Qui sont les représentants d'intêrets? Une classification basée sur une analyse des données de la HATVP

Raphaël Lachièze-Rey*

Résumé

Ce travail se concentre sur une analyse statistique du registre des représentants d'intérêt géré par la Haute Autorité pour la Transparence de la Vie Publique (HATVP). En moyenne, plus de 10 000 commanditaires investissent annuellement un budget total de 215 millions d'euros pour mener des actions d'influence auprès des responsables publics français. Une des conclusions principales révèle que les intérêts privés représentent environ 87 % des actions menées, contre 8 % pour les actions d'intérêt général ou celles des collectivités, et 5 % pour les syndicats de travailleurs. Il est également constaté que les trois secteurs industriels les plus investis sont l'agroalimentaire, l'environnement et la santé.

Sur le plan technique, diverses méthodes statistiques sont employées pour compenser les lacunes et incohérences du registre. Plusieurs méthodologies sont comparées, et la stabilité des résultats témoigne de la solidité de ces analyses. De plus, des sous-analyses par secteur ou par catégorie de décisions ciblées sont également proposées, menant aux mêmes résultats.

Mots-clés : Lobbying, HATVP, Analyse statistique, influence privée

Abstract

This work focuses on a statistical analysis of the register of interest representatives managed by the Haute Autorité pour la Transparence de la Vie Publique (HATVP). On average, over 10,000 clients annually invest a total budget of 215 million euros to conduct influence activities with French public officials. One of the main conclusions reveals that private interests account for approximately 87% of the actions taken, compared to 8% for actions of general interest or those by local authorities, and 5% for labor unions. It is also noted that the three most invested industrial sectors are agri-food, environment, and health.

From a technical perspective, various statistical methods are employed to compensate for the gaps and inconsistencies in the register. Several methodologies are compared, and the stability of the results attests to the robustness of these analyses. Additionally, sub-analyses by sector or by category of targeted decisions are also proposed, leading to the same results.

Keywords : Lobbying, HATVP, Statistical analysis, private influence

*Laboratoire MAP5, Université Paris Cité, et Inria Paris, raphael.lachieze-rey@math.cnrs.fr

1 Introduction

L'objet principal de cette étude est la représentation d'intérêts d'acteurs privés ou publics via des actions d'influence auprès de responsables publics : élus, ministres, agents de l'état ; pratique globalement désignée sous le terme de *lobbying*. La loi Sapin II, promulguée en décembre 2016 en France, vise à renforcer la transparence et l'éthique dans la vie publique ainsi qu'à lutter contre la corruption. Parmi ses dispositions, elle a introduit la création d'un registre des représentants d'intérêts, géré par la HATVP. Cette mesure contraint les lobbyistes, les consultants et autres représentants d'intérêts à s'inscrire et à déclarer leurs activités de lobbying, y compris les rencontres avec les décideurs publics.

La question examinée ici est celle de la représentativité. On peut constater grâce à ce répertoire que différents types d'acteurs sont représentés : entreprises, syndicats, ONG, collectivités, etc... , dans de nombreux secteurs : économie, transports, environnement... La question de la quantification est néanmoins essentielle : plus la ressource dépensée est importante, plus l'empreinte normative sera grande. On cherche en particulier à savoir dans quelles proportion les intérêts privés sont représentés, afin de déterminer dans quelle direction la décision politique pourrait être influencée. La question est posée en ces termes dans Bombardini et Trebbi (2020) : " Est-ce que le lobbying est un mécanisme politique valable pour demander au gouvernement de remédier à des griefs, ou est-ce un outil qui déforme la politique publique en la détournant de l'optimum social, et un mécanisme clé pour transformer le pouvoir économique centralisé en influence politique et en corruption ?"

Un des objectifs finaux de ce registre est de fournir un panorama des influences normatives potentiellement induites par le lobbying. Cependant, l'analyse brute du registre ne permet malheureusement pas d'obtenir une vision satisfaisante. Sa finalité, à savoir la transparence, est parfois obstruée par le manque ou l'omission d'informations cruciales, et noyée dans la masse des données à traiter : 3 200 structures pour 71 000 actions recensées depuis la création du registre en 2016 jusqu'à Février 2024, dernière extraction prise en compte dans cet article. Les données indiquent un budget annuel du lobbying oscillant autour de 215 millions d'euros, impliquant près de 5000 lobbyistes par an. Il convient de noter qu'une portion du lobbying échappe à ce registre, soit par omission, soit par dérogation législative, le décret du 9 mai 2017 autorisant les firmes à ne pas déclarer une interaction dont une entité publique est à l'initiative. L'analyse est donc rendue difficile par une faiblesse en termes d'obligations déclaratives et de qualité concrète des données saisies. Seule une analyse fine mêlant des redressements automatiques basés sur des méthodes statistiques et des corrections manuelles peut permettre une vision débiaisée.

A notre connaissance, il n'existe pas d'autre travail académique portant sur l'ensemble du répertoire, on peut trouver des indicateurs aggrégés fournis par la HATVP elle-même dans des notes ou des rapports annuels [1]. Le premier objectif de cet article est de proposer une catégorisation lisible et pertinente des commanditaires dont les intérêts sont représentés, à l'aune de cette question de la représentativité, et une estimation de la qualité de cette classification ; afin d'en asseoir la légitimité, nous effectuons également un travail statistique d'analyse et de corrections des possibles erreurs du registre. Nous répartissons les commanditaires en deux catégories : d'une part, les intérêts particuliers ou privés (entreprises, groupements d'actionnaires, professions libérales, fédérations d'entreprises, associations et fondations oeuvrant pour des intérêts industriels ou privés, que nous désignerons par l'adjectif englobant "privé"), et, d'autre part, les

1. HATVP. Répertoire des représentants d'intêret, bilan 2023, https://www.hatvp.fr/wordpress/wp-content/uploads/2024/07/HATVP_BILAN_RRI-2023_VF.pdf.

associations œuvrant pour l'intérêt général ou le bien commun, et les collectivités, regroupés dans la catégorie "public". Selon cette classification, notre analyse révèle que 8% des mandants appartiennent à la catégorie "public", et 87% à la catégorie "privé", les 5% restant représentent les syndicats de travailleurs, qui ne rentrent pas de manière satisfaisante dans cette dichotomie. La motivation pour ce découpage, et en particulier les raisons pour lesquelles les catégories natives du registre officiel n'ont pas été utilisées, sont détaillées en Section 2, les outils statistiques utilisés pour établir et évaluer cette classification en Section 3. On donne également en Section 4.1 ces pourcentages secteur par secteur, et on constate que les proportions sont essentiellement conservées, hormis dans les secteurs où l'associatif est naturellement plus actif, comme l'aide sociale, ou les droits de l'homme et libertés fondamentales.

