MODÉLISATION DE L'IMPACT DE FACTEURS SOCIO-ÉCONOMIQUES SUR LA RÉPARTITION DES VOIX LORS D'UNE ÉLECTION

Par Christine THOMAS-AGNAN
et Lukas DARGEL[1]

Ce texte est issu d'un travail collectif effectué lors d'un projet proposé par Alejandro Lara et Olivier Hennebelle de l'entreprise BVA dans le cadre de l'atelier de conseil statistique du master Statistique et Économétrie de l'université Toulouse 1 Capitole. Il a été réalisé d'octobre à mars 2023 à l'aide d'une équipe d'étudiants constituée de Malo Bert, Gaël Charrier, Kyllian James et Claire Lebrun, encadrés par Christine Thomas-Agnan et Lukas Dargel, étudiant en thèse IFRE chez BVA.

L'objectif principal est de construire un modèle statistique visant à approfondir notre compréhension des relations entre les caractéristiques socio-économiques des territoires et les résultats électoraux des différents candidats lors d'une élection telle que le premier tour de l'élection présidentielle 2022. Ce texte permettra d'explorer de manière concrète la manière dont les mathématiques interagissent avec les enjeux sociétaux, à travers une question de sociologie électorale. Il est essentiel de souligner que notre démarche ne vise pas à prédire les résultats de futures élections à partir de données historiques mais plutôt à examiner un outil susceptible d'évaluer comment des phénomènes démographiques ou socio-économiques potentiels pourraient influencer les résultats électoraux. À ce stade, notre travail se trouve encore à la phase de validation conceptuelle au sein de l'entreprise BVA. Dans le domaine de la sociologie électorale, il est courant d'analyser les liens entre deux phénomènes en juxtaposant simplement et en commentant deux cartes représentatives. Afin d'approfondir notre analyse en utilisant des outils de statistique plus avancés, nous allons construire un modèle capable d'établir des relations plus complexes entre ces phénomènes. Bien que les modèles traditionnels de régression linéaire soient couramment utilisés dans de telles situations, il est important de noter qu'ils reposent sur des hypothèses trop restrictives et surtout incompatibles avec la nature de ces données socio-économiques et de ces données d'élection. Les méthodes plus modernes d'apprentissage machine sont excellentes pour prédire, mais peuvent s'avérer moins adaptées pour l'explication et la synthèse, qui sont deux autres objectifs essentiels de la statistique.

---

1 Communication présentée à l'Académie des Sciences, Inscriptions et Belles-Lettres de Toulouse à la séance du 8 juin 2023.

