisque le canton unique

correspondant porte le nom explicite de `Canton fictif`. L'identification d'un canton sur la base de son seul nom est sujette à caution, car celui-ci ne reflète pas l'évolution de ses frontières dans le temps. Ceci est particulièrement vrai pour les cantons correspondant à une partie de commune ou d'agglomération. Or, les cantons ont été largement redécoupés au cours de la réforme des conseils généraux de 2015, et il paraît difficile de déterminer si un même nom correspond au même territoire avant et après la réforme.

Les **circonscriptions législatives** portent quasiment toutes un nom du type `ième circonscription`, où `i` est le numéro de la circonscription. La seule exception est la collectivité de Saint-Pierre-et-Miquelon, qui correspond à un canton unique parfois appelé dans le RNE `1ère circonscription` et parfois simplement `Saint-Pierre-et-Miquelon`. Puisque ces numéros sont relatifs au département concerné, un grand nombre de circonscriptions possède exactement le même nom. On peut faire la même remarque que pour les cantons : ayant été redécoupés en 2010, leur nom n'est pas un bon indicateur pour identifier un territoire et le suivre dans le temps.

Les **EPCI** évoluent parfois au cours du temps en étant rejoints ou quittés par des communes, et il arrive que leur nom soit modifié pour refléter ces changements, comme par exemple la *Communauté de communes du pays de Questembert* qui devient *Questembert Communauté* en 2015. On remarque que 27 EPCI portent exactement les mêmes noms. Il s'agit généralement de noms génériques (ex. `CC DES DEUX VALLEES`) ou faisant référence à des attributs géographiques étendus sur plusieurs départements (ex. `CC DU VAL DE MEURTHE`). Ce type de confusion touche le plus souvent des paires d'EPCI, avec quelques exceptions : par exemple, 5 EPCI bien distincts, appartenant aux départements `14`, `22`, `28`, `51` et `76`, sont nommés `CC DES TROIS RIVIERES`. Tout ceci empêche là encore de suivre l'évolution du territoire en se basant simplement sur le nom.

On trouve aussi certaines variations de noms pour une même **commune**, par exemple `Isigny-le-Buat` (Manche, code INSEE `256`) apparaît également sous les noms `ISIGNY LE BUAT Section 01` et `ISIGNY LE BUAT Section 03`. Ce type d'erreur est cependant très ponctuel pour les communes, et nous rangeons ce problème dans la catégorie des coquilles (P6). Le problème le plus courant pour l'identification des communes est principalement inverse : il existe de nombreuses communes *homonymes* (ex. `Saint-Cyprien` en *Corrèze*, *Dordogne*, *Loire* et *Pyrénées Orientales*). Il arrive également que le nom d'une commune reste le même après une fusion avec une ou plusieurs autres communes (ex. `Magnieu`, Ain, code INSEE `227`). Par conséquent, le seul nom d'une commune ne peut pas être utilisé pour l'identifier de façon unique.

2.1.3 Labels

Nous appelons *labels* les noms donnés à des catégories, et prises par 7 champs du RNE : sexe, nuance politique, nom de profession, motifs de fin de mandat et de fin de fonction, nature du mandat et de la fonction. Au contraire des noms discutés en Sections 2.1.1 et 2.1.2, nous ne détectons aucune variabilité dans les données (chaque label est toujours exactement le même) et aucun problème d'unicité.

Signification. Certains labels expriment **explicitement** les différentes catégories qu'ils représentent. C'est le cas des libellés de **fonction** (ex. `Quatrième vice-président du conseil départemental`), de mandat (ex. `Conseiller municipal`), et de profession (ex. `Entrepreneurs en batiments`). Les libellés de fonction sont listés de façon exhaustive en annexe, dans la Table 30, et ceux de professions dans la Table 32.

Comme indiqué dans la Table 2, le libellé de **mandat** n'apparaît pas dans `CM` et `EPCI`, ce qui n'est pas gênant car il peut être inféré, puisque chaque table de données se concentre sur un seul type de mandat bien spécifique. Pour les tables dans lesquelles il est présent, ce champ ne peut prendre bien sûr qu'une seule valeur : `Conseiller`

`départemental` (`CD`[13]), `Conseiller régional` (`CR`), `Député` (`D`), `Représentant au Parlement Européen` (`DE`), `Conseiller municipal` (`M`), et `Sénateur` (`S`). Les noms de fonctions sont toujours au masculin, indépendemment du genre de l'élu·e, probablement pour des raisons techniques.

Les autres valeurs sont décrites par des **abréviations**, et leur sens est donc parfois moins évident, d'autant plus que la base n'est pas documentée. Les valeurs utilisées pour le **sexe** sont assez simple à interpréter (`F` pour *féminin* et `M` pour *masculin*). À noter qu'on trouve dans `CM` un cas de code sexe erroné (`Frédéric BUISSON`), que nous considérons comme une coquille (P6). Celles utilisée pour les **motifs de fin** sont les mêmes pour les mandats et les fonctions (`AU`, `DC`, `DO`, `DV`, `FM`). La signification estimée de ces abréviations est indiquée de façon exhaustive en annexe, dans la Table 33.

Pour les **nuances politiques**, les valeurs correspondent la plupart du temps aux sigles courants des partis politiques, mais on trouve également des désignations plus génériques regroupant plusieurs partis jugés proches. La signification estimée est là aussi placée en annexe, dans la Table 31. Nous avons détecté trois problèmes mineurs dans le codage des nuances politiques. Premièrement, il existe un symbole `NC` (pour *Non-communiqué*) correspondant à une nuance inconnue (**P13**). Or, on rencontre également des cas de valeurs manquantes, correspondant à une autre façon de représenter la même situation (une nuance politique inconnue), ce qui est au mieux superflu et au pire ambigu. Deuxièmement, deux nuances sont représentées par deux valeurs au lieu d'une seule (**P14**). Il s'agit du *Parti Radical de Gauche* (`PRG` et `RDG`) et de la *Majorité présidentielle* (`M-NC` et `MAJ`). Troisièmement, ces deux derniers labels, qui représentent la majorité présidentielle, présentent la particularité d'être contextuels (**P15**) (leur signification dépend de la situation politique courante), alors que les autres nuances sont absolues (communiste, socialiste, etc.). Enfin, il nous faut indiquer que nous n'avons pas pu déterminer la signification du sigle `PREP`.

Complétude. En théorie, la **nuance politique** devrait être renseignée dans chaque ligne, mais c'est loin d'être le cas en pratique. La nuance est absente (**P16**) de près de 20% des lignes de `CM`, 7% d'`EPCI` et 3% de `M`. Les trois mêmes tables concentrent les manques relatifs au **libellé de profession** (**P17**), avec en plus 2 cas détectés dans `CD`. Remarquons qu'une absence de libellé de profession correspond systématiquement à une absence de code de profession, donc il n'est pas possible d'utiliser l'un pour compléter l'autre.

Le **motif de fin de mandat** n'a pas à être obligatoirement renseigné, puisqu'il est possible que des mandats ne soient pas encore terminés. Nous décomptons donc une valeur manquante seulement si elle est associée à une date de fin de mandat (**P18**). Toutes les tables sont touchées sauf `CD` et `S` ; on obtient (par importance proportionnelle décroissante) : `DE` (60%), `D` (45%), `CR` (31%), `EPCI` (2%), et moins de 1% pour `CM` et `M`.

Le principe est le même pour le **motif de fin de fonction**, à la différence du fait qu'il existe d'autres facteurs à prendre en compte. Nous décomptons une valeur manquante si elle est associée à une date de fin de fonction, ou si des indices indiquent que la fonction existent et que le mandat est fini (car une fonction ne peut pas dépasser la durée d'un mandat). Dans le cas présent, on considère qu'une fonction existe si sa date de début ou son nom est renseigné, et qu'un mandat est terminé si sa date de fin ou son motif de fin son renseignés. Sous ces conditions, on observe que toutes les tables sont touchées par ce problème (**P19**) sauf `S` et `DE` (mais cette dernière ne contient pas d'information relative aux fonctions). En les classant par ordre d'importance proportionnelle décroissante cette fois aussi, on a : `EPCI` (7%), `M` (2%), `CD`, `CM`, `CR` (1% pour toutes les trois), tandis que `D` a moins de 1% de valeurs manquantes inexplicables.

Le code sexe est toujours renseigné, mais il est parfois manifestement faux, comme par exemple pour la sénatrice `Marie-Annick DUCHENE`, qui est décrite par le label `M` alors qu'il

13. Notons l'absence du mandat de `Conseiller général`, alors que la Table 30 montre l'existence de la fonction `Président de Conseil Général`.

s'agit d'une femme d'après le site du Sénat [14].

Précision. Un autre problème est l'hétérogénéité observée dans le degré de précision utilisé pour renseigner certains labels. Considérons d'abord les **libellés de fonction**. Une assemblée comporte un certain nombre de fonctions standard (un président, des vice-présidents, des questeurs, etc.). Cependant, leur traitement dans le RNE n'est pas systématique, et on observe une grande variabilité non seulement en fonction du *type* de mandat, mais aussi du *territoire* et de l'*époque* (**P20**). Les vice-présidents, secrétaires et questeurs ne sont pas toujours numérotés (*Premier secrétaire*, *Deuxième secrétaire*, etc.), auquel cas on ne peut pas les différencier, ce qui pose des problèmes pour la vérification et l'exploitation des données. Le même problème se pose pour les présidents de groupes et de commissions, qui ne sont pas distingués en fonction du groupe ou de la commission concernés. Certaines fonctions ne sont pas du tout renseignées pour certains types de mandats, par exemple `S` et `D` n'incluent que la fonction de président (**P21**). Dans `CD` et `CR`, le niveau de précision dépend du département ou de la région considérés.

Les **motifs de fin** (de mandat comme de fonction) sont également traités de façon hétérogène, dans le sens où la sémantique théoriquement associée aux différents labels (cf. 33) est, en pratique, variable (**P22**). Concrètement, nous avons échantillonné les données pour chaque label possible et avons procédé à des vérifications manuelles. Nous nous sommes concentrés sur les député·es, qui ont tous·tes une page Wikipédia et dont le parcours est donc facilement vérifiable. Les labels `FM` et `DC`, qui signifient respectivement *fin de mandat normal*, et *décès*, semblent utilisés correctement de façon systématique. Une *démission d'office* correspond théoriquement à une décision de justice, ce qui est parfois le cas dans le RNE quand le label `DO` est utilisé. Mais on observe également des cas où ce label est associé à des démissions volontaires, comme par exemple `Thierry BRAILLARD` (député ayant obtenu un poste au gouvernement), `Charles-Ange GINESY` (suppléant rendant son siège à `Christian ESTROSI`), ou `François BAROIN` (démission pour éviter un cumul). Le label `DV` signifie *démission volontaire* et semble utilisé correctement. Nous avons répertorié des élu·es partant au gouvernement (ex. `Carole DELGA`), nommé·es dans l'administration (ex. `Didier MIGAUD`, cours des comptes), ou bien préférant occuper un autre poste électif (ex. `Christian ESTROSI`, mairie de Nice). Enfin, le label `AU` pour *autres* est un peu fourre-tout et empiète manifestement sur les autres labels, puisqu'on le trouve associé lui aussi à des départs au gouvernement (ex. `Xavier BERTRAND`), des suppléant·es rendant leur siège (ex. `Pascale GRUNY` pour le même `Xavier BERTRAND`), et des élu·es préférent un autre mandat (ex. `Jean ROATTA`, `DE` plutôt que `D`). Pour expliquer cette hétérogénéité, nous supposons que les nombreux opérateurs humains qui saisissent les données n'interprètent pas les labels tous de la même façon.

2.2 Dates

Trois types de dates apparaissent dans le RNE : les dates de naissances, qui sont individuelles et supposées identiques pour chaque occurrence d'un·e élu·e (Section 2.2.1) ; les dates de mandats (début et fin), dont celle de début est obligatoire (Section 2.2.2) ; et les dates de fonction (début et fin également), qui sont optionnelles et dépendent du fait que l'élu·e ait occupé ou pas une fonction spécifique au cours de son mandat (Section 2.2.3).

2.2.1 Dates de naissance

Complétude. Nous considérons que la date de naissance est obligatoire à chaque apparition d'un·e élu·e dans le RNE. La Table 3 montre qu'elle n'est pas renseignée dans 10

14. https://www.senat.fr/senateur/duchene_marie_annick11091r.html

lignes, pour l'ensemble du RNE (**P23**). Par exemple, le maire `Claude LAGACHE` n'a pas de date de naissance dans le RNE.

Remarquons aussi que la date `01/01/1900`, qui apparaît en tant que date de naissance à 119 reprises toutes tables confondues, semble être utilisée pour dénoter une absence d'information sur ce champ. Nous n'avons cependant pas pu vérifier cette hypothèse systématiquement, faute de source alternative nous indiquant la date réelle, les cas apparaissant essentiellement dans `CM` et `EPCI`.

Incohérences. La Table 4 résume les résultats des différents tests de validité et de cohérence effectués sur les dates. Nous avons identifié un total de 32 dates de naissance incorrectes (**P24**), en nous basant sur des comparaisons avec les autres dates du RNE et lors de l'intégration des sources secondaires. Les tests réalisés incluent notamment la détection de dates trop précoces (antérieures au 01/01/1900), trop tardives (postérieures à la date d'extraction), ou situées après le début du mandat ou de la fonction. Nous avons également considéré que quand la date de naissance d'un même individu varie au gré des occurrences de celui-ci, une seule d'entre elles est correcte (comme pour les autres informations personnelles telles que le nom de famille, le prénom, et le sexe). On peut par exemple reprendre le cas déjà cité du maire `Robert BLANCHET`, dont la date de naissance est parfois le `22/11/1939` et parfois le `21/11/1939`.

Pb	Description	CD	CM	CR	D	DE	EPCI	M	S
P24	Incohérences de date de naissance	0	5	0	11	0	7	3	7
P26	Incohérences de fin de mandat	0	3	0	0	0	0	1	0
P27	Incohérences de bornes de mandat	81	79	2	0	0	88	3	4
P28	Micro-mandats	158	3 962	20	30	1	1 287	28	7
P29	Mandats se recouvrant	803	–	204	18	82	–	–	65
P30	Mandats couvrant >1 élections	6 189	1 804	1 102	553	167	2 192	1 103	337
P33	Incohérences de début de fonction	0	17	0	0	0	7	12	0
P34	Incohérences de fin de fonction	0	20	0	0	0	3	20	0
P35	Incohérences de bornes de fonction	1	50	0	0	–	4	47	0
P36	Micro-fonctions	1	78	0	0	–	1	67	0
P37	Fonctions se recouvrant	84	10 360	15	0	–	127	2 551	1
P38	Fonctions hors-mandat	56	2 560	0	0	–	1 287	1 758	0
Total		7 373	18 938	1 343	612	250	2 811	5 593	421

Table 4. Cas d'incohérences relatifs aux dates, pour chaque table de donnée. Comme précédemment, les tests qui n'ont abouti à la détection d'aucune erreur sont omis. Les tirets (–) signalent des tests qu'il n'était pas possible de réaliser (cf. le texte pour plus de détails). La colonne *Pb* contient le code du problème correspondant décrit dans le texte.

2.2.2 Dates de mandats

Complétude. Comme nous l'avons déjà mentionné, chaque ligne de l'extraction est supposée représenter un mandat, donc elle doit contenir au moins une **date de début** de mandat. Toute ligne sans cette date peut être considérée comme incomplète (**P25**). La Table 3 montre que cela ne se produit que dans EPCI, à hauteur d'environ 1% des lignes, qui sont bien entendu *inexploitables*.

Pour ce qui est de la **fin de mandat**, la date est optionnelle car certains mandats sont en cours au moment de l'extraction. En d'autres termes, nous considérons une date de fin absente comme l'indication que le mandat ne s'est pas terminé avant la date d'extraction. En revanche, nous faisons l'hypothèse que tout motif de fin de mandat doit s'accompagner d'une date de fin de mandat. La Table 3 révèle que cette erreur n'apparaît pas dans les données.

Compatibilité. Nous avons appliqué aux dates de mandats des tests portant sur leur **précocité** ou **retard** excessif, similaires à ceux déjà décrits pour les dates de naissance (dates pré-01/01/1900, post-extraction). La Table 4 montre que seules 4 dates de fin de mandat sont concernées (**P26**), et aucune date de début.

Un autre test porte sur la cohérence des deux **bornes** du mandat, quand les deux sont renseignées : nous vérifions que la date de début est bien antérieure ou égale à celle de fin (**P27**). Ce n'est pas le cas pour un total de 257 lignes, et cette fois cela constitue une erreur sérieuse. Par exemple, d'après le RNE, la sénatrice `Frédérique ESPAGNAC` a effectué un mandat du `25/09/2011` au `24/09/2011`.

Enfin, il nous faut faire une dernière remarque sur les dates de début et de fin de mandats. On s'attendrait à ce que, pour les mandats standard commençant à une élection donnée et s'achevant à la suivante (par opposition aux mandats plus courts à la suite d'une démission, par exemple), les dates de début et de fin soient **alignées** soit sur ces dates d'élections officielles, soit sur les dates effectives de prise de fonction. Cependant, cela n'est pas fait systématiquement dans le RNE, et les deux approches coexistent. Nous supposons que c'est cette irrégularité qui cause certains problèmes discutés plus loin, et relatifs à l'occupation simultanée d'une position donnée par plusieurs personnes (cf. les *recouvrements*, ci-dessous), et au décompte de mandats (cf. Section 2.4.

Micro-mandats. Nous calculons également la durée de chaque mandat afin de la tester de différentes façons. Tout d'abord, nous avons remarqué l'existence d'un très grand nombre de mandats extrêmement courts dans les données, à tel point qu'ils en sont suspects, et que nous appelons **micro-mandats** (**P28**). Concrètement, nous avons fixé la limite pour leur détection à une semaine ou moins. Nous avons choisi cette durée sur la base d'un examen manuel de la distribution des durées de mandats dans le RNE. On dénombre un nombre conséquent de micro-mandats, notamment dans `CM` et `EPCI`, qui ne représentent cependant qu'une faible proportion des tables, de l'ordre de 1%.

En soi, un micro-mandat n'est pas forcément synonyme d'erreur. Il est possible qu'un mandat si court ait effectivement eu lieu, par exemple à la suite du décès d'un·e élu·e, ou de sa nomination à un poste gouvernemental. Pour déterminer ce qu'il en est, nous avons effectué une vérification manuelle d'un échantillon des cas détectés (vu le nombre, il était impossible en pratique de procéder systématiquement). D'après ces vérifications, il semblerait que dans l'écrasante majorité des cas, un micro-mandat corresponde à une erreur du RNE. Nous avons identifié trois types d'erreurs.

Premièrement, il peut s'agit d'un mandat qui n'*existe pas* en réalité, comme par exemple dans `D` le mandat de `Virginie GALASSO` (`23/03/2014–23/03/2014`, soit une journée) pour la 1ère circonscription du `Jura`, alors qu'elle n'a jamais été députée et que c'était `Jacques PELISSARD` qui occupait ce poste à cette date (comme décrit dans le RNE, par ailleurs). Deuxièmement, le micro-mandat peut être un *doublon* d'un autre mandat déjà décrit dans une autre ligne, mais avec des dates correctes. C'est le cas, par exemple, du conseiller départemental `Michel THIEN`, pour lequel coexistent dans le RNE un mandat d'une journée (`02/04/2015–02/04/2015`) et un autre un peu plus vraisemblable dont la durée dépasse le mois (`02/04/2015–04/05/2015`), et surtout contient temporellement le premier (qui est donc redondant). Troisièmement, le micro-mandat peut se substituer à un mandat complètement différent. Par exemple, le RNE indique deux mandats de députée pour `Nadine MORANO` : `16/06/2002–09/06/2007` et `15/05/2012–19/06/2012`. Le premier est correct, tandis que le second est un micro-mandat (bien qu'il dépasse la limite fixée plus haut, il reste anormalement court). Or, `Nadine MORANO` a bien effectué un deuxième mandat, mais sur la période 2007-2008 : elle n'était pas députée en 2012.

En plus de ces trois types d'erreurs, certains micro-mandats semblent reliés à des *absences* de mandats. Prenons par exemple le cas du député *Xavier BERTRAND*, pour lequel le RNE décrit trois mandats : `16/02/2009–15/12/2010`, `15/05/2012–19/06/2012` et

`20/06/2012–12/1/2016`. Les premier et troisième sont corrects, tandis que le deuxième est un micro-mandat sans réalité. Par contre, ce député a en réalité effectué deux autres mandats : l'un en 2002-2004 et l'autre en 2007. Aucun des deux ne sont présents dans le RNE, et ce bien qu'ils se soient déroulés après 2001. Ce type de problème touche un grand nombre de personnalités politiques de premier plan (`Pierre MOSCOVICI`, `Marisol TOURAINE`, `Geneviève FIORASO`, `Bernard CAZENEUVE`, `Benoît HAMON`…) qui ont tou·tes en commun des passages au gouvernement. On peut donc supposer que le problème sous-jacent est relatif à la gestion de mandats incomplets, que ce soit dans le RNE ou lors de l'extraction.

Recouvrements. Un autre test relatif à la durée est celui de **recouvrement** (P29), qui consiste à vérifier qu'une position élective bien spécifique, comme par exemple *Député de la 1ère circonscription du Vaucluse*, n'est jamais occupée par plusieurs personnes simultanément. L'opération ne peut pas être réalisée pour toutes les tables de données, car l'identification unique d'une position n'est possible que pour certaines d'entre elles.

Il existe ainsi une relation biunivoque entre un·e conseiller·e général·e et son canton (au moins jusqu'en 2015), et entre un·e député·e et sa circonscription. Pour les conseiller·es régional·es, sénateur·rices, et député·es européen·nes, on connaît seulement le *nombre* d'élu·es par *circonscription* (cf. les Tables 36, 37 et 38 en annexe). C'est également le cas pour les conseiller·es départemental·es à partir de 2015, car chaque canton est représenté par un binôme homme/femme. Ceci est insuffisant pour déterminer qu'une position spécifique est occupée par plusieurs personnes à la fois, mais permet au moins de calculer le nombre de mandats *en excès* par rapport au nombre qu'on devrait observer à un moment donné, pour une circonscription donnée. En revanche, le nombre de conseiller·es municipal·es et d'EPCI est très variable d'une institution à l'autre, et dans le temps également, aussi n'a-t-il pas été possible pour des raisons pratiques de réaliser ce test sur `CM`, `EPCI` et `M`.

Les valeurs indiquées dans la Table 4 montrent un fort recouvrement de mandats dans `CD`. Il faut souligner que le recouvrement entre deux mandats concernant la même position n'est pas forcément complet : il peut s'agit parfois de quelques jours à peine. Indépendemment de la durée de recouvrement, il s'agit tout de même d'une erreur, qui peut être notamment causée par le non-alignement des dates sur les élections, comme nous l'avons mentionné précédemment.

Citons par exemple `Kader ARIF` et `Emilienne POUMIROL`, deux cas dont le RNE indique qu'ils ont tous les deux été député de la 10ème circonscription de `Haute-Garonne` sur les périodes respectives du `20/6/2012–22/7/2012` et du `20/6/2012–24/12/2014`. Pour les positions qui ne sont identifiables de façon unique, on peut donner en exemple le conseil régional de `Midi-Pyrénées` qui, d'après le RNE, possédait 92 conseiller·es au `20/07/2007`, alors que le nombre réglementaire à cette date n'était que 91 sièges.

Élections. Le dernier test vise la cohérence relativement aux **élections** (P30). En effet, nous partons du principe que la période couverte par une ligne de la base ne peut contenir qu'*une seule* date d'élection (au plus), qui plus est en *début* de période. Dans le cas contraire, cela signifie que la période indiquée recouvre *plusieurs* mandats consécutifs. Par exemple, une ligne de `CM` couvrant la période 09/03/2008–18/07/2012 est conforme à ce principe, car les élections municipales concernées ont eu lieu les 09/03/2008 et 23/03/2014. Par contre, ce n'est pas vrai de la période 09/03/2008–30/08/2014, puisque celle-ci inclut l'élection de 2008 et celle de 2014. La ligne devrait donc être décomposée en deux pour respecter le principe énoncé ci-dessus.

La détection de ce type de problème n'est pas triviale, et nécessite d'exploiter des ressources extérieures au RNE. Tout d'abord, il est nécessaire de connaître les dates des élections du type de mandat considéré (cf. Tables 34 et Table 35). De plus, certaines institutions (`CD` et `S`) étant renouvelées partiellement, leur traitement requiert d'identifier la *série électorale* concernée (cf. l'Annexe C pour plus de détails). Enfin, le problème de non-alignement des dates de mandats sur les dates d'élection, déjà relevé précédemment, complique encore la

tâche.

Comme indiqué dans la Table 4, ce test révèle un grand nombre de problèmes. Par importance proportionnelle décroissante, on a : `DE` (66%), `CD` (49%), `S` (31%), `D` (22%), `CR` (19%), tandis que `M` et `CM` sont respectivement touchées à hauteur de 1% et moins. Notre interprétation est la suivante : les périodes décrites par ces lignes devraient en réalité faire l'objet de plusieurs lignes séparées (une par mandat réel). Il est donc nécessaire de les décomposer. Le RNE indique par, exemple, que le député européen `Jean-Louis COTTIGNY` a effectué un mandat couvrant la période `13/06/2004`–`30/06/2014`, qui correspond en réalité à deux mandats.

2.2.3 Dates de fonctions

Complétude. Nous avons déjà mentionné que la fonction est optionnelle, dans le sens ou tou·tes les élu·es n'occupent pas forcément de fonction particulière au cours de leurs mandats. Cependant, il est des situations dans lesquelles une date doit apparaître.

La **date de début** doit apparaître si l'existence de la fonction est attestée par ailleurs, que ce soit via le libellé, la date de fin ou le motif de fin. Comme le montre la Table 3, seule `EPCI` est touchée par ce problème (**P31**), qui concerne moins de 1% de ses lignes. On peut citer par exemple `Ambroise CENTONZE SANDRAS`, dont le RNE indique que le mandat commence le `14/01/2017` et est associé à une fonction de vide-président sans date de début.

La **date de fin** de fonction n'est exigée que si l'achèvement de la fonction peut se déduire des autres champs (**P32**). C'est directement le cas si le motif de fin est renseigné. C'est indirectement le cas si la fonction est attestée (voir le paragraphe précédent) et si le mandat qui la contient est terminé, car une fonction doit être (temporellement) contenue dans un mandat. Comme expliqué précédemment, on considère qu'un mandat est terminé si sa date de fin ou son motif de fin sont précisés. Par exemple, le conseiller d'EPCI `Daniel GRAS` a effectué d'après le RNE un mandat du `15/04/2014` au `19/12/2016`, avec une fonction de vice-président commence le `17/04/2014` soit 2 jours après son mandat. Or, aucune date de fin de fonction n'est indiquée, alors que le mandat est terminé. Sous ces conditions, la Table 3 indique que des valeurs manquantes ont été détectées dans toutes les tables de données sauf `D` et `S`, représentant de l'ordre de 7% des lignes pour `EPCI` ; de 1% pour `CD`, `CM` et `M` ; et moins de 1% pour `CR`.

Il faut remarquer que les lignes dont la date de fin n'est pas précisée sont possiblement, mais *pas nécessairement*, les mêmes que celles qui n'ont pas de motif de fin, comme attesté par les valeurs observées par exemple pour `CM` et `EPCI` (plus de motifs manquant que de dates pour l'une, et la réciproque pour l'autre).

Compatibilité. Il existe pour les fonctions certaines incohérences de dates similaires à celles déjà décrites pour les mandats. Le test sur les dates excessivement **précoces** et **tardives** permet d'identifier un problème avec un total de 36 dates de début (**P33**) et 43 dates de fin (**P34**). Il s'agit généralement de coquilles, par exemple `16/03/0201` pour le maire `René DUBOS` (probablement une inversion de caractères à la saisie), et `01/02/3003` pour le maire `Philippe AUPHAN` (sûrement un `3` entré à la place d'un `2`).

Le test sur les **bornes** des fonctions, vérifiant que la date de début est bien antérieure à la date de fin, révèle un total de 102 problèmes (**P35**). Par exemple, le RNE indique que `Françoise ROSSIGNOL` a été maire du `14/03/2008` au `08/03/2008`.

Micro-fonctions. Le test sur les **micro-fonctions** (**P36**), i.e. les fonctions dont la durée ne dépasse pas une semaine, est similaire dans sa nature à celui présenté pour les micro-mandats. On dénombre un total de 147 micro-fonctions. Par exemple, le RNE indique que `Jean CRUSOL` a été maire de `Sainte-Luce` du `16/03/2008` au `16/03/2008`, soit pendant une journée.

Recouvrement. Le principe du test de **recouvrement** (**P37**) est le même que pour les mandats, mais les positions identifiées de façon unique ne sont pas les mêmes quand on considère

les fonctions, et donc les tables de données concernées sont différentes.

Dans `CD`, `CR`, `D`, `EPCI` et `S`, la fonction de président est toujours identifiée de façon unique. Cependant, ce n'est pas toujours le cas de celle de vice-président, qui n'est pas unique. Comme on l'a mentionné en Section 2.1.3, les rangs des vices-présidents sont parfois distingués, et parfois non, ce qui empêche alors la réalisation de ce test. Dans `CM`, la remarque est la même entre les maires et les adjoints, ces derniers n'étant pas systématiquement numérotés. Aucune fonction n'est indiquée dans DE, donc ce test n'y est pas du tout possible.