Il a été largement documenté par Kerleo et Monnery (2022), ainsi que les études qu'ils citent, que le public manifeste une grande défiance envers le lobbying. S'il est largement admis que les entreprises doivent pouvoir informer les décideurs des enjeux de leurs secteurs industriels respectifs, cette analyse met en lumière le poids exercé par les acteurs privés, créant un déséquilibre avec l'intérêt général. Il faudrait encourager une plus grande représentation de la société civile face à l'influence des lobbys privés.

Au niveau européen, le lobbying exercé par des acteurs privés peut avoir un impact plus significatif sur les textes de loi finalement adoptés que celui mené par les ONG [2], accentuant ainsi le déséquilibre. Bien qu'il existe peu de données pour confirmer cette tendance en France, nous proposons en Section 4.2 une analyse secondaire différenciant les types de responsables politiques visés par les actions de lobbying : collaborateur présidentiel, parlementaire, ministre, etc. Un autre biais possible réside dans le fait que les structures associatives et syndicales déclarent souvent plus de ressources que celles réellement dédiées à la représentation d'intérêts, masquant ainsi un déséquilibre encore plus grand.

La Figure 1 illustre l'évolution annuelle de diverses mesures liées aux représentations d'intérêt, notamment le nombre d'actions entreprises, le poids financier déclaré, le poids financier après corrections statistiques, ainsi que les parts respectives du secteur public et privé. On constate une légère augmentation du volume total, probablement attribuable à la prise en compte progressive par les firmes de lobbying de leurs obligations de déclaration. En revanche, la proportion entre le secteur public et le secteur privé demeure très stable. On observe que les montant sont légèrement inférieurs après correction ; une hypothèse est la sur-déclaration des structures associatives expliquée précédemment, les valeurs correspondantes sont alors détectées comme trop élevées et corrigées par l'algorithme.

Nous avons détecté dans notre analyse de nombreuses données incohérentes, manquantes, *anormales* (ou *aberrantes*, terme statistique qui signifie une erreur de saisie probable, *outlier* en anglais) ; et il est clair que les moyens de la HATVP ne lui permettent pas d'exercer de contrôle exhaustif sur les nombreux acteurs de ce secteur. Une étape primordiale de toute analyse est donc d'évaluer les éventuels biais qui pourraient être induits par ces problèmes. Nous avons appliqué des méthodes statistiques de correction de données erronées ou d'imputation de données manquantes, calculé des indicateurs en appliquant ces corrections et constaté que l'écart avec le calcul sans correction était faible. Cela ne fournit pas de garantie théorique absolue, mais c'est un fort indice que les lacunes du registre ne présentent pas de biais a même de fausser significativement les résultats. En l'absence d'autres résultats similaires sur cette base de données, ou de données étiquetées, il ne semble pas qu'une autre méthode mathématique plus élaborée pourrait fournir ce type de garantie.

2. Rozanne Logeart, https://rosannelogeart.github.io/research/

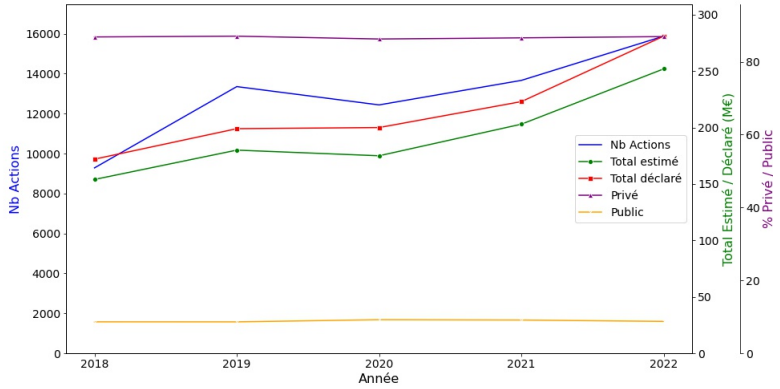


FIGURE 1 – Evolution temporelle de certaines métriques liées à la représentation d'intérêts : *axe de gauche, courbe bleue :* nombre annuel d'actions. *Axe de droite en vert, courbes vertes et rouges :* Budget annuel du lobbying sans correction (déclaré) et avec corrections (estimé). *Axe de droite en violet, courbes jaune et violette :* Pourcentage investi par commanditaires publiques et privés.

Nous avons aussi développé plusieurs méthodologies de calculs d'indicateurs, ou effectué des analyses sur des sous-échantillons, et toutes ces analyses ont mené peu ou prou aux mêmes résultats. Par exemple, les proportions parmi les classes énoncées plus haut se retrouvent dans la plupart des secteurs, et sur chaque année, ce qui permet de vérifier l'affirmation de stabilité via ces sous-échantillons. Par souci de concision, toutes les méthodologies ne sont pas présentées, l'ouverture du script sur un répertoire en ligne permettra à l'utilisateur de tester d'autres paramètres, voir la section suivante. En fin d'article, nous donnons des préconisations pour améliorer la récolte et la lisibilité des données du registre.

1.1 Ressources numériques

Afin d'assurer la reproductabilité des résultats présentés, nous mettons à disposition l'ensemble de nos données et méthodes. Un répertoire public est disponible au dépôt GitHub `plexigraf/lobbys-HATVP` [3]. Notre analyse repose sur un script `Python` appliqué au répertoire des données de la HATVP [4], et une classification des commanditaires [5]. Plusieurs paramètres sont ajustables par l'utilisateur, concernant la manière de traiter les erreurs de saisie et les données manquantes, d'estimer les ressources, ou de filtrer par période, voir le fichier `README`, et la section 5.2. Les paramètres sont modifiables en préambule du script `lobbys-rfap.py` [6]. Il a uniquement été appliqué à l'extraction de données antérieures à Février 2024, mais devrait fonctionner sur de nouvelles données ; il manquerait cependant la classification des clients nouvellement enregistré, qu'il faudrait effectuer indépendamment. Il génère des résultats numériques et plusieurs visualisations des données discutées ici. Dans ce répertoire [7], on peut trouver des dossiers contenant les résultats obtenus avec différents jeux de paramètres, et constater qu'ils

3. https://github.com/plexigraf/lobbys-HATVP
4. https://www.hatvp.fr/le-repertoire/
5. https://github.com/plexigraf/lobbys-HATVP/blob/main/classif_clients.json
6. https://github.com/plexigraf/lobbys-HATVP/blob/main/lobbys-rfap.py
7. https://github.com/plexigraf/lobbys-HATVP/tree/main/resultats

sont similaires.

En parallèle, une visualisation des données du registre sous forme de graphe interactif est disponible à cette adresse [8].