Description des données utilisées

Nous disposons des résultats électoraux des présidentielles 2022 fournis par le ministère de l'Intérieur à l'échelle des 69 682 bureaux de vote, incluant le nombre d'inscrits, de votes exprimés, de votes nuls, et les voix obtenues par chaque candidat. Parallèlement, les données socio-économiques suivantes sont accessibles sur le site de l'INSEE : la population, la répartition en classes d'âge, de sexe et de catégories socio-professionnelles. Elles sont disponibles au niveau des 49 285 ilots regroupés pour l'information statistique (ou IRIS) : il s'agit d'un découpage du territoire en mailles de taille homogène d'environ 2 000 habitants ayant des limites naturelles. Un travail préliminaire de géocodage, affectation de coordonnées géographiques à une adresse, a permis d'imputer des valeurs manquantes, c'est-à-dire d'attribuer des coordonnées géographiques à des informations sans adresse. Avec les douze candidats impliqués dans cette élection, nous avons constitué cinq groupes, chacun composé d'un ou plusieurs candidats. Les trois premiers candidats avec les parts de voix exprimées, les plus élevées, à savoir Macron (27,8%), Le Pen (23,1%) et Mélenchon (21,9%), forment les trois premiers groupes. Les autres candidats sont répartis en deux blocs selon leur orientation politique : un bloc de droite (Zemmour, Pécresse, Lasalle et Dupont-Aignan) et un bloc de gauche (Jadot, Roussel, Hidalgo, Poutou et Arthaud). Afin d'éviter les catégories vides, qui compliqueraient l'application de la méthode que nous allons utiliser, nous avons également regroupé les catégories socio-professionnelles de l'INSEE en quatre groupes. Le premier dénommé CSP (1-2) représente 5,9% du total et rassemble les agriculteurs exploitants, les artisans, les commerçants et les chefs d'entreprise. Le deuxième groupe, CSP (3-4), correspond à 32,7% du total et englobe les cadres, les professions intellectuelles supérieures et les professions intermédiaires. Le troisième groupe, dénommé CSP (5-6), représente 38,5% du total et inclut les ouvriers et les employés. Enfin le dernier groupe, CSP (7-8), réunit les retraités et les personnes sans activité professionnelle. Dans une version ultérieure de cette étude, il serait opportun de subdiviser ce dernier groupe en deux en raison des profils très contrastés de ces deux sous-groupes. L'intégration de ces deux sources d'informations a été opérée à l'échelle des 34 951 communes. Cependant cette démarche est complexe en raison de l'absence de données pour certaines communes des DOM-TOM et pour les Français résidant à l'étranger. Par conséquent, nous avons limité le périmètre de l'étude à la France métropolitaine, entraînant l'exclusion de 425 communes, soit 7% des inscrits ou encore 4% des votants. Les analyses sont ensuite menées au niveau communal ou départemental selon les cas. Par ailleurs, une opération supplémentaire d'imputation de pourcentages nuls par des petites valeurs s'avère nécessaire comme nous l'avons mentionné auparavant.

Une autre particularité de ces données est que les communes ne sont pas des unités d'observation comparables en raison de leurs tailles variées. En effet parmi les 35 000 communes (environ), les six plus grandes en termes d'inscrits (représentant 0,0001 % des communes) rassemblent 6% des inscrits. Pour éviter cet écueil, nous avons stratifié ces communes en quatre niveaux de densité de population, conformément à une classification de l'INSEE en communes denses, intermédiaires, peu denses et très peu denses, et élaboré un modèle distinct pour chaque strate.

Données de composition

Les données de composition (souvent abrégées par le sigle CoDa) se présentent sous forme de collections (que nous désignerons par le terme mathématique de vecteurs) de quantités positives formant des parts d'un tout (également appelé total). Par exemple, dans notre étude, comme nous disposons du nombre de voix pour chaque candidat (ou bloc de candidats) et du nombre total de voix exprimées, nous pouvons calculer les parts de voix correspondantes. Pourquoi remplacer la collection des cinq nombres de voix par les parts et le nombre total de voix ? C'est pour dissocier la taille de la commune de la répartition des voix selon les cinq groupes car c'est cette dernière que l'on souhaite expliquer par les caractéristiques socio-économiques de la commune. On note que chaque part est positive (ou nulle) et que leur somme est égale à un (ou 100% selon la représentation des parts). Ceci est le cas pour les vecteurs de parts qui sont des données de composition : on dit en termes mathématiques que le vecteur de parts appartient à un simplexe. Dans cette étude, outre la collection des parts de voix pour les divers candidats sur un territoire donné, nous disposons également de la collection des parts de population pour chaque tranche d'âge et chaque catégorie socio-professionnelle. Il convient de souligner que ce type de données est en réalité très fréquent dans d'autres domaines tels que la nutrition (parts de calories dues aux protéines, glucides et lipides), le marketing (parts de marché de divers magasins), l'usage des sols, la génomique, la chimiométrie et la géochimie. D'un point de vue mathématique, cet espace abstrait, dit simplexe, peut être muni d'une structure dite d'espace vectoriel, qui partage des similarités avec l'espace euclidien classique, mais qui est adaptée à la présence des contraintes de cet espace et utilise des opérations différentes. On attribue cette construction au mathématicien John Aitchison dont les travaux sont détaillés dans son ouvrage de référence (Aitchison, 1986).

Qu'est-ce qu'un modèle statistique ?