On observe que les cas de recouvrement de fonction sont beaucoup plus fréquents que pour les mandats, en particulier dans `CM` et `M`, même s'ils ne représentent respectivement au final que 1% et 2% des lignes de ces tables. On peut donner comme exemple `Christian JIMENEZ` et `Pierre BERTHET`, dont le RNE indique qu'il sont tous les deux maires de `Bellai` (Ain) respectivement depuis le `30/03/2014` et le `30/01/2015`, sans date de fin de fonction dans les deux cas. On suppose donc que les deux fonctions sont en cours à la date d'extraction. Le RNE indique aussi que le mandat du premier se termine le `20/01/2015`, donc sa fonction s'achève au plus tard à cette date. Celle-ci est elle-même antérieure au `30/01/2015`, début de la seconde fonction. Le problème vient donc, indirectement, du fait qu'aucune date de fonction n'est renseignée pour la première personne (P32).

Il faut souligner que ce n'est pas la seule cause d'erreur. Par exemple, le RNE indique que `Robert DESPLACE` et `Nadine LAVOCAT-DUBUIS` ont été tous les deux maires de `Genouilleux` (Ain) respectivement pour les périodes `10/09/2010–22/03/2014` et `21/03/2008–22/03/2014`. Il ne s'agit donc pas ici d'un problème de fin de fonction non-renseignée, mais d'une erreur dans l'une des dates de début ou de fin de fonction. Le fait que le mandat de `Nadine LAVOCAT-DUBUIS` se termine le `24/07/2010` nous indique que c'est sa date de fin de fonction qui est incorrecte, car elle ne peut dépasser celle de mandat.

Dépassement. Nous ne connaissons pas les dates réglementaires de début et fin de fonction, et il nous est donc impossible de réaliser un test équivalent à celui fait pour les mandats sur les dates d'élection. Par contre, nous effectuons un test spécifique aux fonctions, qui porte sur leur **inclusion** temporelle dans les mandats (**P38**) : la période couverte par la fonction doit être incluse dans celle couverte par le mandat décrit dans la même ligne.

Les tables `CM`, `M` et `EPCI` sont les plus affectées par cette erreur (dans une proportion de moins de 2%, cependant). On peut citer l'exemple de `Jean ODIN`, pour lequel le RNE indique une fonction de maire de `Genilac` (Loire) couvre la période `14/3/2008–22/3/2014` pour un mandat courant sur `09/03/2008–19/01/2011`. Autrement dit, la fonction se terminerait trois ans après le mandat, ce qui n'est bien sûr pas le cas en réalité. On pourrait faire l'hypothèse que cet élu a réalisé un mandat ultérieur lors duquel sa fonction de maire a été renouvelée, et que cela a été mal encodé dans le RNE. Cependant, ce n'est pas le cas : cet élu n'a pas effectué d'autre mandat ensuite, même pas en tant que simple conseiller municipal.

2.3 Identifiants

Comme indiqué dans la Table 2, certains champs du RNE semblent utilisés comme *identifiants*. On s'attend à observer une correspondance bi-univoque, c'est-à-dire qu'un identifiant donné corresponde à une seule entité donnée, et qu'une entité donnée soit toujours désignée par le même identifiant. Cependant, ce n'est pas systématiquement le cas.

2.3.1 Identifiants de lieux

Dans l'énorme majorité des cas, les identifiants de lieux sont renseignés, aussi cette section se concentre-t-elle plus sur l'association entre un territoire et l'identifiant ou le numéro qui lui est associé dans le RNE. La nature de cette association varie en fonction du type de territoire considéré.

Circonscriptions européennes. Avant les élections de 2004, il n'y avait pas de circonscription européenne en France, ou plus exactement une unique circonscription recouvrant tout le pays. À partir de 2004, 8 circonscriptions ont été créées, chacune correspondant à un regroupement de plusieurs régions. Elles ont été abolies pour les élections de 2019, en faveur d'un retour à la circonscription unique.

L'extraction du RNE que nous utilisons décrit uniquement les élections européennes de 2004 à 2014, c'est-à-dire la période pendant laquelle les circonscriptions étaient utilisées. Il y a donc bien une association biunivoque entre les circonscriptions européennes et leurs numéros, et ce en dépit de la réforme des régions qui a eu lieu en 2016. De plus, comme indiqué par la Table 3, ce code est systématiquement indiqué dans chaque ligne de `DE`.

Régions. Les régions sont identifiées dans le RNE grâce à leur numéro INSEE. Celui-ci prend en compte la réforme de 2016, il est défini de manière à ce que le numéro d'une région qui n'existe plus sous la même forme ne soit plus utilisé, et à ce qu'une région nouvellement apparue soit identifiée par un nouveau numéro. Par exemple, l'*Aquitaine* porte le numéro 72 tandis que la *Nouvelle-Aquitaine* a le 75. Le *Poitou-Charentes* porte le numéro 54, et comme cette région a fusionné avec l'*Aquitaine* et le *Limousin* pour former la *Nouvelle-Aquitaine*, elle a disparu. Cependant, son numéro n'est utilisé par aucune autre nouvelle région.

Il y a donc une relation biunivoque entre les régions et leur numéro, qui peut être utilisé comme identifiant. On peut aussi remarquer dans la Table 3 que ce champ ne fait l'objet d'aucune valeur non-renseignée.

Départements. Comme l'indique la Table 3, le code du département de la commune de rattachement est non-renseigné à 4 481 reprises dans EPCI (**P39**). On trouve dans le lot des codes dont la valeur est `0`, ce que nous avons considéré comme une valeur manquante puisqu'aucun département ne porte ni n'a jamais porté ce numéro. Certains EPCI s'étalent sur plusieurs départements, tout en étant rattachés à un département principal : c'est celui-ci qui est supposé renseigné dans le champ *Code département de l'EPCI*. Cependant, 12 EPCI sont incorrectement décrits, puisqu'un département secondaire est parfois renseigné à la place de ce département principal (**P40**). Cette erreur apparaît à 57 reprises.

Pour ce qui est de la correspondance biunivoque entre les départements et leur numéro, elle existe bien pour peu qu'on considère une date donnée. C'est notamment le cas pour notre extraction du RNE. Cependant, cela n'est pas vrai si on considère une période suffisamment longue, car les départements ont évolué au cours du temps : création, suppression, redécoupage, changements de nom, renumérotation, etc. Ne serait-ce que pendant la V^{ème} République, on peut citer le cas de la réorganisation de la région parisienne en 1964, qui a abouti à la disparition des départements de la *Seine* (numéro 75) et la *Seine-et-Oise* (78), et à la création de ceux de *Paris* (75), des *Yvelines* (78), de l'*Essonne* (91), des *Hauts-de-Seine* (92), de la *Seine-Saint-Denis* (93), du *Val-de-Marne* (94) et du *Val-d'Oise* (95). Ces mêmes numéros correspondaient, avant 1962, aux départements d'*Alger* (91), d'*Oran* (92), de *Constantine* (93), et des *Territoires du Sud* (94). La base du Sénat, que nous utilisons comme source secondaire (cf. Section 3), couvre une période remontant jusqu'à certains de ces changements, et il est donc nécessaire d'en tenir compte lors de l'élaboration de notre propre base, en n'utilisant pas le numéro de département comme identifiant (**P41**, cf. aussi Section 3.1.2).

Circonscriptions législatives. La numérotation des circonscriptions législatives est relative à leur département. De plus, comme nous l'avons déjà mentionné, le nom de la circonscription reflète son numéro (`1ère circonscription`, `2ème circonscription`, etc.). Contrairement à d'autres entités comme les départements, ces deux informations sont donc finalement équivalentes.

Les circonscriptions législatives sont relativement stables dans le temps, mais sont tout de même susceptibles de disparaître ou d'être fusionnées, sans que cela soit reflété dans leur numérotation (ni dans leur nom, par conséquent). En conclusion, le couple département–

numéro permet d'identifier une circonscription législative de façon unique dans le pays, mais son utilisation dans une étude longitudinale revient à faire l'hypothèse que ces circonscriptions sont stables. Or, le dernier redécoupage électoral important date de 2009, et il est par conséquent couvert par le RNE. En conclusion, il n'y a pas de relation biunivoque entre les circonscriptions et leurs numéros ou leurs noms (**P42**).

Cantons. À l'instar des circonscriptions législatives, les cantons sont numérotés relativement à leur département. À leur différence, ils portent généralement des noms beaucoup plus explicites. Ceci permet un meilleur suivi de leur évolution dans le temps, et révèle que la numérotation associée est très *ad hoc*. De nombreux cantons changent de numéro au cours du temps, par exemple le canton de `Anzin` dans le département du `Nord` porte le numéro `71` avant 2015 et le `3` après. Les numéros assignés aux cantons dans le RNE, même en les associant au département, ne sont pas suffisants pour les identifier de façon longitudinale.

De plus, il faut remarquer que les cantons ont subi un redécoupage important lors de la réforme des conseils généraux de 2015, un évènement couvert par le RNE. Par conséquent, on peut en conclure qu'il n'y a pas de relation biunivoque entre la réalité géographique et les codes ni même les noms des cantons (**P43**).

EPCI. Les EPCI sont identifiés au moyen d'un numéro SIREN, qui doit respecter une syntaxe bien définie. Cela n'est pas toujours le cas, car on relève des numéros trop courts. Par exemple, la `CC DU VAL DE VOGE` porte le numéro `200006393` mais est parfois désignée sous le numéro `20006393` (un zéro manquant). Incidemment, cette observation laisse supposer que ce numéro n'est pas stocké de manière centralisée dans la base de données interne du RNE.

Le numéro SIREN est supposémment unique et gardé par l'institution tant qu'elle existe. On pourrait donc *a priori* penser qu'un tel code peut être utilisé comme identifiant. Cependant, les données du RNE montrent que certains EPCI utilisent successivement plusieurs numéros SIREN distincts, sans pour autant forcément changer de nom. Par exemple, la `CC Pays de Sault` porte successivement les numéros `241100643` [15] et `248400145` [16]. Certains de ces cas correspondent à des erreurs pures et simples, l'un des numéros étant en réalité assigné à une autre sutructure, tandis que pour d'autres les registres en ligne attestent bien de la réalité de cette multiplicité. On peut supposer que dans ce cas, chaque nouveau numéro correspond à un changement important dans la nature de l'EPCI. Indepéndamment de la cause, il faut souligner que cela empêche d'identifier l'EPCI dans le temps. Il n'y a donc pas de relation biunivoque entre les EPCI et leur numéro SIREN (**P44**). Un traitement supplémentaire important, exploitant des ressources extérieures telles que la BANATIC [17] (Base nationale sur l'intercommunalité) ou la liste de `collectivites-locales.gouv.fr` [18] serait nécessaire pour identifier de façon unique les EPCI.

Communes. Le numéro de la commune de rattachement d'un·e élu·e d'`EPCI` est manquant dans 3 293 lignes (**P45**), souvent les mêmes (mais pas uniquement) que celles ne précisant pas le numéro du département de rattachement.

Chaque commune est associée à un numéro INSEE à 3 chiffres, qui est défini relativement au département. En lui adjoignant le numéro du département, on peut donc identifier une commune de façon unique dans tout le pays. Les communes sont susceptibles d'évoluer : certaines fusionnent (ex. *Bois-Guillaume* et *Bihorel* fusionnent en 2012), un processus généralement encouragé par l'état, tandis que d'autres au contraire se séparent (ex. les mêmes communes se séparent en 2014). Il peut aussi arriver qu'une commune soit créée *ex nihilo*, par exemple *Val-de-Reuil*, créée en 1981 à partir de parcelles issues d'un ensemble de communes contigües. Cependant, le numéro INSEE ne reflète pas forcément ces différentes évolutions.

15. https://www.verif.com/societe/COMMUNAUTE-COMMUNES-PAYS-DE-SAULT-241100643/
16. https://www.verif.com/societe/COMMUNAUTE-COMMUNES-PAYS-DE-SAULT-248400145/
17. https://www.data.gouv.fr/fr/datasets/base-nationale-sur-linter-communalite/
18. https://www.collectivites-locales.gouv.fr/liste-et-composition-des-epci-a-fiscalite-propre

Lors d'une fusion, la commune nouvelle porte le même numéro que la principale commune fusionnée[19]. Parfois, elle porte aussi le même nom, comme par exemple `Carentoir` (département `56`, code INSEE `033`), qui résulte de la fusion des anciennes communes de `Carentoir` et de `Quelneuc` en 2017. Les deux entités (ancienne et nouvelle communes de *Carentoir*) sont donc *indiscernables* dans le RNE. Cette conservation de code se produit même quand la commune nouvelle porte un nouveau nom. Par exemple, les communes aindinoises d'*Arbignieu* (`015`) et de *Saint-Bois* (`340`) ont fusionné en 2015 pour donner naissance à la commune nouvelle d'*Arboys en Bugey* (`015`). Ces observations suffisent à établir que la relation entre les communes et leur code INSEE n'est pas bi-univoque, car un même numéro peut désigner deux entités différentes (au moins en partie).

De plus, on peut observer que le RNE ne représente pas lui-même ce type de changement (**P46**). Ainsi, des trois communes mentionnées dans notre exemple aindinois, seules les deux dernières apparaissent dans la table `CM` du RNE, qui couvre pourtant une période suffisamment étendue pour contenir les trois communes, en théorie. *Arboys en Bugey* apparaît car elle est la plus récemment créée, *Saint-Bois* en raison de son code INSEE différent du sien, tandis qu'*Arbignieu* est écrasée par *Arboys en Bugey* à cause de son code identique. Preuve en est les mandats du conseiller municipal `Daniel GIRARDET`, qui ont lieu avant 2015, à une époque où *Arboys en Bugey* n'existait pas encore, mais qui sont quand même rattachés à cette commune (au lieu d'*Arbignieu*).

Nous faisons l'hypothèse que la structure de la base interne du RNE ne modélise pas l'évolution des communes. Autrement dit, une commune est représentée par une unique entrée, qui quand elle est modifiée affecte toutes les occurrences de la commune dans la base. Ceci transparaît également quand on considère la population des communes, qui est toujours exactement la même quelle que soit la date du mandat considéré. Ce faisceau de présomptions nous amène à penser que ce type de table est mis à jour régulièrement, et que l'extraction reflète les dernières valeurs disponibles (dernière population recensée, dernier nom de commune associé à un numéro INSEE, etc.). Il faut souligner qu'il s'agit d'un traitement différent de celui des cantons par exemple, car les différents noms associés à un même numéro de canton sont conservés dans le RNE.

Le dernier point concernant les numéros associés aux communes est leur hétérogénéité (**P47**). Tous ne présentent pas la forme de trois chiffres décrite précédemment. Certains codes incluent 4 caractères de plus, deux lettres et deux chiffres, qui précisent une partie de la commune. Par exemple, le code `013SN03` correspond à la section 3 d'`Arutua` en *Polynésie française*, le code `055AR02` au deuxième arrondissement de `Marseille`, et le code `342CD01` à une subdivision d'`Isigny-sur-Mer` dans le *Calvados*. Cependant, ces formes longues (qui ne correspondent techniquement plus à des codes INSEE de communes) ne sont pas utilisés systématiquement.

2.3.2 Identifiants de personnes

Le numéro associé aux différents élu·es dans le RNE semble être conçu comme un identifiant. Cependant, on relève plusieurs problèmes relatifs à la nature bijective de la relation entre élu·es et numéros (ou plutôt à son absence constatée).

Plusieurs numéros pour un·e même élu·e. Un examen des données révèle qu'il arrive qu'une même personne soit associée à plusieurs numéros distincts (**P48**). Nous réalisons un test consistant à comparer les quadruplets (*nom*, *prénom*, *date de naissance*, *sexe*) afin d'identifier ceux qui sont exactement **identiques** tout en étant associés à des numéros **distincts**. Un ensemble de tests supplémentaires vient compléter les cas détectés, notamment l'identification des noms d'usage (déjà mentionnée en Section 2.1.1), et la comparaison approximative de noms et prénoms pour détecter les coquilles qui y sont parfois présentes. Ce traitement

19. https://www.insee.fr/fr/information/2549968

produit une liste d'homonymes (ou quasi) susceptibles d'être en réalité la même personne portant plusieurs numéros.

Nous recoupons ensuite manuellement les lignes concernées dans le RNE, en considérant leurs aspects spatiaux, temporels, socio-professionnels et réglementaires. Ainsi, une même personne ne peut pas occuper simultanément plusieurs mandats du même type (par exemple, être conseiller·e municipal dans deux communes différentes en même temps), ou des mandats de types différents dans des lieux incompatibles (maire dans un département et conseiller·e général·e dans un autre). Il est peu probable que deux homonymes habitant à l'autre bout de la France, ou n'ayant pas du tout la même profession, soient la même personne, même s'ils ne présentent pas d'incompatibilité de mandats par ailleurs. Ce traitement manuel nous permet de distinguer les élu·es qui sont effectivement désigné·es par plusieurs numéros de ceux qui sont effectivement des homonymes (deux personnes distinctes portant le même nom).

Parmi les 911 206 numéros d'élu·es présents dans le RNE, nous détectons ainsi 5 388 cas d'élu·es associés à plusieurs numéros distincts. Le plus généralement il s'agit de 2 numéros, plus rarement 3 (à 11 reprises), jamais plus. Cela représente donc plus de 1% des numéros de la base. On décompte également 179 paires d'homonymes confirmés manuellement, c'est-à-dire des personnes de même prénom, nom, et date de naissance, mais d'identifiants différents et pour lesquels une vérification a confirmé qu'il ne s'agit probablement pas de la même personne. Il faut souligner que notre liste d'élu·es possédant plusieurs numéros n'est pas forcément exhaustive, puisqu'une partie du traitement est réalisée manuellement. Si le lieu de naissance était indiqué dans le RNE, comme c'est le cas dans la base de données du Sénat (cf. Section 3.2), cette information nous permettrait d'améliorer le niveau d'automatisation de la tâche consistant à distinguer les homonymes des doublons.

Un examen des cas concernés par ces numéros multiples semble indiquer qu'ils se caractérisent souvent par une **coquille** dans leur nom de famille ou prénom, ce qui fait que l'on n'est pas en présence d'homonymes au sens strict. Par exemple, la conseillère municipale `Leonarda AZERONDE` (numéro `3682`) est également désignée sous le nom `Léonarda AZERONDE` (numéro `1395483`), avec un accent. On peut supposer que le problème est survenu à la saisie du deuxième mandat de cette élue, qui a vu l'opérateur orthographier le prénom légèrement différemment. Même dans l'hypothèse où un système d'autocomplétion automatique est fourni, si celui-ci procède de manière exacte, alors l'entrée existant dans le RNE ne sera pas proposée à l'utilisateur, qui créera une nouvelle entrée. La coquille peut également porter sur la date, par exemple la conseillère municipale `Karine FETTUCIARI` apparaît toujours exactement sous ce nom-ci, mais elle est décrite par deux dates de naissance différentes : `14/5/1968`, avec laquelle l'élue est associée au numéro `845248`, et `4/5/1968`, pour laquelle il s'agit de `1178596`. Les dates diffèrent seulement d'un chiffre, cela correspond clairement à une simple coquille.

Plusieurs élu·es pour un même numéro. L'erreur réciproque à la précédente correspond au cas où un même numéro est utilisé pour désigner plusieurs élu·es distinct·es (**P49**). Nos tests ne révèlent aucun numéro utilisé pour référer à des élu·es portant des noms différents, donc le cas échéant il ne pourrait s'agir que d'homonymes. Cependant, cela rend leur détection beaucoup plus difficile, voire impossible. En effet, si comme nous le supposons les informations personnelles d'un·e élu·e sont bien stockées dans une table spécifique de la base interne du RNE, alors toute mention de cet·te élu·e entraînera l'apparition des mêmes nom, prénom et date de naissance. Cette erreur revient donc à rechercher des mandats ou fonctions qui ont été attribués au même numéro, alors qu'il y a impossibilité physique ou légale qu'ils soient effectués par la même personne. En d'autres termes, cela revient à chercher des mandats ou fonctions qui ont été attribués à une autre personne qu'à l'élu·e véritablement concerné·e.

Bien que n'ayons pas trouvé de moyen d'automatiser ce test et donc d'évaluer l'amplitude de ce type d'erreur, nous avons trouvé manuellement quelques cas qui montrent que de tels

problèmes apparaissent bel et bien dans le RNE. Citons d'abord le cas de deux élus distincts, tous les deux nommés `Jean-Michel PETIT`, l'un portant le numéro `154538` et l'autre le `1165263`. Tous les deux ont été conseillers municipaux, le premier à `Glonville` dans la `Meurthe-et-Moselle` de `03/03/2003` à `08/03/2008`, et le second à `Courrières` dans le `Pas-de-Calais` de `09/03/2008` à `22/03/2014`. Outre leur numéro, ils ont des professions différentes : `Retraités salariés privés` vs. `Professeurs du secondaire et techn.` Le second a effectué un second mandat à `Courrières` à partir du `23/03/2014`, mais par erreur celui-ci a été attribué au premier `Jean-Michel PETIT`. Il est très difficile de détecter automatiquement ce type d'erreur. Tout au plus peut on identifier des cas où un·e élu·e identifié·e par un certain numéro occupe deux positions incompatibles (par exemple député·e de deux circonscriptions différentes en même temps). Mais d'une part la cause de ce type d'incohérence n'est pas forcément une confusion d'homonyme (par exemple l'un des deux mandats peut simplement être inexistant, et non pas simplement attribué à la mauvaise personne), et d'autre part des confusions d'homonymes peuvent amener à des situations où il n'y a pas d'incohérence (par exemple si les deux homonymes n'occupent pas simultanément des positions incompatibles).

Le sénateur `Christian COINTAT` constitue un autre cas intéressant. Le RNE lui attribue deux mandats de sénateur : l'un couvre la période `09/10/2001`–`20/09/2008` et l'autre `26/9/2004`–`30/9/2014`. On peut déjà remarquer un problème dans les dates, puisque les deux mandats se recouvrent partiellement. La base du Sénat indique que la seconde période correspond à une activité réelle de sénateur pour cet élu, mais que la première est erronée. En y regardant de plus près, la date de naissance varie dans les deux lignes du RNE (ce qui n'est bien entendu pas normal) : `01/08/1922` vs. `11/07/1943`. Après vérification, la seconde (celle du mandat réel) est correcte, tandis que la première est celle d'un autre sénateur : `Paul D'ORDANO`, dont `Christian COINTAT` était le suppléant. La première ligne correspond à un mandat de `Paul D'ORDANO` lors duquel celui-ci a été remplacé par `Christian COINTAT`, et qui a été improprement attribué en intégralité à ce dernier dans le RNE.

Pour finir, on peut mentionner le cas d'un sénateur appelé `Nouvel ESSAI`. Il s'agit probablement de l'intégration accidentelle dans les données mêmes du RNE d'un message d'erreur ou de journalisation, ou bien d'un résidu de test réalisé par le ministère. Toujours est-il qu'un recoupement avec la base du Sénat nous a permis de déterminer que cet élu fictif se substitue à une sénatrice bien réelle nommée `Dinah DERYCKE`. Il faut noter que la date de naissance de l'élu fictif ne correspond pas à la date réelle. C'est notamment ce type d'observation qui nous a amené à recouper systématiquement le RNE avec des sources secondaires, afin d'évaluer de façon plus objective la prévalence de cette forme d'erreur.

2.4 Autres

Certains problèmes détectés dans le RNE ne sont pas directement connectés à un champ particulier, mais sont plutôt relatifs au recoupement de plusieurs champs et/ou de plusieurs lignes. La Section 2.4.1 aborde le cas des lignes compatibles, qui constituent un cas de redondance d'information. La Section 2.4.2 se concentre sur les territoires non-renseignés et élu·es absent·es du RNE, et la Section 2.4.3 sur les institutions qui n'y sont pas représentées. Les Sections 2.4.4 et 2.4.5 traitent des mandats manquants et surnuméraires.

2.4.1 Lignes compatibles

Nous qualifions de **compatibles** deux lignes dont les champs sont soit exactement identiques, soit non-renseignés dans au moins l'une des deux lignes. Par exemple, on trouve dans `CM` les deux lignes suivantes décrites dans la Table 5 (n°839940 et 839941). Elles sont exactement similaires, à l'exception du champ de motif de fin de fonction, qui est renseigné dans la première mais pas dans la seconde. Il ne s'agit pas là d'une erreur à proprement parler, mais plutôt d'une redondance, puisqu'ici la seconde ligne ne fait que répéter la

première sans ajouter d'information (**P50**).

Département		Commune			Élu			
Libellé	Code	Libellé	Code	Pop.	Nom	Prénom	Sexe	Naissance
PUY DE DOME	63	Brassac-les-Mines	050	3308	CROZE	Yves-Serge	M	10/04/1950
PUY DE DOME	63	Brassac-les-Mines	050	3308	CROZE	Yves-Serge	M	10/04/1950

Profession		Mandat		
Libellé	Code	Date de début	Date de fin	Motif de fin
Retraités des professions libérales	60	28/03/2014	05/02/2017	FM
Retraités des professions libérales	60	28/03/2014	05/02/2017	FM

Fonction				Nuance politique	Numéro
Libellé	Date de début	Date de fin	Motif de fin		
Maire	28/03/2014	05/02/2017	FM	DVD	46764
Maire	28/03/2014	05/02/2017	--	DVD	46764

Table 5. Exemple de deux lignes compatibles, extraites de `CM` : la seule différence est leur motif de fin de fonction (indiqué en rouge).

On rencontre également des cas où il n'y a pas entre les lignes de hiérarchie claire, au contraire de cet exemple, dans lequel la première inclut la seconde. Par exemple, supposons que la nuance politique ne soit pas renseignée dans la seconde ligne de la Table 5 : alors on aurait toujours deux lignes compatibles, selon notre définition. Par contre, si la nuance était `DVD` dans la première ligne et `UMP` dans la seconde, on aurait alors une différence rendant ces deux lignes incompatibles, et ce malgré leur similarité sur les autres champs.

Pb	Description	CD	CM	CR	D	DE	EPCI	M	S
P50	Lignes compatibles	20	825	6	0	2	290	23	0
P58	Mandats débutant avant 2001	110	1 326	14	4	22	205	978	82
Total		130	2 151	20	4	24	495	1 001	82

Table 6. Décompte de cas problématiques divers, pour chaque table de données. La colonne *Pb* contient le code du problème correspondant décrit dans le texte.

Les nombres de *paires* de lignes compatibles trouvées dans chaque table sont indiqués dans la Table 6. Nous verrons plus tard (cf. Section 3) que ce type de compatibilité peut être exploité pour simplifier nos données d'une partie de ces redondances.

2.4.2 Territoire incorrect et élus manquants

Lors de vérifications réalisées manuellement, nous avons pu observer ponctuellement des erreurs portant sur le territoire associé à un mandat, et une absence pure et simple du RNE pour certain·es élu·es. En l'absence d'une référence externe au RNE, il semble impossible d'automatiser cette tâche, car ces cas ne peuvent pas être détectés simplement en recoupant les seules informations du RNE. Nous n'avons donc pas pu évaluer la fréquence de ce type de problèmes.

Territoire incorrect. Certains mandats sont correctement renseignés pour ce qui concerne les dates et l'élu·e concerné·e, mais la circonscription associée n'est pas la bonne, et parfois même n'existe pas à l'époque du mandat (**P51**). On peut par exemple citer la conseillère départementale `Emmanuelle ANTHOINE`, dont le RNE indique qu'elle est élue en 2017 dans le canton de `Valence-4`, alors qu'elle représente en réalité celui de `Drôme des collines`.

Parmi les député·es européen·nes, mentionnons `Patricia LALONDE`, dont le RNE indique qu'elle est élue en 2017 dans la circonscription `SUD-EST`, alors qu'il s'agit en réalité de celle d'`ILE-DE-FRANCE`. Le RNE indique aussi que `Pervenche BERÈS` est élue en 1999 dans la circonscription de l'`ILE-DE-FRANCE`, alors qu'à cette date le scrutin européen utilisait une circonscription unique. À noter que dans ce cas précis, la députée a enchaîné deux mandats,

et que ceux-ci sont fusionnés dans le RNE qui indique la période `1999/06/13–2014/06/30`. La circonscription incriminée pour le premier mandat est correcte pour le second, et cette situation pourrait donc être la conséquence du problème P30 déjà identifié (mandat couvrant plusieurs élections).

Élu·es manquants. Certain·es élu·es sont complètement absent·es du RNE (**P52**). Il ne s'agit pas d'un mandat oublié dans une table donnée, mais de l'ensemble des mandats de l'élu·e, indépendemment du type de mandat. Nous avons tout d'abord décelé ce problème en examinant le cas de Nicolas Sarkozy. Son mandat de président de la République n'est bien entendu pas représenté dans le RNE, puisqu'il n'est pas supposé couvrir ce type de mandat. Mais cet élu a également été conseiller général des *Hauts-de-Seine* en 2004–2007 (et même président de ce conseil), et député européen en 2002–2004. De plus, il a occupé une position de député à plusieurs reprises, y compris sur des périodes supposées couvertes par le RNE, en 2002 et 2005.

Il faut souligner que cette absence n'est pas liée au fait d'avoir été président de la république. D'une part, d'autres présidents sont recensés dans le RNE, par exemple `François HOLLANDE`, député en 2002 et 2007. D'autre part, ces absences touchent des personnes n'ayant pas été présidentes, comme Roselyne Bachelot, qui a été députée notamment en 2002 et 2007, et députée européenne en 2004–2007, mais n'apparaît pas non plus dans le RNE.