2 Classification

Le registre propose nativement une catégorisation des firmes exerçant du lobbying, mais sa pertinence est limitée. Le champ *catégorie* est parfois non-renseigné, et des acteurs de natures très diverses peuvent se retrouver dans la même catégorie ; l'exemple emblématique étant la catégorie *Associations* qui comporte syndicats agricoles, associations d'actionnaires de multinationales, ONG caritatives... Plus important, cette catégorisation concerne les entités effectuant le lobbying plutôt que les commanditaires de ces actions. Le registre présente une lacune manifeste en n'offrant aucune information détaillée sur ces derniers, si ce n'est le nom ou un acronyme parfois ambigu, ni aucune donnée sur les frais qu'ils engagent dans les actions qui les concernent. Comparativement, la législation américaine (Lobbying Disclosure Act) impose le dépôt d'une fiche distincte pour chaque client.

Toute subdivision précise se heurte à cette absence d'informations officielle, les frontières étant floues et subjectives entre de nombreuses catégories pertinentes, aussi nous avons choisi initialement un nombre de classes minimal, à savoir deux. Dés lors, les catégories "public" et "privé" semblent les plus pertinentes à l'aune de la question de la représentativité. Il importe de préciser que la classification est fondée sur les intérêts finaux que poursuit une entité. Même si l'objectif d'une entreprise peut être la protection de l'environnement, son modèle économique demeure soumis à l'impératif de rentabilité. A l'inverse, il a été documenté, par Peng (2016), Bertrand et al. (2021) que certaines ONG peuvent être influencées par des intérêts privés via des donations.

D'autres classes sont envisageables, mais l'avantage de cette classification est qu'il y a peu de recouvrement et d'ambiguité pour la plupart des acteurs, la proportion de cas discutables reste faible. Une catégorie hybride est celle des syndicats : la frontière est floue entre un syndicat de travailleurs, un syndicat d'employés d'un secteur spécifique, un syndicat agricole rassemblant des petits exploitants et des grands employeurs, et un syndicat de professions libérales. En raison du manque de détails disponibles sur les commanditaires, il est ardu de concevoir une classification plus fine basée sur ces informations. Une sous-catégorisation par secteur est présentée en section 4.1. Par ailleurs, pour réduire la complexité de l'analyse, une catégorie "collectivité" a été fusionnée avec la catégorie "public" en raison de sa taille insuffisante (environ 0,2% du total), et donc de l'importance négligeable sur les résultats finaux.

3 Méthodologie et robustesse

Le principe de notre méthodologie est relativement simple. Expliquons d'abord la structure du fichier. Le répertoire liste des *firmes*, qui chaque année déclarent un *exercice*, et pour chaque exercice des *actions de représentations d'intérêt* pour le compte de *tiers* (désigné ici comme *commanditaires*), sans préciser de manière nominative les responsables publics concernés. Il se limite à la catégorisation de ces derniers (*députés, membres de cabinet, agents de l'État*, etc.). La première tâche est d'assigner à chaque commanditaire une classe parmi : *intêrets privés*, *syndicats, interêt public*, comme discuté précédemment.

8. https://plexigraf.github.io/graph-v3.html?filename=hatvplobbys/lobbys.json

3.1 Classification

Les commanditaires ont été classifiés à l'aide de trois outils. La classification résultante utilisée pour notre analyse est disponible dans ce fichier [9].

Catégorie HATVP. Le premier outil est la catégorie HATVP, pertinente dans le cas où la firme représente ses propres intérêts ("Lobbying direct") et a une entrée dans le registre sous une catégorie non-ambigue. A titre d'exemple, les actions effectuées par *Google France* sont comptabilisés en *privé* car cette firme est catégorisée *Société commerciale* dans le registre.

Base INSEE. Le second outil est la base INSEE, qui répertorie les entreprises [10] et associations françaises [11], et fournit éventuellement des informations supplémentaires sur leurs activités et leur fonctionnement. Nous avons interrogé cette base via le nom de chaque commanditaire, seul élément disponible dans le registre. Les informations récoltées sont disponibles dans le dépôt GitHub [12]. Lorsque des données ont été trouvées, les firmes ont été classées en *privé* si leur *objet social*, leur *activité principale*, leur *forme juridique* ou leur *famille* est dans une certaine liste pré-établie (*commerce de gros, gestion immobilière, cabinet d'avocats*, ...). Le même traitement a été appliqué pour éventuellement classifier une firme en *public*, via des mots-clés comme *Association déclarée reconnue d'utilité publique, Établissement public national à caractère industriel ou commercial, Commune, ...*), en *syndicat*, en *syndicat agricole* ou en *mutuelle*, avec les mots-clés appropriés. A l'issue de cette première phase, les entités restantes ont été classées comme privées si leur dénomination comporte des mots spécifiques (*SARL, SAS, HOLDING, ...*). De nombreuses entités se situent à l'intersection de plusieurs catégories, ou les informations disponibles ne permettent pas une classification claire. Pour ces cas-là, nous attribuons une proportion à chaque classe afin de mieux refléter leur positionnement. Les financements de syndicats agricole ont ainsi été répartis à 50% en "privé", et 50% en "syndicat", en raison de la diversité de leurs membres. Ce chiffre peut paraître arbitraire, mais comme la part de syndicats agricole est d'environ 5%, choisir une autre valeur entre 0 et 100% ne peut mathématiquement pas modifier les proportions finales de plus de 2%, il ne sera pas difficile pour un utilisateur de répertoire Github d'observer les effets d'une modification de ce paramètre. Selon le même principe, les assurances de type mutuelle sont classées à 50% en intérêt général et 50% en intérêts privés car elles sont en général à but non-lucratif et oeuvrent souvent pour des améliorations globales dans le secteur du bien-être, mais elles représentent in fine les intérêts de leurs adhérents, et sont de plus parfois à but lucratif (cette propriété est difficile à déceler sans une étude approfondie), là encore, une autre distribution modifiera les proportions finales d'au plus 1%. Ces catégories d'entités illustrent toute l'ambiguité de ce type de classification, l'important étant de ne pas adopter de biais particulier dans leur traitement. Les classifications établies à ce stade sont dites "manuelles", ce qui est également indiqué dans le fichier [16].

Pour les firmes non classées à l'issue de cette phase, nous recherchons des correspondances avec des entités dont le nom est fortement similaire à celui d'une entité déjà notée [13] , à laquelle nous attribuons alors la même note. Cette information est également disponible dans le dépôt. A noter que des erreurs subsistent dans cette première phase en raison d'erreurs de saisie de la

9. https://github.com/plexigraf/lobbys-HATVP/blob/main/classif_clients.json
10. https://annuaire-entreprises.data.gouv.fr/
11. https://sirene.fr/sirene/public/accueil, https://www.data-asso.fr/
12. https://github.com/plexigraf/lobbys-HATVP/blob/main/firms_suppl_info.json
13. 90% pour la distance de Levenshtein, voir documentation du package `fuzzywuzzy` https://pypi.org/project/fuzzywuzzy/

part des entreprises/associations ; des acteurs du monde professionnel renseignent par exemple parfois à tort comme “activité principale” une “actions sociale” dans la base INSEE), ou en raison de structures aux dénominations très proches mais aux finalités divergentes.