Un modèle statistique est bâti à partir de données récoltées sur des unités d'observation qui dans notre cas sont les communes ou les départements. Le modèle se construit en établissant une relation entre des variables (caractéristiques mesurées sur les unités d'observation) : d'une part les variables à expliquer, généralement notées Y, et d'autre part les variables explicatives, souvent désignées par X. Ici les variables à expliquer sont les vecteurs de parts de voix tandis que les variables explicatives sont les vecteurs de répartition de la population en catégories socio-professionnelles. La relation entre Y et X (c'est-à-dire le modèle) s'exprime par une équation mathématique comportant des paramètres. La forme de la relation (équation) définit le modèle ; par exemple si l'équation est linéaire on parle d'un modèle linéaire (notion que nous allons explorer plus en détails ci-dessous). Les paramètres quant à eux sont des valeurs initialement inconnues que l'on détermine à partir des données grâce à des méthodes dites d'ajustement telles que la célèbre méthode des moindres carrés dans le cas du modèle linéaire. L'objectif essentiel d'un tel modèle est de synthétiser l'information, puisque l'on passe d'un grand nombre de données à un petit jeu de paramètres, ce qui rend la relation entre variables plus accessible et compréhensible. Un modèle peut également servir à des fins prédictives, mais cela n'est pas notre objectif ici. Voyons maintenant le modèle le plus élémentaire, à savoir le modèle linéaire, dans le cas où X et Y sont de simples grandeurs réelles. Dans ce cas, l'équation $Y = a + bX + e$ exprime que la

variable Y est la somme d'une partie dite expliquée a + bX, où a et b sont des paramètres, et d'une partie inexpliquée e désignée sous le terme d'erreur, et supposée de grandeur insignifiante. Plus simplement, sans évoquer formellement une équation, on peut expliquer que le modèle est linéaire lorsque l'on peut supposer qu'un accroissement de la variable X entraîne un accroissement de Y proportionnel à celui de X (avec un facteur de proportionnalité égal au paramètre b).

Modèles de régression pour vecteurs de parts

Il est peu vraisemblable que les relations entre des vecteurs de parts soient linéaires, ne serait-ce que compte tenu de la contrainte de positivité inhérente à ces parts. Pour renforcer cette conviction, on peut réaliser des graphiques exploratoires. Les vecteurs de parts peuvent intervenir dans un modèle statistique de différentes manières : en tant que variable à expliquer, en tant que variables explicatives, ou même les deux simultanément : c'est ce dernier cas auquel nous sommes confrontés dans notre étude puisque le vecteur des parts de voix doit être expliqué, alors que celui des parts des catégories socio-professionnelles fait partie des variables explicatives de la répartition des voix. Un modèle statistique conçu pour ce type de données doit tenir compte de leur spécificité, c'est-à-dire leur positivité et leur contrainte de somme égale à 1. Ceci entraîne en particulier qu'il faut redéfinir les notions usuelles de moyenne, variance et corrélation pour de telles variables, conduisant à une analyse statistique adaptée et des modèles non linéaires dits modèles de régression CoDa. Le schéma général des méthodes pour données de composition, dont font partie les modèles CoDa, consiste en trois étapes résumées dans la figure 1. La première phase consiste en l'application d'une transformation aux vecteurs de parts visant à éliminer en quelque sorte les contraintes, envoyant ainsi les données dans un autre espace appelé espace des coordonnées. La deuxième étape repose sur l'application d'une méthode classique aux données ainsi transformées. La dernière étape revient alors à l'espace d'origine par la transformation inverse de celle employée à la première étape : en effet, le résultat de la deuxième étape est souvent difficile à interpréter et dépend de la transformation particulière utilisée à la première étape dont le choix est arbitraire. Il est crucial de s@'assurer à la fin que le résultat dans l'espace d'origine demeure indépendant du choix initial de cette transformation. Quelles transformations utiliser ? On peut montrer mathématiquement que la transformation doit être fonction d'un certain nombre de rapports de parts. En pratique, bien que ce ne soit pas une exigence absolue, on privilégie plutôt le logarithme de ces rapports de parts. En effet, cette approche présente d'abord l'avantage d'aboutir à des valeurs avec une gamme de variations symétrique. De plus le logarithme transforme les produits en sommes, ce qui peut s'avérer utile : en effet, il est souvent plus pertinent de modifier une part en lui appliquant un facteur multiplicatif plutôt qu'un accroissement additif.

