Rapport Statistiques Bayésiennes
Is Tax Money used wisely according to Americans ?

Luc YAO, Léos Coutrot, Adèle Berger, Richard Cheam, Loic Xu

30 mars 2025

Table des matières

1 Introduction

Ce projet [1] vise à analyser la perception du gaspillage des fonds publics en explorant les réponses à deux variables clés :

— `wastecent` : estimation du pourcentage des taxes perçu comme étant gaspillé par le gouvernement.

— `wastethink` : associations spécifiques faites par les répondants lorsqu'ils pensent au gaspillage gouvernemental.

Nous chercherons à comprendre les déterminants de cette perception : est-elle influencée par l'affiliation politique, le revenu, le fait de payer ou non des impôts, ou encore par d'autres facteurs ?

L'analyse repose sur un jeu de données [2] issu d'une enquête menée auprès de 1000 adultes américains entre le 5 et le 19 novembre 2014 via un panel opt-in de Qualtrics.

2 Présentation du jeu de données

Les données utilisées dans cette étude combinent des enquêtes quantitatives et des entretiens qualitatifs afin d'explorer les perceptions des Américains sur les impôts et les dépenses publiques. Une enquête menée en novembre 2014 auprès de 1 000 adultes américains a inclus des questions ouvertes, permettant aux répondants d'exprimer librement leurs opinions sur le sujet. En complément, 49 entretiens approfondis ont été réalisés avec des participants diversifiés à travers les États-Unis, offrant une perspective détaillée sur les perceptions individuelles [2].

Ces données comprennent des informations détaillées sur les répondants, incluant des caractéristiques sociodémographiques (revenu, niveau d'éducation, statut d'emploi), des attitudes politiques (idéologie, participation électorale, connaissance politique) et des comportements liés à l'impôt (statut de contribuable, perception du système fiscal).

Les résultats de cette enquête révèlent que les Américains considèrent généralement les impôts comme un devoir civique et une obligation morale, mais montrent également des malentendus courants sur la répartition de la charge fiscale et les bénéficiaires des dépenses publiques. En particulier, bien que les Américains soutiennent généralement les dépenses locales, ils se montrent souvent sceptiques vis-à-vis des dépenses perçues comme bénéficiant à des groupes extérieurs, tels que les immigrants et les bénéficiaires de l'aide sociale.

Dans cette étude, nous examinerons spécifiquement les liens entre `wastecent` et `wastethink` avec ces différentes variables explicatives afin d'identifier les principaux facteurs influençant la perception de l'efficacité de la dépense publique.

La description complète de l'ensemble des variables utilisées dans cette étude est disponible en **Annexe 9**.

3 Pré-processing et analyse exploratoire des données

3.1 Pré-processing

Dans le cadre de notre étude, nous disposons de 41 variables, sans compter la variable `wastecent`. Parmi ces variables, certaines sont de type texte et nous avons décidé de les exclure de nos modèles d'analyse.

Les variables supprimées sont :

— `statename`
— `firstthought`
— `recent`
— `tpfeel`
— `biggest`
— `paystub`
— `depend`
— `glad`
— `upset`
— `benefit`
— `wastethink`
— `eitcexp`
— `eitcthink`

Ces variables ont été écartées car, à première vue, elles ne sont pas directement exploitables dans les modèles que nous souhaitons développer. Bien qu'il soit possible d'utiliser des techniques de traitement du langage naturel (NLP) pour en extraire des informations exploitables, cela ne correspond pas aux objectifs principaux de notre projet. Nous avons donc privilégié les variables numériques et catégorielles pouvant être directement intégrées dans nos modèles.

Après avoir supprimé les variables textuelles, on a 192 lignes du dataset qui ont des valeurs manquantes dont 4 sur `wastecent`. On a supprimé ces 4 lignes.

3.2 Analyse des données

Pour l'analyse des données, on a commencé par distinguer les variables catégorielles et les variables continues.

Les variables catégorielles sont les suivantes :

— `gender`
— `educ`
— `partyid`
— `taxpayer`

— wagesal
— eitcself
— eitcother
— labforce
— polideo
— polinffreq
— regvote
— voted
— discusspol
— poleffic
— polvol
— polknow1
— polknow2
— polknow3
— marital
— ownhome
— raceeth
— hhinc
— child

Les variables continues sont les suivantes :

— feelfedgov_1
— percenttp
— yearbirth

Afin d'explorer l'influence des variables catégorielles sur wastecents, nous avons tracé des boxplots. Ceux-ci permettent d'analyser la répartition de wastecents pour chaque groupe des variables catégorielles. On va présenter dans la suite les plus significatifs.

3.2.1 Variable cible : wastecent

La variable wastecents reflète la perception qu'ont les individus du gaspillage des fonds publics, en indiquant le nombre de centimes qui, selon eux, sont gaspillés pour chaque dollar d'impôt versé par le gouvernement. Avec une moyenne de 55,03 cents et une médiane de 59 cents, les personnes interrogées ont tendance à penser qu'une proportion substantielle des impôts est mal utilisée. L'écart-type de 26,58 suggère une diversité d'opinions, certaines estimations allant de 0 à 100 cents. Les histogrammes révèlent des pics occasionnels et indiquent également des perceptions communes, notamment autour de 60 cents. Ces données montrent une méfiance générale à l'égard de l'efficacité des dépenses publiques, avec des variations notables selon les individus. Ces perceptions peuvent influencer l'attitude des citoyens à l'égard de la fiscalité et de la gouvernance, soulignant l'importance pour les gouvernements de communiquer sur l'utilisation des fonds publics afin de rétablir la confiance

Statistic	wastecents
Count	996
Mean	55.03
Std	26.58
Min	0.00
25%	36.00
50% (Median)	59.00
75%	80.00
Max	100.00