2.4.3 Institutions particulières

Plusieurs territoires disposent de différents statuts particuliers, leur accordant différents types d'assemblées bénéficiant elles-mêmes de prérogatives différentes des conseils standard. Ces assemblées sont gérées de différentes façons dans le contexte du RNE : certaines sont assimilées à des conseils départementaux ou régionaux, d'autres sont simplement absentes des données. La Table 7 résume la situation de ces institutions.

Pb	Territoire	CD	CM	CR	EPCI	M
–	Conseils municipaux de Paris/Lyon/Marseille	✗	✓	–	✓	✓
P53	Conseils d'arrondissement de Paris/Lyon/Marseille	✗	✗	–	–	✗
–	Conseils départementaux de Guadeloupe/Réunion	✓	–	–	–	–
–	Conseils régionaux de Guadeloupe/Réunion	–	✓	–	–	–
P54	Assemblée de Corse	✗	–	✗	–	–
–	Conseils départementaux de Martinique/Guyane (*avant* 2016)	✓	–	–	–	–
–	Conseils régionaux de Martinique/Guyane (*avant* 2016)	–	–	✓	–	–
P55	Assemblées de Martinique/Guyane (*après* 2016)	✗	–	✗	–	–
–	Conseil départemental de Mayotte	✓	–	✗	–	–
P56	Assemblée territoriale de Wallis-et-Futuna	✗	–	✗	–	–
–	Conseil territorial de Saint-Pierre-et-Miquelon	✓	–	✗	–	–
P56	Conseil territorial de Saint-Barthélemy	✗	–	✗	–	–
P56	Conseil territorial de Saint-Martin	✗	–	✗	–	–
P56	Assemblée de la Polynésie française	✗	–	✗	–	–
P57	Congrès de la Nouvelle-Calédonie	✗	–	✗	–	–
P57	Assemblées des provinces néo-calédoniennes	✗	–	✗	–	–

Table 7. Institutions relatives aux territoires dotés d'un statut spécial, et description de leur représentation dans le RNE. La colonne *Pb* contient le code du problème correspondant décrit dans le texte.

Paris, Lyon & Marseille. Le *Conseil de Paris* est créé le 01/01/1968 à la suite de la réorganisation de la région parisienne. Il cumule les compétences de conseil municipal de la commune de Paris et de conseil général du département de Paris. Ces deux structures sont fusionnées le 01/01/2019 en une collectivité à statut particulier appelée *Ville de Paris*. Le Conseil de Paris

comporte 163 conseiller·es municipal·es, tandis que 354 (puis 364 après 2014) autres élu·es sont seulement conseiller·es d'arrondissements. Seul·es les conseiller·es de Paris apparaissent dans le RNE, en tant que conseiller·es municipal·es et d'EPCI (pour la *Métropole du Grand Paris*). Les conseiller·es d'arrondissement semblent complètement absent·es du RNE (**P53**).

Depuis la loi PLM du 31/12/1982, le conseil municipal de *Lyon* se compose de 73 conseiller·es municipal·es et 148 conseiller·es d'arrondissement, et celui de *Marseille* comporte 101 conseiller·es municipal·es et 174 conseiller·es de secteur. Comme pour Paris, il semblerait que les conseiller·es d'arrondissement ou de secteur soient complètement absent·es du RNE (P53), alors que les conseiller·es municipal·es apparaissent dans `CM`, `M` et `EPCI`. À noter que si la *Métropole d'Aix-Marseille-Provence* est bien un EPCI, en revanche la *Métropole de Lyon* est une collectivité à statut particulier, à l'instar de la Ville de Paris. Comme cette dernière, cette structure possède les compétences du département sur son territoire. Ses conseiller·es, théoriquement au nombre de 165, apparaissent cependant dans la table `EPCI` (comme pour Paris).

Guadeloupe & Réunion. La *Guadeloupe* et *La Réunion* sont à la fois des départements et des régions, et à ce titre possèdent chacune un conseil départemental et un conseil régional. Les deux types d'élu·es sont répertoriés respectivement dans les tables `CD` et `CR` du RNE.

Corse. Les deux départements corses ont été (re)créés le 01/01/1976. Le 08/08/1982, la Corse est dotée d'une *Assemblée de Corse* au statut proche de celui d'un conseil régional. Le 01/01/2018, l'*Assemblée de Corse* récupère les pouvoirs des deux conseils départementaux (Haute-Corse et Corse-du-Sud) qui sont supprimés. Elle se compose de 63 membres élus les 03/12/2017 et 10/12/2017.

L'extraction du RNE que nous utilisons [20] ne contient pas d'élu·es de l'Assemblée de Corse (**P54**). La table `CR` contient seulement 3 mandats, qui s'achèvent en 2010. Les conseiller·es départemental·es, en revanche sont présent·es dans `CD` jusqu'à la dissolution de ces conseils départementaux.

Martinique, Guyane & Mayotte. La *Martinique* et la *Guyane* étaient également dans ce cas-là, avant de devenir des collectivités territoriales uniques en 2016. Ce changement de statut a vu la création des *Assemblée de Martinique* (51 membres) et *Assemblée de Guyane* (51 membres également), qui se sont substituées à leurs conseils généraux et régionaux respectifs. Les élu·es martiniquais·es et guyanais·es présents dans le RNE apparaissent dans les tables `CD` et `CR` pour les mandats antérieurs à cette date. Le RNE ne semble pas contenir d'élu·es de l'Assemblée de Martinique ni de celle de Guyane (**P55**).

Mayotte est également une collectivité territoriale unique, et son conseil départemental possède à la fois les attributions d'un conseil départemental et d'un conseil régional. Dans le RNE, ses élu·es sont listé·es dans la table `CD`.

Collectivités d'outre-mer. *Wallis-et-Futuna* est une collectivité d'outre-mer régie par l'*Assemblée territoriale de Wallis-et-Futuna*. Ses 20 membres n'apparaissent pas dans le RNE (**P56**), bien que celui-ci contienne pourtant des député·es et sénateur·rices wallis-et-futunien·nes.

Saint-Pierre-et-Miquelon était un département d'outre-mer jusqu'à 1985, avant de devenir une collectivité d'outre-mer. Son conseil territorial (19 membres) cumule les attributions des conseils départementaux et régionaux, et ses membres apparaissent dans la table `CD` du RNE.

Saint-Barthélémy et *Saint-Martin* étaient des communes du département de Guadeloupe jusqu'au 15/07/2007, avant de devenir deux collectivités d'outre-mer. Leurs conseils territoriaux se composent respectivement de 19 et 23 membres. Ceux et celles-ci ne sont pas enregistré·es dans le RNE (P56), qui ne contient que les sénateur·rices et député·es de ces collectivités (pas de conseiller·es municipal·es non plus).

20. Cependant, certaines extraction photographiques contiennent une table dédiée à l'Assemblée de Corse.

La *Polynésie française* est une collectivité d'outre-mer régie par l'*Assemblée de la Polynésie française* qui compte 57 membres. Ceux et celles-ci ne sont pas présent·es dans le RNE (P56), à la différence des député·es, sénateur·rices et conseiller·es municipal·es polynésien·nes.

Nouvelle-Calédonie. La *Nouvelle-Calédonie* est une collectivité d'outre-mer à statut spécifique, de même que les trois provinces qui la constituent. La Nouvelle-Calédonie est régie par le *Congrès de la Nouvelle-Calédonie* (54 membres), et chaque province par une *Assemblée de province* (76 membres au total), dont les membres n'apparaissent pas dans le RNE (**P57**). En revanche, les autres élu·es néo-calédonien·nes y sont présent·es (conseiller·es municipal·es, sénateur·rices, député·es).

2.4.4 Couverture temporelle

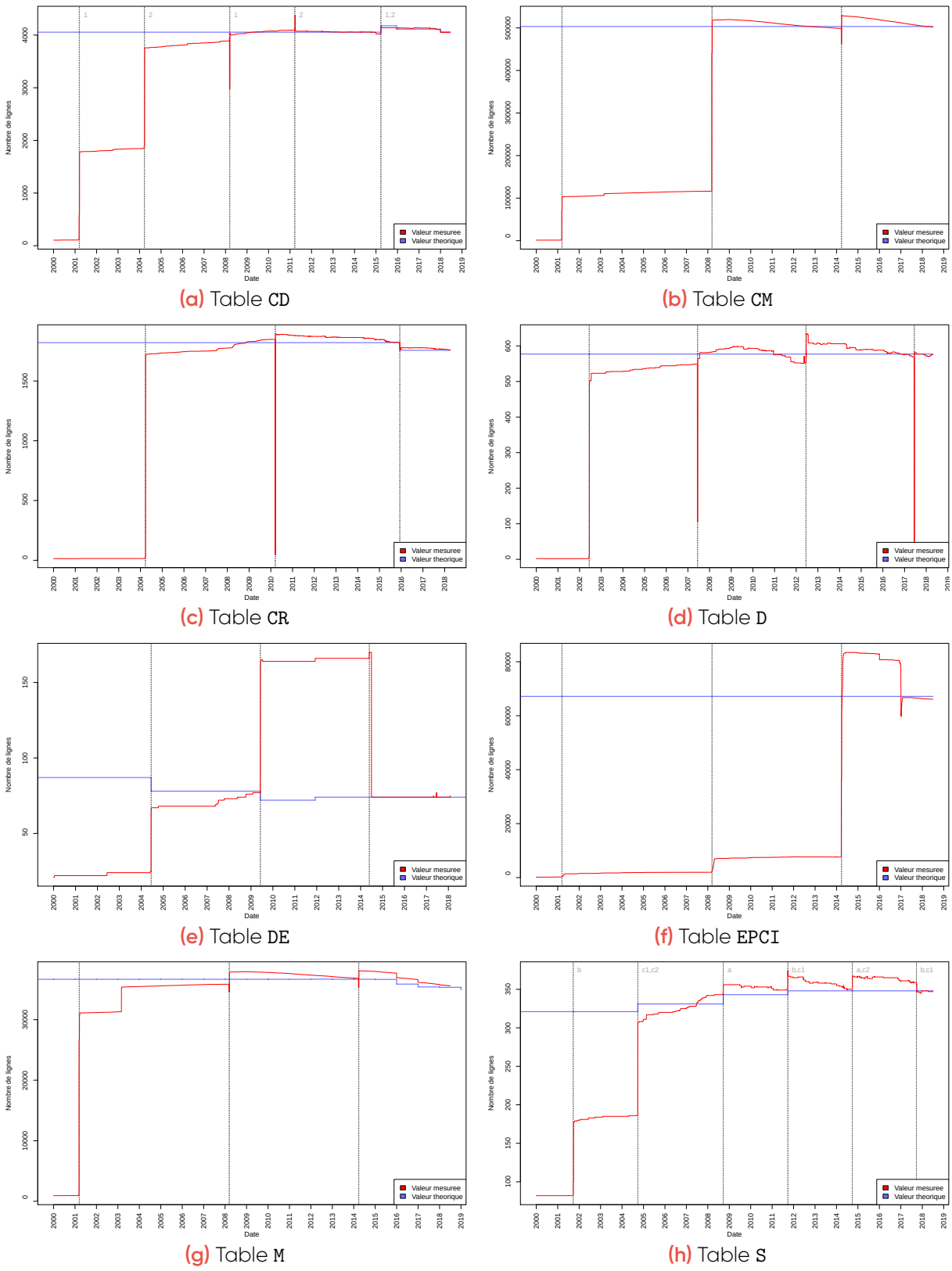
La Figure 1 montre, pour chaque table de données, l'évolution du nombre de mandats en cours, en fonction du jour considéré. Le nombre de mandats (i.e. de lignes) simultanés décomptés dans les données est affiché en rouge, tandis que le nombre réglementaire est représenté en bleu. Il faut souligner que certaines de ces valeurs réglementaires sont des *approximations*, comme détaillé en Annexe C.2. De plus, ces valeurs réglementaires sont calculées de manière à tenir compte de la manière dont le RNE inclut ou pas les institutions particulières décrites en Section 2.4.3. Les lignes verticales en pointillés correspondent aux dates d'élection, la série étant précisée en haut le cas échéant (cf. les Tables 34 et 35 en annexe pour le détail des séries).

On peut constater de façon globale que le nombre de mandats est parfois significativement en deçà ou au delà de la limite réglementaire. On se concentre ici sur la période temporelle que l'on peut raisonnablement estimer comme couverte par le RNE. Les mandats manquants dans la période couverte et les dépassements de limite réglementaire sont discutés plus loin (cf. Section 2.4.5).

Hypothèse sur la saisie. Comme indiqué dans la Section 1.1, le RNE est supposé recenser tous les élus à partir de 2001. Cependant, les figures montrent qu'il n'est pas exhaustif dès sa création, et que son niveau de complétude a crû progressivement. Notre hypothèse est que le RNE ne contient que les mandats qui ont débuté *pendant* ou *après* l'année 2001, par opposition aux mandats en cours à cette date, et ayant donc début *avant*. En effet, on constate pour toutes les tables que le nombre de lignes décomptées démarre bien en deçà de la limite réglementaire, puis augmente avec le temps. On voit aussi que ces augmentations ont tendance à se produire au moment des élections, ce qui conforte notre hypothèse.

Cependant, cela n'est pas très régulier et dépend de la table considérée. Ainsi, on observe que les maires, conseiller·es régional·es, député·es et député·es européen·nes atteignent un niveau proche de la limite réglementaire en une seule élection. On peut néanmoins constater un saut anormal pour les maires en 2003. Pour les conseiller·es général·es, cela prend deux élections, ce qui y est cohérent avec le fait que ces assemblées étaient à l'époque renouvelées par moitié à chaque élection. En revanche, c'est aussi le cas des conseiller·es municipal·es (deux élections pour atteindre approximativement la limite réglementaire), alors qu'une seule élection devrait suffire, à l'instar des maires. Les sénateur·rices, pour leur part, devraient nécessiter trois élections pour atteindre la limite réglementaire, puisque cette assemblée était renouvelée par tiers à cette époque. Pourtant, c'est le cas en seulement deux élections, et on peut constater que les sénateur·rices de la série A sont intégré·es prématurément (avant l'élection de 2009 qui les concerne). La table `EPCI` est très en-dessous de la valeur réglementaire, même après deux élections. Mais cela peut être dû au fait qu'il s'agit d'une estimation de la limite réglementaire datant de 2019, et ne s'appliquant pas forcément à toute la période considérée.

En opposition à notre hypothèse, on peut aussi remarquer que chaque table de données



(a) Table `CD` (b) Table `CM`

(c) Table `CR` (d) Table `D`

(e) Table `DE` (f) Table `EPCI`

(g) Table `M` (h) Table `S`

Figure 1. Évolution du nombre de mandats au cours du temps, pour chaque table de données du RNE. En rouge et bleu les nombres de mandats mesurés et réglementaires, en noir les dates d'élections.

comporte des mandats dont la date de début est antérieure à 2001. On peut citer par exemple le cas du conseiller municipal `Arthur ALBERT`, pour lequel le RNE décrit un mandat couvrant la période `11/05/1953–08/03/2008`, soit près de 55 ans. La Table 6 indique plus de mille cas pour *CM* ; des centaines pour *CD*, *EPCI* et *M* ; et des dizaines pour *CR*, *DE* et *S*, pour un total de 2 741 cas. Bien que ces nombres soient négligeables en regard de la taille de ces tables, cela indique tout de même que le ministère a intégré des données couvrant une période antérieure à celle que le RNE a pour but de couvrir. On peut alors se demander pourquoi cela n'est pas systématique, et ne concerne que certains mandats. Sur la base de nos observations précédentes, notamment dans la Section 2.2.2, on pourrait formuler l'hypothèse que seul·e·s les élu·es ayant effectué des mandats successifs dont un en cours en 2001 sont concerné·e·s. Mais celle-ci semble invalidée par le faible nombre d'élu·es concerné·es (2 741).

Processus de purge. Lors de la réunion au ministère du 08/10/2019 mentionnée dans la Section 1.2.1, nous avons appris que les services du ministère procédaient régulièrement à la suppression de certaines lignes du RNE. En 2008 d'abord, il s'agissait des mandats relatifs aux conseils municipaux des communes de moins de 3 500 habitants et concernant des élu·es n'ayant effectué que ce seul mandat électif dans leur carrière. En 2014 ensuite, ce seuil a été baissé à 1 000 habitants. La raison de cette suppression ne nous est pas très claire. Elle pourrait être technique, et la suppression aurait alors pour but d'alléger la base interne. Cependant, vu de l'extérieur, la taille finalement assez réduite du RNE en regard des quantités de données manipulées à l'époque du big data ne semble pas justifier une telle gestion des données. Il pourrait également s'agir d'enjeux réglementaires, concernant notamment les données à caractère personnel. Le ministère ne conserverait pas, par sécurité, les données non-essentielles pour remplir sa mission d'identification de cas de cumul de mandats.

Ceci pourrait expliquer plusieurs des anomalies relevées ou des observations incompatibles avec notre hypothèse (celle de saisie restreinte aux mandats post-2001). Tout d'abord, cela justifie au moins en partie que le nombre de conseiller·es municipal·es n'atteigne pas la limite réglementaire après une élection. S'il s'agit de la seule cause à cela, alors cela signifierait que la purge effectuée par le ministère concerne un grand nombre de lignes de `CM` (près des trois-quarts). Cela serait aussi cohérent avec le fait que `M` n'est pas touchée par ce phénomène (la purge ne concerne pas les conseiller·es municipal·es ayant été maires). Et enfin, cela pourrait expliquer en partie l'évolution de la courbe d'`EPCI`, puisque les conseiller·es communautaires sont avant tout des conseiller·es municipal·es. Par contre, cela ne permet pas de justifier la croissance trop rapide (selon notre hypothèse) observée pour le Sénat.

Indépendemment de ces causes possibles, le premier enseignement de la Figure 1 est que la période supposée couverte par le RNE l'est de façon inégale, surtout au début de son existence (**P58**). En l'état, nous ne pourrions exploiter ces données qu'à partir de 2008.

2.4.5 Excès et défaut de mandats

La deuxième anomalie qui ressort de la Figure 1 est que même quand les courbes rouges s'approchent de la limite réglementaire, on observe parfois une différence assez significative, se chiffrant au moins en dizaines de postes au moment de l'élection, et ce aussi bien en excès qu'en défaut. La Table 8 donne les différences entre le nombre de mandats décomptés dans le RNE et la valeur attendue, pour chaque élection.

Défaut de mandats. Considérons les premières dates auxquelles le RNE devrait être exhaustif, même dans le cas où notre hypothèse (enregistrement progressif des mandats post-2001) serait correcte. Si on ne tient compte que des tables n'ayant pas fait l'objet de la purge décrite en Section 2.4.4, on constate un important défaut de mandats (**P59**) : 299 mandats manquant de `CD` (élections de 2004), 104 de `CR` (2004), 77 de `D` (2002), 12 de `DE` (2004), 5 986

de M (2001) et 4 de S (2004). Dans le même ordre d'idée, une mise en correspondance de EPCI avec CM révèle que 277 conseiller·es communautaires n'ont pas de mandat associés de conseiller·es municipal·es (en revanche, on retrouve dans CM tous les maires de M).

Le problème des élu·es manquant·es (P52) tient probablement un rôle dans cette observation, mais pour aboutir à un effet aussi marqué, la raison principale est vraisemblablement plutôt un problème d'ordre systématique. On peut supposer que les données qui sont remontées des préfectures vers le ministère étaient incomplètes. Une raison à cela peut être la relative nouveauté du RNE à cette époque et le manque d'habitude des opérateurs préfectoraux, qui se sont estompés par la suite (et avec eux le nombre d'erreurs).

Pb	Table	Élection	Valeur décomptée	Limite réglementaire	Différence
P58	CD	11-18/03/2001	1 757	4 055	−2 298
P59		21-28/03/2004	3 756	4 055	−299
P59		09-16/03/2008	4 038	4 055	−17
P60		20-27/03/2011	4 376	4 055	+321
P59		22-29/03/2015	4 145	4 177	−32
P59	CR	21-28/03/2004	1 717	1 821	−104
P60		14-21/03/2010	1 889	1 821	+68
P60		06-13/12/2015	1 765	1 757	+8
P59	D	09-16/06/2002	500	577	−77
P59		10-17/06/2007	564	577	−13
P60		10-17/06/2012	585	577	+8
P59		11-18/06/2017	525	577	−52
P59	DE	13/06/2004	66	78	−12
P60		07/06/2009	164	72	+92
P60		25/05/2014	170	74	+96
P59	M	11-18/03/2001	30 691	36 677	−5 986
P60		09-16/03/2008	37 790	36 681	+1 109
P60		23-30/03/2014	38 078	36 681	+1 397
P59	S	23/09/2001	100	322	−222
P59		26/09/2004	307	311	−4
P60		21/09/2008	355	343	+12
P60		25/09/2011	374	348	+26
P60		28/09/2014	354	348	+6
P60		24/09/2017	357	348	+9

Table 8. Mandats décomptés aux dates d'élections, pour chaque table dont les limites réglementaires sont jugées suffisamment fiables. Cette table reprend les valeurs correspondant aux droites verticales dans la Figure 1. La colonne *Pb* contient le code du problème correspondant décrit dans le texte.

Mandats surnuméraires. Outre ce défaut de mandats, la Figure 1 et la Table 8 mettent en évidence un autre problème, de nature opposée : on observe parfois un *excès* de mandats (**P60**).

À un moment ou à un un autre, cela se produit dans toutes les tables. Un examen manuel des tables révèle la présence de mandats sans réalité, comme par exemple Paule Helyette PELTIER, dont le RNE indique qu'elle a été conseillère générale de La Réunion à deux reprises, alors que ce n'est pas le cas. Mais là encore, au vu de l'ampleur du problème, nous pensons que la cause réelle est plus systématique. Dans certains cas, on observe des pics ponctuels, et proches des élections (ex. CD en 2011), que nous supposons être dus à une mauvaise saisie des dates de début/fin de mandat, entraînant un recouvrement entre des périodes de temps qui devraient être strictement successives. On observe aussi le contraire, des chutes brutales (CD en 2008, CR en 2010, D en 2007, M en 2008 & 2014), qui peuvent s'expliquer par l'erreur inverse (laisser une vacance entre deux mandats successifs).

Mais il y a également dans certaines tables des périodes d'excès de mandats importantes

et prolongées : c'est le cas pour `CM`, `D`, `EPCI`, `M`, `S`, et surtout `DE`. Dans cette dernière, on excède le nombre de mandats attendu de plus du double, ce qui dénote un problème systématique dans les données. Il pourrait s'agit d'un problème relatif aux dates de fin des mandats commençant avant 2004, dont l'année de fin est fréquemment 2014 (au lieu de 2004, peut être une coquille systématique).

Variation progressive. En théorie, pour un type de mandat donné, il ne devrait jamais y avoir plus de mandats que de sièges disponibles. Il est possible qu'il y en ait moins, mais de façon temporaire. Donc on attend de ces courbes qu'elles aient une allure quasi-horizontale, avec des variations mineures ne dépassant jamais le maximum réglementaire et ne s'en éloignant pas trop non plus.

Pourtant, en plus des défauts/excès ponctuels/durables de mandats discutés précédemment, on observe un troisième type de comportement : une croissance ou décroissance progressive du nombre de mandats entre deux élections (**P61**). Ainsi, dans `CD` on passe de 3 756 mandats après les élections de 2004 à 3 951 la veille de l'élection suivante, soit un déficit réduit de 299 à 104 grâce à une croissance à peu près continue sur l'ensemble de la période. On peut faire la même observation pour toutes les autres tables de données. Dans le cas de `M`, la croissance fait même l'objet d'un pic en 2003, que nous n'expliquons pas. Ces observations semblent indiquer un problème dans les dates saisies pour décrire les mandats dans le RNE. En effet, rien ne justifie qu'autant de débuts de mandats s'échelonnent sur toute la période séparant deux élections.

On constante également le phénomène inverse dans `CM`, `CR`, `D`, `M` et `S` : un nombre de mandats supérieur à la limite réglementaire à la suite d'une élection, puis une décroissance progressive sur toute la durée du mandat amenant à une valeur proche de la limite juste avant l'élection suivante. Si cette décroissance se produisait peu de temps avant l'élection suivante, on pourrait la justifier par le fait qu'il n'est plus possible d'organiser d'élections partielles pour remplacer un élu·e quittant son poste, ce qui conduit à une diminution du nombre de postes occupés. Mais ce n'est pas le cas : la décroissance démarre dès le début du mandat, ce qui est difficile à expliquer autrement que par des erreurs dans les dates de début des mandats.

2.5 Bilan des problèmes

Dans cette section, nous revenons rapidement sur les principaux problèmes que nous avons pu détecter dans l'extraction historique du RNE (Section 2.5.1), et nous tentons d'avancer quelques hypothèses visant à expliquer ces problèmes (Section 2.5.2).

2.5.1 Principaux problèmes détectés

La Table 9 dresse le bilan des problèmes que nous avons détectés dans le RNE, et que l'on peut regrouper en différentes catégories.

Forme, validité & complétude. Les problèmes de forme concernent la façon dont l'information est représentée dans les tables de données, ceux de validité portent sur la nature même de cette information, et la notion de complétude renvoie aux valeurs non-renseignées.

Ces problèmes apparaissent assez fréquemment dans la base, et bien qu'ils puissent paraître triviaux, ils entraînent néanmoins des complications parfois assez importantes lors de la comparaison automatique des données. Pour cette raison, il est nécessaire de les résoudre en premier lieu.

Code	Description	Section
P1	Signes diacritiques irréguliers	2.1
P2	Espaces consécutives	2.1
P3	Ponctuation irrégulière	2.1
P4	Noms d'usage irréguliers	2.1.1
P5	Prénoms non-renseignés	2.1.1
P6	Coquilles	2
P7	Casse irrégulière	2.1.2
P8	Numérotation des noms de cantons	2.1.2
P9	Mention `archivé·e` dans les noms de communes/EPCI	2.1.2
P10	Utilisation d'articles dans les noms de lieux	2.1.2
P11	Abréviation de `Saint` dans les noms de lieux	2.1.2
P12	Noms de commune non-renseignés	2.1.2
P13	Nuance politique `NC`	2.1.3
P14	Doublon de nuance	2.1.3
P15	Nuance relative	2.1.3
P16	Nuances non-renseignées	2.1.3
P17	Professions non-renseignées	2.1.3
P18	Motifs de fin de mandat non-renseignés	2.1.3
P19	Motifs de fin de fonction non-renseignés	2.1.3
P20	Hétérogénéité des libellés de fonction	2.1.3
P21	Libellés de fonction incomplets	2.1.3
P22	Sémantique variable du motif de fin	2.1.3
P23	Dates de naissance non-renseignées	2.2.1
P24	Dates de naissance incohérentes	2.2.1
P25	Dates de début de mandat non-renseignées	2.2.2
P26	Dates de fin de mandat incohérentes	2.2.2
P27	Bornes de mandat incohérentes	2.2.2
P28	Micro-mandats	2.2.2
P29	Mandats se recouvrant	2.2.2
P30	Mandats couvrant plusieurs élections	2.2.2
P31	Dates de début de fonction non-renseignées	2.2.3
P32	Dates de fin de fonction non-renseignées	2.2.3
P33	Dates de début de fonction incohérentes	2.2.3
P34	Dates de fin de fonction incohérentes	2.2.3
P35	Bornes de fonction incohérente	2.2.3
P36	Micro-fonctions	2.2.3
P37	Fonctions se recouvrant	2.2.3
P38	Fonctions hors-mandats	2.2.3
P39	Numéros de département non-renseignés	2.3.1
P40	Numéros de département incorrects	2.3.1
P41	Non-unicité des numéros de départements	2.3.1
P42	Unicité des numéros de circonscriptions législatives	2.3.1
P43	Non-unicité des numéros de cantons	2.3.1
P44	Non-unicité des numéros d'EPCI	2.3.1
P45	Numéros de commune non-renseignés	2.3.1
P46	Non-unicité des numéros de commune	2.3.1
P47	Hétérogénéité des numéros de commune	2.3.1
P48	Plusieurs numéros pour un·e seul·e élu·e	2.3.2
P49	Plusieurs élu·es associé·es au même numéro	2.3.2
P50	Présence de lignes compatibles	2.4.1
P51	Circonscriptions incorrectes	2.4.2
P52	Élu·es manquant·es	2.4.2
P53	Absence des conseiller·es d'arrondissement	2.4.3
P54	Absence des conseiller·es de Corse	2.4.3
P55	Absence des conseiller·es de Martinique/Guyane	2.4.3
P56	Absence des conseiller·es territorial·es	2.4.3
P57	Absence des conseiller·es de Nouvelle-Calédonie	2.4.3
P58	Couverture temporelle incomplète	2.4.4
P59	Mandats manquants	2.4.5
P60	Mandats surnuméraires	2.4.5
P61	Variation progressive du nombre de mandats	2.4.5

Table 9. Principaux problèmes détectés dans les données du RNE, listés par ordre d'apparition dans le texte.

Cohérence. Les problèmes de **cohérence** entre valeurs touchent principalement les dates. Un bon nombre d'entre eux peuvent être au moins détectés, sinon résolus, par des moyens automatiques, en recoupant les informations contenues dans la base, et en la confrontant à des ressources externes du type de celles décrites en annexe (dates d'élection, nombre de sièges par institution etc.).

Relations biunivoques. La nature **biunivoque** attendue des relations entre entités et identifiants, fait défaut à bon nombre des champs concernés. Là encore, l'utilisation de ressources externes devrait permettre de résoudre une partie du problème, en particulier celle concernant les territoires. Mais la résolution des problèmes d'identifiant des élu·es, qui constituent l'entité centrale de la base, apparaît comme plus difficile.