Comment mesurer les impacts de variables explicatives dans un modèle de régression CoDa ?

Revenons à l'enjeu initial de notre problème. Nous avons pour objectif de mesurer l'impact de la variation d'une variable explicative sur la variable à expliquer dans un modèle de régression CoDa. Plus particulièrement ici, nous cherchons à comprendre et mesurer l'effet d'une fluctuation de la répartition des catégories socio-professionnelles

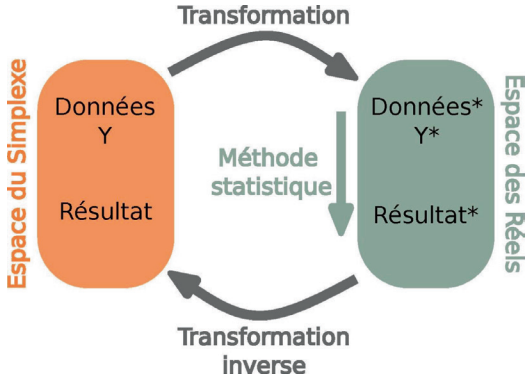

Fig. 1 : principe des modèles de régression CoDa.

sur la répartition des votes. Dans cette perspective, deux questions se posent naturellement : comment réaliser une telle variation et quel sens lui attribuer ? Des études antérieures (Morais et *al*., 2018), (Morais et Thomas-Agnan, 2021) et (Dargel et Thomas-Agnan, 2023) ont permis de répondre à ces interrogations. Voici maintenant l'illustration sur notre cas d'étude. Pour de petites variations, la notion à utiliser est celle d'élasticité qui mesure le changement relatif d'une part de la variable explicative sur une part de la variable dépendante. Joanna Morais, Michel Simioni et Christine Thomas-Agnan (2018) montrent que celles-ci s'expriment aisément à partir des paramètres du modèle. Cependant, il est important de noter que ces élasticités dépendent de l'unité d'observation considérée (en l'occurrence la commune). C'est pourquoi Lukas Dargel et Christine Thomas-Agnan (2023) plaident pour l'utilisation des différences d'élasticités qui n'en dépendent pas et constituent donc un meilleur résumé du phénomène dans un tel modèle. Pour des variations de plus grande envergure il est aisé d'observer les courbes de variations comme l'expliquent Lukas Dargel et Christine Thomas-Agnan (2023).

Nous allons générer des scénarios hypothétiques, susceptibles de se réaliser dans le futur, afin d'observer les prédictions du modèle dans de tels scénarios. Il ne s'agit donc pas, comme évoqué précédemment, de prédictions basées sur le passé. Pour illustrer cela, imaginons une situation où, pour une ville donnée, on souhaite observer l'effet d'un changement de la composition des catégories socio-professionnelles sans modification de ses autres caractéristiques. Mathématiquement, nous définissons des familles de changements hypothétiques compatibles avec la structure linéaire du simplexe et paramétrés par une direction et un nombre réel, ce dernier permettant de moduler leur intensité : on peut les représenter par des courbes dans le simplexe, qui sont en fait des droites au sens de la structure introduite par John Atchison. Cependant la contrainte de somme égale à 1 implique que la variation d'une composante du vecteur entraîne nécessairement des variations dans les autres composantes : concrètement ici, on ne peut pas modifier la part des CSP (1-2) sans modifier la part des autres CSP, ce qui fait qu'il n'est plus possible d'appliquer le paradigme statistique usuel « toutes choses égales par ailleurs ». Un scénario de ce type, pour une intensité et une direction choisies, est représenté dans la figure 2. Sur le graphique de gauche (dit diagramme en barres), la valeur initiale des CSP correspond à la hauteur des barres de couleur vert foncé, le pourcentage étant converti en une valeur entre 0 et 1. La couleur vert clair indique que la CSP (1-2) augmente dans ce scénario alors que la couleur rose indique une diminution des trois autres catégories afin de maintenir une somme égale à 1 : le fait que cette