TABLE 1 – Summary statistics for wastecents

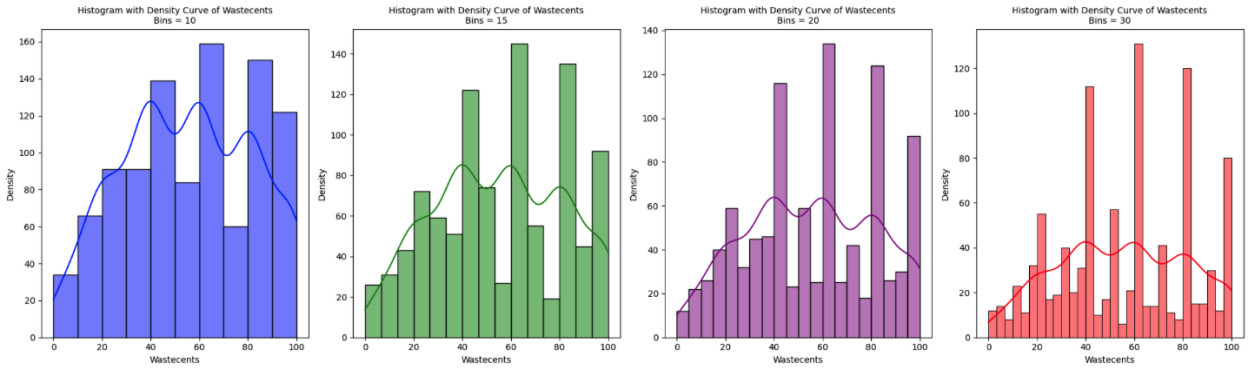

FIGURE 1 – Histogramme et densité de la variable `wastecent`

3.2.2 Gender

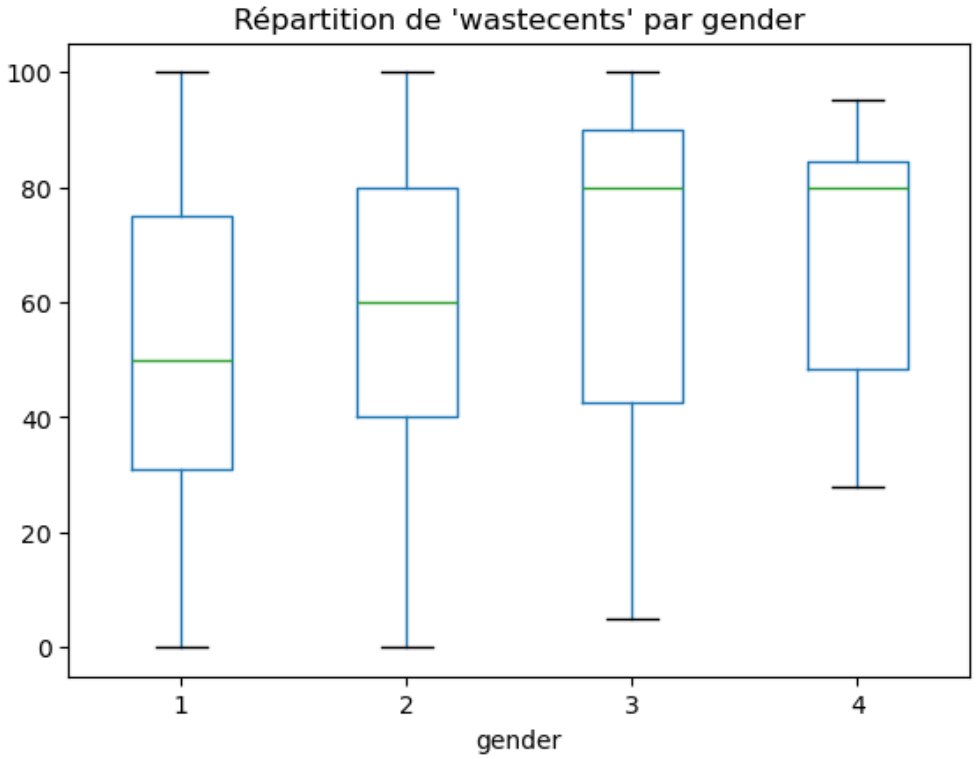

FIGURE 2 – Boxplot wastecents par rapport à gender

gender	count	mean	std	min	25%	50%	75%	max
1	489.0	52.451943	26.440279	0.0	31.0	50.0	75.0	100.0
2	497.0	57.364185	26.412557	0.0	40.0	60.0	80.0	100.0
3	3.0	61.666667	50.083264	5.0	42.5	80.0	90.0	100.0
4	7.0	67.000000	25.625508	28.0	48.5	80.0	84.5	95.0

TABLE 2 – Statistiques descriptives de `wastecents` par rapport à `gender`

On remarque que les hommes (gender=1) pensent en moyenne qu'il y a moins de gâchis d'impots que les femmes (gender=2). On ne peut rien dire sur le reste car pas assez d'effectif.

3.2.3 Education

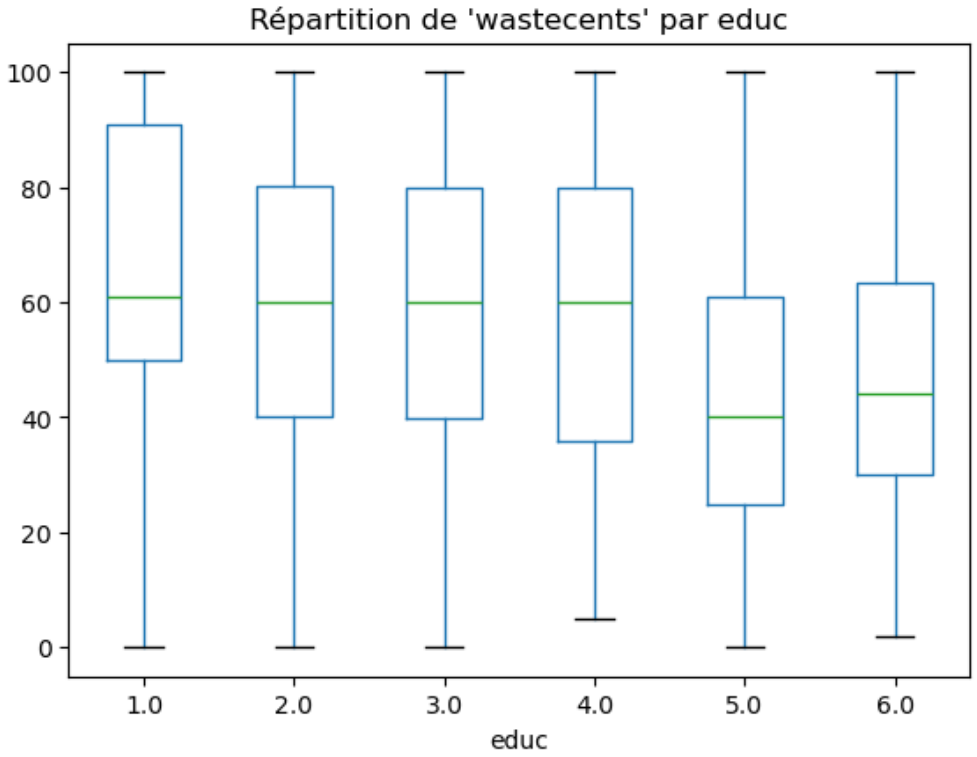

FIGURE 3 – Boxplot wastecents par rapport à `educ`

educ	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
1.0	23.0	65.521739	28.108345	0.0	50.00	61.0	91.00	100.0
2.0	388.0	59.863402	26.013501	0.0	40.00	60.0	80.25	100.0
3.0	192.0	56.552083	26.492267	0.0	39.75	60.0	80.00	100.0
4.0	78.0	57.641026	27.856765	5.0	36.00	60.0	80.00	100.0
5.0	207.0	45.613527	24.653431	0.0	25.00	40.0	61.00	100.0
6.0	102.0	48.960784	25.562551	2.0	30.00	44.0	63.25	100.0

TABLE 3 – Statistiques descriptives de `wastecents` par rapport à `educ`

Les personnes ayant une licence et au-delà (educ = 5-6) sont plus enclines à penser que l'on gaspille moins, contrairement à celles qui n'ont pas de diplôme de l'enseignement supérieur (educ = 1-4). Toutefois, il convient de prendre les résultats pour educ = 1 avec précaution, car seulement 23 réponses ont été recueillies.

3.2.4 Revenues

La variable hhinc représente le revenu total du ménage avant impôts pour l'année 2013, catégorisé en différents groupes de revenu. Par exemple, une valeur de 10,24 dans le tableau statistique correspond à une catégorie de revenu située entre 35 000 et 39 999 dollars par an. La médiane de 11,00 indique que la moitié des ménages se situent dans des catégories de revenu inférieures ou égales à 40 000 à 49 999 dollars.

Statistic	hhinc
Count	944
Mean	10.24
Std	5.15
Min	1.00
25%	7.00
50% (Median)	11.00
75%	14.00
Max	19.00

TABLE 4 – Summary statistics for hhinc

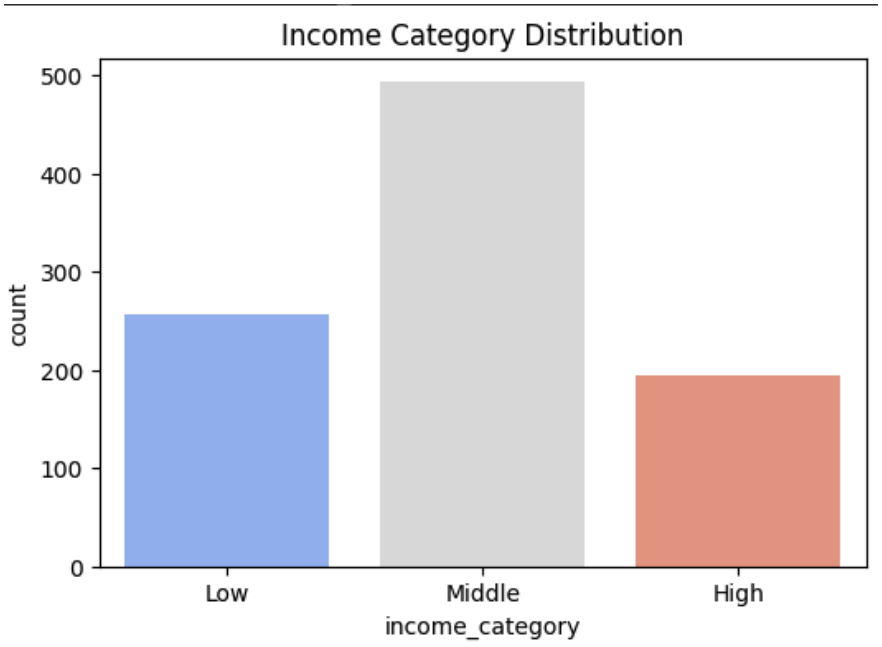

FIGURE 4 – Histogramme de la variable Income

Le diagramme en boîte montre des différences notables entre les catégories de revenu en ce qui concerne la perception du gaspillage des fonds publics. Les ménages à revenu faible et moyen estiment en moyenne que 60 cents de chaque dollar d'impôt sont gaspillés par le

gouvernement, avec une variabilité significative, reflétant des opinions diverses au sein de ces groupes. En revanche, les ménages à revenu élevé perçoivent un gaspillage médian légèrement inférieur, autour de 50 cents, avec moins de variation dans leurs réponses.