Fiabilité des mandats. Le fait que d'une part, il **manque** des mandats dans le RNE, et que d'autre part, de nombreux mandats soient **erronés**, constitue un problème critique et d'une difficulté majeure. Son traitement nécessitera l'exploitation de sources secondaires pouvant être recoupées au moins partiellement avec le RNE, ne serait-ce que pour évaluer ses niveaux d'incomplétude et d'erreur.

2.5.2 Causes potentielles

Nous voyons trois causes possibles principales aux problèmes détectés : le processus de saisie des données au niveau des préfectures, la structure même de la base interne du ministère, et le processus qui a abouti à l'extraction de la base historique sur laquelle nous travaillons.

Processus de saisie. Le processus de saisie ne semble pas homogène sur l'ensemble des opérateurs chargés de cette tâche dans les préfectures. Tous ne comprennent visiblement pas les champs de la même façon, et les renseignent par conséquent différemment. Par exemple, certains opérateurs interprètent visiblement les champs relatifs à la fin de mandat comme concernant le mandat *précédent* celui décrit dans la ligne concernée, alors qu'il s'agit bien de ce dernier. Il en résulte des dates de fin de mandat ou de fonction antérieures à celles de début.

L'interface graphique utilisée est elle-même une source d'erreur potentielle, quoi que cela soit difficile à confirmer sans y avoir accès. Nous supposons qu'elle incorpore une forme de complétion automatique pour certains champs et/ou lors de la sélection de l'élu·e concerné·e. Il est possible que certains des mandats incorrects aient ainsi été attribués à la mauvaise personne simplement en sélectionnant un homonyme dans une liste d'élu·es déjà présent dans la base et proposés par l'interface graphique à l'utilisateur. Inversement, certains élu·es possèdent plusieurs identifiant dans le RNE, ce qui indique qu'il·elles ont été recréé·es plusieurs fois, probablement parce que l'interface graphique n'a pas su avertir l'opérateur. On constate également des inconsistances de date qui paraissent pourtant facilement évitables à la saisie, en informant par exemple l'opérateur que la date de fin qu'il vient d'entrer est antérieure à celle de début.

Lors de la réunion évoquée dans la Section 1.2.1, les agents du ministère ont mentionné des contraintes ergonomiques visant à limiter ce type de vérification afin de ne pas décourager les opérateurs d'effectuer les saisies, ce qui peut expliquer cette absence de contrôle.

Structure de la base interne. La deuxième cause potentielle de problème est la structure de la base interne, que nous ne pouvons que deviner à travers les extractions dont nous disposons et nos discussions avec les agents du ministère. Il semblerait que celle-ci ne soit pas conçue pour stocker certains aspects des données de façon longitudinale, ce qui engendre notamment les problèmes relatifs aux identifiants de certains territoires, et empêche plus généralement de représenter l'évolution de leurs caractéristiques (ex. population des communes).

Processus d'extraction. Enfin, nous soupçonnons que le processus d'extraction, qui a pour objectif de produire les tables dont nous disposons à partir de la base interne, occasionne lui-même certaines des erreurs que nous avons détectées dans les données. En effet, comme nous l'avons mentionné, il semble peu probable que certaines informations comme la date de naissance d'un·e élu·e soit stockée plus d'une fois dans la base interne. Pourtant, nous avons détecté une certaine variabilité dans ce type de données, qui pourrait être due à un bug lors de l'extraction.

Remarques diverses. On peut associer bon nombre des erreurs détectées dans les données à des **mandats partiels**, i.e. dus à des départs prématurés (décès, démissions, départs au gouvernement ou vers d'autres responsabilités, etc.), ou à des retours en poste avant la fin du mandat (fin d'un poste gouvernemental et retour à l'assemblée, élection partielle, etc.). Nous avons observé que les dates de mandats et/ou de fonctions sont souvent incorrectes dans ces cas-là, que ce soit celles de début et/ou de fin, engendrant des recouvrement de mandats et/ou fonction. Ces cas sont aussi souvent associés à des micro-mandats, des mandats manquants dans le RNE, ou des mandats présents dans le RNE mais inexistants en réalité. Cependant, il est difficile d'identifier la cause de ce type de problème, qui peut être aussi bien d'origine humaine (lors de la saisie), liée à la structure de la base, provoquée par un bug lors de l'extraction, ou même issue de la conjonction de plusieurs de ces facteurs.

Si on compare la **qualité des tables** composants l'extraction historique, il ressort clairement que c'est celle relative aux `EPCI` qui apparaît comme la plus fragile, que ce soit en termes de données manquantes, incohérentes, ou erronées. Ceci est d'autant plus problématique qu'il s'agit de la table la plus difficile à recouper avec des sources secondaires, avec `CM`, et donc la plus difficile à corriger. Si on considère le problème en termes de proportions et non de valeurs absolues, c'est alors `DE` qui est la table la moins fiable, ne serait-ce qu'en raison du nombre élevé de mandats surnuméraires (plus du double du nombre réglementaire) sur la période 2009–2014.

3 Traitement des données

Cette section décrit le traitement mis en oeuvre pour résoudre autant que possible les problèmes identifiés dans la Section 2. Le programme déjà mentionné, écrit en Langage R[12] et disponible en ligne[1], applique ce traitement aux données brutes. Les modifications que nous apportons au RNE sont donc non-destructives, dans le sens où il est possible de les réaliser sélectivement, ou de les réviser par la suite. Le traitement comporte une part importante de modifications des données du RNE, décrites dans la Section 3.1).

Cependant, un certain nombre des problèmes relevés dans la Section 2 ne peuvent pas être résolus simplement sur la base des données présentes dans le RNE, et requièrent d'effectuer un recoupement avec des bases externes. La difficulté est alors d'identifier des sources secondaires à la fois fiables et suffisamment complètes, afin d'introduire leur contenu (ou une partie) dans notre propre base sans diminuer la qualité de l'information qu'elle contient. Dans un premier temps, nous avons exploité trois sources externes, dont nous détaillons l'intégration : la base de données publique du Sénat (Section 3.2), celle de l'Assemblée Nationale (Section 3.3) et une conjonction de Wikipédia et du site officiel du Parlement Européen (Section 3.4).

Une fois ces sources secondaires intégrées, nous fusionnons les différentes tables (Section 3.5) afin d'obtenir une table unique, qui fait ensuite elle-même l'objet de certains des traitements décrits en Section 3.1.

3.1 Opérations sur les données du RNE

Le traitement se décompose en un certain nombre d'étapes que nous détaillons dans cette section. La plupart sont génériques, et sont appliquées séparément à chaque table constituant le RNE, tandis que certaines sont spécifiques à une ou plusieurs tables en particulier. Une fois chaque table traitée, notre programme les fusionne pour obtenir une table unique réunissant l'ensemble des données. Certaines étapes sont appliquées à cette table fusionnée également (cf. Section 3.5).

Code	Description	Problèmes traités
E1	Normalisation des valeurs	P1, P2, P3, P7, P8, P9, P11
E2	Correction des identifiants d'élu·es	P48
E3	Corrections *ad hoc*	P4, P5, P6, P10, P12, P18, P19, P20, P17, P23, P24, P25, P26, P27, P29, P31, P32, P33, P34, P35, P37, P39, P44, P45, P51, P52, P58, P59, P60
E4	Corrections systématiques diverses	P4, P12, P13, P14, P20, P25, P29, P31, P32, P39, P40, P45, P47
E5	Ajouts de colonnes manquantes	P41, P43
E6	Fusion des lignes compatibles	P50
E7	Ajustement des dates de fonction	P38
E8	Alignement aux dates d'élection	P30
E9	Suppression des micro-mandats/fonctions	P28, P36
E10	Fusion des mandats se recouvrant	P29, P37
E11	Division des mandats longs	P30
E12	Résolution des recouvrements de positions	P29, P37
E13	Révision des motifs de fin	P18, P19

Table 10. Étapes du traitement appliqué aux données du RNE, et problèmes résolus par chacune d'entre elles.

La Table 10 liste les différentes étapes, ainsi que les problèmes que chacune permet de

résoudre (eux-mêmes détaillés dans la Table 9). Il faut souligner que certains problèmes sont traités au cours de plusieurs étapes, qu'une étape ne règle pas forcément un problème complètement, et que certains problèmes n'ont simplement pas pu être traités du tout, pour différentes raisons expliquées plus loin. La Section 3.1.14 fait le bilan des modifications effectuées lors de cette partie du traitement.

3.1.1 Normalisation des valeurs

La première étape (**E1**) du traitement consiste à effectuer une normalisation des valeurs qui vise à résoudre plusieurs problèmes relevés précédemment. Les attributs les plus concernés sont ceux représentant des noms, dont la normalisation inclut les opérations suivantes :

- Suppression des mots `archivé` (`EPCI`) et `archivée` des noms d'EPCI et de communes (P9).
- Suppression des signes diacritiques (P1).
- Remplacement de la ponctuation par des espaces (P3).
- Suppression des espaces consécutives (P2), et de celles situées en début et fin de nom.
- Transformation des capitales en minuscules (P7)

Pour les toponymes en particulier, la normalisation comporte deux opérations supplémentaires :

- Remplacement des abréviations de `Saint` et ses variantes par leur forme complète (P11).
- Ré-écriture des nombres exprimés en chiffres romains par leur équivalent utilisation la numérotation indo-arabes (P8).

Un traitement complémentaire, appliqué à tous les attributs (pas seulement les noms) consiste à remplacer toutes les cellules vides par le symbole explicite `NA` utilisé en langage R pour dénoter une absence d'information.

3.1.2 Correction des identifiants d'élu·es

La deuxième étape (**E2**) vise à résoudre le problème des élu·es associé·es à plusieurs identifiants distincts (P48). Nous faisons l'hypothèse que l'identifiant de plus petite valeur est aussi le plus ancien, et c'est donc celui que nous conservons. La deuxième étape consiste à substituer cet identifiant de plus petite valeur aux autres identifiants, qui sont superflus. La Table 11 donne un exemple de correction, pour un cas fictif.

Identifiant	Nom	Prénom	Naissance	Sexe	Début	Fin	Motif
12345	Marcel	DUPONT	07/07/1957	M	06/05/2003	06/12/2007	FM
987654	Marcel	DUPONT	07/07/1957	M	03/04/2002	05/05/2003	DV
987654	Marcel	DUPONT	07/07/1957	M	08/12/2004	03/02/2008	FM
12345	Marcel	DUPONT	07/07/1957	M	06/05/2003	06/12/2007	FM
12345	Marcel	DUPONT	07/07/1957	M	03/04/2002	05/05/2003	DV
12345	Marcel	DUPONT	07/07/1957	M	08/12/2004	03/02/2008	FM

Table 11. Exemple de correction réalisée à l'Étape E2. Les trois premières lignes représentent l'état original des données : le même élu y est désigné par deux identifiant distincts. Les trois dernières lignes correspondent aux résultat de la correction : on ne conserve que l'identifiant le plus petit (les modifications sont indiquées en rouge).

Les élu·es et identifiants concernés ont été identifiés au préalable, lors de l'analyse des données du RNE, comme expliqué en Section 2.3.2. On décompte plus de 5 000 cas d'élu·es associés à plusieurs identifiants, et environ 200 cas d'homonymes réels, i.e. personnes de même nom, prénom, sexe et date de naissance, et donc correctement désignés par des identifiants distincts. La distinction a été effectuée manuellement pour les cas litigieux, en

donnant la préférence à l'existence d'homonymes dans les cas où même un examen manuel n'a pas permis de trancher.

3.1.3 Corrections *ad hoc*

La troisième étape (**E3**) consiste à appliquer un ensemble de corrections prédéfinies manuellement, et visant à résoudre des problèmes ponctuels, ne nécessitant pas d'implémenter un traitement spécifique. Ce type de corrections inclut notamment :

- Correction de noms, prénoms, dates de naissance et sexes d'élu·e (P4, P5, P6, P23, P24).
- Correction de libellés et codes de profession (P17).
- Correction de dates de début et fin de mandats (P6, P25, P26, P27, P29).
- Correction de dates de début et fin de fonctions (P6, P31, P32, P33, P34, P35, P37).
- Correction de libellés et motifs de fin de fonction et mandat (P18, P19, P20).
- Correction de libellés et codes de territoires (P10, P12, P39, P44, P45, P51).
- Suppression de lignes jugées incorrectes ou redondantes (P60).
- Ajout de certains mandats manquants (P52, P58, P59).

Cette étape aborde donc un grand nombre de problèmes distincts, mais généralement de façon très incomplète, car elle consiste à appliquer des corrections définies manuellement. De plus, celles-ci concernent des erreurs détectées au gré de l'exploration des données, ou bien signalées automatiquement par les tests décrits en Section 2 mais peu fréquentes. Nous traitons les erreurs à grande échelle de façon plus systématique et automatique.

Champ/Modification	CD	CM	CR	D	DE	EPCI	M	S	Total
Code de la circonscription européenne	0	0	0	0	11	0	0	0	11
Libellé de la circonscription européenne	0	0	0	0	11	0	0	0	11
Libelle de la région	0	0	2	0	0	0	0	0	2
Code du département	0	0	2	0	0	38	0	0	40
Libelle du département	0	0	2	0	0	0	0	0	2
Libelle de la circonscription législative	0	0	0	2	0	0	0	0	2
Code du canton	4	0	0	0	0	0	0	0	4
Libellé du canton	22	0	0	0	0	0	0	0	22
Numéro SIREN de l'EPCI	0	0	0	0	0	64	0	0	64
Libellé de l'EPCI	0	0	0	0	0	10	0	0	10
Code INSEE de la commune	0	5	0	0	0	36	0	0	41
Libellé de la commune	0	22	0	0	0	64	0	0	86
Date de naissance de l'élu·e	7	26	4	11	0	4	9	7	68
Nom de l'élu·e	9	33	9	2	2	8	13	1	77
Prénom de l'élu·e	5	44	2	1	0	8	8	1	69
Code de sexe	0	3	0	0	0	0	1	2	6
Code de profession	0	2	0	1	0	1	1	0	5
Libellé de profession	0	2	0	1	0	1	1	0	5
Date de début de mandat	19	32	4	92	58	42	12	66	325
Date de fin de mandat	95	154	12	61	160	102	22	18	624
Motif de fin de mandat	2	32	4	1	12	2	14	3	70
Libellé de fonction	5	50	0	0	0	9	0	0	64
Date de début de fonction	2	184	0	0	0	47	129	1	363
Date de fin de fonction	4	104	16	1	0	17	55	0	197
Motif de fin de fonction	1	36	16	0	0	1	10	0	64
Suppression de ligne	16	175	3	12	2	3	55	13	279
Ajout de ligne	3	1	1	4	792	0	1	0	802
Total	194	905	71	187	1 026	419	331	112	3 245

Table 12. Nombre de corrections *ad hoc* effectuées dans chaque table de données au cours de l'Étape E3.

La Table 12 résume les modifications effectuées pour chaque table de données du

RNE. Toutes les lignes correspondent à des corrections apportées à des lignes existantes à l'exception des deux dernières, qui décomptent respectivement le nombre de lignes supprimées et le nombre de lignes insérées.

3.1.4 Corrections systématiques diverses

La quatrième étape (**E4**) porte sur l'application automatique de corrections diverses, certaines spécifiques à une ou plusieurs tables en particulier. On y réalise notamment les opérations suivantes, dans l'ordre indiqué ci-dessous :

- Normalisation des noms d'usage (P4), en systématisant la forme : *Nom de naissance* puis *Nom d'usage*.
- Normalisation des identifiants de communes (P47), en ne gardant que les trois chiffres de tête.
- Remplacement de la nuance `NC` par la valeur `NA` (P13).
- Remplacement des nuances `RDG` et `M-NC` respectivement par `PRG` et `MAJ` (P14).
- Pour la table `EPCI` :
  - Suppression des codes de département et de municipalité valant zéro (P39, P45).
  - Correction des codes de département incorrects (P40).
  - Ajout du nom de municipalité quand il manque mais que les codes de département et de municipalité sont présents (P12).
  - Ajout des nom/code de municipalité / code de département manquants, en exploitant les informations présentes dans `EPCI` et `CM` (P12, P39, P45).
- Pour la table `M` :
  - Alignement de certaines dates de début de mandat (P29), qui sont systématiquement fausses dans certains départements et pour certains scrutins.
- Pour la table `D` :
  - Correction du libellé de la fonction de président de l'Assemblée Nationale (P20).
- Suppression des lignes sans dates de mandat ni de fonction (P25).
- Ajout de la date de début/fin de fonction quand elle manque (P31, P32), en utilisant celle de début/fin de mandat.

3.1.5 Ajout des colonnes manquantes

Lors de la cinquième étape (**E5**), nous rajoutons plusieurs colonnes à certaines tables, pour traiter différents problèmes.

Pour résoudre le problème d'identification des départements (P41), nous créons un nouveau champ `ID departement`, chargé d'identifier un département de façon unique indépendemment des changements de nom, changements de numéro, et plus généralement de l'évolution des départements (disparition/réapparition). Cet identifiant a été élaboré à partir de sources extérieures, et est de la forme `xxxx_yyy`, où `xxxx` est l'année de première création du département, et `yyy` est un code de 2 ou 3 caractères correspondant au premier code historiquement attribué au département. Par exemple, l'identifiant du département des *Côtes-d'Armor*, créé en 1790 sous le nom de *Côtes-du-Nord* sous le numéro 21, est le `1790_21` alors que son code actuel est le `22`.

Nous créons également un code pour désigner les cantons de façon unique (P43). Comme on l'a expliqué en Section 2.3.1, la combinaison du département et du nom semble être l'information la plus pertinente pour identifier un canton de façon unique dans le RNE. Elle n'est pas parfaite car un même nom peut recouvrir différents territoires au fil du temps, mais il s'agit de la meilleure approximation possible sans faire appel à des données externes. Nous élaborons un nouvel identifiant appelé `ID canton`, de la forme `xx_yy`, où `xx` représente le numéro du département et `yy` est celui du canton dans ce département. Par exemple, l'identifiant du canton de `MARSEILLE 1` est `13_35`.La partie `yy` ne correspond pas forcément

au numéro indiqué dans le RNE. Il est obtenu en numérotant les cantons placés dans l'ordre alphabétique, ce qui permet au moins d'associer toujours le même identifiant au même nom.

Nous avons soulevé le même type de problème pour les circonscription législatives en Section 2.3.1 (P42). Cependant, aucun nom explicite ne leur est associé dans le RNE, et la même opération ne peut donc pas être réalisée. Leur suivi dans le temps nécessite l'exploitation d'une source de données externe au RNE. On peut faire la même remarque pour les codes désignant les communes (P46). Pour ce qui est des EPCI (P44), une partie du problème est résolu à l'Étape E3, qui traite manuellement les cas où plusieurs numéros SIREN sont associés *exactement* au même nom. Les autres cas requièrent l'utilisation d'une source externe.

Outre ces identifiants, nous créons également deux nouvelles colonnes afin de tracer les changements apportées aux données :

- `Sources` : nature de la source ou des sources originales pour la ligne considérée (à ce stade du traitement, seulement le RNE).
- `Correction dates` : indique si au moins l'une des dates contenues dans la ligne été modifiée, ou si une date manquante a été rajoutée.
- `Correction autres` : indique si un champ autre que temporel a été modifié.

Enfin, nous réordonnons les colonnes afin qu'elles soient dans le même ordre dans toutes les tables, ce qui n'est pas le cas dans les données du RNE.

3.1.6 Fusion des lignes compatibles

La sixième étape (**E6**) porte sur les lignes compatibles (P50), telles qu'elles sont définies dans la Section 2.4.1, à savoir des lignes dont les champs sont soit identiques, soit non-renseignés pour au moins l'un d'entre eux.

Identifiant	Nom	Prénom	Début	Fin	Motif	Profession
12345	DUPONT	Marcel	06/05/2003	06/12/2007	FM	NA
12345	DUPONT	Marcel	06/05/2003	NA	NA	Enseignant
12345	DUPONT	Marcel	06/05/2003	06/12/2007	FM	Enseignant

Table 13. Exemple de fusion entre deux lignes compatibles, effectuée lors de l'Étape E6. La partie supérieure de la table contient les deux lignes compatibles, et la partie inférieure est le résultat de leur fusion. Les valeurs indiquées en couleur correspondent aux champs fusionnés.

Cette situation redondante est traitée en fusionnant les deux lignes. Les champs dont les valeurs sont identiques sont bien sûr conservés tels quels, et dans le cas où l'une des valeurs est non renseignée pour un champ, on conserve l'autre valeur. La Table 13 donne un exemple simplifié du résultat de la fusion de deux lignes compatibles (fictives).

3.1.7 Ajustement des dates de fonction

La septième étape (**E7**) vise à corriger les dates de fonction qui ne sont pas contenues dans le mandat défini sur la même ligne (P38). Notre méthode de résolution consiste à substituer aux dates de fonction incorrectes celles de mandat. Ainsi, si la date de début de fonction est antérieure à celle de début de mandat, on effectue la substitution. De même, si la date de fin de fonction est postérieure à celle de fin de mandat, on réalise aussi cette substitution.

Cela touche également les dates manquantes. Par exemple, si aucune date de fin de fonction n'est indiquée (i.e. valeur `NA`) alors qu'il y a une date de fin de mandat, cette dernière est substituée au `NA`. De plus, le motif de fin de fonction est lui aussi aligné sur celui du mandat. La Table 14 donne un exemple fictif de l'ajustement réalisé au cours de cette étape.

Mandat			Fonction		
Début	Fin	Motif	Début	Fin	Motif
03/04/2002	06/05/2007	FM	01/02/2002	06/05/2007	FM
03/04/2002	06/05/2007	FM	18/04/2002	03/11/2008	AU
03/04/2002	06/05/2007	FM	18/04/2002	NA	NA
03/04/2002	06/05/2007	FM	03/04/2002	06/05/2007	FM
03/04/2002	06/05/2007	FM	18/04/2002	06/05/2007	FM
03/04/2002	06/05/2007	FM	18/04/2002	06/05/2007	FM

Table 14. Exemples d'ajustement de dates de fonction relativement aux dates de mandat effectué à l'Étape E7. Les 3 premières lignes correspondent aux données originales et les 3 dernières au résultat de l'ajustement. Les couleurs indiquent les valeurs modifiées.

3.1.8 Arrondissement aux dates d'élection

La huitième étape (**E8**) effectue une mise à jour des bornes de certains mandats et fonctions. L'analyse réalisée dans la Section 2.2.2 a révélé que les dates utilisées pour marquer ces bornes ne correspondaient pas systématiquement aux dates d'élections, mais pas forcément non plus aux dates de prise officielle de fonction. Or, nous n'avons pas trouvé de source exhaustive listant ces dernières. Nous avons donc opté pour l'approximation consistant à utiliser les dates d'élection, qui sont elles trouvables en ligne assez facilement (cf. Annexe B), pour marquer les changements de mandats dans les données.

Lorsque une date de début de mandat se trouve à 7 jours ou moins après une date d'élection, cette borne est modifiée pour prendre la valeur de cette date d'élection. Selon le même principe, une date de fin de mandat se trouvant à 7 jour ou moins avant une date d'élection est remplacée par la veille de cette date d'élection. En d'autres termes, on suppose qu'un mandat s'achève au plus tard la veille de la date d'élection la plus proche, et qu'un mandat commence le jour de l'élection la plus proche. La table 15 donne des exemples d'application de ces règles.

Début de mandat		
Date originale	Diff.	Date alignée
01/06/2002	−8	01/06/2002
08/06/2002	−1	**09**/06/2002
11/06/2002	+2	**09**/06/2002
14/06/2002	−2	**16**/06/2002
20/06/2002	+3	**16**/06/2002
28/06/2002	+12	28/06/2002

Début de mandat		
Date originale	Diff.	Date alignée
01/06/2002	−8	01/06/2002
08/06/2002	−1	08/06/2002
11/06/2002	+2	**08**/06/2002
14/06/2002	−2	**15**/06/2002
20/06/2002	+3	**15**/06/2002
28/06/2002	+12	28/06/2002

Table 15. Exemple d'alignement de dates effectué lors de l'Étape E8, pour les élections législatives ayant eu lieu les 09/06/2002 & 16/06/2002. Les parties en gras correspondent aux modifications apportées lors de l'alignement. La colonne *Diff* indique la différence à la date d'élection la plus proche, exprimée en jours.

La durée limite de 7 jours a été déterminée empiriquement afin que la tolérance soit suffisamment élevée pour traiter tous les cas ciblés, tout en n'étant pas trop longue pour ne pas traiter des mandats jugés courts sans pour autant être des micro-mandats.

Les dates de fonctions sont également modifiées quand c'est nécessaire, c'est-à-dire dans les cas où l'alignement des dates de mandats les rend incompatibles, et occasionnent le problème de dépassement relevé précédemment (P38), i.e. la fonction commence (resp. finit) avant (resp. après) le mandat.

Nous avons conscience que cette étape introduit une certaine approximation dans les dates. Cependant, celle-ci n'est pas dommageable à l'analyse de séquences destinées à être effectuée sur ces données, qui ne nécessite pas une telle précision. En contrepartie, cette étape permet de résoudre certains cas particuliers de mandats recouvrant plusieurs élections

(P30). Plus important, l'approximation réalisée facilite grandement un certain nombre de vérifications et de corrections sur les mandats, effectués lors des étapes suivantes.

3.1.9 Suppression des micro-mandats/fonctions

Comme discuté dans les Section 2.2.2 et 2.5.2, les micro-mandats, que nous définissons comme des mandats durant au plus 7 jours, constituent un problème important du RNE (P28), pas vraiment en eux-mêmes mais plutôt en raison du fait qu'ils signalent généralement la présence d'erreurs ou de mandats manquants. Certaines de ces erreurs sont corrigées par ailleurs, et les mandats manquants sont traités au moyen de source secondaires décrites plus loin.

Cette neuvième étape (**E9**) vise plutôt à traiter les micro-mandats eux-mêmes, simplement en supprimant la ligne correspondante. Les micro-fontions posent également problème (P36), cependant elles n'amènent pas à la suppression de la ligne, car le mandat associé n'est pas forcément un micro-mandat. On se contente donc de supprimer la fonction de la ligne (non seulement les dates, mais aussi le libellée, le code, et le motif de fin le cas échéant).

3.1.10 Fusion des mandats se recouvrant

Le RNE contient un certain nombre de mandats du même type, occupés par la même personne, mais correspondant pourtant à des périodes qui se recouvrent au moins partiellement (P29). Le but de cette dixième étape (**E10**) est de fusionner les lignes correspondant. L'opération est réalisée seulement si les lignes concernées sont *compatibles*, c'est à dire que, outre l'élu·e, elles concernent la même circonscription et la même fonction (si la fonction est renseignée).

Quand de telles lignes sont détectées, elles sont fusionnées de manière à obtenir l'union des périodes temporelles qu'elles décrivent. On effectue la même opération en cas de recouvrement de fonction (P37). Il faut souligner que grâce aux étapes mises en oeuvre avant celle-ci, deux fonctions qui se recouvrent appartiennent forcément à des lignes dont les mandats se recouvrent également, et ce au moins sur la période commune aux deux fonctions, ce qui rend cette opération cohérente.

Élu			Mandat			Fonction		
Id.	Nom	Prénom	Début	Fin	Motif	Début	Fin	Motif
26490	LEBLANC	Martine	05/06/2002	06/06/2007	FM	15/06/2002	06/06/2007	FM
26490	LEBLANC	Martine	18/07/2005	05/06/2012	FM	18/07/2005	08/12/2009	DV
12345	DUPONT	Marcel	05/06/2002	06/06/2007	FM	NA	NA	NA
12345	DUPONT	Marcel	05/06/2002	05/06/2012	FM	18/06/2002	12/03/2003	DV
12345	DUPONT	Marcel	07/06/2007	05/06/2012	FM	18/012/2004	05/06/2012	FM
26490	LEBLANC	Martine	05/06/2002	05/06/2012	FM	15/06/2002	08/12/2009	DV
12345	DUPONT	Marcel	05/06/2002	05/06/2012	FM	18/06/2002	12/03/2003	DV
12345	DUPONT	Marcel	07/06/2007	05/06/2012	FM	18/012/2004	05/06/2012	FM

Table 16. Exemples de fusion de mandats se recouvrant, effectuées lors de l'Étape E10. Les 5 premières lignes correspondent aux données originales et les 3 dernières au résultat de la fusion. Les couleurs indiquent les modifications. La toute dernière ligne n'est pas modifiée.

La Table 16 donne deux exemples fictifs de ce type de fusion. La première élue (Martine LEBLANC) est décrite par deux lignes dont les mandats et les fonctions se recouvrent partiellement. La ligne fusionnée regroupe les unions de ces périodes. Le second élu (Marcel DUPONT) est décrit par trois lignes. Les deux premières sont fusionnables car les mandats se recouvrent et la première ne contient pas du tout de fonction. Cependant, ce n'est pas le cas de la troisième : les mandats se recouvrent encore, mais pas les fonctions, qui correspondent

à deux périodes discontinues. Il est donc nécessaires de conserver deux lignes distinctes dans la base pour représenter ces deux fonctions séparées dans le temps.

3.1.11 Division des mandats longs

La onzième étape (**E11**) a pour objectif de diviser les mandats jugés trop longs car ils englobent au moins une élection (P30). Cela permet de traiter à la fois les lignes qui étaient dans ce cas déjà dans les données de départ, mais aussi les cas pouvant résulter de l'étape précédente (i.e. E10).