Large Language Model.
Les classifications issues de la phase précédentes concernent 9 500 clients. Pour les commanditaires restants, environ 1 500, nous avons effectué une classification à l'aide d'un Large Language Model (spécifiquement GPT-3.5 d'OpenAI), basé uniquement sur le nom, et l'objet social si disponible, avec un *prompt* basé sur les principes expliqués en Section 2, et comportant des exemples étiquetés manuellement. Cet algorithme a attribué une répartition entre nos trois catégories. Cette méthode présente l'avantage de ne pas nécessiter d'entraînement statistique préalable. Bien qu'elle puisse produire des erreurs, celles-ci sont limitées à un faible pourcentage [14] , ce qui ne devrait pas introduire de biais significatif dans les résultats (la proportions de cas traités par un LLM se calcule à partir de ce fichier). Si certaines classifications spécifiques présentent des ambiguïtés pouvant être discutées, il est plus difficile de démontrer un biais global susceptible d'influencer les résultats finaux.

Le classement global est disponible sur le dépôt github [15]avec une explication brève pour chaque item. Si un utilisateur souhaite lui-même évaluer le taux d'erreur, l'execution du script `lobbys-rfap.py` fournit automatiquement en sortie console des échantillon aléatoires issus de chaque classe, il est donc aisé de comparer avec un étiquetage manuel de cet échantillon.

3.2 Prise en compte des ressources

Chaque action de lobbying ne bénéficie pas du même financement, et il convient donc de pondérer son influence par les ressources effectivement dépensées par la firme. Cette imputation est rendue difficile par le fait que les firmes ne sont pas tenues de rapporter leur budget par action, mais par exercice, ce qui pose un problème de granularité, la durée typique d'un exercice étant d'un an. Chaque firme doit déclarer une liste de *clients*, et en plus pour chaque action une liste de *tiers*, mais ces deux listes sont parfois incohérentes, notamment pour les actions remontant à plusieurs années ; il est donc plus pertinent de se baser sur les tiers désignés pour chaque action. Les clients comme les tiers sont désignés par le terme unique *commanditaires* dans cet article. De nombreuses actions sont effectuées “en propre”, donc pour le compte de la firme, mais il s'agit aussi souvent de ne pas mettre de liste de tiers, parfois par négligence. Lorsque les tiers d'une action ne sont pas mentionnés et que la mention *en propre* n'est pas présente, nous avons automatiquement affecté l'action aux clients de la firme s'ils sont indiqués, et sinon à la firme elle-même.

Calcul du coût d'une action. Le parti pris dans cette étude est de diviser les ressources annuelles d'une firme entre toutes les actions menées durant l'exercice concerné. Comme les actions ne sont pas non plus financées à parts égales, nous avons affecté un score s_a à chaque action a, égal au nombre d'items qu'elle présente (nombre de tâches à effectuer, de catégories de responsables visés, de tiers concernés, de décisions politiques). Le budget b_a alloué à l'action

14. Nous avons empiriquement constaté moins de 5% d'erreurs. Comme cette classification LLM représente 15% des entrées, la proportion de misclassifications serait de moins de 1%, et il ne semble pas que les erreurs se produisent plus dans un sens que dans l'autre, n'imprimant donc pas de biais sur les résultats.
15. https://github.com/plexigraf/lobbys-HATVP/

est alors

$$b_a = B_e \frac{s_a}{\sum_{a=1}^{N_e} s_a}$$

où N_e est le nombre d'actions de l'exercice e, s_a le score de la a-ème action de cet exercice, et B_e est le budget total de la firme sur l'exercice e, renseigné dans le registre. Pour savoir comment le budget b_a va effectivement être ventilé entre les classes, il faut partir des classifications des tiers (dans ce fichier [16]). On note $c_{i,j} \in [0,1]$ la proportion du tiers i imputé à la classe j, pour $j \in \{\text{privé, public, syndicat}\}$ et $1 \leqslant i \leqslant N_a$ où N_a est le nombre de tiers de l'action a (on n'a d'autre choix que de considérer que tous les tiers ont la même influence). En définitive, étant donné une classe j, la contribution $C_{j,a}$ à la classe j de l'action a menée lors de l'exercice e est

$$C_{j,a} = b_a \times \sum_{i=1}^{N_a} c_{i,j}.$$

Comme $\sum_j c_{i,j} = 1$, on vérifie bien que le budget réparti entre toutes les classes est bien le budget total de l'action $\sum_j C_{j,a} = b_a$. La quantité totale de ressources allouées à la classe j sera la somme sur toutes les actions répertoriées

$$C_j = \sum_a c_{j,a}.$$

Les pourcentages annoncés en introduction du budget total affecté à chaque classe sont les proportions respectives de ces montants pour les différentes classes, e.g. pour la classe privé

$$\frac{C_{\text{privé}}}{\sum_{j \in \{\text{privé,public,syndicats}\}} C_j} = 0,87.$$

Autres méthodologies. Nous avons également testé d'autres manières d'imputer un budget à une action. Par exemple, au sein d'un exercice, toutes les actions reçoivent un budget équivalent $b_a = B_e/N_e$. Une autre manière de faire est de considérer que, indépendamment du cabinet et de ses ressources, chaque action du répertoire reçoit la même ressource $b_e = c$ où c est le coût moyen d'une action estimé à partir de l'ensemble du répertoire. Ces deux paradigmes semblent moins pertinents, leurs résultats peuvent être consultés dans les résultats présents sur le dépôt GitHub [17]. Les résultats obtenus sont dans tous les cas similaires (corrélation supérieure à 95%), ce qui permet de s'assurer que les erreurs de saisie au niveau des budgets influencent marginalement les résultats finaux. D'autres formules peuvent être expérimentées en modifiant le script.

4 Analyses transversales

4.1 Ventilation par secteur et robustesse

Nous présentons dans cette section une répartition des dépenses de lobbying par thème ; le détail se trouve dans le tableau ci-dessous. Quatre catégories semblent se détacher comme plus

16. https://github.com/plexigraf/lobbys-HATVP/blob/main/classif_clients.json
17. https://github.com/plexigraf/lobbys-HATVP/

représentées sur la période 2017 - 2024 et regroupent près de la moitié des investissements : *Environnement et développement durable* (141,9M€), *Santé et pharmaceutique* (131,3M€), *Economie et Finances Publiques* (126,1M€), *Agro-alimentaire* (121,6M€), les autres secteurs oscillant entre 27,45M€ (*Gouvernance et institutions publiques*) et 76,0M€ (*Construction, logement, aménagement du territoire*), une visualisation des montants est disponible à la figure 2-(a). Les proportions de chaque classe sont indiquées dans le tableau ci-dessous.