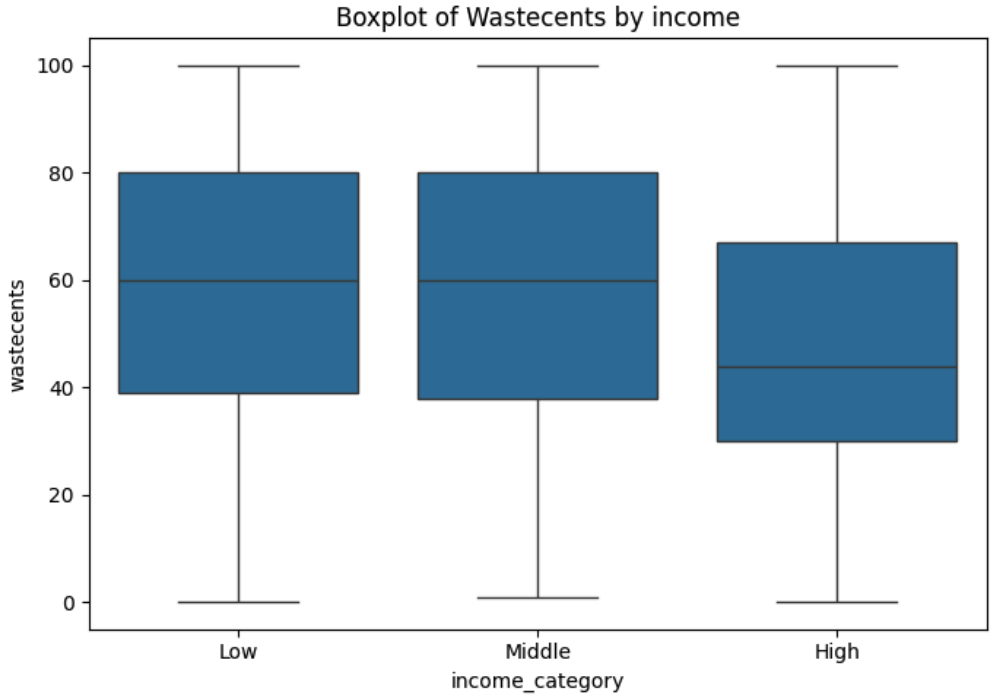

FIGURE 5 – Boxplot wastecents par rapport à `hhinc`

Ces résultats suggèrent que les perceptions du gaspillage des fonds publics peuvent varier en fonction du niveau de revenu, les ménages plus aisés se montrant peut-être plus optimistes ou ayant une confiance différente dans l'efficacité du gouvernement. Ces informations sont précieuses pour comprendre les attitudes envers la fiscalité et la dépense publique.

3.2.5 Affiliation politique (partyid et polideo)

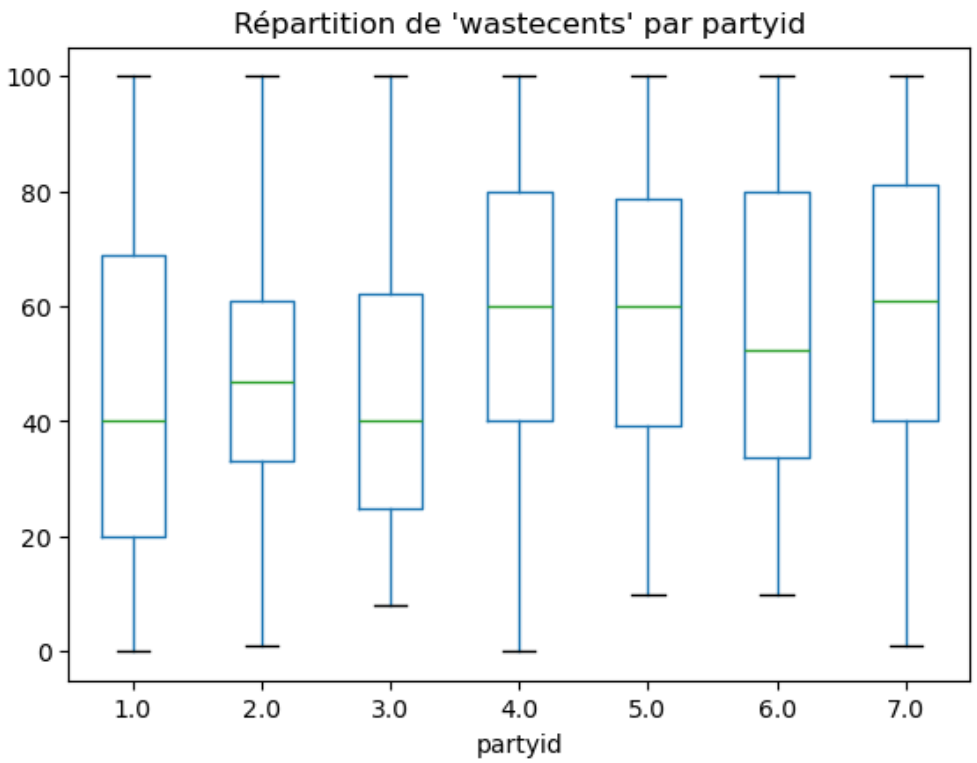

FIGURE 6 – Boxplot wastecents par rapport à `partyid`

partyid	count	mean	std	min	25%	50%	75%	max
1.0	165.0	46.206061	28.857407	0.0	20.00	40.0	69.0	100.0
2.0	127.0	48.748031	23.259055	1.0	33.00	47.0	61.0	100.0
3.0	57.0	47.842105	24.352170	8.0	25.00	40.0	62.0	100.0
4.0	329.0	59.465046	25.935653	0.0	40.00	60.0	80.0	100.0
5.0	66.0	57.787879	24.947030	10.0	39.25	60.0	78.5	100.0
6.0	84.0	53.940476	24.813206	10.0	33.75	52.5	80.0	100.0
7.0	130.0	60.123077	25.802474	1.0	40.00	61.0	81.0	100.0