Lorsqu'un mandat long est détecté, on le découpe au niveau des dates d'élection qu'il contient. Quand une élection comporte deux tours (et donc deux dates), on ne découpe le mandat que sur la première des deux dates. Si nécessaire, les périodes des fonctions sont découpées en même temps, selon le même principe. Des motifs de fin de mandat ou de fonction normales (valeur `FM`) sont insérées dans toutes les nouvelles lignes, sauf la dernière (pour laquelle on conserve les motifs déjà existants).

Mandat			Fonction		
Début	Fin	Motif	Début	Fin	Motif
02/03/1997	10/12/2008	DV	10/04/2000	03/06/2005	AU
02/03/1997	24/05/1997	FM	NA	NA	NA
25/05/1997	08/06/2002	FM	10/04/2000	08/06/2002	FM
09/06/2002	09/06/2007	FM	09/06/2002	03/06/2005	AU
10/06/2007	10/12/2008	DV	NA	NA	NA

Table 17. Exemple de découpage d'un mandat long selon les dates d'élection, lors de l'Étape E11. La première ligne montre le mandat original, tandis que les quatre lignes suivantes résultent du découpage. Les couleurs permettent de visualiser les dates conservées. Les dates des élections législatives utilisées dans cet exemple sont 25/05/1997, 09/06/2002, et 10/06/2007.

La Table 17 donne un exemple illustrant les différents cas de figure. La période originale contient 3 élections, donc elle est divisée 3 fois, ce qui produit en tout 4 lignes distinctes. Le motif de fin `FM` est inséré quand il est approprié, et le motif `DV` original est conservé uniquement pour la dernière ligne. En ce qui concerne la fonction, celle-ci démarre après la première élection et se termine avant la troisième, donc elle n'apparaît ni dans la première ni dans la dernière ligne. Là aussi, on insère `FM` à la fin de la période obtenue sur la division de la deuxième date d'élection, et on conserve le motif de fin original `AU` pour la dernière partie.

3.1.12 Résolution des recouvrements de position

Comme on l'a vu en Section 2.2.2, certains types de mandats peuvent être associés à une positon unique. C'est le cas des député·es (un·e par circonscription), et des conseiller·es général·es (un·e par canton). Cela signifie qu'il est possible de détecter les cas de recouvrement de mandats pour une position donnée (P29), i.e. les moments où celle-ci est incorrectement occupée par deux personnes simultanément.

Si les deux mandats concernés impliquent la même personne, le problème aura été résolu lors des étapes précédentes. Le but de cette douzième étape (**E12**) est de traiter le reste des cas, i.e. ceux mettant en jeu des personnes différentes. Notre traitement vérifie d'abord l'ampleur du recouvrement, et se concentre sur ceux inférieurs à 8 jours. Nous choisissons arbitrairement d'ajuster la date de fin de celui qui commence le plus tôt, afin qu'il se termine un jour avant l'autre mandat.

Lorsque la période de recouvrement dépasse 8 jours, nous avons remarqué qu'il arrive fréquemment que les lignes concernées partagent soit la même date de début, soit la même date de fin. Dans le premier cas, on ajuste la date de début du mandat finissant le

Identifiant	Nom	Prénom	Avant la résolution		Après la résolution	
			Début	Fin	Début	Fin
12345	DUPONT	Marcel	06/06/2002	08/09/2004	06/06/2002	**03**/09/2004
26490	LEBLANC	Martine	04/09/2004	04/06/2007	04/09/2004	04/06/2007
12345	DUPONT	Marcel	06/06/2002	08/09/2004	06/06/2002	08/09/2004
26490	LEBLANC	Martine	06/06/2002	04/06/2007	**09/09/2004**	04/06/2007
12345	DUPONT	Marcel	06/06/2002	04/06/2007	06/06/2002	**17/07/2005**
26490	LEBLANC	Martine	18/07/2005	04/06/2007	18/07/2005	04/06/2007

Table 18. Trois exemples de résolution de recouvrement de position, telle que réalisée à l'Étape E12. Les modifications sont indiquées en gras. Le premier cas correspond à un recouvrement mineur (8 jours ou moins), et les deux autres à des recouvrement majeurs, avec une date de début commune pour le premier et une date de fin pour le second.

plus tard, de manière à ce qu'il commence un jour après la fin de l'autre. À l'inverse, quand c'est la date de fin qui est commune, on ajuste la date de fin du mandat commençant le plus tôt, afin qu'il se termine un jour avant le début de l'autre.

La Table 18 illustre ces différent cas de figure. Dans tous les cas, les dates de fonction sont adaptées selon la même approche, afin d'être compatibles avec le mandat supposé les contenir. Lorsque la période de recouvrement dépasse 8 jours et que les mandats n'ont aucune date en commun, nous considérons qu'il s'agit d'une erreur nécessitant une intervention manuelle (Étape E3).

Un traitement similaire est appliqué pour résoudre le problème analogue rencontré pour les fonctions (P37). Toutes les fonctions ne sont pas uniques cependant, par exemple il n'y a qu'un seul président du Sénat, mais plusieurs vice-présidents indifférenciés. Le traitement ne peut être appliqué qu'aux fonctions uniques.

3.1.13 Révision des motifs de fin

La treizième étape (**E13**) consiste à compléter les motifs de fin manquant (P18, P19) et à réviser les motifs existant :

- Suppression des motifs associés à une date de fin de mandat/fonction manquante. On part du principe qu'un motif de fin n'a pas de sens si le mandat ou la fonction n'est pas effectivement terminée.
- Ajout du symbole `FM` (fin normale) quand la date de fin est alignée avec une élection. On suppose que quand un mandat ou une fonction se termine à proximité d'une élection, il s'agit d'une fin normale.

3.1.14 Ampleur des modifications effectuées

Les Tables 19 et 20 donnent un aperçu de l'ampleur des modifications réalisées sur les données brutes. La première se concentre sur les étapes susceptibles de changer le nombre de lignes dans la table de données traitée, que ce soit en plus ou moins. Elle indique le nombre de lignes présentes dans la table à l'issue de chacune de ces étapes. La dernière ligne montre la différence entre les données brutes et celles issues du traitement. Proportionnellement, les tables dont la balance est la plus grande sont `DE` (+2,8%), `EPCI` (−2,1%), `D` (−1,5%) et `CM` (−1,2%). En comparaison, les modifications des autres tables touchent moins de 1% des lignes. On constate aussi que l'Étape E10 supprime de nombreuses lignes en fusionnant les mandats qui se recouvrent, mais que l'Étape E11 en recrée ensuite une grande partie lors de la division des mandats longs. Ceci peut sembler inutile, mais il faut souligner que le but de ces deux étapes est principalement de procéder à un redécoupage des mandats permettant d'éviter les recouvrements (et non pas d'écarter des lignes).

La Table 20 est plus générale, dans le sens où elle montre les étapes qui peuvent modifier

Étape		CD	CM	CR	D	DE	EPCI	M	S
–	Données brutes	12 728	1 239 413	5 818	2 488	253	127 940	113 522	1 076
E3	Corrections *ad hoc*	12 712	1 239 241	5 815	2 477	251	127 938	113 468	1 063
E4	Corrections systém.	12 712	1 239 241	5 815	2 477	251	126 710	113 468	1 063
E6	Lignes compatibles	12 702	1 239 128	5 813	2 477	250	126 515	113 467	1 063
E9	Micro-mandats	12 616	1 234 183	5 795	2 450	250	125 166	113 453	1 063
E10	Mandats recouvrants	10 011	1 002 827	5 571	2 078	197	122 085	110 150	780
E11	Mandats longs	12 768	1 224 345	5 804	2 452	260	125 230	114 616	1 167
E9	Micro-mandats	12 766	1 224 316	5 805	2 451	260	125 230	114 616	1 066
Différence		+38	−15 097	−14	−37	+7	−2 710	+1 094	−10

Table 19. Nombre de lignes dans chaque table à l'issue de chaque opération susceptible d'affecter ce nombre. La dernière ligne indique la différence entre la taille initiale (i.e. données brutes du RNE) et la taille finale de chaque table.

les valeurs, quelle que soit la nature de cette modification. Elle contient les nombres de modifications/ajouts/suppressions d'une valeur/ligne réalisés à chaque étape pour chaque table. L'Étape E1 n'est pas représentée, puisqu'elle implique la normalisation de tous les noms d'élu·es, et donc elle s'applique par définition à toutes les lignes sans exception.

On peut constater que la plupart des lignes originales sont modifiées d'une façon ou d'une autre pour toutes les tables à l'exception de `M` et `S`, avec des taux de modification s'élevant à 100% pour `DE` et 90% pour `CD`. Pour `CD`, `CR`, `D` et `M`, l'essentiel des modifications est réalisé lors de l'Étape E8 (alignement sur les dates d'élection), donc cela ne concerne pas des erreurs à proprement parler, puisque cet alignement est réalisé dans l'objectif de faciliter le reste des corrections (comme expliqué en Section 3.1.8). Dans le cas de `CR`, les corrections sont probablement dues à certains problèmes de cohérence observés sur les élections de 2015. En effet, celles-ci étaient couplées à la réforme des régions, et les dates de passation de pouvoir n'ont semble-t-il pas été traitées de la même manière pour toutes les régions.

Étape		CD	CM	CR	D	DE	EPCI	M	S
E2	Identifiants d'élu·es	65	5 603	49	3	3	1 442	559	2
E3	Corrections *ad hoc*	195	733	1 173	162	255	4 176	271	104
E4	Corrections systématiques	488	481 778	296	30	2	35 095	8 820	18
E6	Lignes compatibles	10	113	2	0	1	195	1	0
E7	Ajustement des dates	55	2 499	16	0	0	1 050	1 704	0
E8	Alignement des dates	10 330	10 632	1 837	1 274	209	17 393	1 486	500
E9	Micro-mandats	87	5 021	19	27	0	1 364	48	0
E10	Mandats recouvrants	2 605	231 356	224	372	53	3 081	3 303	283
E11	Mandats longs	2 757	221 518	233	374	63	3 145	6 874	387
E12	Recouvrements de position	296	2 810	4	97	–	213	251	0
E9	Micro-mandats	2	54 884	0	1	0	24	428	101
E13	Motifs de fin	5	58	2	0	0	40	5	0
Valeurs globales		11 492	736 630	3 539	1 419	260	73 198	20 063	865
		90%	63%	61%	58%	100%	57%	18%	12%

Table 20. Nombre de lignes modifiées lors de chaque étape susceptible d'affecter ce nombre, exceptée E1 qui modifie toutes les lignes. La dernière ligne n'est pas la somme de ces nombres, mais le nombre de lignes modifiées au cours de l'intégralité du traitement (certaines lignes peuvent être modifiées à plusieurs reprises).

Pour les mandats locaux `CM`, `M` et `EPCI`, c'est l'Étape E4 (corrections systématiques) qui domine, sans que nous puissions identifier une raison particulière à cela. La table `DE` est surtout modifiée lors de l'Étape E3 (corrections *ad hoc*), ce qui reflète le fait qu'une bonne partie des dates de fin de mandat étaient incorrectes (cf. Section 2.4.5) et on nécessité un ajustement manuel.

Enfin, il faut remarquer que l'Étape E9 (suppression des micro-mandats/fonctions) est

appliquée à deux reprise, car l'Étape E11 (division des mandats longs) est susceptible de produire de nouveaux micro-mandats (ce qui est le cas, d'après la table).

Le reste de cette section porte sur l'intégration de sources de données secondaires à celles du RNE, et au recoupement des mandats décrits afin d'estimer la fréquence des erreurs et l'importance de l'incomplétude des données de ce dernier.

3.2 Données du Sénat

Nous décrivons d'abord la base de données publiques du Sénat (Section 3.2.1) avant d'expliquer comment nous l'avons combinée au données du RNE (Section 3.2.2), et de faire le bilan des modifications qui en ont résulté (Section 3.2.3).

3.2.1 Présentation

La base de données du Sénat [21] contient des informations sur les sénateur·rices actuel·les et passé·es, sous licence ouverte `data.gouv.fr`. Elle se présente sous deux formes : d'une part un dump PostgreSQL contenant l'intégralité des données sous forme de base de données relationnelle, et d'autre part un ensemble de fichiers aux formats CSV, JSON et XML semblant correspondre à l'extraction des différentes tables formant cette base de données. Nous nous sommes concentrés sur les fichiers CSV, dont le format était plus accessible dans le contexte de l'état de nos scripts R à ce stade. Cependant, un bref examen a révélé que l'information semble plus complète dans le fichier PostgreSQL, donc il s'agit de la source à privilégier dans le futur.

Les différents fichiers CSV disponibles décrivent l'activité des sénateur·rices, que l'on peut répartir en plusieurs catégories :

- **Général**
  - Informations générales sur les sénateur·rices (notamment état-civil), dans leur dernier état connu ;
  - Appartenances courantes et historiques aux groupes politiques ;
- **Commissions**
  - Appartenances courantes et historiques aux commissions permanentes, spéciales, des affaires européennes, d'enquête, et aux missions d'information, à l'exception des commissions mixtes paritaires ;
  - Appartenances en cours aux commissions permanentes, spéciales, des affaires européennes, d'enquête, et aux missions d'information, à l'exception des commissions mixtes paritaires
- **Offices parlementaires et délégations**
  - Appartenances courantes et historiques aux offices parlementaires, délégations et à des instances internationales ;
  - Appartenances courantes aux offices parlementaires, délégations et à des instances internationales ;
- **Mandats**
  - Mandat cantonaux et départementaux, courants et historiques ;
  - Mandats de Député, actuels et historiques ;
  - Mandats divers : syndicat des eaux, etc., actuels et historiques ;
  - Mandats européens, actuels et historiques ;
  - Mandats métropolitains, actuels et historiques ;
  - Mandats régionaux, actuels et historiques ;
  - Mandats sénatoriaux, actuels et historiques ;
  - Mandats territoriaux, actuels et historiques ;
  - Mandats municipaux, actuels et historiques ;

21. https://data.senat.fr/les-senateurs/

- **Autres**
  - Appartenances aux groupes d'études (actuels et historiques) ;
  - Appartenances aux Groupes Interparlementaires d'Amitié (actuels et historiques) ;
  - Sénateur·rices siégeant dans les organismes extraparlementaires (actuels et historiques) ;
  - Données sur les Organismes Extraparlementaires dans lesquels siègent des sénateur·rices (actuel·les et historiques).

Le but de notre traitement est de déterminer l'ampleur des problèmes présents dans le RNE (notamment les informations erronées et manquantes) au moyen de cette source extérieure. Pour cette raison, nous nous concentrons seulement sur les tables relatives à la description des élu·es et des mandats du type de ceux déjà présents dans le RNE. Un examen des données nous révèle assez rapidement que si les données relatives aux mandats de sénateur·rice sont particulièrement fiables, cela ne semble pas être le cas pour les autres mandats, qui paraissent de qualité moindre. En particulier, de nombreuses dates de début ou de fin de mandat ne sont pas précisées. Elles sont parfois remplacées par l'année, mais ce n'est pas systématique. De plus, il existe des conflits entre dates de mandats et de fonctions. Une vérification manuelle de certains cas a révélé que certaines dates de mandats et/ou de fonctions sont incorrectes, nous faisant soupçonner un remplissage déclaratif de cette partie de la base. Pour cette raison, nous exploitons finalement uniquement la table contenant les données personnelles des sénateur·rices, et celle décrivant leurs mandats de sénateur·rice. Nous remettons l'exploitation des autres données à une date ultérieure.

La base n'est pas documentée, ce qui rend difficile l'estimation de son périmètre. Les plus anciens mandats sénatoriaux contenus remontent à 1948, mais ils sont peu nombreux, ce qui semble indiquer que la base n'est pas exhaustive quand on remonte trop loin dans le temps. On décompte 314 mandats en 1959 (pour 309 sièges possibles à cette époque), ce qui laisse supposer que la base contient tous les mandats à partir des élections du 26/04/1959, soit le début de la V^{ème} République.

3.2.2 Mise en correspondance et intégration

En préliminaire, on applique aux données du Sénat une normalisation similaire à l'Étape E1 afin qu'elles soient comparables à celles du RNE. Bien que ces données soient de bonne qualité, elles ne sont pas exemptes de toute erreur, aussi réalisons nous également environ 80 corrections *ad hoc* similaires à celles de l'Étape E3.

Les identifiants utilisés dans la base du Sénat pour identifier les sénateur·rices de façon unique sont différents de ceux du RNE, donc la première étape de notre traitement consiste à réaliser la mise en correspondance entre ces deux nomenclatures. La deuxième consiste à mettre en relation les lignes de `S` et celles de la table des mandats sénatoriaux, sur la base de leurs dates et de la correspondance entre identifiants. Enfin, nous pouvons réaliser l'intégration proprement dite.

Mise en correspondance des élu·es. Pour la mise en correspondance entre les élu·es, nous nous basons sur une comparaison de leurs nom, prénom, sexe, et date de naissance. Cette opération permet de mettre en évidence 10 divergences entre les informations décrivant les élu·es dans les deux bases (date de naissance, nom, prénom, sexe). Une vérification manuelle nous a permis de déterminer que c'était systématiquement le RNE qui était erroné, donc nous modifions l'Étape E3 afin de rectifier et obtenir ainsi des données compatibles dans les deux bases.

Une fois la mise en correspondance des identifiants effectuée, nous pouvons associer chaque sénateur·rice du RNE à un·e sénateur·rice de la base du Sénat, et bien entendu cette base possède également de nouvel·les sénateur·rices absent·es du RNE car elle remonte plus loin dans le temps. Nous ne pouvons cependant pas attribuer un identifiant de type RNE

à ces nouveaux sénateur·rices, car nous ne contrôlons pas la génération de ces ID, et toute valeur que nous créerions pourrait entrer en conflit avec un identifiant issu d'une extraction future du RNE. Pour cette raison, nous décidons de créer notre propre numéro d'identification, nommé `ID universel` dans la base, en procédant de la façon suivante. Si l'élu existe déjà dans le RNE, nous utilisons son numéro dans le RNE, en le reformatant : nous lui ajoutons le préfixe `RNE_` et portons son nombre de chiffres à 7 afin d'obtenir une longueur constante, en complétant avec des zéros de tête quand c'est nécessaire. Par exemple, un élu dont l'identifiant RNE est `724` obtiendra l'identifiant unifié `RNE_0000724`. Pour un élu apparaissant dans la base du Sénat mais *pas* dans le RNE, nous concaténons le préfixe `SEN_0` à son ID du Sénat, dont la longueur est déjà fixée à 7 caractères. Ainsi, un·e sénateur·rice d'identifiant `57826T` obtiendra l'identifiant unifié `SEN_057826T`.

Mise en correspondance des mandats. L'étape suivante consiste à faire correspondre les mandats de la base du Sénat à ceux déjà présents dans le RNE, afin d'identifier les différences et les manques. Pour cela, nous utilisons les ID des sénateur·rices concerné·es et les dates de début de mandat. Pour chaque ligne de `S`, nous cherchons la ligne correspondante dans la base du Sénat. Il faut noter que la base du Sénat fait explicitement la différence entre date d'élection et date de prise de fonction, ce qui n'est pas le cas du RNE.

Il apparaît assez rapidement que souvent, les dates ne correspondent pas exactement entre les deux tables, mais sont similaires à quelques jours près. Nous introduisons donc une tolérance de 14 jours afin de réaliser notre mise en correspondance. La période entre deux tours d'élection étant de 7 jours, cette durée de 14 jours permet d'assimiler une date proche d'un tour à la date exacte de ce tour. La date exacte à laquelle un mandat commence officiellement est difficile à déterminer : en fonction de la BD et même du mandat considéré, celle-ci peut commencer le jour de l'élection, mais aussi quelques jours plus tard. Pour simplifier le traitement, nous arrondissons les dates approximativement égales à une date d'élection, comme nous l'avons déjà fait pour les données du RNE à l'Étape E8. Cela permet de les comparer plus facilement.

Malgré la comparaison relaxée, 48 lignes du RNE ne trouvent pas d'équivalent dans la base du Sénat, car leur date de début est incorrecte de plus de 14 jours (et même plutôt de un an, voire de plusieurs années). Nous complétons la liste de corrections *ad hoc* réalisées à l'Étape E3 afin de les traiter. On décompte 15 autres lignes du RNE qui ne correspondent à aucun mandat réel, d'après la base du Sénat, ce qui représente 1% de la table `S`. Si on fait abstraction des mandats antérieurs à 2001 et postérieurs à la date d'extraction du RNE, on détecte 50 lignes présentes dans la base du Sénat mais absentes du RNE (soit 5%).

Intégration. L'intégration des données issues de sources secondaires constitue la quatorzième étape de notre traitement (**E14**). En ce qui concerne la base du Sénat, nous procédons en deux opérations. Tout d'abord, lorsque la mise en correspondance a pu être faite, les informations du Sénat sont utilisées pour mettre à jour la ligne correspondante dans notre base. Quand les deux sources divergent sur une date de début de mandat, une vérification manuelle révèle que celle de la base du Sénat est généralement la date correcte, aussi écrasons-nous celle du RNE en cas de désaccord entre les deux sources. La date d'extraction des données du Sénat étant postérieure à celle du RNE, il arrive qu'un mandat décrit comme en cours dans le RNE soit terminé dans la base de Sénat, auquel cas nous réalisons la mise à jour.

Lorsqu'aucune correspondance n'a pu être trouvée entre les deux tables, deux situations sont possibles. S'il s'agit d'un mandat antérieur à 2001, il est normal que celui-ci soit absent du RNE ; et sinon cela signifie que le mandat est manquant, et c'est donc une erreur du RNE. Dans les deux cas, nous rajoutons la ligne dans notre base pour palier cette omission. Le champ `Sources` introduit à l'Étape E5 est mis à jour de manière à identifier si une ligne provient du RNE, de la base du Sénat, ou des deux (i.e. donnée RNE corrigée grâce au Sénat).

La base du Sénat contient un certain nombre de champs absents du RNE, en particulier

la date éventuelle de décès du sénateur ou de la sénatrice, que nous rajoutons dans notre propre base. Les circonscriptions sénatoriales apparaissant dans la base du Sénat ne correspondent pas forcément à des départements actuels. On voit notamment apparaître les *Comores*, et des circonscriptions algériennes. Cela justifie notre décision à l'Étape E5 de renuméroter les départements de façon unique afin de tenir compte de leur évolution dans le temps.

Il faut souligner que certaines informations sont codées différemment dans la base du Sénat. C'est notamment le cas de la profession déclarée, qui est décrite par trois champs de niveau hiérarchique croissant dans une typologie de d'INSEE, au lieu d'une seule catégorie pour le RNE. Aucun des ces trois niveaux ne correspond exactement à la typologie du RNE, aussi avons nous décidé de conserver l'information du niveau le plus proche. La nuance politique n'est pas représentée en tant que telle dans la base du Sénat. L'information s'en rapprochant le plus est le groupe parlementaire auquel le sénateur ou la sénatrice appartient. Le motif de fin de mandat est beaucoup plus détaillé dans la base du Sénat, et semble de plus beaucoup plus fiable que dans le RNE. Il est en outre complété d'un commentaire en texte libre précisant d'éventuelles circonstances extraordinaires. Nous avons décidé de ne recoder aucun de ces champs pour l'instant (une tâche restant à réaliser) et de ne pas conserver ces commentaires.

Une autre différence importante est que certaines informations qui sont associées à une ligne dans le RNE (et donc à une période temporelle), sont représentées de façon statique dans la base du Sénat. Autrement dit, ces informations sont associées à un·e sénateur·rice sans indication de date et de façon unique (une seule valeur par sénateur·rice). C'est notamment le cas de la circonscription sénatoriale, du groupe politique, de l'éventuelle fonction occupée au bureau du Sénat, et de la profession. Ceci est une limitation de cette base, car cela suppose implicitement que ces attributs ne sont pas susceptibles d'évoluer dans le temps, ce qui est incorrect : un·e sénateur·rice peut changer de groupe, être élu·e successivement dans des circonscriptions différentes, occuper plusieurs fonctions, et changer de profession. En l'occurrence, nous n'avons accès qu'à la dernière valeur de ces champs [22]. Nous l'utilisons donc faute de mieux lors de l'intégration, mais n'écrasons pas la valeur originale du RNE (si celle-ci est déjà renseignée). Pour les fonctions, la conséquence de cette représentation liée à l'individu plutôt qu'au mandat est que la base du Sénat ne contient pas les dates de début ni de fin.

À l'issue de l'intégration, nous soumettons la table `S` ainsi complétée à un certain nombre des étapes déjà subies lors du traitement des données du RNE, afin de supprimer les éventuels problèmes que l'intégration aurait pu occasionner.

3.2.3 Bilan de l'intégration

La Table 21 montre les résultats obtenus lors de l'étape d'intégration des données secondaires et de la répétition des étapes déjà appliquées aux données brutes du RNE. Bien sûr elle se concentre uniquement sur la table `S`.

On peut observer que l'intégration introduit 2 075 lignes (soit +195%) dans la table issue du traitement de données brutes du RNE. Par la suite, l'alignement sur les dates d'élection (Étape E7) touche la plupart des lignes (72%), car les dates de la base du Sénat correspondent à la prise de fonction et non pas à l'élection. L'opération de fusion de mandats revouvrant (Étape E10) et l'opération réciproque de division des mandats long (Étape E11) sont également appliquées à de nombreuses lignes, et ont pour effet d'augmenter la taille de la table ($+487$, soit +16%). Ceci est dû au fait que dans la base du Sénat aussi, certains mandats consécutifs sont représentés de façon fusionnée, sur une seule ligne.

Au contraire, l'intégration des données du Sénat n'introduit aucune fusion de lignes

22. Il est possible que la version PostgreSQL de la base du Sénat ne souffre pas du même problème.

Étape		Modifications	Nbr. lignes	Différence
–	Données brutes	–	1 076	–
E1–E13	Traitement des données brutes	865	1 066	−10
E14	Intégration de la base du Sénat	3 141	3 141	+2 075
E6	Fusion des lignes compatibles	0	3 141	0
E7	Ajustement des dates de fonction	1	3 141	0
E8	Alignement aux dates d'élection	2 277	3 141	0
E9	Suppression des micro-mandats/fonctions	7	3 136	−5
E10	Fusion des mandats se recouvrant	1 235	1 901	−1 235
E11	Division des mandats longs	1 720	3 621	+1 720
E12	Résolution des recouvrements de position	0	3 621	0
E9	Suppression des micro-mandats/fonctions	349	3 272	−349
E13	Révision des motifs de fin	218	3 272	0
Valeurs globales		3 095 95%	3 272 –	+2 196 +204%

Table 21. Pour chaque étape appliquée sur S après le traitement des données brutes, cette table indique le nombre de lignes modifiées (ajouts et suppressions inclus), le nombre de lignes total, et l'évolution du nombre de lignes (par rapport à l'étape précédente). La dernière ligne indique le nombre de lignes modifiées au cours de l'intégralité du traitement, le nombre de lignes obtenu à la fin du traitement, et la différence par rapport aux données brutes du RNE.

compatibles (Étape E6), ni de résolution des recouvrements de position (Étape E12), ce qui signifie que ce type d'erreur est complètement absent de ces données.

La comparaison entre S et la base du Sénat, restreinte à la période couverte par le RNE, révèle quelles lignes du RNE sont incorrectes et quelles lignes en sont simplement absentes. Cela nous permet donc d'estimer les taux d'erreur et de complétude du RNE. Comme indiqué en Section 3.2.2, 15 de ses lignes (soit 1% de la table S) sont erronées, et il y manque 50 lignes (soit 5%) pour la période couverte.

3.3 Données de l'Assemblée Nationale

Comme pour le Sénat, nous décrivons d'abord la base de données publiques de l'Assemblée Nationale (Section 3.3.1), puis nous décrivons comment nous l'avons intégrée à notre base (Section 3.3.2), avant de résumer en quoi cela a modifié nos données (Section 3.3.3).

Les données de l'Assemblée Nationale sont disponibles à travers deux bases distinctes : l'une prend la forme d'extractions tabulaires et SQL, tandis que l'autre, appelée Sycomore, n'est accessible que très indirectement, via une interface Web. Nous nous intéressons ici à la première, et nous revenons plus loin sur la seconde (cf. Section 4.3.2).

3.3.1 Présentation

La base de données de l'Assemblée Nationale [23] donne des renseignements sur l'état-civil des député·es et sur leurs mandats, mais à la différence de celle du Sénat, elle se consacre uniquement aux mandats de député·es et aux activité qui leur sont directement liées (assemblées parlementaires internationales, organismes extra-parlementaires).

Comme pour le Sénat, la base n'est pas documentée. Son descriptif indique qu'elle est supposée couvrir toutes les législatures depuis la XI^{ème}, soit juin 1997. Cependant, en pratique les mandats les plus anciens qu'elle contient commencent en 2002 (XII^{ème} législature). On trouve 560 mandats débutant en 2002, ce qui correspond à peu près au nombre de député·es pour cette période (577) et indique donc que la base peut être supposée exhaustive à partir de cette date.

23. http://data.assemblee-nationale.fr/acteurs/historique-des-deputes

Concrètement, elle prend la forme d'une collection de fichiers disponibles aux formats XML et JSON. Cette collection se divise en deux parties, l'une décrivant les député·es et l'autre les organes dans lesquels il·elles interviennent. Les fichiers décrivant les député·es contiennent différentes informations personnelles, et la liste des mandats et fonctions occupés dans le cadre de leur députation, avec une référence aux organes concernés. Les fichiers décrivant les organes donnent un ensemble d'informations sur chacun, telles que leur nom, description, période d'activité, site Web, etc. Nous nous sommes concentrés exclusivement sur l'exploitation des mandats de député·e, mais les autres ressources sont intéressantes, et devraient être intégrées à la BRÉF dans le futur.