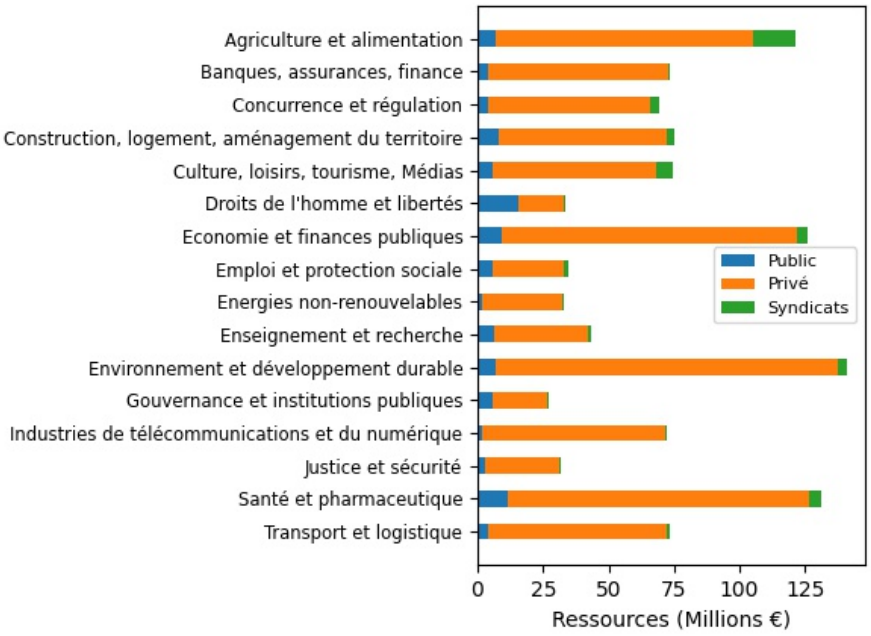
Thème	Privé (M€)	Syndicats (M€)	Public (M€)	Total
Agriculture et alimentation	98.27 (81%)	16.36 (13%)	6.95 (6%)	121.64
Banques, assurances, finance	68.54 (94%)	0.28 (0%)	4.24 (6%)	73.11
Concurrence et régulation	61.88 (89%)	3.38 (5%)	3.99 (6%)	69.33
Construction, logement, aménagement du territoire	64.33 (85%)	3.15 (4%)	7.86 (10%)	760
Culture, loisirs, tourisme, Médias	62.78 (84%)	6.42 (9%)	5.54 (7%)	74.87
Droits de l'homme et libertés	17.47 (52%)	0.53 (2%)	15.44 (46%)	33.45
Économie et finances publiques	112.92 (90%)	3.93 (3%)	90 (7%)	1269
Emploi et protection sociale	27.55 (79%)	1.59 (5%)	5.60 (16%)	34.77
Énergies non-renouvelables	30.59 (92%)	0.56 (2%)	1.93 (6%)	33.14
Enseignement et recherche	36.22 (83%)	0.93 (2%)	6.20 (14%)	43.71
Environnement et développement durable	130.41 (93%)	3.41 (2%)	71 (5%)	140.94
Gouvernance et institutions publiques	21.20 (77%)	0.42 (2%)	5.55 (20%)	27.45
Industries de télécommunications et du numérique	69.99 (97%)	0.68 (1%)	1.50 (2%)	72.41
Justice et sécurité	28.26 (89%)	0.32 (1%)	36 (10%)	31.83
Santé et pharmaceutique	114.48 (87%)	52 (4%)	11.75 (9%)	131.35
Transport et logistique	68.61 (93%)	0.99 (1%)	3.87 (5%)	73.55

Le privé reste très majoritaire (plus de 77%) dans chaque secteur, et les proportions globales ne sont pas bouleversées, à l'exception de *Droits de l'homme et liberté*, où le public est à 49,4%, et l'agro-alimentaire, où les syndicats représentent 13% (on rappelle que les syndicats agricoles sont ventilés à parité entre *privé* et *syndicats*). Il est étonnant que même le thème *Droits de l'Homme et libertés* représente 17,5M€ d'investissements privés (soit 52%) qui peuvent venir par exemple de cabinets de conseil en stratégie. Pour comprendre le mécanisme correspondant d'extrapolation de données de notre méthodologie, développons l'exemple de COMYA France, un cabinet d'avocats. Ce dernier déclare parmi ses 29 secteurs privilégiés les *Questions migratoires*, et n'ajoute pas d'autres *domaines d'intervention* dans le détail de ses actions, ce qui fait que son budget est réparti automatiquement entre tous ses secteurs déclarés, le thème *Droits de l'homme et libertés* se voit donc affecter 1/29ème de son budget.

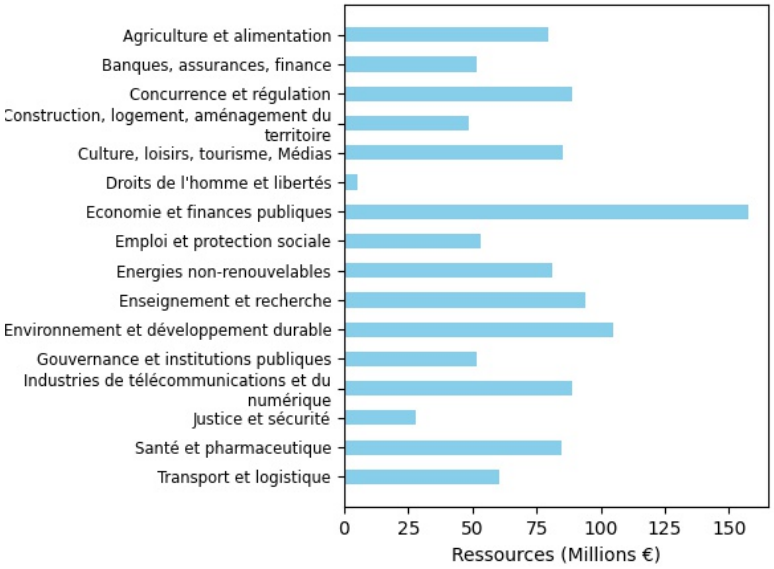
Secteurs et domaines. En ce qui concerne la répartition par secteurs, il y a aussi une ambiguité dans le registre. D'une part, chaque firme déclare des "secteurs" privilégiés. D'autre part, pour chaque action effectuée, une liste de "domaines d'intervention" est (parfois) indiquée. La liste de tous les "secteurs" possibles n'est pas la même que celle de tous les "domaines" possibles, et ces deux listes sont de toute façon trop grandes pour une visualisation, c'est pourquoi nous avons créés les thèmes plus englobants du tableau 6, les secteurs et domaines correspondants sont dans la colonne de droite. On dispose donc de deux manières de comptabiliser les actions par thèmes : soit en tenant compte des "secteurs" déclarés par les firmes (méthode NF), soit en tenant compte des "domaines" déclarés pour chaque action (méthode AC), c'est cette dernière méthode qui sera retenue.