TABLE 5 – Statistiques descriptives de `wastecents` par rapport à `partyid`

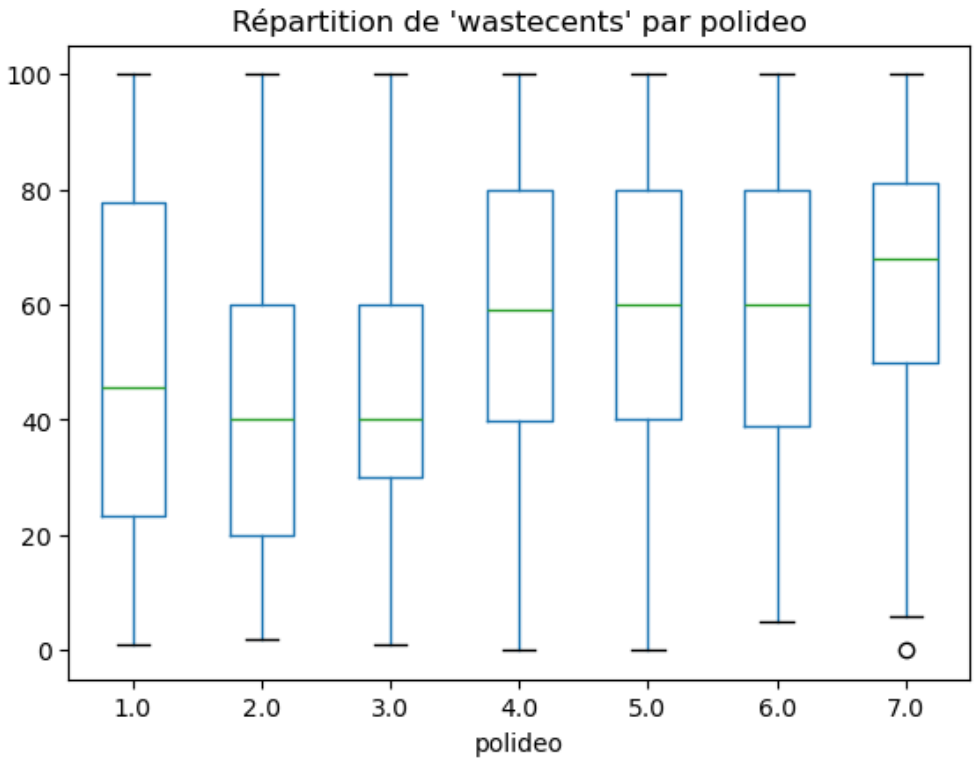

FIGURE 7 – Boxplot wastecents par rapport à `polideo`

polideo	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
1.0	54.0	50.129630	31.431295	1.0	23.25	45.5	77.75	100.0
2.0	123.0	43.170732	25.264407	2.0	20.00	40.0	60.00	100.0
3.0	68.0	44.661765	21.201669	1.0	30.00	40.0	60.00	100.0
4.0	316.0	55.579114	25.479841	0.0	39.75	59.0	80.00	100.0
5.0	90.0	58.277778	25.766465	0.0	40.00	60.0	80.00	100.0
6.0	188.0	58.191489	25.290304	5.0	39.00	60.0	80.00	100.0
7.0	85.0	64.929412	25.323990	0.0	50.00	68.0	81.00	100.0

TABLE 6 – Statistiques descriptives de `wastecents` par rapport à `polideo`

On remarque les démocrates (partyid = 1 à 3) estiment qu'il y a moins de gaspillage que les républicains (partyid = 5 à 7). Cette tendance se reflète également dans les positions politiques, où les individus libéraux (polideo = 1 à 3) rapportent des résultats similaires à ceux des démocrates, tandis que les personnes plus conservatrices (polideo = 4 à 7) affichent des résultats proches de ceux des républicains.

3.2.6 Propriétaire ou locataire (ownhome)

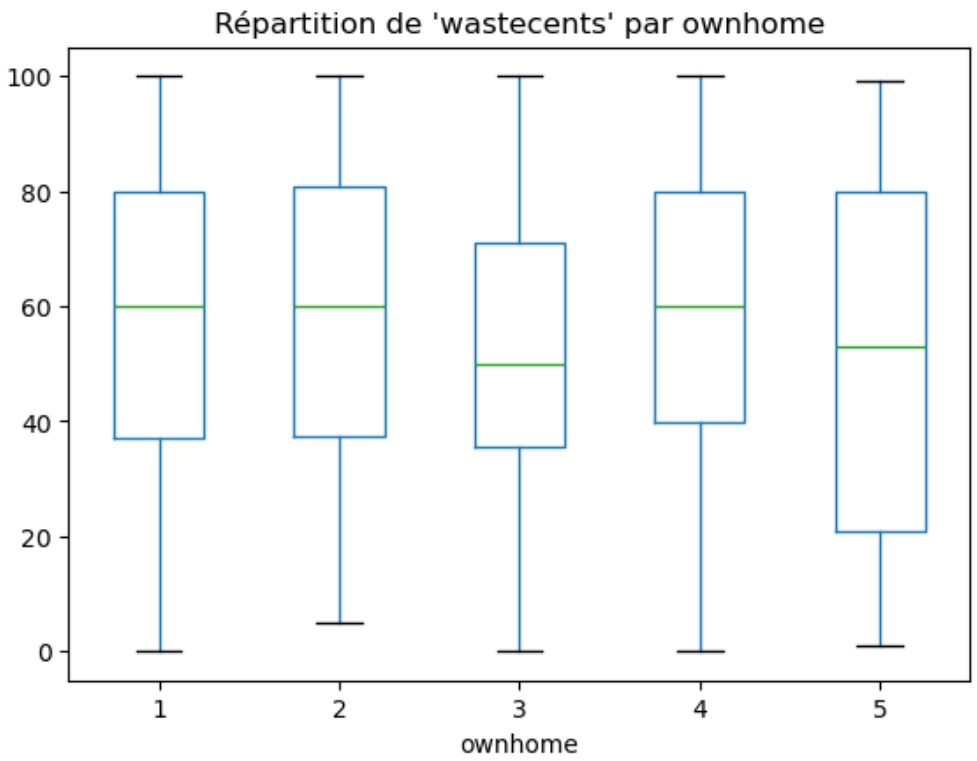

FIGURE 8 – Boxplot wastecents par rapport à `ownhome`

ownhome	count	mean	std	min	25%	50%	75%	max
1	622.0	55.961415	26.143573	0.0	37.00	60.0	80.00	100.0
2	22.0	56.909091	30.358037	5.0	37.25	60.0	80.75	100.0
3	283.0	52.427562	26.618614	0.0	35.50	50.0	71.00	100.0
4	52.0	58.288462	27.880935	0.0	39.75	60.0	80.00	100.0
5	17.0	52.058824	32.340900	1.0	21.00	53.0	80.00	99.0

TABLE 7 – Statistiques descriptives de `wastecents` par rapport à `ownhome`

Les propriétaires de logement (ownhome = 1) estiment qu'il y a davantage de gaspillage que les locataires (ownhome = 3). Toutefois, il est difficile de tirer des conclusions pour les autres catégories en raison des effectifs trop faibles.

3.2.7 Ressenti envers le gouvernement

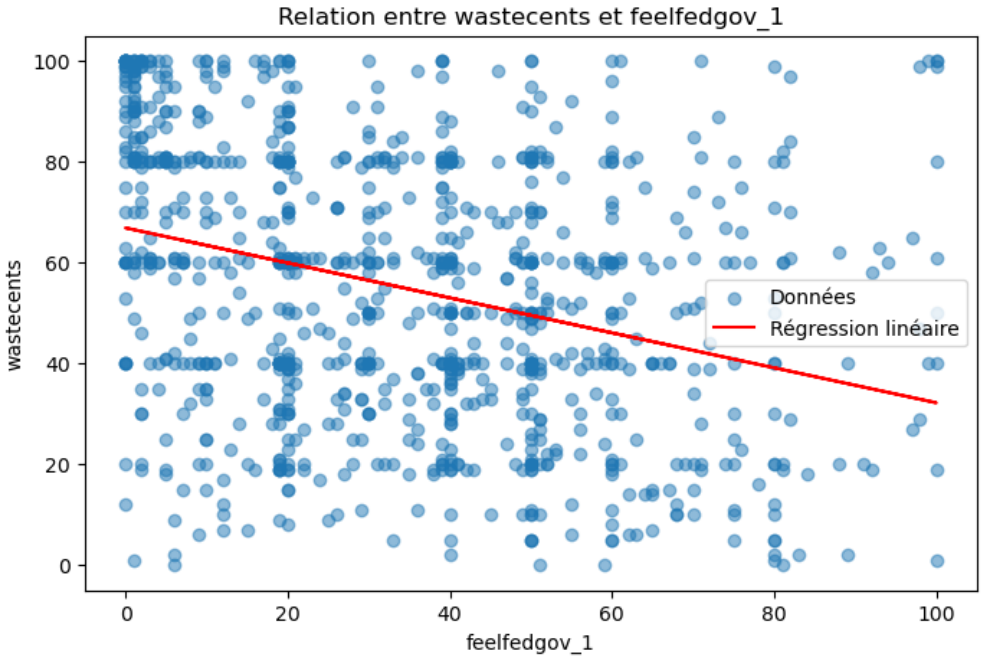

FIGURE 9 – Wastecents en fonction du ressenti envers le gouvernement

On remarque que plus on apprécie le gouvernement moins on pense qu'il y a un gachis d'impôt. C'est assez logique.

3.2.8 Wastethink

Pour étudier la variable `WasteThink`, il est nécessaire de la comparer en fonction de la variable `WasteCents`. Selon plusieurs sources, la valeur de `WasteCents` en 2014 était estimée à 51% [3]. Nous prenons donc cette valeur comme étant la "valeur réelle".

Ainsi, les réponses situées dans un intervalle de 45% à 55% sont considérées comme des réponses proches de la vérité, que nous qualifions de "bonne réponse". En revanche, les réponses dont les valeurs sont inférieures à 45% sont classées comme des "sous-estimations", tandis que celles supérieures à 55% sont considérées comme des "sur-estimations".