diminution ne soit pas égale pour les autres catégories correspond au schéma de variation linéaire choisi, qui impose que les ratios entre les autres composantes restent constants lors de la variation. Le graphique de droite présente la variation de ces trois catégories en fonction de la CSP (1-2) et le segment vertical marque la valeur à l'origine pour notre ville hypothétique. On observe clairement que l'accroissement de la CSP (1-2) se fait surtout au détriment de la CSP (7-8). Un tel scénario n'est qu'un exemple et l'on peut façonner ces scénarios à loisir sous la contrainte de linéarité.

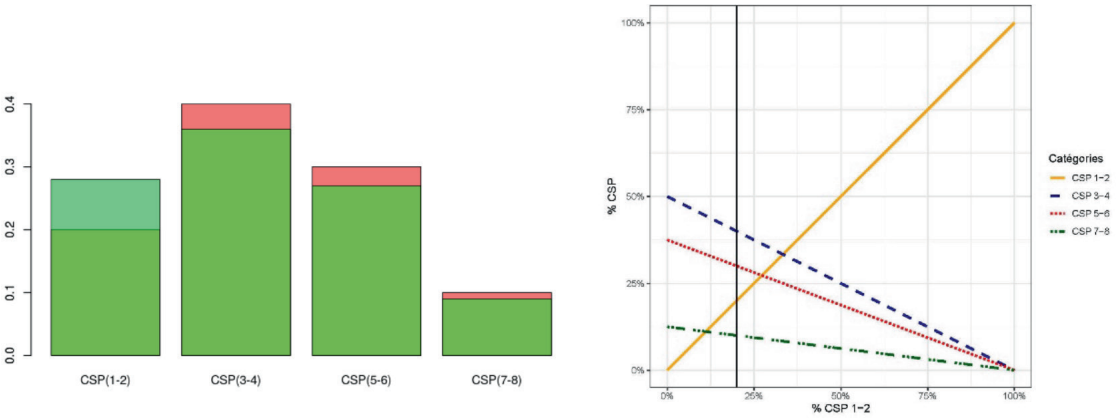

Fig. 2 : scénario de variation de la part de CSP (1-2)

Le nombre de paramètres nécessaires au calcul des élasticités ici est de 20 résultant de la multiplication de 4 catégories socio-professionnelles par 5 candidats ou blocs. Grâce aux élasticités, on peut déjà donner une estimation chiffrée de l'impact d'un scénario. Par exemple, si l'on envisage une augmentation de 5% de la part des cadres, professions intellectuelles supérieures et intermédiaires dans la ville de Toulouse, le tableau suivant présente les répercussions de cette variation en termes d'élasticité mais également (pour le lecteur profane) en termes de variations en pourcentage, en points de pourcentage et en nombre de voix pour chaque candidat ou bloc. Le signe des élasticités indique si la variation de 5% de cette CSP va résulter en une augmentation ou une diminution de la part de voix correspondant à la colonne considérée. La variation en pourcentage des parts de voix de chaque candidat dans la ligne 2, est obtenue en appliquant une formule de Taylor et la valeur de l'élasticité. Pour faciliter l'interprétation, elle est ensuite convertie en utilisant la valeur initiale de la part en variation exprimée en nombre de points de pourcentage. On note sur cette ligne que les variations positives et négatives se compensent automatiquement pour conduire à une somme nulle, ce qui maintient un total de 100% des composantes du nouveau vecteur de parts. Ce phénomène indique que les voix perdues par un bloc politique sont redistribuées vers un autre, et que les pertes d'un candidat sont réparties entre d'autres, bien qu'il soit évident que cet équilibre global ne renseigne par exactement sur le processus de reports. Finalement, en multipliant par le nombre d'inscrits de la commune, la dernière ligne donne la variation en nombre de voix. Un tel tableau est relatif à une commune donnée et à une variation choisie d'une part de CSP choisie ; il s'agit donc d'un zoom sur un point précis d'intérêt et nous allons nous tourner vers des résumés plus globaux.