3.3.2 Mise en correspondance et intégration

Une analyse rapide révèle que les données de la base de l'Assemblée sont globalement moins fiables que celles du Sénat. On identifie ainsi 2 cas d'élu·es associés à des identifiants distincts, un grand nombre de champs personnels non-renseignés (1014/2270), des mandats partiels (par exemple suite à un remplacement) dont les dates correspondent à un mandat complet, et qui interfèrent donc avec une autre entrée de la base (par exemple la personne remplacée), voire dédoublant un mandat partiel finissant/commençant en même temps. Pour ces raisons, nous appliquons aux données de l'Assemblée les étapes pertinentes du traitement déjà réalisé sur le RNE et la base du Sénat, et décrites en Section 3.1.

La base de l'Assemblée Nationale possède ses propres identifiants, différents de ceux du Sénat et du RNE. Nous procédons donc sur le même principe que pour le Sénat (cf. Section 3.2.2), à savoir : mise en relation des identifiants d'élu·es, puis mise en relations des lignes (i.e. mandats), et enfin intégration proprement dite.

Mise en correspondance des élu·es. Nous procédons comme pour le Sénat, en nous basant sur la comparaison des nom, prénom, sexe et date de naissance. Nous détectons 12 divergences, essentiellement sur la date de naissance, pour lesquelles une vérification manuelle révèle que c'est la version de l'Assemblée Nationale qui est correcte au détriment du RNE.

Pour les identifiants des député·es qui ne sont pas déjà présents dans notre base (qu'ils aient été obtenus du RNE ou de la base du Sénat), nous créons un identifiant universel à partir de l'identifiant de la base de l'Assemblée. Celui-ci est de la forme `PAxxxxxx`, où `xxxxxx` est une valeur numérique constituée de 4 à 6 chiffres. Nous complétons la valeur numérique avec des zéros pour obtenir 6 chiffres le cas échéant, et remplaçons le préfixe `PA` par `ASN_0` pour garder la trace de la source de l'identifiant. Ainsi, un·e député·e identifié·e par le code `PA1001` dans la base de l'assemblée obtiendra le code `ASN_0001001` dans notre base.

Mise en correspondance des mandats. La mise en correspondance des mandats est effectuée de la même façon que pour le Sénat. La base de l'Assemblée ne distingue pas date d'élection et date de prise de fonction, elle indique seulement une date de début de mandat. Nous conservons la même tolérance qu'indiqué en Section 3.2.2 lors de la comparaison des dates, et nous alignons également les dates de début/fin sur les dates d'élection (E8).

Le traitement nécessite la correction de 123 lignes du RNE dont les dates se révèlent trop différentes de celles de l'Assemblée pour pouvoir être considérées comme similaires, même en tenant compte de la tolérance. Nous détectons 11 lignes de `D` (soit moins de 1%) ne correspondant à aucun mandat dans la base de l'Assemblée, et une vérification manuelle confirme qu'il s'agit de mandats fictifs. À l'inverse, 115 lignes (soit 5% de `D`) présentes dans la base de l'Assemblée sont complètement absentes du RNE. La période 2001-2002 n'étant pas couverte par la base de l'Assemblée, cette valeur sous-estime le nombre lignes manquant du RNE.

Intégration. Le principe de l'intégration des données de l'Assemblée est le même que pour celles du Sénat. Comme pour le Sénat, la date d'extraction de la base est ultérieure à celle du RNE, et certains mandats représentés comme en cours dans le RNE sont terminés dans

la base de l'Assemblée, auquel cas nous mettons à jour nos données de manière à tenir compte de cette évolution. À noter que les données de l'Assemblée présentent des cas d'élu·es cumulant plusieurs fonctions parlementaires simultanément.

Cette base contient elle aussi certains champs absents du RNE : date de décès, lieu de naissance (commune, département, pays), que nous conservons. Le codage de certains champs diffère de celui du RNE (et du Sénat) : profession et motif de fin de mandat. La profession utilise une troisième nomenclature, différente de celle du RNE et du Sénat, que nous conservons telle quelle pour l'instant. De plus, elle est indiquée dans la table décrivant le·a député·e et non pas le mandat, ce qui signifie que nous ne disposons que de la profession en cours. Le motif de fin de mandat est lui aussi représenté différemment par rapport au RNE et au Sénat, et il est plus précis que dans le RNE. Ni le parti politique ni le groupe parlementaire ne sont disponibles.

Comme pour le Sénat, à l'issue de l'intégration nous soumettons la table `D` complétée au traitement déjà appliqué aux données brutes du RNE, afin de supprimer les problèmes éventuellement apparus lors de l'intégration.

3.3.3 Bilan de l'intégration

La Table 22 contient les résultats obtenus lors de l'étape d'intégration des données de l'Assemblée dans la table `D`. On peut voir que la base de l'Assemblée apporte bien moins d'information supplémentaire que celle du Sénat : +161 lignes, soit +7%. Ceci est dû au fait que la base du Sénat couvre toute la V^{ème} république, alors que celle de l'Assemblée décrit la même période que le RNE.

Étape		Modifications	Nbr. lignes	Différence
–	Données brutes	–	2 488	–
E1–E13	Traitement des données brutes	1 419	2 451	−37
E14	Intégration de la base de l'Assemblée	2 520	2 616	+165
E6	Fusion des lignes compatibles	0	2 616	0
E7	Ajustement des dates de fonction	35	2 616	0
E8	Alignement aux dates d'élection	1 924	2 616	0
E9	Suppression des micro-mandats/fonctions	1	2 615	−1
E10	Fusion des mandats se recouvrant	446	2 169	−446
E11	Division des mandats longs	386	2 555	+386
E12	Résolution des recouvrements de position	113	2 555	0
E9	Suppression des micro-mandats/fonctions	6	2 549	−6
E13	Révision des motifs de fin	0	2 549	0
Valeurs globales		2 543	2 549	+784
		100%	–	+32%

Table 22. Pour chaque étape appliquée sur `D` après le traitement des données brutes, cette table indique le nombre de lignes modifiées (ajouts et suppressions inclus), le nombre de lignes total, et l'évolution du nombre de lignes (par rapport à l'étape précédente). La dernière ligne indique le nombre de lignes modifiées au cours de l'intégralité du traitement, le nombre de lignes obtenu à la fin du traitement, et la différence par rapport aux données brutes du RNE.

L'application des étapes duales de fusion des mandats se recouvrant (Étape E10) et de division des mandats longs (Étape E11) ne change pas le nombre de lignes (ce qui ne veut pas dire qu'elles ne modifient pas les dates délimitant les mandats). Par rapport à l'intégration des données du Sénat, on remarque un bon nombre de modifications dues à des recouvrements de position (Étape E12), une étape qui ne s'applique pas à `S` puisque les positions n'y sont pas identifiées de façon unique (cf. Section 3.1.12).

Pour mémoire, comme indiqué en Section 3.3.2, l'estimation des taux d'erreur et de complétude du RNE pour `D` sont respectivement de 1% (11 lignes) et 5% (115 lignes), ce qui est

du même ordre que pour **S**.

3.4 Données du Parlement Européen

En plus des bases de données issues du Sénat et de l'Assemblée, nous intégrons à notre base des données provenant de sources non-structurées. Nous nous concentrons sur les mandats de député·es européen·nes, car **DE** souffre d'un bon nombre de problèmes (cf. Section 2.5.2), et le contingent de député·es européen·nes est suffisamment petit pour pouvoir être traité manuellement.

3.4.1 Processus d'intégration

Nous exploitons le site officiel du Parlement Européen [24] (PE) et les pages Wikipédia décrivant les résultats des élections européennes [25]. Nous récupérons les informations personnelles équivalentes à celles déjà présentes dans le RNE (nom, prénom, date de naissance, etc.) ainsi que les dates de mandats. Nous n'incluons pas les fonctions, car celles-ci sont originellement absentes de **DE**, mais cela pourrait être fait dans le futur, à partir des informations du site du PE.

Comme pour le Sénat et l'Assemblée, ces informations personnelles sont ensuite utilisées pour mettre en correspondance les élu·es puis leurs mandats. Pour ce qui est de la période couverte, en raison du faible nombre d'élu·es, nous remontons jusqu'à la première législature européenne, et intégrons donc tous et toutes les parlementaires français·es. Certain·es d'entre eux ne sont présent·es ni dans le RNE ni dans les bases de l'Assemblée ou du Sénat, et ne disposent donc d'aucun identifiant. Comme précédemment, nous en créons un à partir du système utilisé par le PE pour numéroter ses propres élu·es. Ce code prend la forme d'une valeur numérique d'au plus 6. Nous complétons avec des zéros à gauche pour atteindre 7 chiffres, et accolons le préfixe **EUR_** pour obtenir notre identifiant universel.

Une fois les nouveaux mandats intégrés à **DE**, nous appliquons à nouveau le traitement standard (Section 3.1) au résultat de cette fusion. Si on se concentre uniquement sur la période supposémment couverte par le RNE (i.e. de 2001 à 2018), 52 mandats (soit un taux d'incomplétude de 20% de **DE**) sont absentes du RNE, tandis que 96 mandats n'ont pas de réalité (soit un taux d'erreur de 37%). Plus globalement, les données intégrées couvrent l'ensemble des mandats de député·es européen·nes élu·es au suffrage universel direct soit depuis 1978.

3.4.2 Bilan de l'intégration

La Table 23 contient les résultats obtenus lors de l'étape d'intégration des données secondaires dans la table **DE**. La modification est importante, d'abord par le nombre de lignes insérées : 792, soit 318% de **DE** à l'issue du traitement des données du RNE. Les données insérées sont constituées manuellement, donc les corrections apportées à la suite de leur insertion sont minimes.

Les étapes duales de fusion des mandats se recouvrant (Étape E10) et de division des mandats longs (Étape E11) concernent comme toujours un bonne proportion de la table, et contribuent globalement à réduire le nombre de lignes. Rappelons que pour la période originalement couverte pour le RNE, le taux d'erreur est de 37% et celui d'incomplétude de 20%.

24. https://www.europarl.europa.eu/
25. https://fr.wikipedia.org/wiki/Liste_des_députés_européens_de_France_de_la_9e_législature

Étape		Modifications	Nbr. lignes	Différence
–	Données brutes	–	253	–
E1–E13	Traitement des données brutes	260	260	+7
E14	Intégration de la source secondaire	792	1052	+792
E6	Fusion des lignes compatibles	33	1 019	−33
E7	Ajustement des dates de fonction	0	1 019	0
E8	Alignement aux dates d'élection	6	1 019	0
E9	Suppression des micro-mandats/fonctions	2	1 017	−2
E10	Fusion des mandats se recouvrant	297	720	−297
E11	Division des mandats longs	173	893	+173
E12	Résolution des recouvrements de position	0	893	0
E9	Suppression des micro-mandats/fonctions	0	893	0
E13	Révision des motifs de fin	1	893	0
Valeurs globales		481 54%	893 –	+640 +253%

Table 23. Pour chaque étape appliquée sur `DE` après le traitement des données brutes, cette table indique le nombre de lignes modifiées (ajouts et suppressions inclus), le nombre de lignes total, et l'évolution du nombre de lignes (par rapport à l'étape précédente). La dernière ligne indique le nombre de lignes modifiées au cours de l'intégralité du traitement, le nombre de lignes obtenu à la fin du traitement, et la différence par rapport aux données brutes du RNE.

3.5 Fusion des tables

L'intégration des sources secondaires a permis de résoudre au moins en partie un certain nombre de problèmes ouverts : P21 (fonctions manquantes dans certaines tables), P52 (élus manquants), P58 (couverture incomplète), P59 (mandats manquants), et P60 (mandats surnuméraires).

Le traitement suivant consiste à fusionner les différentes tables préparées lors des étapes précédentes, et de soumettre la table unique ainsi obtenue à un certain nombre de tests et de modifications additionnelles.

3.5.1 Description du traitement

À ce stade du traitement, nous disposons d'une version nettoyée et éventuellement complétée de chacune des huit tables correspondant aux types de mandats du RNE (`CD`, `CM`, `CR`, `D`, `DE`, `EPCI`, `M`, `S`), via l'application des opérations décrites dans les Sections 3.1 à 3.4. Comme nous allons le voir à présent, le processus de fusion ne consiste pas simplement à les concaténer.

Fusion des tables municipales. Comme mentionné en Section 1.2.2, nous soupçonnons que les données constituant `M` sont également présentes dans `CM`. Nous ne pouvons donc pas simplement concaténer ces deux tables, au risque de produire de nombreux doublons. Nous n'avons pas réalisé de comparaison exhaustive des deux tables lors de nos tests, car les imperfections présentes dans les données rendaient cette tâche plus difficile et peu fiable. Cela n'est plus le cas maintenant qu'elles ont été traitées, ce qui nous permet de mettre en place un processus de fusion spécifique aux tables municipales (**E15**).

Nous nous concentrons d'une part sur les lignes de `CM` relatives à des fonctions de maire, et de l'autre sur l'intégralité de `M`. Nous réalisons d'abord la mise en relation de chaque ligne de `M` avec celles retenues dans `CM`, puis le contraire (chaque ligne retenue de `CM` dans `M`). Nous effectuons d'abord cette mise en relation de façon exacte, avec succès pour 88% des lignes. Autrement dit, ces lignes sont parfaitement identiques dans les deux tables. En autorisant une tolérance de 7 jours lors de la comparaison des dates, nous en mettons en relation 11% de plus. Les lignes restantes (1%) sont toutes mises en relation en relaxant encore les critères

de comparaison, et en identifiant ainsi des périodes qui se recouvrent temporellement. Au final, nous obtenons une relation biunivoque parfaite entre les deux ensembles de lignes.

À noter que les différences existant entre certaines lignes (12% en tout) peuvent être aussi bien dues aux données brutes du RNE qu'aux différentes transformations que nous leur appliquons au cours de notre propre traitement. Enfin de réconcilier les deux versions divergentes d'une même ligne (celle de `M` et celle de `CM`), nous appliquons à cette paire de lignes l'Étape de fusion des mandats se recouvrant (E10). Nous réintégrons dans `CM` les lignes ainsi fusionnées, et supprimons purement et simplement `M`, dont nous savons maintenant avec certitude qu'elle est redondante.

Concaténation. Après avoir réalisé le traitement de fusion spécifique aux données municipales, nous construisons une table unique en concaténant les 7 tables de données restant (**E16**) : `CD`, `CM`, `CR`, `D`, `DE`, `EPCI`, et `S`. La taille de la table fusionnée, en termes de nombre de lignes, est donc la somme des tailles de ces sept tables.

Lorsqu'une colonne est présente dans une table mais pas dans une autre, comme par exemple le code du canton dans `CD`, qui n'apparaît nulle part ailleurs, alors une colonne vide (i.e. contenant uniquement des `NA`) est rajoutée à cette dernière avant la concaténation. Au total, la table fusionnée contient 47 colonnes.

Mandats présidentiels. Après la fusion, nous rajoutons à la table obtenue les mandats de président de la république (**E17**), la seule position élective absente du RNE. Elle concerne très peu de personnes, aussi la traitons-nous manuellement. Nous exploitons la page Wikipédia listant les présidents [26], ainsi que leurs pages personnelles dans cette encyclopédie.

Nous procédons comme pour les autres types de mandats déjà complétés (`S`, `D`, `DE`). Les élu·es n'apparaissant dans aucune autre source déjà traitée sont désigné·es par un identifiant universel de la forme `PRF_xxxxxxx` où `xxxxxxx` est un nombre à 7 chiffres attribué arbitrairement dans l'ordre d'insertion des élu·es.

Opérations complémentaires. Toutes les étapes décrites dans la Section 3.1 ont déjà été appliquées à chaque table séparément, et parfois plusieurs fois. Pour la plupart d'entre elles, il n'y aurait pas d'intérêt de les effectuer une fois de plus, car la présence des lignes des autres tables n'apporterait pas de nouveaux cas à traiter. De ce point de vue, l'Étape 6 est une exception, car il est possible que des lignes provenant de tables différentes soient compatibles, et doivent donc être fusionnées. Par mesure de sécurité, nous appliquons donc cette étape de fusion des lignes compatibles à la table fusionnée et complétée.

Enfin, nous réalisons un traitement visant à compléter les identifiants manquants (**E18**). En effet, certain·es élu·es occupent plusieurs types de mandats, et sont donc susceptibles d'avoir un identifiant d'un type donné dans certaines de leurs occurrences mais pas dans d'autres, en fonction de la table d'origine. Par exemple, un·e député·e-sénateur·rice aura un identifiant du Sénat issu de `S`, mais apparaîtra dans `D` sans cet identifiant. Par souci de complétude, nous insérons les identifiants manquants pour chaque élu·e, en nous basant sur notre propre identifiant unique pour réaliser la mise en correspondance. Nous réalisons la même opération avec les informations personnelles susceptibles d'être présentes dans certaines sources et pas d'autres, notamment le lieu de naissance et la date de décès (**E19**).

3.5.2 Bilan de la fusion

Les différentes étapes de la fusion sont décrites dans la Table 24. La première ligne correspond à l'ensemble des lignes contenues dans les tables brutes, pour référence. La deuxième (E15) résume la fusion des deux tables municipales `CM` et `M`, qui aboutit concrètement à la mise à jour de certaines lignes de `CM` sur la base de celles de `M`, mais n'augmente pas le nombre de lignes de `CM`. Vient ensuite la concaténation du reste des tables traitées (E16)

26. https://fr.wikipedia.org/wiki/Président_de_la_République_française#Cinquième_République

ainsi que des données présidentielles (E17), qui donnent à la table fusionnée sa taille finale. En effet, l'étape de fusion des lignes compatibles (E6) n'entraîne aucune modification. La complétion des identifiants (E18) modifie un nombre important de lignes, mais ne change pas leur nombre puisqu'il s'agit uniquement de donner une valeur à certains champs vides.

Étape		Modifications	Nbr. lignes	Différence
–	Données brutes	–	1 503 238	–
E15	Fusion des données municipales	13 278	1 224 192	–
E16	Concaténation des 7 tables	–	1 374 706	+150 514
E17	Intégration des mandats présidentiels	13	1 374 719	+13
E18	Complétion des identifiants manquants	8 195	1 374 719	0
E19	Complétion des données personnelles	7 410	1 374 719	0
E6	Fusion des lignes compatibles	0	1 374 699	0
Valeurs globales		916 663 67%	1 374 719 –	−128 519 −8%

Table 24. Pour chaque étape appliquée sur la table des données fusionnées, cette table indique le nombre de lignes modifiées (ajouts et suppressions inclus), le nombre de lignes total, et l'évolution du nombre de lignes (par rapport à l'étape précédente). La dernière ligne indique le nombre de lignes modifiées au cours de l'intégralité du traitement, le nombre de lignes obtenu à la fin du traitement, et la différence par rapport aux données brutes du RNE.

4 Résultat du traitement

Dans cette section, nous décrivons la base de données obtenue à l'issue de notre traitement (Section 4.1), nous comparons ses limitations avec celles du RNE original (Section 4.2), et nous discutons les points restant à traiter (Sections 4.3 et 4.4).

4.1 Description de la table fusionnée

Après l'application du traitement décrit dans la Section 3, la table fusionnée contient un total de 1 374 699 lignes. La répartition entre les différents types de mandats est décrite dans la Table 25, qui la compare à celle observée dans les données brutes (avant traitement). En raison de la fusion de `M` dans `CM` (cf. Section 3.5.1), la colonne `CM` inclut les maires : 227 044 sans traitement et 114 194 avec.

Les types de mandats les plus affectés par le traitement, en termes de nombre de lignes, sont `DE` et `S`, en raison des nombreux ajouts. Les suppressions qui affectent `EPCI` correspondent essentiellement aux lignes si incomplètes qu'elles en sont rendues inexploitables pour nous. Le grand nombre de lignes supprimées de `CM` sont pour la plupart celles des maires, qui apparaissaient également dans `M` et étaient donc redondantes.

Type de mandat	CD	CM	CR	D	DE	EPCI	S	**Tout**
Données brutes	12 728	1 352 935	5 818	2 488	253	127 940	1 076	1 503 238
Table fusionnée	12 766	1 224 316	5 805	2 549	893	125 230	3 272	1 374 719
Différence	+39	−128 766	−12	+61	+640	−2 710	+2 196	−128 519
	+0,3%	−9%	−0,2%	+2%	+253%	−2%	+204%	−9%

Table 25. Répartition des types de mandat dans les données brutes et dans la table fusionnée issue du traitement.

La table fusionnée contient tous les champs du RNE, listés dans la Table 2, plus ceux décrits en Table 26, qui sont insérés au cours du traitement, en particulier l'intégration des sources secondaires. La plupart d'entre eux ne sont pas renseignés systématiquement, car soit ils concernent certains types de mandats et pas les autres, soient ils sont issus de sources secondaires incluant des mandats n'apparaissant pas originellement dans le RNE.

Champ	Nature	Description
Identifiant du département	ID	Code unique attribué pour résoudre P41
Identifiant du canton	ID	Code unique attribué pour résoudre P43
Identifiant AN de l'élu·e	ID	Code unique attribué par l'Assemblée Nationale
Identifiant Sénat de l'élu·e	ID	Code unique attribué par le Sénat
Identifiant PE de l'élu·e	ID	Code unique attribué par le Parlement Européen
Identifiant interne de l'élu·e	ID	Code unique spécifique à la BRÉF
Libellé de la commune de naissance	Nom	Information issue de certaines sources secondaires
Code du département de naissance	ID	Information issue de certaines sources secondaires
Pays de naissance	Nom	Information issue de certaines sources secondaires
Date de décès	Date	Information issue de certaines sources secondaires
Nationalité de l'élu·e	Nom	Information issue de certaines sources secondaires
Sources	Catégories	Liste des sources documentant la ligne
Corrections de date	Booléen	Indique si des dates de la ligne ont été modifiées
Autres corrections	Booléen	Indique si d'autres champs ont été modifiés

Table 26. Champs présents dans la table fusionnée, en plus de ceux existant dans le RNE et déjà listés dans la Table 2.

Ainsi, les identifiants de département et de canton ne concernent pas `DE`. Nous les avons constitués en préalable au traitement, afin de distinguer des entités portant le même code dans les RNE et/ou les sources secondaires (P41, P43). Chaque source attribue son propre identifiant à un·e élu·e, aussi nous les stockons tous et toutes dans la BRÉF. Un·e élu·e a

au moins un identifiant, dit *interne*, que nous construisons pour la BRÉF. Mais il peut avoir jusqu'à 4 autres identifiants utilisés dans le RNE et les bases du Senat, de l'Assemblée ou du Parlement Européen. Les champs relatifs à la naissance et au décès des élu·es sont présents dans certaines sources secondaires, et nous avons décidé de les inclure dans la BRÉF. Bien entendu, ils ne sont pas renseignés pour les élu·es n'apparaissant pas dans ces sources-là.

Enfin, le champ *Sources* permet de stocker les différentes sources utilisées pour constituer une ligne de la BRÉF. Ainsi, pour une ligne présente dans le RNE puis mise à jour en utilisant les données du Sénat (ou même existant à l'identique dans ces données), ce champ contiendra ces deux informations. Les deux derniers champs sont booléens et indiquent le fait que la ligne ait subi deux types de correction distincts durant le traitement : correction d'une date (n'importe laquelle) et correction d'un champ autre qu'une date (là encore, n'importe lequel). Ces deux champs sont conçus dans le but de faciliter le débogage de la BRÉF.

4.2 Erreurs résiduelles

Le traitement décrit dans la Section 3 a permis de résoudre un grand nombre des erreurs repérées dans la Section 2. Cependant, tous les problèmes n'ont pas pu être résolus, et l'insertion de données issues de sources secondaires a même parfois introduit de nouvelles erreurs. Nous nous intéressons d'abord aux problèmes concernant les valeurs prises séparément (Section 4.2.1), puis à ceux impliquant de prendre en compte des lignes (Section 4.2.2), avant de faire le bilan des problèmes ouverts (Section 4.2.3).

4.2.1 Valeurs manquantes ou erronées

La Table 27 fait le décompte des erreurs non-résolues relatives aux champs renseignés de façon incorrecte ou pas du tout, et elle est à comparer aux Tables 3 (valeurs manquantes), 4 (dates incohérentes) et 6 (valeurs problématiques diverses) de la Section 2. Les lignes *Total* et *Différence* montrent que si un grand nombre d'erreurs a été corrigé, en revanche pour certaines tables l'intégration de nouvelles données a aussi introduit de nouvelles erreurs.

Informations personnelles. Les élu·es pour lesquel·les le RNE ne précisait pas de date de naissance (P23) ou une date incohérente (P24) ont été complètement traités. De même pour la cinquantaine d'élu·es sans prénom (P5).

Le nombre de nuances politiques manquantes (P16) a augmenté pour deux raisons. La première, majeure, est que pour résoudre le problème P13, la valeur *NC* a été remplacée par *NA*. La seconde, mineure, est que certaines sources secondaires ne contiennent pas cette information.

Nous n'avons pas mis en oeuvre de traitement spécifique pour compléter les codes et libellés de profession manquant (P17). Par conséquent, leur nombre dans `CM` et `M` n'a quasiment pas changé. Il a même augmenté dans les autres tables, en particulier `D`, `DE` et `S`, en raison de l'intégration de sources secondaires ne contenant que le libellé (pas de code) et un libellé relevant d'une typologie différente qui plus est.

Pour ce qui est des incorrections présentes dans les valeurs renseignées, la plupart des problèmes (P1, P2, P3, P4, P6, P7, P14) sont complètement résolus par notre traitement (du moins toutes les occurrences qui ont pu être détectées). Le problème concernant la nature relative de certaines nuances (P15) reste quant à lui ouvert : son traitement nécessite le recodage général des nuances, y compris les valeurs intégrées à partir des sources secondaires.

Les cas des élu·es associés à plusieurs identifiants distincts (P48) ont tous été traités, du moins tous ceux que nous avons pu détecter. C'est aussi le cas pour les occurrences du problème inverse, i.e. celui des identifiants associés à plusieurs élu·es distinct·es (P49). Cependant, à la différence du premier problème, il n'est pas possible d'automatiser la détec-

tion du second. Il est probable que notre examen manuel est passé à côté de nombreuses occurrences.

Pb	Champ	CD	CM	CR	D	DE	EPCI	M	S
P39	Code du département	0	0	0	0	–	342	0	0
P45	Code INSEE commune	–	0	–	–	–	341	0	–
P12	Libellé de la commune	–	0	–	–	–	341	0	–
P16	Nuance politique	56	640 079	12	78	6	23 002	8 847	0
P17	Code de profession	3	3 154	1	83	532	8	152	1 987
P17	Libellé de profession	3	3 154	1	8	532	8	152	0
P18	Motif de fin de mandat	2	481	2	3	0	2 467	3	0
P31	Date début fonction	0	0	0	0	–	0	0	324
P32	Date de fin de fonction	0	0	0	0	–	0	0	299
P19	Motif de fin de fonction	76	8 501	15	71	–	8 960	2 297	303
P29	Mandats se recouvrant	0	–	184	0	0	–	–	0
P37	Fonct. se recouvrant	10	651	2	0	–	27	0	0
Total		148	656 186	218	243	1 070	35 492	11 450	2 913
Différence		−7 475	+363 753	−3 064	−1 502	+667	−15 342	−1 063	+2 492

Table 27. Erreurs relevées dans les Tables 3, 4 & 6, et n'ayant pas été résolues par le traitement. Les lignes présentes dans ces tables-là mais absentes de celle ci-dessus correspondent à des problèmes complètement résolus par le traitement. Les tirets (–) signalent les cas où un champ donné est complètement absent d'une table de données. La colonne *Pb* contient le code du problème correspondant décrit dans le texte. La partie supérieure de la table concerne les valeurs manquantes, tandis que la partie inférieure (les deux dernières lignes) porte sur les valeurs incohérentes, comme expliqué en Section 2. La ligne dénotée *Différence* correspond à l'écart entre les données *avant* et *après* l'application de notre traitement.

Territoires. Les codes de département manquants (P39) ont été complétés en intégralité pour `CD` et en grande partie pour `EPCI` (92% des cas). Les codes restants à traiter correspondent à des lignes insuffisamment renseignées pour que notre stratégie de recoupement de l'information fonctionne (cf. Section 3.1.4). Il s'agit des mêmes cas pour lesquels le code INSEE de la commune (P45) et son nom (P12) n'ont pas non plus pu être complétés. Ces cas restant pourraient être traités manuellement, à un coût temporel assez élevé bien que l'information soit facilement accessible.

Comme pour les informations personnelles, la plupart des erreurs détectées dans les valeurs renseignées sont corrigées par notre traitement (P1, P2, P3, P6, P7, P8, P9, P10, P11, P40, P41, P44, P47). Les problèmes d'unicité des numéros de circonscriptions législatives (P42), de cantons (P43) et communes (P46) restent ouvert et requièrent un recoupement avec des données secondaires longitudinales et suffisamment détaillées.

Mandats et fonctions. Les 1 269 lignes d'`EPCI` ne contenant aucune date (ni début ni fin, ni mandat ni fonction, P25) ont été purement et simplement supprimées car jugées inexploitables en l'état (E4). Cela résout le problème des dates de début de mandat manquantes (P25), qui ne touchait qu'`EPCI`. Une approche plus précautionneuse consisterait à recouper les lignes d'`EPCI` concernées avec `CM` afin d'y récupérer une approximation raisonnable des dates de mandats. Certains motifs de fin de mandat manquants (P18) ont pu être complétés automatiquement (E13), mais le cas restants requièrent d'exploiter une source secondaire appropriée.