Cette classification permet d'analyser les perceptions des répondants concernant le gaspillage gouvernemental, en les comparant à la valeur réelle estimée de 51%. Cette approche permet de mieux comprendre comment les individus perçoivent les dépenses publiques et de catégoriser leurs réponses en fonction de leur proximité avec la valeur réelle de `WasteCents`.

La Figure 10 représente la distribution des réponses de `WasteCents` par rapport à la valeur de 51%. On observe que la majorité des répondants tendent à sur-estimer la valeur réelle du

gaspillage gouvernemental, ce qui suggère une perception exagérée des dépenses publiques. En effet, la courbe montre une concentration plus importante de réponses supérieures à 51%, ce qui indique une tendance générale à percevoir un gaspillage plus élevé que la réalité.

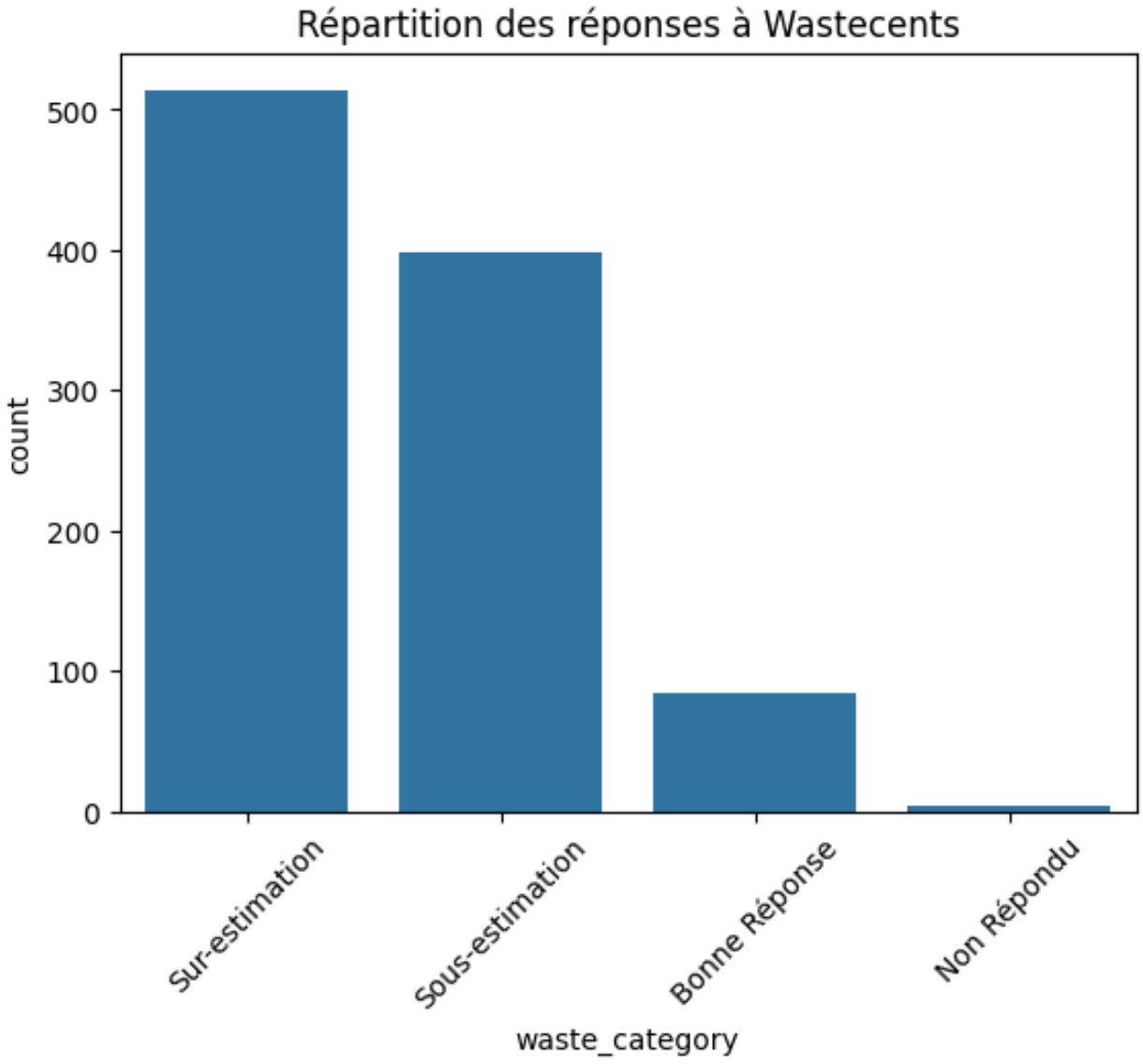

FIGURE 10 – Répartition des réponses à Wastecent

Les Figures 11, 12, et 13 montrent les nuages de mots pour chaque catégorie de réponses : "Bonne Réponse", "Sous-estimation", et "Sur-estimation". Ces nuages de mots illustrent les termes les plus fréquemment associés à chaque type de perception du gaspillage gouvernemental.

— **Bonne Réponse :** Dans le nuage de mots pour la "bonne réponse" (Figure 11), des termes tels que *education*, *space*, *payment*, *federal*, *work*, *studies*, *research*, et *use*

apparaissent fréquemment. Ces mots suggèrent que les répondants qui estiment que le gaspillage est proche de la réalité tendent à se concentrer sur des thèmes liés à l'éducation, la recherche, et les programmes fédéraux.

— **Sous-estimation :** Dans le nuage de mots pour les "sous-estimations" (Figure 12), des termes comme *projects*, *congress*, *nothing*, *fraud*, *research*, *war*, *defense*, et *salaries* ressortent souvent. Cela indique que les personnes qui sous-estiment le gaspillage gouvernemental mettent davantage l'accent sur des projets gouvernementaux, la fraude, ainsi que sur des préoccupations militaires et de défense.

— **Sur-estimation :** Enfin, le nuage de mots pour les "sur-estimations" (Figure 13) met en avant des mots tels que *benefit*, *politician*, *don't*, *vacation*, *help*, *countries*, et *dollar*. Cela suggère que ceux qui estiment que le gaspillage est plus élevé ont tendance à se concentrer sur des termes liés aux bénéfices, aux politiciens, ainsi qu'à l'aide étrangère et aux dépenses en vacances.