Stabilité. Nous avons donc comparé plusieurs méthodologies pour effectuer cette classification, qui amènent à des répartitions proches. Les résultats des deux méthodologies NF et AC sont représentées en Figure 2. En (a), on compte les investissement par action, avec correction des valeurs anormales, et en (b) on compte les investissements par firmes, les deux résultats ont une corrélation de 0,71, la disparité principale se trouve au niveau des secteurs de la Santé et l'Agro-alimentaire, sur-représentés, cela dénote vraisemblablement des erreurs récurrentes de saisie de données dans ces secteurs. On donne ci-dessous les corrélations des distributions par secteurs avec d'autres méthodologies :

AC Investissements par actions, avec corrections statistiques des budgets
NF Nombre de firmes d'un secteur donné, avec corrections
AU Investissement égal pour chaque action
PEAC Actions uniforme sur chaque exercice, avec corrections



(a) Poids des investissements en actions de lobbying, après correction des valeurs anormales, croisés avec les classes des commanditaires



(b) Nombre de firmes déclarant un secteur donné, pondéré par le budget.

FIGURE 2 – Deux méthodologies de comptage par thème

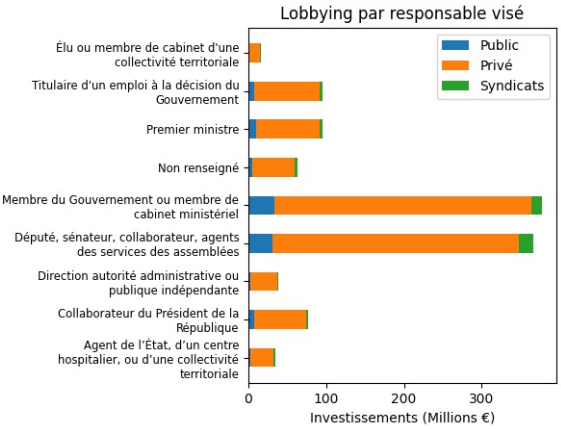
PESC Actions uniforme sur chaque exercice, sans corrections
2018 Sur une année spécifique

Corrélation	AC	NF	AU	PEAC	PESC	2018	2019	2020	2021	2022
Avec corrections	1	0,71	0,93	1	0,98	0,96	0,99	0,99	0,97	0,98
Par nombre de firmes	0,71	1	0,57	0,7	0,62	0,99	1,0	1,0	0,99	0,99

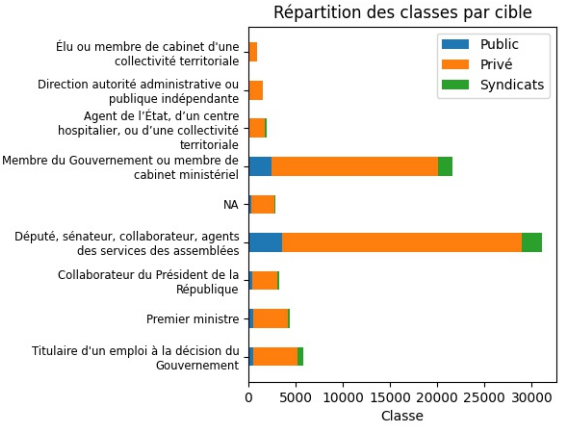
On observe que les quatre méthodes AC, AU, PEAC, PESC donnent des résultats très proches (corrélation $> 0,93$), et qu'on a une forte stabilité des proportions de chaque secteur au fil des années (corrélation $> 0,98$). C'est la méthodologie principale retenue dans cet article, les chiffres qui en sont issus devraient fidèlement représenter les actions concernées par le registre. Le comptage des secteurs par nombre de firmes (méthode NF) semble un peu moins corrélé, mais la stabilité temporelle est tout de même conservée.

4.2 Responsables politiques visés

Les firmes de lobbying indiquent pour chaque action les décisions politiques concernées et les responsables politiques visés. Nous visualisons ici un croisement entre les catégories de responsables visés, et la classification des commanditaires représentés. Il en ressort que les parlementaires et membres du gouvernement sont de loin les plus sollicités avec respectivement 31% et 32% des investissements. Les distributions des catégories *public* et *privé* sont essentiellement conservées au sein des classes.



(a) Avec correction des données et imputation des ressources



(b) Sans imputation des ressources (actions uniformes)

FIGURE 3 – Investissement des ressources de lobbying par type de responsable visé.

Pour comparaison, on a représenté sur la figure 3-(b) les répartitions si l'on ne prend pas en compte les ressources. La principale différence tient à ce que dans la figure a, le budget d'une firme est distribué entre toutes les actions de l'exercice concerné, alors que dans la figure b, chaque action reçoit le même budget, indépendamment de la firme et de l'exercice. La corrélation entre les deux distributions est de 97,6%, ce qui indique une grande stabilité des résultats lorsque la méthodologie varie, même si les parlementaires et membres du gouvernement

semblent plus dominants dans le second cas. Si le scénario b est en soi peu pertinent, il nous informe sur le fait que les résultats changent peu si l'on ne corrige pas les erreurs, indiquant que celles-ci ne présentent pas de biais significatif.

5 Précaution d'usage et préconisations

5.1 Valeurs anormales

De nombreuses erreurs ou omissions existent dans les données. Il semble aussi que certaines entités ne déclarent pas correctement. Un biais récurrent est de déclarer toutes les ressources humaines ou financières d'un syndicat ou d'une association et pas seulement celles dévolues à la représentation d'interêt, biais qui a également été relevé au niveau du registre européen [18]. Cela concerne essentiellement les structures associatives et syndicales, ce qui indiquerait typiquement une proportion du public plus basse que celle annoncée en introduction. Plusieurs options sont possibles ; on peut choisir d'ignorer ces données, et donc toutes les informations rentrées par les firmes concernées, affecter une valeur "standard" indifférente des autres données de la firme, ou tenter de trouver la meilleure valeur de remplacement (imputation) selon certains critères. Nous optons ici pour la dernière possibilité, car même si une valeur est manquante ou aberrante, par exemple le budget, il est pertinent de prendre en compte les autres informations (nombre de salariés et quantité d'actions menées), pour par exemple le calcul global du budget du lobbying.