La figure 2 constitue un exercice illustratif destiné à clarifier le principe des variations considérées. À présent, pour illustrer les variations des parts de voix correspondant à

	MACRON	LE PEN	MELENCHON	BLOC DROITE	BLOC GAUCHE
élasticité	0,019	- 0,036	0,030	-0,016	0,034
Variation en %	0,165	- 0, 298	0,254	-0,137	0,281
Variation en points de %	0,043	-0,085	0,042	-0,027	0,027
Variation en nombre de voix	84	-166	82	- 52	52

Tableau 1 : effet d'une augmentation de 5% de la part de CSP (3-4) à Toulouse.

toute une gamme de valeurs de l'intensité des variations de CSP, on obtient des courbes de variation telles que celles de la figure 3. Sur celle-ci la commune d'origine est la ville de Paris (segment vertical) et la catégorie qui subit la variation est le pourcentage de cadres (CSP (3-4)). La courbe révèle notamment que l'augmentation de la proportion de cadres et professions intellectuelles supérieures et intermédiaires à Paris serait bénéfique pour le candidat Macron, mais préjudiciable au candidat Le Pen. Toutefois elle n'aurait qu'un effet modéré sur les autres candidats ou blocs. La non-linéarité du phénomène se manifeste principalement aux abords des valeurs extrêmes de la part en question, c'est-à-dire quand celle-ci approche zéro ou un. Un exemple intéressant d'exploitation de ce type de graphique pourrait être le suivant : si, au lieu de se concentrer sur les voix exprimées comme on l'a fait dans ce projet préliminaire, on avait également inclus la part des abstentions comme une composante additionnelle, il aurait été envisageable de fournir des recommandations aux candidats. Ces recommandations auraient eu pour objectif d'orienter leurs discours de campagne vers les CSP dont la variation engendrerait le plus de bénéfice pour leur part de voix dans un lieu donné, dans l'espoir de persuader ces abstentionnistes-là à voter en leur faveur.

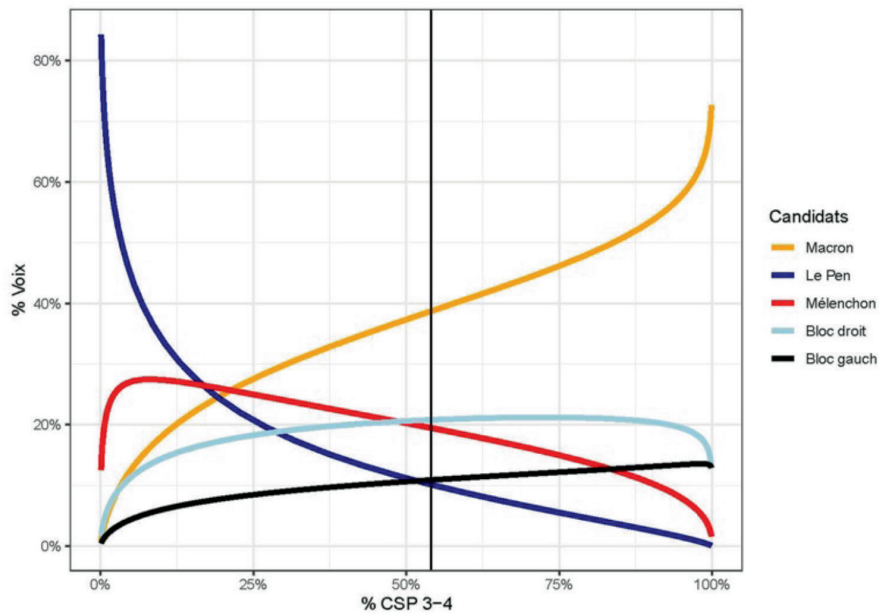

Fig. 3 : impact de la variation de la CSP (3-4) à Paris sur les parts de voix.