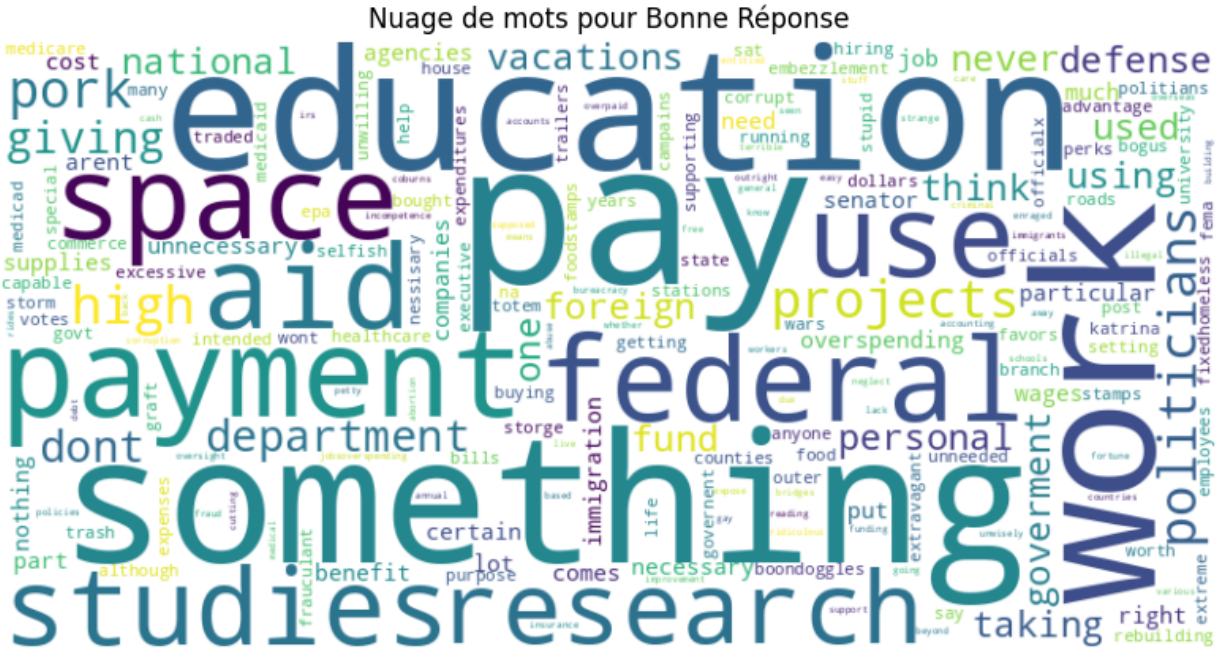

FIGURE 11 – Nuage de mot pour Bonne Réponse

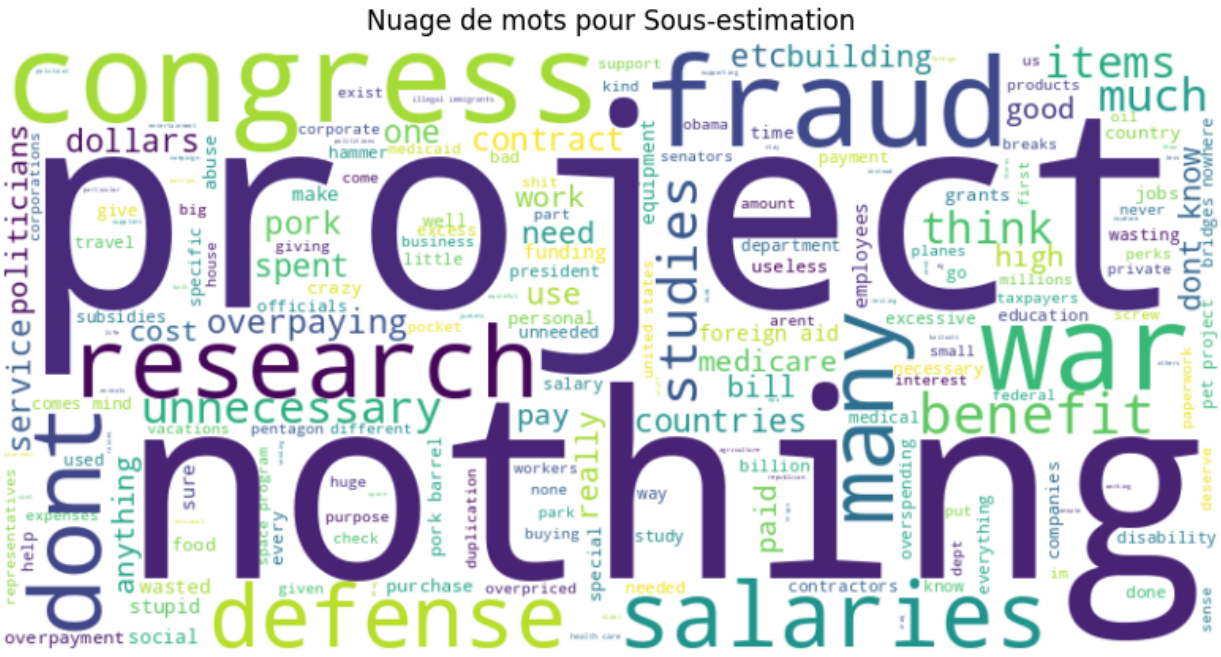

FIGURE 12 – Nuage de mot pour Sous Estimation

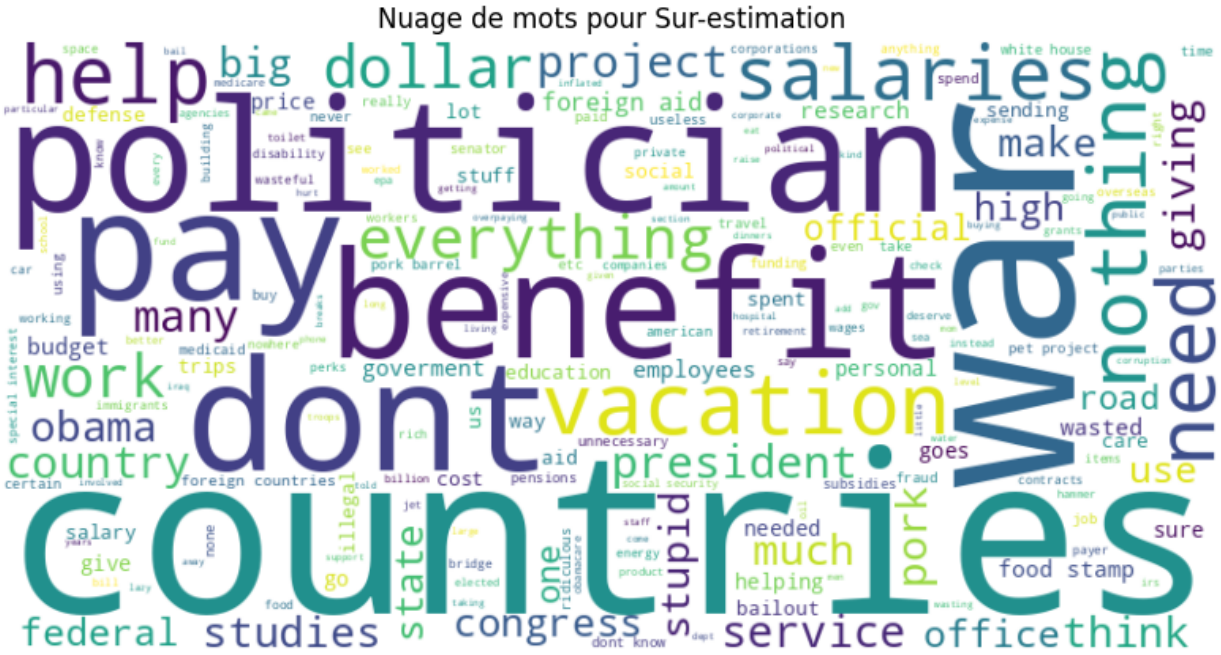

FIGURE 13 – Nuage de mot pour Sur Estimation

Les résultats suggèrent une tendance générale à sur-estimer le gaspillage gouvernemental, avec des perceptions qui varient en fonction des thèmes évoqués. Les répondants qui estiment

que le gaspillage est proche de la réalité mettent l'accent sur des thèmes tels que l'éducation et la recherche, tandis que ceux qui sous-estiment ou sur-estiment les dépenses se concentrent davantage sur des préoccupations politiques et militaires.

4 Description des Modèles

Dans ce projet, plusieurs modèles bayésiens ont été utilisés pour analyser les relations entre les variables explicatives et la variable `wastecents`. Les modèles varient par le type de distribution et de lien utilisé pour modéliser la variable cible, mais partagent des éléments communs dans leur construction, tels que l'imputation des valeurs manquantes, l'utilisation de priors vagues, et l'échantillonnage via la méthode MCMC.

4.1 Modèle de Régression Linéaire Bayésienne

Le premier modèle utilisé est une régression linéaire bayésienne [4], qui suppose une relation linéaire entre les prédicteurs et la variable cible `wastecents`. Ce modèle est basé sur une approche probabiliste où les coefficients sont échantillonnés à partir de distributions normales, et les observations sont modélisées à l'aide d'une distribution normale.

Dans ce modèle, nous avons utilisé les variables explicatives disponibles dans le jeu de données, à l'exception de la variable cible `wastecents`. Les étapes de construction du modèle sont les suivantes :

- **Imputation des valeurs manquantes :** Avant de commencer l'analyse, les valeurs manquantes dans les colonnes numériques ont été imputées par la médiane de chaque colonne.
- **Définition des priors :** Un prior vague a été assigné aux coefficients du modèle. Les coefficients de régression ainsi que l'intercept sont tous modélisés à partir de distributions normales avec une moyenne de 0 et un écart type de 10.
- **Modèle linéaire :** La relation entre les variables explicatives et la variable cible `wastecents` est supposée linéaire. Cette relation est exprimée comme :

$$\mu = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

  où β_0 est l'intercept, β_i les coefficients des prédicteurs x_i.
- **Modèle de vraisemblance :** La distribution de `wastecents` est modélisée par une distribution normale avec une moyenne μ et une variance σ^2 inconnue.
- **Échantillonnage :** L'échantillonnage a été réalisé à l'aide de la méthode MCMC (Markov Chain Monte Carlo), avec un échantillonnage de 4000 itérations et une phase de burn-in de 1000 itérations.