Nous avons du appliquer là encore une procédure statistique pour détecter les valeurs manquantes, les remplacer par des valeurs vraisemblables, détecter les données anormales, et les remplacer également, comme expliqué à la section ci-dessous.

5.2 Algorithme des plus proches voisins

Nous avons employé une démarche statistiques basée sur l'algorithme des plus proches voisins pour détecter les valeurs anormalement basses ou élevées, et pour remplacer les valeurs manquantes, en particulier afin de tester la robustesse des résultats. Cette méthodologie a été choisie pour sa simplicité, d'autres alternatives impliquant des regressions plus complexes ou de l'apprentissage auraient pu être choisies, mais au vu du peu d'écart entre les résultats obtenus avec ou sans correction, le gain avec une méthode plus élaborée serait marginal.

Le principe est le suivant. Etant donné un exercice e d'une firme, soit n_e le nombre de lobbyistes déclarés pour l'exercice, et supposons que le montant dépensé m_e ne soit pas renseigné ; il faut alors trouver une valeur de remplacement pour estimer les ressources totales. On calcule S_e la somme de tous les scores des actions menées lors de l'exercice e, calculés comme à la Section 3.2. Nous allons affecter à l'exercice e une valeur fictive m_e^* pour le montant, et pour faire cela on va choisir une valeur proche du montant moyen des exercices d'autres firmes aux valeurs similaires pour S_e, n_e. On cherche donc d'autres exercices $e_1, \dots, e_{10}$ (éventuellement d'autres firmes) tels que les valeurs $n_{e_1}, \dots, n_{e_{10}}$ et les scores de leurs actions $S_{e_1}, \dots, S_{e_{10}}$ soient proches respectivement de n_e et S_e. Mathématiquement, on cherche les 10 plus proches voisins $\{(S_1, e_1), \dots, (S_{10}, \dots, e_{10})\}$ de (S_e, n_e) dans $\mathbb{R}^2$. Il faut également que pour ces firmes soient renseignés les montants $m_{e_1}, \dots, m_{e_{10}}$. On décide alors d'imputer la valeur m_e^* à l'exercice e comme la moyenne de ces exercices proches :

$$m_e^* = \frac{1}{10} \sum_{i=1}^{10} m_{e_i}.$$

Le chiffre 10 est arbitraire et résulte d'un tâtonnement. La méthode est similaire pour une firme dont le montant est renseigné et le nombre de salariés pas renseigné ; on cherche les plus proches voisins en termes de score d'actions et de montants, et on impute la moyenne des nombres de salariés correspondants ; à noter que ce chiffre peut ne pas être entier. Finalement, si une firme ne renseigne ni montant ni nombre de salariés, mais déclare des actions pour un score total S_e, alors on fait simplement les moyennes pour les exercices ayant des valeurs S_{e_i} proches de S_e.

Expliquons désormais comment détecter automatiquement qu'une valeur d'un montant m_e est anormale au regard du nombre de salariés n_e et du score S_e. Selon la procédure précédente, on collecte les valeurs $m_{e_1}, \dots, m_{e_{10}}$ des montants pour d'autres firmes ayant déclaré des exercices proches, et on compare la valeur m_e à l'échantillon $M = (m_{e_1}, \dots, m_{e_{10}})$. Pour cela, on calcule la moyenne m_e^* et la déviation standard σ_e^* de l'échantillon M, et on

18. https://www.integritywatch.eu/organizations.php

déclare m_e comme anormale si sa déviation à la moyenne m_e^* est plus grande qu'un multiple $\alpha\sigma_e^*$ de la déviation standard, c'est-à-dire si

$$\frac{|m_e - m_e^*|}{\sigma_e^*} > \alpha,$$

où α est un seuil calculé empiriquement sur l'ensemble des données par tâtonnement. Plusieurs valeurs de α sont proposés dans le script, une valeur α_{strict} plus petite qui déclare donc plus facilement qu'une valeur est anormale, et une valeur plus conservatrice $\alpha_{\text{conserv}} > \alpha_{\text{script}}$. [19].

5.3 Stabilité des résultats

On présente ici les résultats de divers jeux de paramètre. La première colonne redonne les résultats avec les outils statistiques présentés ici. La colonne *actions uniformes* donne les résultats sans prendre en compte les ressources des firmes, i.e. chaque action a le même poids financier. La colonne *Nombre de commanditaires* donne les pourcentages en comptant le nombre de clients d'une classe donnée. Les 4 dernières colonnes donnent les résultats sur des années spécifiques. Les corrélations entre les différentes classifications sont supérieures à $0,99$, ce qui indique une excellente adéquation, les résultats sont donc essentiellement conservés en variant de méthodologie.

	Avec correction	Actions uniformes	Nombre de commanditaires	2018	2019	2020	2021	2022
prive %	87,01	86,33	82,44	86,81	87,02	86,24	86,55	86,90
public %	8,55	8,31	10,56	8,6	8,6	9,2	9,1	8,7

Dans une méthodologie très pauvre où on ne prendrait pas du tout en compte les ressource déclarées par les firmes, ce qui revient à dire que chaque action aurait le même poids financier, on obtient des pourcentages de respectivement 82% et 11% pour le privé et le public. D'une manière générale, en testant plusieurs jeux de paramètres possibles, on observe une variation maximale globale de 4% sur les pourcentages calculés, ce qui reste raisonnable et montre que l'ampleur des erreurs de saisie est relativement contrôlée. Ces problèmes ne semblent également pas affecter les résultats globaux, sauf peut-être dans certains secteurs sur- ou sous-représentés si l'on ne corrige pas ces valeurs. Voir le dépôt GitHub [21] pour plus de détails.

5.4 Préconisations

Effectuons quelques recommandations simples sur la saisie des données qui permettront d'avoir accès à un panorama plus précis et donc plus de confiance dans les résultats obtenus. Ces recommandations sont essentiellement techniques, mais d'autres recommandations structurelles ou juridiques ont été énoncées dans Kerleo (2023); Kerleo et Monnery (2022). La HATVP publie également ses propres préconisations juridiques, voir par exemple p.10 du dernier bilan de la HATVP [22]

- Exiger plus d'informations sur les tiers ; au minimum, les identifiants nationaux type RNA et SIREN afin d'identifier chaque structure de manière non-ambigue et pouvoir récupérer des informations pertinentes, ou leurs pays d'affiliation pour les entités étrangères.
- De manière évidente, la granularité actuelle des budgets est trop vague, il serait bon d'indiquer la ventilation du budget annuel pour chaque action, et indiquer les contributions des tiers concernés au financement.
- Rendre cohérents la rubrique "clients" de la firme, avec les listes de "tiers" (commanditaires) pour les actions déclarées, des clients n'étant enregistrés comme tiers pour aucune action. La pertinence d'avoir ces deux rubriques n'est pas claire, les incohérences rencontrées sont listées à l'entrée `incohérences clients/tiers` de chaque résultat dans le dépôt github [23].
- Dans le même esprit, rendre cohérents les "secteurs" déclarés sur la fiche d'une entreprise, et les "domaines d'intervention" qui accompagnent chaque action. Les listes de secteurs proposés dans ces deux