Les dates de début et fin de fonction manquantes (P31 & P32) ont été complétées pour toutes les tables sauf `S`. En effet, pour cette dernière la source secondaire exploitée contenait le libellé des fonctions occupées, mais pas leurs dates, ce qui a augmenté le nombre de valeurs. Nous n'avons pas mis en oeuvre de traitement particulier pour compléter les motifs de fin de fonction manquants (P19), aussi leur nombre n'a-t-il pas diminué significativement. Il a même augmenté pour `S`, pour les mêmes raisons que précédemment.

Pour ce qui est des valeurs renseignées, notre traitement résout la plupart des problèmes. En ce qui concerne la cohérence des dates de mandats et fonctions, il s'agit de P26, P27, P28, P30, P33, P34, P35, P36, P38, tandis que ceux portant sur les mandats et fonctions se recouvrant (P29 & P37) ne sont pas complètement résolus. La table `CR` contient encore 185 mandats se recouvrant, tandis que les tables `CD`, `CM`, `CR`, et `EPCI` contiennent des fonctions se recouvrant. Il faut souligner que ces nombres sont assez réduits, en particulier pour `CM` le traitement résout 94% des occurrences de problème. Le traitement des cas résiduels nécessite une coûteuse intervention manuelle : l'information décrivant la composition historique de certaines de ces institutions est difficile à trouver (conseils régionaux, communautés de communes). Ceci est plus marqué encore pour les fonctions, car les erreurs restant concernent des fonctions peu documentées comme *secrétaire* ou *6ème vice-président*.

En ce qui concerne les libellés, notre traitement ne résout que très ponctuellement le problème d'hétérogénéité des fonctions (P20), aussi on peut le considérer comme ouvert. L'intégration des données du Sénat et de l'Assemblée nationale ont permis de traiter le problème des fonctions incomplètes (P21), au moins pour `D` et `S`. Cependant il reste entier pour `CD` et `CR`. La situation est la même pour le problème de la sémantique variable des labels de motif de fin de mandat/fonction (P22), qui nécessiterait un recoupement à grande échelle avec une source extérieure pour être complètement éradiqué.

4.2.2 Autres erreurs

Problèmes divers. Le traitement fusionne toutes les lignes compatibles, donc ce problème (P50) peut être considéré comme réglé.

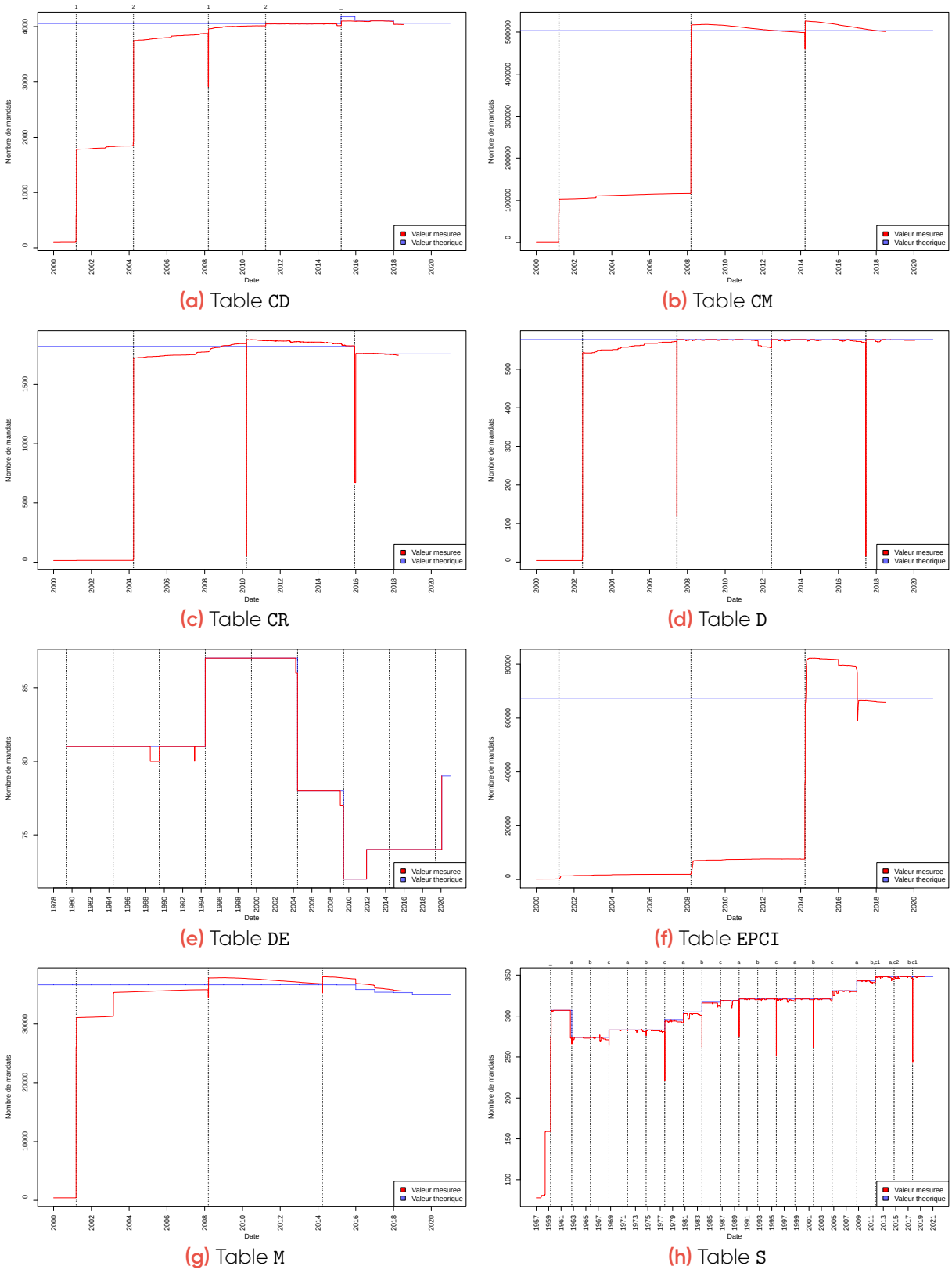
Un certain nombre de territoires incorrects (P51) font l'objet de corrections *ad hoc* (E3). Le problème est complètement résolu pour ces cas détectés manuellement, mais en l'absence de méthode automatique pour les identifier, il est impossible de déterminer si d'autres erreurs de ce type subsistent dans notre base.

La situation est la même pour les élu·es manquant·es (P52). Un certain nombre a été détecté et traité manuellement, tandis que l'intégration des données du Sénat, de l'Assemblée Nationale, et du Parlement Européen, a permis d'en détecter et d'en insérer un nombre plus conséquent dans `S`, `D` et `DE`. Un recoupement avec des sources plus complètes ou pourtant sur les autres tables est nécessaire pour s'assurer que le reste des données est complet.

Aucun traitement particulier n'a été réalisé pour les types particuliers de conseiller·es absents du RNE (P53, P54, P55, P56, P57) et détaillés en Section 2.4.3 (Assemblée de Corse, Martinique, Guyane, Nouvelle-Calédonie, conseiller·es territorial·es, conseiller·es d'arrondissement). Ces problèmes restent donc ouverts, et leur résolution nécessite l'identification et l'exploitation de sources secondaires appropriées.

Couverture temporelle. Comme détaillé dans la Section 2.4.4, la couverture temporelle du RNE est incomplète (P58). La Figure 2 montre l'évolution du nombre de mandats dans chaque table après le traitement, et elle est à comparer avec la Figure 1 qui se concentre sur les données brutes (non-traitées). Attention, si la Figure 1 se concentre systématiquement sur la période 2001–2018, supposée couverte par le RNE, en revanche la Figure 2 montre parfois des périodes démarrant avant 2001, afin de refléter le cas échéant, le résultat de notre traitement.

La Figure 2 montre que l'ensemble du traitement effectué, ainsi que l'intégration des données issues des sources secondaires, ont permis de grandement diminuer l'étendu du Problème P58, au moins dans certaines tables. Pour les tables qui ne sont pas directement concernées par l'intégration des sources secondaires, et/ou pour lesquelles il est difficile de trouver des informations permettant de corriger le problème, c'est à dire `CD`, `CM`, `CR`, `M` et `EPCI`, le traitement n'a que peu d'effet. Cela est également valable pour les cas de mandats manquants (P59) ou surnuméraires (P60) et sur l'absence de stabilité du nombre de mandats



(a) Table `CD` (b) Table `CM`

(c) Table `CR` (d) Table `D`

(e) Table `DE` (f) Table `EPCI`

(g) Table `M` (h) Table `S`

Figure 2. Évolution du nombre de mandats au cours du temps, pour les données *traitées*. En rouge et bleu les nombres de mandats mesurés et réglementaires, en noir les dates d'élections. Cette figure est à comparer à la Figure 1 correspondant aux données brutes (non-traitées).

(P61). Pour `CD`, il faut néanmoins remarquer que les données traitées ne dépassent plus ponctuellement la limite réglementaire représentée en bleu (P60), comme c'était le cas avec les données brutes.

Le résultat du traitement est bien plus marqué sur `D`, `S` et `DE`. Pour les député·es, le recoupement avec les données de l'Assemblée Nationale a permis non seulement de corriger les erreurs qui causaient de fréquents dépassements de la limite réglementaire, mais aussi de compléter une partie des mandats manquants pour la période 2002–2007 (P59). Cependant, on voit que le nombre réglementaire n'est pas atteint (P61), ce qui indique que même la base de l'Assemblée Nationale n'est pas complète pour cette période. Pour les sénateur·rices, la base du Sénat a permis de couvrir l'intégralité de la V^{ème} république (et même la fin de la IV^{ème}, en partie), et de résoudre à la fois les cas de dépassement du nombre de mandats réglementaire (P60), de mandats manquants (P59), et de variation progressive du nombre de mandats (P61). Enfin, pour les député·es européen·nes, la source secondaire a permis de couvrir l'intégralité des mandats depuis 1978, en résolvant là aussi les problèmes de mandats surnuméraires et manquants, qui étaient extrêmement fréquents dans cette table-là.

Il faut souligner que certains types de mandats pourraient bénéficier d'une couverture supérieure en incluant les mandats relativement similaires qui les ont précédés. C'est notamment le cas des député·es européen·nes : la période couverte correspond à la mise en oeuvre du mandat universel direct, mais le PE existe depuis 1962, et il a été précédé par l'*Assemblée Parlementaire* de la CEE de 1957 à 1962, et avant cela par l'*Assemblée Commune* de la CECA de 1951 à 1957. Jusqu'en 1979, les député·es européen·nes représentant la France étaient des délégué·es désigné·es par l'Assemblée Nationale. De même pour les conseil régionaux, qui ont été précédés par les *Établissements Publics Régionaux* de 1972 à 1986. Ceux-si se composaient des parlementaires de la région et de représentants issus des conseils généraux et des principales municipalités. Il serait donc légitime d'intégrer tous ces mandats dans la BRÉF, bien qu'ils ne soient pas électoraux.

4.2.3 Problèmes ouverts

Statut des problèmes. La Table 28 fait le bilan du traitement des problèmes détectés en Section 2. La plus grande partie des problèmes est complètement résolue par les différentes opérations que nous avons fait subir aux données du RNE, couplée à l'intégration de données issues de sources secondaires.

Statut	Problèmes
Complètement traités	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, P11, P13, P14, P23, P24, P26, P27, P28, P30, P33, P34, P35, P36, P38, P41, P44, P47, P48, P50, P51
Partiellement traités	P12, P16, P17, P18, P19, P20, P21, P25, P29, P25, P31, P32, P37, P39, P42, P43, P45, P49, P52, P58, P59, P60, P61
Pas du tout traités	P15, P22, P46, P53, P54, P55, P56, P57

Table 28. État des problèmes relevés dans la Section 2, à l'issue du traitement décrit dans la Section 3.

Cependant, on trouve également un grand nombre de problèmes restés ouverts car traités seulement de façon partielle. Il s'agit exclusivement de problèmes nécessitant soit un traitement manuel, mais trop fréquents pour que nous l'ayons fait pour l'instant, soit un recoupement avec une source secondaire. Les problèmes qui n'ont fait l'objet d'aucun traitement correspondent soit à des erreurs nécessitant un recodage poussé de certains champs (une tâche que nous avons choisi de reporter), soit nécessitant là aussi l'identification et l'intégration de sources secondaires appropriées.

Taux d'incomplétude et d'erreur. Pour finir, l'un des points importants du recoupement avec des sources externes est l'évaluation des taux d'incomplétude (mandats manquants) et d'erreur (mandats présents dans le RNE, mais sans existence réelle) du RNE. La Table 29 récapitule les taux et décomptes indiquées dans la Section 3 pour `S`, `D` et `DE`, et extrapole les décomptes pour les autres tables de données.

Taux ou nombre	CD	CM	CR	D		DE		EPCI	M	S	
Erreur	127	12 394	58	11	1%	96	37%	1 279	1 135	15	1%
Incomplétude	636	61 971	291	115	5%	52	20%	6 397	5 676	50	5%

Table 29. Nombres et taux d'erreur (mandats inexistant en réalité) et d'incomplétude (mandats manquants) mesurés pour `S`, `D` et `DE`, et nombres estimés pour les autres tables de données brutes du RNE. Les valeurs estimées sont précédées d'une astérisque.

Ces valeurs sont à interpréter avec précaution pour deux raisons. Tout d'abord, notre évaluation a montré que les données de l'Assemblée Nationale n'étaient elles-mêmes pas complètes pour la période 2002–2007, ce qui signifie que les valeurs mesurées pour `D` sont inférieures à la réalité. Par ailleurs, si les taux obtenus pour `S` et `D` sont comparables, en revanche ceux de `DE` leur sont supérieurs d'un ordre de grandeur. Au vu des différentes erreurs relevées dans la Section 2, les tables `CM`, `M` et `EPCI` nous semblent être les moins fiables du RNE après `DE` (cf. Section 2.5.2), donc les taux s'appliquant à celles-ci sont probablement entre ceux obtenus pour `D` et `S` et ceux de `DE`. À titre indicatif, nous avons estimés les nombres de lignes erronées/manquantes pour le reste des tables sur la base des taux de 1% et 5% obtenus pour `D` et `S`. Ces valeurs sont indiquées dans la Table .

Règles de cumul. La vérification des règles de cumul constitue un test qui nous reste encore à implémenter, et qui est susceptible de révéler certaines incohérences, et de potentiellement nous demander d'effectuer de nouvelles corrections. L'ensemble des règles concernant les cumuls est décrit en annexe dans la Table 40.

4.3 Sources exploitables

Le bilan que nous avons effectué sur les problèmes ouverts en Section 4.2.3 indique que les deux moyens principaux de leur résolution sont un traitement manuel, généralement mis de côté pour l'instant car coûteux en temps, et le recoupement avec des sources de données secondaires. Au cours du traitement, nous avons identifié un certain nombre de sources pouvant remplir ce rôle. Nous avons déjà exploité en partie certaines de ces sources (cf. Sections 3.2 à 3.4), tandis que nous n'avons pas encore tiré parti des autres.

4.3.1 Base du Sénat

Nous avons déjà introduit cette base de données dans la Section 3.2, et avons expliqué comment nous n'avions exploité qu'une partie des informations disponibles : les mandats de sénateur·rice. Toutefois, la base décrit également d'autres types de mandats électifs et non-électifs occupés par des sénateur·rices, et susceptibles d'être intégrés ultérieurement dans la BRÉF.

Outre les données ouvertes du Sénat, qui prennent la forme d'une base de données proprement dite et couvrant la V^{ème} République[21], il existe visiblement une autre base dédiée aux sénateurs du Second Empire [27], des III^{ème} et IV^{ème} républiques [28, 29] et la Communauté [30]. Cependant, celle-ci n'est pas accessible directement via une API, mais seulement à travers une interface Web manifestement dévolue à un usage manuel.

27. http://www.senat.fr/senateurs-2nd-empire/index.html
28. http://www.senat.fr/senateurs-3eme-republique/index.html
29. http://www.senat.fr/senateurs-4eme-republique/index.html
30. http://www.senat.fr/senateurs-communaute/index.html

4.3.2 Base de l'Assemblée Nationale

Comme expliqué dans la Section 3.3, nous avons également exploité partiellement la base de données ouvertes de l'Assemblée Nationale. Là aussi, nous nous sommes concentrés dans un premier temps sur les mandats auxquels cette base est dédié (donc ceux de député), mais celle-ci inclut également d'autres aspects de l'activité parlementaire qui pourraient être intégrés à la BRÉF. Cependant, à la différence de la base du Sénat, il faut souligner que la base de l'Assemblée ne contient pas d'autre type de mandat électif.

À l'instar du Sénat, outre les données ouvertes présentées sous forme de base de données proprement dite[23], il existe également une base décrivant l'activité historique de l'Assemblée Nationale. Elle se nomme Sycomore [31], et est elle aussi accessible uniquement via une interface Web. Sycomore remonte à la révolution française, et semble intégrer toutes les institutions analogues à l'Assemblée actuelle qui se sont succédées au cours des différents régimes. Pour mémoire, la base que nous avons exploitée pour l'instant ne couvre pas les mandats antérieurs à 2002.

4.3.3 Site France Politique

Nous avons également identifié des sources non-institutionnelles prometteuses. Le site Web *France Politique* [32] a été créé en 2001 par Laurent de Boissieu, journaliste politique au journal *La Croix*, qui en est encore aujourd'hui le seul administrateur. Les données décrivent les résultats des élections pour une bonne partie de la V^{ème} république (cela dépend du type de mandat). Cependant, il semble s'agir uniquement de données agrégées, et non pas de listes nominatives qui nous permettraient d'identifier les élus et leurs mandats.

4.3.4 Site Politiquemania

Politiquemania est un autre site Web [33], créé en 2009 par deux personnes dont Vincent Lefebvre, avec pour objectif de mettre en place une base de données unifiée sur la vie politique française. Les données y ont été insérées et vérifiées manuellement par les deux administrateurs, d'abord à partir des résultats électoraux publiés dans la presse, puis sur la base de fichiers contenant des résultats électoraux, fournis par Brigitte Hazart du Ministère de l'Intérieur (voir aussi Section 1.2). Cette base a donc été constituée à partir de données distinctes de celles du RNE.

Elle semble couvrir la V^{ème} République de façon très complète. Cependant, à la différence du RNE, les conseiller·es municipal·es et d'EPCI ne sont pas représenté·es. Ceci rend les deux bases très complémentaires. D'après les vérifications manuelles que nous avons pu réaliser en corrigeant certains cas problématiques détectés dans le RNE, les données de Politiquemania semblent généralement plus fiables, notamment pour ce qui concerne les mandats partiels (remplacements). De plus, les motifs de fin de mandats sont catégorisés de façon plus fine. Outre les mandats eux-mêmes, ces données intègrent également les résultats des élections exprimés en termes de voix recueillies, ainsi que l'identité des candidats malheureux.

4.3.5 Site MairesGenWeb

MairesGenWeb [34] est un site Web maintenu par FranceGenWeb [35], une association dont l'activité porte sur la généalogie. Elle a constitué manuellement, grâce à la participation volontaire de ses membres, une base de données contenant les maires des communes

31. http://www2.assemblee-nationale.fr/sycomore/recherche
32. https://www.france-politique.fr/
33. https://www.politiquemania.com/
34. http://www.francegenweb.org/mairesgenweb/
35. http://www.francegenweb.org/

françaises pour une période s'étendant jusqu'à la Révolution. La méthode de constitution n'étant pas systématique, le niveau de complétude de la base est difficile à évaluer. De plus, elle porte sur les maires à l'exclusion des autres conseiller·es municipal·es. Elle contient néanmoins des données qui pourraient venir compléter la BRÉF.

À noter l'existence d'une base de données complémentaire, CommunesGenWeb [36], qui recense les communes et leur évolution (création, suppression, fusion, etc.) depuis 1790 (par comparaison, la base INSEE démarre en 1940).

4.3.6 Bases des EPCI

Pour compléter les données manquantes relatives aux mandats de type EPCI, il serait possible d'exploiter différentes bases qui les recensent : BANATIC [37] (Base nationale sur l'inter-communalité) et la page de `collectivites-locales.gouv.fr` [38]. Cependant, elles ne contiennent visiblement que des informations sur les EPCI eux-mêmes, et non pas sur les mandats de leurs conseiller·es.

4.4 Mises à jour du RNE

Outre les erreurs et l'exploitation des bases secondaires décrites en Sections 4.2 et 4.3, il existe un autre problème ouvert, qui concerne le traitement des mises à jour à venir du RNE, et leur intégration dans la BRÉF.

Pour mémoire, la BRÉF est basée sur une extraction historique, et le BdE effectue (approximativement) chaque trimestre une extraction photographique mise à disposition du public sous forme de données ouvertes. Nous comptons mettre à jour la BRÉF au moyen de ces données. Cependant ce point n'est pas trivial, tout d'abord parce que l'appauvrissement des données qui a coïncidé avec le passage du RNE en données ouvertes, décrit en Section 1.1, complique significativement la tâche. De plus, la résolution temporelle de trois mois implique forcément des approximation sur les dates des mandats.

Identifiants manquants. Le premier problème vient de l'absence de tout identifiant d'élu·e dans les extraction désormais fournies sur `data.gouv.fr`, à la différence de ce qui se faisait auparavant. Cela signifie qu'il n'est pas possible de mettre en relation les élu·es déjà présents dans la BRÉF et ceux décrits dans les extractions à venir sur la base d'une information non-ambigüe.

Une façon de contourner ce problème est d'appliquer les mêmes méthodes que lors du traitement que nous avons réalisé sur les données brutes pour détecter les doublons, à savoir : des recoupements basés sur les nom, prénom, et date de naissance de l'élu·e (ainsi que sur des éléments temporels et géographiques). Cependant, ce type de recoupement est bien moins rigoureux et bien moins fiable qu'en passant par des identifiants uniques, et il est quasi-certain que des erreurs seront introduites dans la BRÉF à chaque mise à jour.

Pour limiter ces erreurs, on pourrait envisager de croiser les mises à jour du RNE avec des sources secondaires comme nous l'avons fait ici, mais cela nécessite évidemment un travail supplémentaire non-négligeable et, qui plus est, difficilement automatisable. De plus, les informations concernant les élu·es locaux sont particulièrement difficiles à trouver.

Extractions photographiques. Le deuxième problème est relatif à la nature exclusivement photographique (et non pas historique) des extractions publiées sur `data.gouv.fr`. Le traitement nécessaire pour intégrer ces données à la BRÉF est donc différent de celui décrit dans ce document, qui portait sur une extraction historique, et requiert le développement d'un logiciel adapté.

36. http://www.francegenweb.org/communes/accueil.php
37. https://www.data.gouv.fr/fr/datasets/base-nationale-sur-linter-communalite/
38. https://www.collectivites-locales.gouv.fr/liste-et-composition-des-epci-a-fiscalite-propre

De plus, cette intégration des mises à jour doit être réalisée de façon *trimestrielle*, afin de suivre le rythme de publication adopté sur `data.gouv.fr`. Bien sûr, on peut supposer que la plupart des mandats décrits à un instant t seront également présents dans la photographie prise à l'instant $t + 3$ (temps exprimé en *mois*), pour peu qu'aucune élection n'ait lieu entre-temps. Cependant, ce traitement régulier des mises à jours est nécessaire, afin de ne pas passer à côté des mandats qui apparaîtraient ou disparaîtraient d'une extraction à la suivante. Pour certains types de mandats, en particulier `CM`, il est vraisemblable que cela représente un nombre significatif de lignes.

Il faut souligner que même alors, la résolution temporelle des données est extrêmement plus faible que précédemment, puisqu'elle est seulement de l'ordre d'un trimestre. Autrement dit, si un mandat disparaît des données entre t et $t + 3$, nous savons qu'il a pris fin au cours du semestre concerné, mais nous ne connaissons pas la date exacte. Là aussi, le croisement avec des sources de données externes peut permettre de résoudre certains cas problématiques et d'accroître la précision temporelle.

Autres attributs manquants. Par rapport aux extractions historiques en notre possession, les données publiées sur `data.gouv.fr` n'incluent pas les motifs de fin de mandat et de fonction, et surtout les nuances politiques des élu·es (un champ renseigné par les préfectures). Ce dernier champ était particulièrement précieux pour l'analyse du RNE Science Politique. Encore une fois, un recoupement avec des sources secondaires peut être envisagé pour contourner le problème, avec les limites de coût et de fiabilité déjà discutées.

Perspectives. En sollicitant directement le BdE, nous avons pu obtenir une version *augmentée* de ces extractions photographiques, pour la période allant de notre extraction historique (juillet 2018) à l'automne 2019. Ces extractions-là contiennent l'identifiant et la couleur politique associés à chaque mandat. Pour cette période, le problème des attributs absents ne se pose donc pas (à la différence de celui de la résolution temporelle). Après cette date, les données que nous avons pu obtenir directement du BdE contiennent les couleurs politiques, mais plus les identifiants.

Nous comptons sur la possibilité à venir d'un accès à la *nouvelle version* du RNE (RNE2), actuellement développée par le BdE, et si possible dans sa version historique, afin de pouvoir efficacement actualiser le contenu de la BRÉF en y intégrant les mandats futurs.

A Valeurs des labels du RNE

Table	Valeur
CD	Président du Conseil Général / Président du Conseil Départemental
	Vice-président du Conseil Départemental
	Premier à Quinzième vice-président du Conseil Départemental
	Vice-président délégué du Conseil Départemental
	Questeur
	Premier à Sixième secrétaire
	Président de groupe
	Président de commission
	Autre membre commission permanente
	Autre membre
CM & M	Maire
CM	Maire délégué
	Premier à Vingt-septième adjoint au maire
	Premier à Vingt-et-unième adjoint au maire de Lyon
	Premier à Vingt-septième adjoint au maire de Marseille
	Premier à Vingt-septième adjoint au maire de Paris
CR	Président du Conseil Régional
	Premier à Quinzième vice-président du Conseil Régional
	Président de commission
	Autre membre commission permanente
D	Président de l'Assemblée Nationale
EPCI	Président d'EPCI
	Vice-président d'EPCI
S	Président du Sénat

Table 30. Fonctions rencontrées dans les différentes tables de données ; DE n'apparaît pas car elle ne détaille pas les fonctions des député·e européen·ne qu'elle décrit.

Valeur	Signification
ALLI	Alliance Centriste
AUT	Autres
CEN	Le Centre pour la France
COM	Parti Communiste Français
CPNT	Chasse, Pêche, Nature et Tradition
DIV	Divers
DL	Démocratie Libérale
DLF	Debout la France
DVD	Divers Droite
DVG	Divers Gauche
ECO	Autre Écologistes
EXD	Extrême Droite
EXG	Extrême Gauche
FG	Front de Gauche
FI	France Insoumise
FN	Front National
LCR	Ligue Communiste Révolutionnaire
LO	Lutte Ouvrière
LR	Les Républicains
M	Majorité présidentielle autre
M-NC	Majorité présidentielle
MAJ	Majorité présidentielle
MDC	Mouvement des Citoyens

Valeur	Signification
MDM	Mouvement Démocrate
MNR	Mouvement National Républicain
MODM	MoDem
MPF	Mouvement pour la France
NCE	Nouveau Centre
NC	Non-communiqué
PG	Parti de Gauche
PREP	Signification inconnue
PRG	Parti Radical de Gauche
PRS	Parti Républicain Socialiste
PRV	Parti Radical
RDG	Parti Radical de Gauche
REG	Régionalistes
REM	La République En Marche
RPF	Rassemblement pour la France
RPR	Rassemblement pour la République
SOC	Parti Socialiste
UDF	Union pour la Démocratie Française
UDFD	UDF / Mouvement Démocrate
UDI	Union Démocrates Indépendants
UMP	Union pour un Mouvement Populaire
VEC	Europe Écologie Les Verts

Table 31. Valeurs du champ *Nuance politique* du RNE, avec la signification des abréviations correspondantes.

Code et Libellé	
1	Agriculteurs propriétaires exploit.
2	Salariés agricoles
3	Marins (patrons)
4	Marins (salariés)
5	Industriels-Chefs entreprise
6	Administrateurs de sociétés
7	Agents d'affaires
8	Agents immobiliers
9	Commerçants
10	Artisans
11	Entrepreneurs en batiments
12	Propriétaires
13	Ingénieurs
14	Agents technique[39] et techniciens
15	Contremaitres
16	Représentants de commerce
17	Agents d'assurances
18	Cadres supérieurs (secteur privé)
19	Autres cadres (secteur privé)
20	Employés (secteur privé)
21	Ouvriers (secteur privé)
22	Assistantes sociales
23	Salariés du secteur médical
24	Médecins
25	Chirurgiens
26	Dentistes
27	Vétérinaires
28	Pharmaciens
29	Avocats
30	Notaires
31	Huissiers
32	Conseillers juridiques
33	Agents généraux d'assurances
34	Experts comptables

Code et Libellé	
35	Ingénieurs conseils
36	Architectes
37	Journalistes et autres médias
38	Hommes de lettres et Artistes
39	Autres professions libérales
40	Etudiants
41	Professeurs de faculté
42	Professeurs du secondaire et techn.
43	Enseignants 1er deg.-directeurs école
44	Professions rattachées à enseignt.
45	Magistrats
46	Grands corps de l'état
47	Fonctionnaires de catégorie A
48	Fonctionnaires de catégorie B
49	Fonctionnaires de catégorie C
50	Cadres sup. (entreprises publiques)
51	Cadres (entreprises publiques)
52	Employés (autres entrep. publiques)
53	Agents subalternes (entr.publiques)
54	Permanents politiques
55	Ministres du culte
56	Autres professions
57	Sans profession déclarée
58	Retraités agricoles
59	Retr.artis.commerc.chefs d'entrep.
60	Retraités des professions libérales
61	Retraités salariés privés
62	Retraités de l'enseignement
63	Retraités fonct.publique (sf enseig.)
64	Retraités des entreprises publiques
65	Autres retraités

Table 32. Professions rencontrées dans le RNE, avec le code associé.