4.2 Modèle de Régression Poisson Bayésienne

Le second modèle que nous avons utilisé est une régression Poisson bayésienne, qui est adaptée aux variables de comptage [4][5]. Ce modèle suppose que les observations suivent une distribution de Poisson, et la relation entre les prédicteurs et la variable cible `wastecents` est log-linéaire.

Les étapes pour construire le modèle Poisson sont similaires à celles du modèle linéaire, à l'exception de la distribution de la variable cible et du lien entre les prédicteurs et la variable cible :

- **Imputation des valeurs manquantes :** Les valeurs manquantes dans les colonnes numériques sont également imputées par la médiane de chaque colonne.
- **Définition des priors :** Comme pour le modèle linéaire, des priors vagues sont assignés aux coefficients et à l'intercept, en utilisant une distribution normale avec une moyenne de 0 et un écart type de 10.
- **Lien log-linéaire :** La relation entre les variables explicatives et `wastecents` est maintenant modélisée de manière log-linéaire, c'est-à-dire :

$$\lambda = \exp(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)$$

  où λ est le taux de la distribution de Poisson, et β_0, β_i sont les paramètres à estimer.
- **Modèle de vraisemblance :** La distribution de `wastecents` suit une distribution de Poisson avec un taux λ, qui dépend des prédicteurs.
- **Échantillonnage :** L'échantillonnage est effectué avec MCMC de manière similaire au modèle linéaire.

4.3 Modèle de Régression Bayésienne avec Prior Laplace

Le modèle utilisé est une régression linéaire bayésienne avec un prior Laplace, qui est la version bayésienne de la régression LASSO. Ce modèle suppose une relation linéaire entre les prédicteurs et la variable cible, `wastecents`, tout en incorporant un prior de type Laplace sur les coefficients de régression, ce qui permet d'obtenir des solutions plus parcimonieuses. L'approche bayésienne permet d'échantillonner les coefficients à partir d'une distribution postérieure, et d'introduire des incertitudes dans les estimations des coefficients. Dans ce modèle, les variables explicatives disponibles dans le jeu de données sont utilisées, à l'exception de la variable cible `wastecents`. Les étapes de construction du modèle sont les suivantes :

- **Imputation des valeurs manquantes :** Avant de commencer l'analyse, les valeurs manquantes dans les colonnes numériques ont été imputées par la médiane de chaque colonne.
- **Définition des priors :** Un prior Laplace a été assigné aux coefficients du modèle. Ce prior est une distribution de type Laplace, qui favorise des coefficients proches de zéro, et permet ainsi une régularisation du modèle. L'intercept est modélisé à partir

d'une distribution normale avec une moyenne de 0 et un écart-type de 1, tandis que les coefficients des prédicteurs sont soumis à un prior Laplace avec un paramètre de régularisation λ.

— **Modèle linéaire :** La relation entre les variables explicatives et la variable cible `wastecents` est supposée linéaire. Cette relation est exprimée comme :

$$\mu = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

où β_0 est l'intercept, β_i les coefficients des prédicteurs x_i.

— **Modèle de vraisemblance :** La distribution de `wastecents` est modélisée par une distribution normale avec une moyenne μ et une variance σ^2 inconnue.

— **Échantillonnage :** L'échantillonnage des coefficients a été réalisé à l'aide de la méthode MCMC (Markov Chain Monte Carlo), avec un échantillonnage de 4000 itérations et une phase de burn-in de 1000 itérations.

5 Évaluation des Performances des Modèles

L'évaluation des performances des modèles bayésiens de régression, comme ceux que nous avons construits, diffère des modèles traditionnels tels que la régression linéaire classique ou la régression logistique. Au lieu de simplement calculer l'erreur entre les valeurs observées et les valeurs prédites, les modèles bayésiens prennent en compte l'incertitude dans les estimations des paramètres du modèle à travers les distributions postérieures. Par conséquent, plusieurs approches sont utilisées pour évaluer la qualité du modèle dans un cadre bayésien.

5.1 R-hat (Convergence des Échantillons)

Une des principales métriques pour évaluer la qualité de l'échantillonnage dans les modèles bayésiens est l'indicateur de convergence R-hat, également appelé *Gelman-Rubin statistic*. Cette métrique permet de vérifier si les chaînes de Markov ont convergé vers une distribution stationnaire, c'est-à-dire si les échantillons de la chaîne sont représentatifs de la véritable distribution a posteriori.

Un R-hat proche de 1 (en général, un seuil inférieur à 1.1) indique une bonne convergence des chaînes de Markov. Si R-hat est supérieur à 1.1, cela suggère que l'échantillonnage n'a pas encore convergé et que les résultats du modèle peuvent être biaisés.

Dans le contexte de notre modèle, nous avons calculé R-hat pour vérifier que l'échantillonnage a bien convergé avant de tirer des conclusions sur les paramètres du modèle. Par exemple, les valeurs de R-hat pour les coefficients du modèle sont vérifiées pour garantir la stabilité des estimations.

5.2 Intervalle de Crédibilité à 95% (HDI)

Un autre indicateur crucial dans un modèle bayésien est l'Intervalle de Crédibilité à 95% (HDI, pour *Highest Density Interval*). Ce concept remplace les intervalles de confiance traditionnels utilisés dans les modèles fréquentistes. L'HDI est un intervalle dans lequel les valeurs des paramètres ont la plus haute densité a posteriori, avec une probabilité cumulative de 95%. En d'autres termes, il s'agit de l'intervalle dans lequel les valeurs des coefficients ont le plus de chances d'être, étant donné les données observées.

L'HDI permet de quantifier l'incertitude sur les paramètres du modèle. Des intervalles larges suggèrent une plus grande incertitude sur la valeur des coefficients, tandis que des intervalles plus étroits indiquent une estimation plus précise.

5.3 WAIC

Le *Watanabe-Akaike Information Criterion* (WAIC) est une autre mesure de la qualité du modèle bayésien, qui est souvent utilisée pour évaluer l'ajustement du modèle en prenant en compte la complexité du modèle. Contrairement à l'AIC et au BIC, qui sont basés sur la vraisemblance marginale, le WAIC utilise les valeurs de log-vraisemblance des données pour chaque échantillon de la chaîne et les intègre pour obtenir une estimation du compromis entre la qualité de l'ajustement et la complexité du modèle.

Un WAIC plus faible indique un meilleur modèle. Le WAIC est particulièrement utile lorsque l'on compare plusieurs modèles bayésiens, car il prend en compte l'incertitude dans les paramètres estimés.

5.4 LOO (Leave-One-Out Cross Validation)

La validation croisée *Leave-One-Out* (LOO) est une méthode d'évaluation de la performance des modèles en utilisant chaque observation dans les données comme un point de validation, tandis que le reste des données est utilisé pour l'apprentissage du modèle. LOO est particulièrement adapté aux modèles bayésiens, car il permet d'évaluer la performance du modèle tout en tenant compte de l'incertitude des paramètres estimés.

Un score LOO plus faible indique une meilleure capacité de généralisation du modèle. LOO est généralement utilisé pour comparer la capacité de prédiction des différents modèles.

6 Simulation de Données Factices

Afin d'évaluer la robustesse de notre approche et de tester notre méthodologie avant d'appliquer les analyses aux données réelles, nous utilisons une simulation de données factices. Cette technique permet de générer artificiellement un jeu de données respectant certaines

hypothèses, notamment en ce qui concerne la distribution des variables et l'effet d'un traitement.

6.1 Principe de la Simulation

L'idée de la simulation de données factices repose sur la génération d'un échantillon artificiel dans lequel les variables sont définies selon des distributions spécifiques [4]. Nous attribuons ensuite aléatoirement un traitement aux observations et analysons les résultats obtenus. Ce processus permet de :

- Vérifier que les méthodes statistiques employées sont capables d'identifier correctement un effet lorsqu'il existe.
- Évaluer la précision des estimations et la variabilité des résultats.
- Anticiper les éventuels biais pouvant découler de la structure des données ou du mode d'attribution du traitement.

6.2 Génération des Données Factices

Nous créons un jeu de données factices avec 100 observations et deux variables explicatives x_1 et x_2. La variable de traitement z est également incluse, et la variable cible, wastecents, est générée en appliquant un effet linéaire des prédicteurs x_1, x_2 et du traitement z, avec un bruit ajouté.