19. Le paramètre `STRICT` peut être réglé dans le script `lobbys-rfap.py` [20] pour choisir quelle valeur utiliser. On peut aussi régler le paramètre `CORRECT_OUTLIERS` à `FALSE` pour laisser telles quelles les valeurs anormales. L'algorithme des plus proches voisins est implementé par le package `sklearn`

21. https://github.com/plexigraf/lobbys-HATVP/

22. HATVP. Répertoire des représentants d'intêret, bilan 2023, https://www.hatvp.fr/wordpress/wp-content/uploads/2024/07/HATVP_BILAN_RRI-2023_VF.pdf.

23. https://github.com/plexigraf/lobbys-HATVP/tree/main/results

catégories ne sont pas les mêmes, et elles sont parfois contradictoires pour une firme donnée. Ici encore, on pourrait se contenter des domaines déclarés sur chaque action afin d'éviter de la redondance, de l'incohérence, et du travail inutile pour les firmes.

— Appliquer des filtres statistiques de détections de valeurs anormales lors de la saisie (comme celui de la Section 5.2), afin de demander des confirmations aux agents entrant ces données en cas d'anomalies ; et demander des budgets annuels exacts plutôt que des fourchettes.

6 Conclusion

Le registre HATVP présente des imperfections et des incohérences. Malgré cela, il remplit sa fonction principale en fournissant des informations extrêmement précieuses pour comprendre la nature du lobbying en France. Il est particulièrement notable que de nombreuses catégories de la population sont insuffisamment représentées auprès des décideurs publics, tandis que l'influence privée est financièrement environ dix fois supérieure à celle d'autres acteurs, mis à part les syndicats.

Références

Monique Bertrand, Matilde Bombardini, Raymond Fisman, Brad Hackinen, et Francesco Trebbi. Hall of mirrors : Corporate philanthropy and strategic advocacy. *The Quarterly Journal of Economics*, 136(4) :2413–2465, 2021.

Matilde Bombardini et Francesco Trebbi. Empirical models of lobbying. *Annual Review of Economics*, 12 : 391–413, 2020.

Jean-François Kerleo. Pour un contrôle de la représentation d'intérêts efficace et transparent. Note 31, Observatoire de l'éthique publique, 2023.

Jean-François Kerleo et Benjamin Monnery. Probité et transparence au parlement : bilan et leçons d'une décennie de changements autour de la hatvp. *Revue Française de l'Administration Publique*, 4(184) :1097–1113, 2022.

Victora Peng. Astroturf campaigns : Transparency in telecom merger review. *University of Michigan Journal of Law Reform*, 49(2) :521, 2016.

Appendice : correspondance thèmes, domaines et secteurs

Thème	Domaines et secteurs
Construction, logement, aménagement du territoire	Occupation des sols, Patrimoine, Développement des territoires, Bâtiments et travaux publics, Construction, logement, aménagement du territoire, Logement, Construction
Emploi et protection sociale	Assurance chômage, Handicap, Droit du travail, Dialogue social, Famille, Emploi, solidarité
Enseignement et recherche	Recherche et innovation, Enseignement supérieur, recherche, innovation, Éducation, enseignement, formation, Enseignement supérieur, Formation professionnelle, Statistiques
Droits de l'homme et libertés	Aide au développement, Humanitaire, Principe de précaution, Egalité femmes/hommes, Bien-être animal, Egalité des chances, Asile, Immigration, Discriminations, Questions migratoires, Droits et libertés fondamentales, Liberté d'expression et d'information, Droits des victimes, Laïcité
Industries de télécommunications et du numérique	Accès à l'Internet, Accès aux moyens de télécommunications, Protection des données, E-commerce, Numérique, Télécommunications, Infrastructures de télécommunications, Marché du numérique, E-commerce
Santé et pharmaceutique	Médicaments, Remboursements, Soins et maladies, Prévention, Santé, Système de santé et médico-social, Soins et maladies
Banques, assurances, finance	Banques, assurances, secteur financier, Protection des marques, Banques, assurances, secteur financier et extra financier, Finances, Banques, Assurances
Gouvernance et institutions publiques	Outre-mer, Français de l'étranger, Collectivités territoriales, Pouvoirs publics et institutions, Institutions européennes, Institutions des outre-mer, Ordre administratif, Moralisation/Transparence, Coopération internationale, Accords internationaux, Economie des outre-mer, Développement économique des outre-mer
Concurrence et régulation	Secret des affaires / Secret professionnel, Aides aux entreprises, Brevet, Droit de la concurrence, Marchés réglementés, PME/TPE, Normes de production, Professions réglementées, Entreprises et professions libérales, Concurrence, consommation, Gouvernance d'entreprise, Commerce extérieur
Economie et finances publiques	Société, Taxes, Politique industrielle, Fonction publique, Retraites, Partenariats public/privé, Finances publiques, Economie, Budget, Taxation, Impôts
Energies non-renouvelables	Energie nucléaire, Energie, Ressources naturelles, Ressources minières, Energies fossiles
Environnement et développement durable	Chasse, Eaux, Produits chimiques, Accidents et catastrophes naturelles, Impact des transports individuels, Impact des transports marchands et collectifs, Qualité de l'eau, Environnement, Déchets, Energies renouvelables, Impact de l'activité industrielle, Dépollution, Forêt
Justice et sécurité	Défense, Institutions pénitentiaires, Justice, Institutions judiciaires, Justice civile, Justice pénale, Défense, sécurité, Sécurité nationale, Sécurité routière, Espionnage et surveillance
Culture, loisirs, tourisme, Médias	Médias, Tourisme/hôtellerie, Droit d'auteur, Publicité, Sports, Propriété intellectuelle, Sports, loisirs, tourisme, Accès à la culture, Arts, culture, Musique, Spectacle vivant, Cinéma, Presse écrite, Audiovisuel, Jeux d'argent, Jeux-vidéo, Livre
Transport et logistique	Industrie aérospatiale, Infrastructures, Industrie aéronautique, Aéronautique, aérospatiale, Transports, logistique, Transport de voyageurs, Transport de fret, Transports alternatifs, Services postaux
Agriculture et alimentation	Sécurité et normes alimentaires, Appellations, Agriculture, agroalimentaire, Agriculture, Industrie agroalimentaire, Pêche