De nombreuses autres exploitations de ces outils de calcul sont envisageables. Pour conclure, prenons un dernier exemple mettant en œuvre la cartographie pour représenter le gain relatif du candidat Macron suite à une augmentation de 5% de chaque CSP individuellement au niveau des départements. La figure 4 illustre de manière éloquente qu'un changement de la composition des CSP influence particulièrement la part de voix en faveur de Macron, surtout lorsque c'est la part de cadres (et professions intellectuelles supérieures et intermédiaires) qui varie alors qu'une variation des deux autres CSP a un impact négligeable.

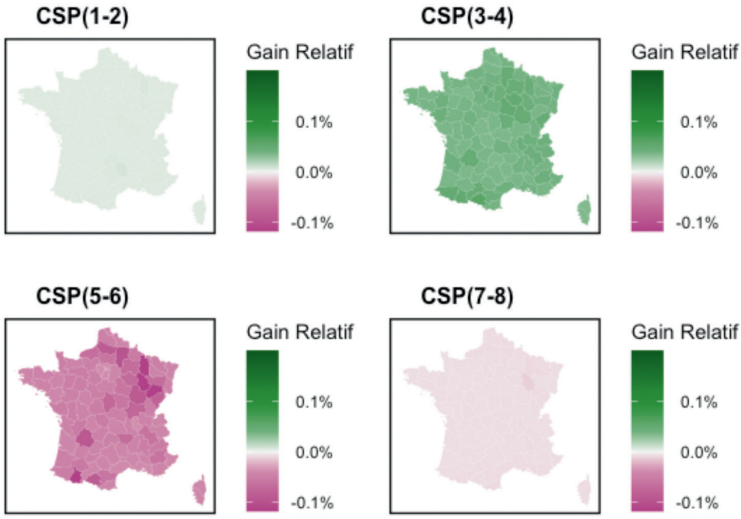

Fig. 4 : effet sur le gain relatif de Macron d'une augmentation de 5% de chaque CSP par département.

Conclusion

À travers cet exemple, nous avons démontré la capacité des modèles de régression CoDa, spécialement conçus pour traiter les vecteurs de parts, à analyser de manière appropriée les données de composition, en tenant compte de leurs caractéristiques spécifiques. Cette illustration a ainsi confirmé l'utilité de telles techniques, notamment dans le contexte des questions de sociologie politique. Nous avons démontré comment le modèle de régression CoDa est un outil précieux pour expliquer le lien entre les parts de voix et la répartition des catégories socio-professionnelles en mesurant l'impact d'une variation cohérente de cette dernière. Nous avons également expliqué comment ce modèle est capable de résumer ce lien à travers quelques paramètres permettant de calculer les indicateurs que sont les élasticités. Il est important de souligner que l'étude exposée ici, comme nous l'avons mentionné, constitue une première approche qui est perfectible à divers égards et qui sera poursuivie à l'avenir dans cette optique. Plusieurs pistes d'amélioration sont envisageables telles qu'une restructuration des catégories socio-professionnelles et la prise en compte de la catégorie abstention pour les parts de voix. Nous espérons aussi au travers de cet exposé avoir apporté un éclairage sur la possibilité d'interaction entre les techniques statistiques modernes et les enjeux sociétaux.

Bibliographie

Aitchison, John, *The statistical analysis of compositional data*, Londres, Chapman and Hall, 1986.

Morais, Joanna et Thomas-Agnan, Christine, « Impact of covariates in compositional models and simplicial derivatives », *Austrian Journal of Statistics* 50(2), pages1-5, 2021.

Morais, Joanna, Simioni, Michel et Thomas-Agnan, Christine, « Interpretation of eplanatory variables impacts in compositional regression models », *Austrian Journal of Statistics* 47, pages 1-25, 2018.

Dargel, Lukas, et Thomas-Agnan, Christine, « Share-ratio interpretations of compositional regression models », *Working paper* N° 23-1456, Toulouse School of Economics (TSE), 2023.