Champ	Valeur	Signification
Code de sexe	F	Féminin
	M	Masculin
Motif de fin de mandat/fonction	AU	Autre
	DC	Décès
	DO	Démission d'office
	DV	Démission volontaire
	FM	Fin de mandat normale

Table 33. Valeurs des champs *Code de sexe* et *Motif de fin* du RNE, avec la signification des abréviations correspondantes.

B Dates d'élections

Conseils Généraux/Départementaux (CD)		
Tour 1	Tour 2	Série
20/04/1958	27/04/1958	1
04/06/1961	11/06/1961	2
08/03/1964	15/03/1964	1
24/09/1967	01/10/1967	2
23/09/1970	30/09/1970	1
23/09/1973	30/09/1973	2
07/03/1976	14/03/1976	1
18/03/1979	25/03/1979	2
14/03/1982	21/03/1982	1
10/03/1985	17/03/1985	2
25/09/1988	02/10/1988	1
22/03/1992	29/03/1992	2
20/03/1994	27/03/1994	1
15/03/1998	22/03/1998	2
11/03/2001	18/03/2001	1
21/03/2004	28/03/2004	2
09/03/2008	16/03/2008	1
20/03/2011	27/03/2011	2
22/03/2015	29/03/2015	1 & 2

Conseils Municipaux (CM & M)	
Tour 1	Tour 2
08/03/1959	15/03/1959
14/03/1965	21/03/1965
14/03/1971	21/03/1971
13/03/1977	20/03/1977
06/03/1983	13/03/1983
12/03/1989	19/03/1989
11/06/1995	18/06/1995
11/03/2001	18/03/2001
09/03/2008	16/03/2008
23/03/2014	30/03/2014
15/03/2020	2020/30/22

Conseils Régionaux (CR)	
Tour 1	Tour 2
16/03/1986	
22/03/1992	
15/03/1998	
21/03/2004	28/03/2004
14/03/2010	21/03/2010
06/12/2015	13/12/2015

Table 34. Dates des élections pour les mandats *locaux*. Les élections régionales ne comportaient qu'un seul tour de 1986 à 1998. Les séries correspondent au renouvellement par moitié des Conseils Généraux, abandonné en 2015 lors du passage aux Conseils Départementaux.

Assemblée Nationale (D)	
Tour 1	Tour 2
23/11/1958	30/11/1958
18/11/1962	25/11/1962
05/03/1967	12/03/1967
23/06/1968	30/06/1968
04/03/1973	11/03/1973
12/03/1978	19/03/1978
14/06/1981	21/06/1981
16/03/1986	
05/06/1988	12/06/1988
21/03/1993	28/03/1993
25/05/1997	01/06/1997
09/06/2002	16/06/2002
10/06/2007	17/06/2007
10/06/2012	17/06/2012
11/06/2017	18/06/2017

Sénat (S)	
Date	Série
26/04/1959	A, B, C1 & C2
23/09/1962	A
26/09/1965	B
22/09/1968	C1 & C2
26/09/1971	A
22/09/1974	B
25/09/1977	C1 & C2
28/09/1980	A
25/09/1983	B
28/09/1986	C1 & C2
24/09/1989	A
27/09/1992	B
24/09/1995	C1 & C2
27/09/1998	A
23/09/2001	B
26/09/2004	C1 & C2
21/09/2008	A
25/09/2011	B & C1
28/09/2014	A & C2
24/09/2017	B & C1

Parlement Européen (DE)
07/06/1979
17/06/1984
18/06/1989
12/06/1994
13/06/1999
13/06/2004
07/06/2009
25/05/2014
26/05/2019

Table 35. Dates des élections pour les mandats *nationaux*. Les élections sénatoriales et européennes ne comportent qu'un seul tour. C'était aussi le cas des législatives en 1986. Les séries correspondent au renouvellement partiel du Sénat, d'abord par tiers jusqu'en 2004, puis par moitié.

C Nombres de sièges

C.1 Par circonscription

La vérification des mandats décrite en Section 2.2 passe par le recoupement de leurs dates avec un certain nombre d'informations obtenues par ailleurs. Cela inclut les dates d'élection, décrites dans les Tables 34 et 35 ; les nombres de sièges accordés à certaines divisions électorales, décrits dans les Tables 36, 37 et 38 ; et les séries définies pour certaines institutions renouvelées de façon partielle à chaque élection, dont la description est répartie sur les Tables 34, 35, et 38. Il faut souligner que les cantons sont également concernés par ce dernier point. Cependant, vu leur nombre élevé, il n'était pas pratique de les lister dans ce document. Cette liste, avec la série à laquelle appartient chaque canton, peut néanmoins être obtenue en consultant les ressources exploitées par notre code source[1].

Deux types de mandats font ou ont fait l'objet d'un renouvellement partiel à chaque élection, dans le sens où seulement une partie des mandats fait l'objet du scrutin. C'est le cas du Sénat et des Conseils Généraux. Jusqu'en 2015, les Conseils Généraux étaient renouvelés par *moitié*. Chaque canton appartenait à une série numérotée 1 ou 2 dans la Table 34, et les élections cantonales ne concernaient qu'une série à la fois. Après la réforme de 2013 et la transformation en Conseils Départementaux, la règle est passée à un renouvellement complet de ces assemblées. Une autre différence est la mise en place d'un scrutin binominal (et non plus uninominal) visant à obtenir la parité sexuelle. Lors des élections départementales, chaque canton élit dorénavant 2 conseiller·es, une femme et un homme. Un redécoupage des cantons a également eu lieu en 2014, notamment pour prendre ce changement en compte.

Jusqu'en 2004, les sénateur·rices étaient renouvelé·es par *tiers*. Chaque circonscription sénatoriale appartenait à l'une des trois séries notées A, B et C dans les Tables 35 et 38. Après la réforme de 2003, on passe à un renouvellement du Sénat par *moitié*. La série C est divisée en deux parties notées $C1$ et $C2$: l'une est fusionnée avec la série A et l'autre avec la série B.

Circonscription européenne	Nombre de sièges	Date de début	Date de fin
Circonscription unique	81	07/06/1979	11/06/1994
Circonscription unique	87	12/06/1994	12/06/2004
Circonscription unique	74	26/05/2019	31/01/2020
Circonscription unique	79	01/02/2020	–
Élus par l'Assemblée Nationale	2	07/12/2011	25/05/2014
Est	10	13/06/2004	06/06/2009
Est	9	07/06/2009	–
Île-de-France	14	13/06/2004	06/06/2009
Île-de-France	13	07/06/2009	24/05/2014
Île-de-France	15	25/05/2014	–
Massif central-Centre	6	13/06/2004	06/06/2009
Massif central-Centre	5	07/06/2009	–
Nord-Ouest	12	13/06/2004	06/06/2009
Nord-Ouest	10	07/06/2009	–
Ouest	10	13/06/2004	06/06/2009
Ouest	9	07/06/2009	–
Outre-Mer	3	13/06/2004	–
Sud-Est	13	13/06/2004	–
Sud-Est	10	13/06/2004	–

Table 36. Nombre de sièges de député·es européen·nes pour chaque circonscription européenne, pour la période indiquée.

Région	Nombre de sièges	Date de début	Date de fin
Alsace	47	16/03/1986	05/12/2015
Aquitaine	85	16/03/1986	05/12/2015
Auvergne	47	16/03/1986	05/12/2015
Auvergne-Rhône-Alpes	204	13/12/2015	–
Basse-Normandie	47	16/03/1986	05/12/2015
Bourgogne	57	16/03/1986	05/12/2015
Bourgogne-Franche-Comté	100	13/12/2015	–
Bretagne	83	16/03/1986	–
Centre	77	16/03/1986	–
Champagne-Ardenne	49	16/03/1986	05/12/2015
Corse	51	16/03/1986	02/12/2017
Assemblée de Corse	63	03/12/2017	–
Franche-Comté	43	16/03/1986	05/12/2015
Grand Est	169	13/12/2015	–
Guadeloupe	41	16/03/1986	–
Guyane	31	16/03/1986	05/12/2015
Assemblée de Guyane	51	13/12/2015	–
Haute-Normandie	55	16/03/1986	05/12/2015
Hauts-de-France	170	13/12/2015	–
Île-de-France	209	16/03/1986	–
La Réunion	45	16/03/1986	–
Languedoc-Roussillon	67	16/03/1986	05/12/2015
Limousin	43	16/03/1986	05/12/2015
Lorraine	73	16/03/1986	05/12/2015
Martinique	41	16/03/1986	05/12/2015
Assemblée de Martinique	51	13/12/2015	NA
Midi-Pyrénées	91	16/03/1986	05/12/2015
Nord-Pas-de-Calais	113	16/03/1986	05/12/2015
Normandie	102	13/12/2015	–
Nouvelle-Aquitaine	183	13/12/2015	–
Occitanie	158	13/12/2015	–
Pays de la Loire	93	16/03/1986	–
Picardie	57	16/03/1986	05/12/2015
Poitou-Charentes	55	16/03/1986	05/12/2015
Provence-Alpes-Côte d'Azur	123	16/03/1986	–
Rhône-Alpes	157	16/03/1986	05/12/2015

Table 37. Nombre de sièges de conseiller·es régional·es pour chaque région, pour la période indiquée. Cette liste inclut des régions pré- et post-réforme de 2015.

C.2 Total national

L'élaboration des graphiques de la Figure 1 nécessite de connaître l'évolution du nombre de sièges au niveau national pour chaque type de mandat. La Table 39 résume les informations que nous avons obtenues et intégrées à cette fin. Il faut souligner que le nombre de positions électives existant à une date donnée s'est révélé très difficile à trouver pour certains types de mandats moins bien documentés que d'autres, ce qui explique que nous ayons parfois utilisé des estimations à défaut de valeurs exactes.

Pour `CD`, les sources varient quand au nombre exact de conseiller·es, notamment en

Circonscription sénatoriale	S.	N.	Date de début	Date de fin
Ain	A	2	26/04/1959	20/09/2008
Ain	A	3	21/09/2008	–
Aisne	A	3	26/04/1959	–
Allier	A	2	26/04/1959	–
Alpes-de-Haute-Provence	A	1	26/04/1959	–
Hautes-Alpes	A	1	26/04/1959	–
Alpes-Maritimes	A	3	26/04/1959	27/09/1980
Alpes-Maritimes	A	4	28/09/1980	20/09/2008
Alpes-Maritimes	A	5	21/09/2008	–
Ardèche	A	2	26/04/1959	–
Ardennes	A	2	26/04/1959	–
Ariège	A	1	26/04/1959	–
Aube	A	2	26/04/1959	–
Aude	A	2	26/04/1959	–
Aveyron	A	2	26/04/1959	–
Bouches-du-Rhône	A	6	26/04/1959	27/09/1980
Bouches-du-Rhône	A	7	28/09/1980	20/09/2008
Bouches-du-Rhône	A	8	21/09/2008	–
Calvados	A	3	26/04/1959	–
Cantal	A	2	26/04/1959	–
Charente	A	2	26/04/1959	–
Charente-Maritime	A	3	26/04/1959	–
Cher	A	2	26/04/1959	–
Corrèze	A	2	26/04/1959	–
Corse	A	2	26/04/1959	27/09/1980
Corse-du-Sud	A	1	28/09/1980	–
Haute-Corse	A	1	28/09/1980	–
Côte-d'Or	A	2	26/04/1959	27/09/1980
Côte-d'Or	A	3	28/09/1980	–
Côtes-d'Armor	A	3	26/04/1959	–
Creuse	A	2	26/04/1959	–
Dordogne	A	2	26/04/1959	–
Doubs	A	2	26/04/1959	27/09/1980
Doubs	A	3	28/09/1980	–
Drôme	A	3	26/04/1959	–
Eure	A	2	26/04/1959	27/09/1980
Eure	A	3	28/09/1980	–
Eure-et-Loir	A	3	26/04/1959	–
Finistère	A	4	26/04/1959	–
Gard	A	2	26/04/1959	27/09/1980
Gard	A	3	28/09/1980	–
Haute-Garonne	A	3	26/04/1959	27/09/1980
Haute-Garonne	A	4	28/09/1980	20/09/2008
Haute-Garonne	A	5	21/09/2008	–
Gers	A	2	26/04/1959	–
Gironde	A	4	26/04/1959	27/09/1980
Gironde	A	5	28/09/1980	20/09/2008
Gironde	A	6	21/09/2008	–
Hérault	A	4	26/04/1959	–
Ille-et-Vilaine	A	3	26/04/1959	27/09/1980
Ille-et-Vilaine	A	4	28/09/1980	–
Indre	A	2	26/04/1959	–
Indre-et-Loire	B	2	26/04/1959	24/09/1983
Indre-et-Loire	B	3	25/09/1983	–
Isère	B	3	26/04/1959	24/09/1983
Isère	B	4	25/09/1983	24/09/2011
Isère	B	5	25/09/2011	–
Jura	B	2	26/04/1959	–
Landes	B	2	26/04/1959	–
Loir-et-Cher	B	2	26/04/1959	–
Loire	B	4	26/04/1959	–
Haute-Loire	B	2	26/04/1959	–
Loire-Atlantique	B	4	26/04/1959	24/09/1983
Loire-Atlantique	B	5	25/09/1983	–
Loiret	B	2	26/04/1959	24/09/1983
Loiret	B	3	25/09/1983	–
Lot	B	1	26/04/1959	24/09/1983
Lot	B	2	25/09/1983	–
Lot-et-Garonne	B	2	26/04/1959	–
Lozère	B	1	26/04/1959	–
Maine-et-Loire	B	4	26/04/1959	–
Manche	B	3	26/04/1959	–
Marne	B	3	26/04/1959	–
Haute-Marne	B	2	26/04/1959	–
Mayenne	B	2	26/04/1959	–
Meurthe-et-Moselle	B	3	26/04/1959	24/09/1983
Meurthe-et-Moselle	B	4	25/09/1983	–
Meuse	B	2	26/04/1959	–
Morbihan	B	3	26/04/1959	–
Moselle	B	4	26/04/1959	24/09/1983
Moselle	B	5	25/09/1983	–
Nièvre	B	2	26/04/1959	–
Nord	B	10	26/04/1959	24/09/1983
Nord	B	11	25/09/1983	–
Oise	B	4	26/04/1959	–
Orne	$C1$	2	26/04/1959	–
Pas-de-Calais	B	6	26/04/1959	24/09/1983
Pas-de-Calais	B	7	25/09/1983	–
Puy-de-Dôme	B	3	26/04/1959	–
Pyrénées-Atlantiques	B	3	26/04/1959	–
Hautes-Pyrénées	B	2	26/04/1959	–
Pyrénées-Orientales	B	2	26/04/1959	–
Bas-Rhin	$C2$	5	26/04/1959	–
Haut-Rhin	$C2$	4	26/04/1959	–
Rhône	$C2$	5	26/04/1959	24/09/1977
Rhône	$C2$	7	25/09/1977	–
Haute-Saône	$C2$	2	26/04/1959	–
Saône-et-Loire	$C2$	3	26/04/1959	–
Sarthe	$C2$	3	26/04/1959	–
Savoie	$C2$	2	26/04/1959	–
Haute-Savoie	$C2$	2	26/04/1959	24/09/1977
Haute-Savoie	$C2$	3	25/09/1977	–
Paris	$C2$	12	22/09/1968	–
Seine-Maritime	$C2$	5	26/04/1959	24/09/1977
Seine-Maritime	$C2$	6	25/09/1977	–
Seine-et-Marne	$C2$	3	26/04/1959	24/09/1977
Seine-et-Marne	$C2$	4	25/09/1977	25/09/2004
Seine-et-Marne	$C2$	6	26/09/2004	–
Yvelines	$C2$	4	22/09/1968	24/09/1977
Yvelines	$C2$	5	25/09/1977	25/09/2004
Yvelines	$C2$	6	26/09/2004	–
Deux-Sèvres	$C2$	2	26/04/1959	–
Somme	$C2$	3	26/04/1959	–
Tarn	$C2$	2	26/04/1959	–
Tarn-et-Garonne	$C2$	2	26/04/1959	–
Var	$C2$	4	26/04/1959	–
Vaucluse	$C2$	3	26/04/1959	–
Vendée	$C2$	2	26/04/1959	24/09/1977
Vendée	$C2$	3	25/09/1977	–
Vienne	$C2$	2	26/04/1959	–
Haute-Vienne	$C2$	2	26/04/1959	–
Vosges	$C2$	2	26/04/1959	–
Yonne	$C2$	2	26/04/1959	–
Territoire de Belfort	A	1	26/04/1959	–
Essonne	$C1$	3	22/09/1968	24/09/1977
Essonne	$C1$	5	25/09/1977	–
Hauts-de-Seine	$C1$	7	22/09/1968	–
Seine-Saint-Denis	$C1$	5	22/09/1968	24/09/1977
Seine-Saint-Denis	$C1$	6	25/09/1977	–
Val-de-Marne	$C1$	5	22/09/1968	24/09/1977
Val-de-Marne	$C1$	6	25/09/1977	–
Val-d'Oise	$C1$	3	22/09/1968	24/09/1977
Val-d'Oise	$C1$	4	25/09/1977	25/09/2004
Val-d'Oise	$C1$	5	26/09/2004	–
Seine-et-Oise	$C1$	8	26/04/1959	21/09/1968
Seine	$C1$	22	26/04/1959	21/09/1968
Guadeloupe	$C1$	3	26/04/1959	–
Martinique	$C1$	2	26/04/1959	–
Guyane	A	?	?	?
La Réunion	B	2	26/04/1959	24/09/1983
La Réunion	B	3	25/09/1983	24/09/2011
La Réunion	B	4	25/09/2011	–
Mayotte	$C1$	?	?	?
Nouvelle-Calédonie	B	?	?	?
Polynesie francaise	A	?	?	?
Saint-Pierre-et-Miquelon	$C1$	1	26/04/1959	–
Saint-Martin	A	1	21/09/2008	–
Wallis et Futuna	A	1	23/09/1962	–
Saint-Barthelemy	A	?	?	?
Français de l'étranger	?	––		

Table 38. Nombre de sièges de sénateur·rice pour chaque circonscription sénatoriale, pour la période indiquée. Certaines valeurs manquantes sont indiquées par des points d'interrogation (?). Les colonnes *S.* et *N.* désignent respectivement la série à laquelle la circonscription appartient, et son nombre de sièges. La circonscription des *Français de l'étranger* n'est pas associée à une série en particulier (chacun de ses sièges l'est indépendemment des autres).

raison du statut spécial de certaines institutions comme les conseils de Martinique, Guyane et Corse, ou ceux des villes de Paris et Lyon. Nous avons d'abord récupéré le décompte de sièges pour chaque département, puis avons tracé son évolution dans le temps, et enfin

sommé pour obtenir les valeurs affichées dans la Table 39. Il faut souligner que ce décompte tient compte des spécificités du RNE décrites en Section 2.4.3 relativement aux institutions incluses ou pas dans `CD`.

Tab.	Nombre de sièges	Date de début	Date de fin
CD	3 988	24/09/1967	06/03/1976
	4 048	07/03/1976	09/03/1985
	4 055	10/03/1985	21/03/2015
	4 177	22/03/2015	15/12/2015
	4 113	16/12/2015	31/12/2017
	4 061	01/01/2018	–
D	579	23/11/1958	17/11/1962
	482	18/11/1962	04/03/1967
	487	05/03/1967	03/03/1973
	490	04/03/1973	11/03/1978
	491	12/03/1978	15/03/1986
	577	16/03/1986	–
EPCI	67 159	?	?
DE	81	07/06/1979	11/06/1994
	87	12/06/1994	12/06/2004
	78	13/06/2004	06/06/2009
	72	07/06/2009	11/07/2011
	74	12/07/2011	24/05/2019
	79	01/02/2020	–
S	307	26/04/1959	22/09/1962
	274	23/09/1962	21/09/1968
	283	22/09/1968	24/09/1977
	295	25/09/1977	27/09/1980
	305	28/09/1980	24/09/1983
	317	25/09/1983	27/09/1986
	319	28/09/1986	23/09/1989
	321	24/09/1989	22/09/2001
	321	23/09/2001	25/09/2004
	331	26/09/2004	20/09/2008
	343	21/09/2008	24/09/2011
	348	25/09/2011	–

Tab.	Nombre de sièges	Date de début	Date de fin
CR	1 829	16/03/1986	05/12/2015
	1 757	13/12/2015	–
CM	503 305	?	?
M	38076	01/01/1962	31/12/1967
	37823	01/01/1968	31/12/1974
	36407	01/01/1975	31/12/1981
	36547	01/01/1982	31/12/1984
	36614	01/01/1985	31/12/1989
	36664	01/01/1990	31/12/1993
	36673	01/01/1994	31/12/1998
	36679	01/01/1999	31/12/1999
	36680	01/01/2000	31/12/2000
	36677	01/01/2001	31/12/2001
	36679	01/01/2002	31/12/2002
	36678	01/01/2003	31/12/2003
	36682	01/01/2004	31/12/2004
	36684	01/01/2005	31/12/2005
	36685	01/01/2006	31/12/2006
	36683	01/01/2007	31/12/2007
	36681	01/01/2008	31/12/2008
	36682	01/01/2009	31/12/2009
	36682	01/01/2010	31/12/2010
	36680	01/01/2011	31/12/2011
	36700	01/01/2012	31/12/2012
	36681	01/01/2013	31/12/2013
	36681	01/01/2014	31/12/2014
	36658	01/01/2015	31/12/2015
	35885	01/01/2016	31/12/2016
	35416	01/01/2017	31/12/2017
	35357	01/01/2018	31/12/2018
	34968	01/01/2019	–

Table 39. Nombre de sièges total au niveau national pour les différents types de mandats. Certaines valeurs sont des approximations, cf. le texte.

Pour `CR`, nous avons procédé comme pour `CD`, c'est à dire : répertorier le nombre de conseiller·es dans chaque région au fil de leurs évolutions, et en particulier la réforme de 2015, et sommer en comptant seulement les institutions considérées comme des régions dans le RNE. Notre source principale ici était Wikipédia [40].

Pour `CM` et `EPCI`, nous avons utilisé une estimation parue dans la presse [41], qui se base sur des données du ministère de l'intérieur.

Pour `M`, nous nous sommes basés sur l'évolution du nombre de municipalités indiqué sur Wikipédia [42].

Pour le nombre de député·es et de sénateur·rices, nous nous sommes basés sur une étude d'impact de l'Assemblée Nationale [43], dont les valeurs diffèrent parfois assez de celles

40. https://fr.wikipedia.org/wiki/Conseil_régional_(France)
41. https://www.ledauphine.com/france-monde/2019/01/21/la-france-a-t-elle-trop-d-elus
42. https://fr.wikipedia.org/wiki/Nombre_de_communes_en_France
43. http://www.assemblee-nationale.fr/dyn/15/textes/l15b0977_etude-impact

indiquées sur Wikipédia [44] pour les sénateur·rices. En revanche, les deux sources concordent pour les député·es. Le nombre de député·es européen·nes semble bien documenté sur Wikipédia [45], aussi avons nous utilisé ces valeurs.

44. https://fr.wikipedia.org/wiki/Nombre_de_parlementaires_sous_la_Cinquième_République_française
45. https://fr.wikipedia.org/wiki/Député_européen#Compléments

D Contre-exemples

Un certain nombre de cas particuliers ne se conforment pas au comportement attendu, que celui-ci corresponde à un cadre réglementaire ou à une hypothèse issue de l'identification de régularités dans les données. Ils constituent donc des *contre-exemples* dans le contexte de nos tests, dans le sens où ils ne les satisfont pas tout en correspondent néanmoins à une description fidèle de la réalité. Nous listons ici certains de ceux que nous avons pu identifier, à des fins de référence.

D.1 Micro-mandats

Lorsqu'une ligne décrit une période de mandat très courte, nous qualifions celui-ci de micro-mandat. Comme indiqué en Section 2.2, nous considérons cela comme un problème, car une vérification manuelle a révélé que ces mandats étaient erronés dans l'écrasante majorité des cas.

Toutefois, il existe également des micro-mandats correspondant à des situations bien réelles. On peut citer `Max ROUSTAN`[46], qui a occupé le poste de sénateur du Gard seulement 6 jours, du 25/09/2017 au 30/09/2017, puis a démissioné. Il remplaçait `Jean-Paul FOURNIER`, lui-même démissionnaire.

D.2 Cumul de mandats identiques

On considère lors de nos tests qu'une personne ne peut pas occuper plusieurs positions électives de *même type* en *même temps* (voir aussi la Table 40).

Cela s'est pourtant déjà produit, notamment pour le conseiller général `Thierry ROBERT`[47], qui a été élu dans deux cantons en même temps et a pu occuper les deux postes. Il a d'abord été elu le 21/03/2008 dans le canton de `Saint-Leu-1` (La Réunion), puis dans celui de `Saint-Leu-2` le 5/09/2009 sans pour autant démissionner du premier. Le tribunal administratif de Saint-Denis l'a autorisé à cumuler pendant plus d'un an, avant qu'il ne démissionne de son poste de `Saint-Leu-1` le 28/12/2010.

D.3 Recouvrement de mandats identiques

On considère généralement qu'il ne peut y avoir de recouvrement de mandats, i.e. qu'une position élective identifiée de façon unique ne peut pas être occupée par plusieurs personnes en même temps. Dans le cas des conseils départementaux, la règle est différente, puisqu'à partir de 2015 deux conseiller·es représentent chaque canton, avec une contrainte de parité sexuelle. Nos tests considèrent qu'il y a recouvrement si les deux personnes sont du même sexe.

Or, cette situation s'est déjà produite, dans le canton de `Villeneuve-sur-Yonne`. Aux élections de 2015, Mme `Erika ROSET` et M `Claude THION` sont élus dans ce canton au second tour[48], c'est-à-dire le 29/03/2015. Cependant, `Claude THION` démissionne le 20/05/2015. Son suppléant `André FISCHER` refuse de le remplacer. Une élection partielle a lieu les 05/06/2015 et 12/06/2015, qui voit `Elisabeth FRASSETTO` être élue. On a donc bien un binôme de deux femmes (`Erika ROSET` & `Elisabeth FRASSETTO`), et par là-même un recouvrement.

46. https://www.senat.fr/senateur/roustan_max16431b.html
47. https://fr.wikipedia.org/wiki/Thierry_Robert_(homme_politique)
48. https://fr.wikipedia.org/wiki/Canton_de_Villeneuve-sur-Yonne

E Cumul des mandats

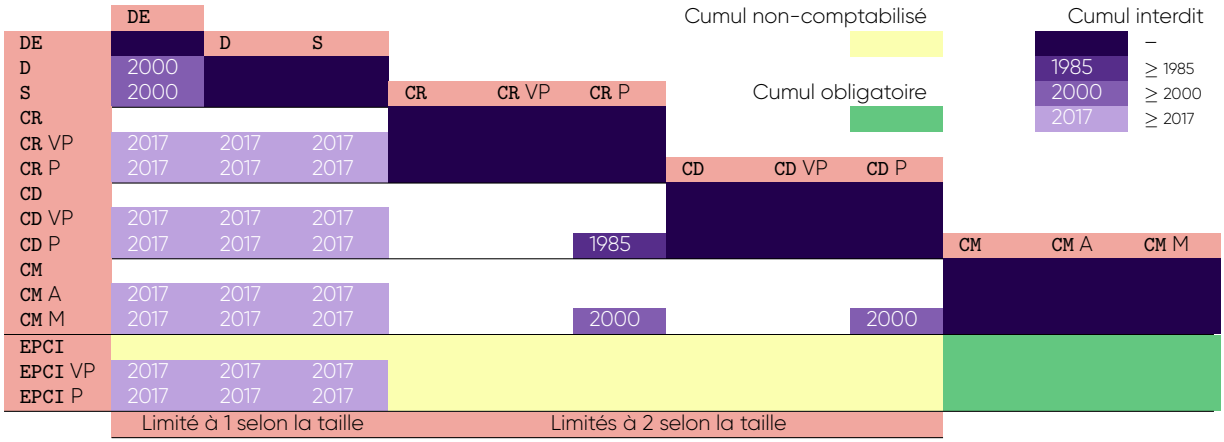


Table 40. Règles de cumul des mandats et fonctions exécutives. La signification des couleurs est la suivante : violet (cumul interdit, à partir de l'année indiquée), jaune (cumul non-comptabilisé), blanc (cumul autorisé), et vert (cumul obligatoire). Les fonctions concernées sont *président* (P), *vice-président* (VP), *maire* (M), et *adjoint au maire* (A).

F Versions de BRÉF

Version	Date	Description
`1`	24/06/2020	Données intégrées et révisées du RNE, Sénat, Assemblée, et Parlement Européen.
`1.0.1`	08/07/2020	Correction de deux erreurs : 1) un prénom manquant ; et 2) ID et noms incorrects pour certains départements d'`EPCI` Ajout d'une colonne manquante : ID des départements d'EPCI.
`1.0.2`	18/07/2020	Correction de deux erreurs concernant les EPCI : 1) certains EPCI distincts mais de même nom étaient incorrectement considérés comme un seul établissement ; et 2) le département renseigné pour certains EPCI étendus sur plusieurs départements ne correspondait pas au département principal.

Table 41. Versions successives de la base de données décrite dans ce document.

Liste des figures

Liste des tables