La génération de ces données suit le modèle suivant :

$$\text{wastecents}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \gamma z_i + \epsilon_i$$

où :

- β_0 est l'intercept,
- β_1 et β_2 sont les coefficients des variables explicatives x_1 et x_2,
- γ est l'effet du traitement z,
- ϵ_i est un terme d'erreur tiré d'une distribution normale.

6.3 Modélisation Bayésienne avec PyMC

Pour modéliser cette relation de manière bayésienne, nous utilisons PyMC pour définir des priors sur les coefficients et l'intercept, ainsi qu'une distribution normale pour le bruit. Voici les étapes du modèle :

- Nous définissons des priors pour les coefficients β_0, β_1, β_2, et γ, typiquement en utilisant une distribution normale avec une moyenne de zéro et un écart-type élevé pour refléter une incertitude large.
- Nous spécifions la relation linéaire entre les variables explicatives x_1, x_2 et z et la variable cible wastecents.
- Nous procédons à l'échantillonnage pour obtenir les distributions postérieures des paramètres.

7 Performances sur les Données Factices

7.1 Performances de la Régression Linéaire Bayésienne

La régression linéaire bayésienne a montré des résultats intéressants, avec des coefficients estimés de manière précise. En examinant les valeurs des paramètres, nous pouvons faire plusieurs observations importantes.

Le modèle a bien appris les relations entre les prédicteurs (‘x1‘, ‘x2‘, ‘z‘) et la variable cible (‘wastecents‘). L'intercept est estimé à 14.541, ce qui suggère qu'en l'absence de tout effet des prédicteurs, la valeur moyenne de ‘wastecents‘ serait d'environ 14.5. Les coefficients des variables explicatives sont également intéressants :

- Le coefficient de ‘x1‘ est 0.474, ce qui signifie que pour chaque augmentation d'une unité de ‘x1‘, ‘wastecents‘ augmente de 0.474 en moyenne, en tenant compte des autres variables.
- Le coefficient de ‘x2‘ est -0.403, indiquant que pour chaque augmentation d'une unité de ‘x2‘, la variable cible diminue de 0.403 en moyenne.
- Enfin, le coefficient pour ‘z‘ est de 5.107, ce qui montre que le traitement (variable binaire ‘z‘) a un impact considérable sur ‘wastecents‘. Cela suggère que les observations traitées ont des valeurs de ‘wastecents‘ beaucoup plus élevées.

La convergence du modèle est confirmée par les valeurs de R-hat, qui sont toutes égales à 1.0 pour tous les paramètres. R-hat est une statistique de convergence qui permet de vérifier si les chaînes MCMC ont convergé vers la distribution cible. Une valeur de R-hat proche de 1.0 indique que l'échantillonnage a bien convergé, ce qui est le cas ici pour tous les paramètres.

Trace des échantillons de données factices pour l'intercept et les coefficients du modèle bayésien Linéaire

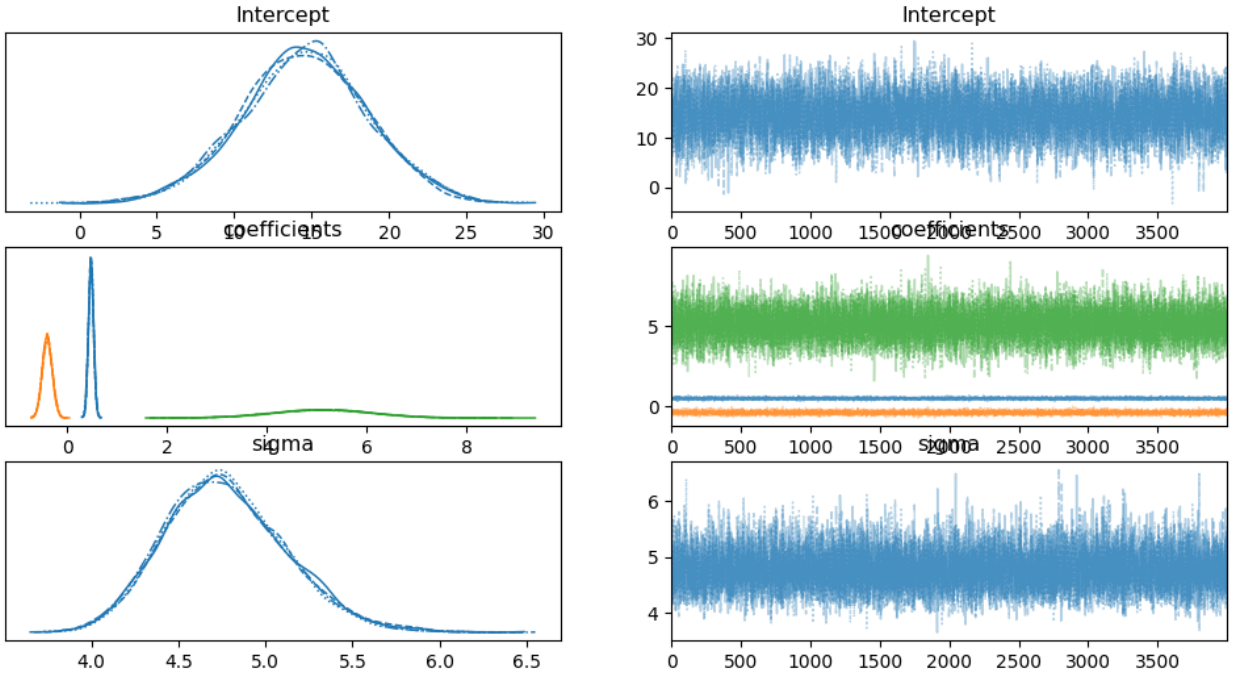

FIGURE 14 – Graphique des traces pour la régression linéaire bayésienne.

Les graphiques de trace pour le modèle linéaire bayésien (voir Figure 14) montrent des coefficients qui suivent une distribution normale bien centrée autour de la moyenne, ce qui indique que les estimations des paramètres sont stables et fiables. Voici l'analyse détaillée :

- **Distribution des coefficients** : Les coefficients suivent une distribution normale bien centrée autour de la moyenne, ce qui indique que les estimations des paramètres sont stables et fiables.
- **Variabilité des résidus (sigma)** : La variabilité des résidus est généralement faible, ce qui suggère que le modèle ajuste bien les données. Toutefois, certaines valeurs de sigma plus élevées peuvent indiquer des périodes d'incertitude ou des points aberrants dans les données.
- **Distribution des résidus** : Les résidus sont distribués autour de zéro, ce qui montre que le modèle capture bien la structure sous-jacente des données sans introduire de biais systématique.
- **Ajustement du modèle** : Le modèle linéaire semble bien ajusté aux données, avec des coefficients stables et des résidus centrés autour de zéro. La faible variabilité dans les observations renforce la fiabilité du modèle.
- **Pics de sigma élevés** : Les pics de sigma plus élevés pourraient signaler des périodes de plus grande incertitude, nécessitant un examen plus approfondi pour mieux comprendre les facteurs sous-jacents.

En somme, ce modèle est robuste et fournit des prédictions fiables pour les données analysées.

7.2 Performances de la Régression Poisson Bayésienne

En ce qui concerne la régression de Poisson bayésienne, nous avons également obtenu des résultats intéressants, avec des coefficients et une estimation de la variance qui sont tous significatifs.

Les résultats montrent que le modèle de Poisson a bien capturé les relations entre les prédicteurs et la variable cible ‘wastecents‘, qui suit une distribution de comptage. Voici quelques points à noter :

- Le coefficient pour ‘x1‘ est de 0.016, indiquant une relation positive modérée entre ‘x1‘ et ‘wastecents‘.
- Le coefficient de ‘x2‘ est légèrement négatif (-0.017), ce qui suggère que ‘x2‘ a un effet inverse sur ‘wastecents‘.
- Le coefficient de ‘z‘ est de 0.174, ce qui montre que les observations traitées ont, en moyenne, un nombre d’événements (‘wastecents‘) supérieur à celles du groupe contrôle.

Comme pour la régression linéaire, la convergence du modèle de Poisson est validée par les valeurs de R-hat égales à 1.0 pour tous les paramètres, ce qui indique que les chaînes MCMC ont convergé. De plus, les graphiques des traces (voir *Figure 15*) confirment que les échantillons sont bien mélangés et que le modèle a exploré l’espace des paramètres de manière adéquate, sans tendances visibles dans les traces.

Trace des échantillons de données factices pour l'intercept et les coefficients du modèle bayésien Poisson

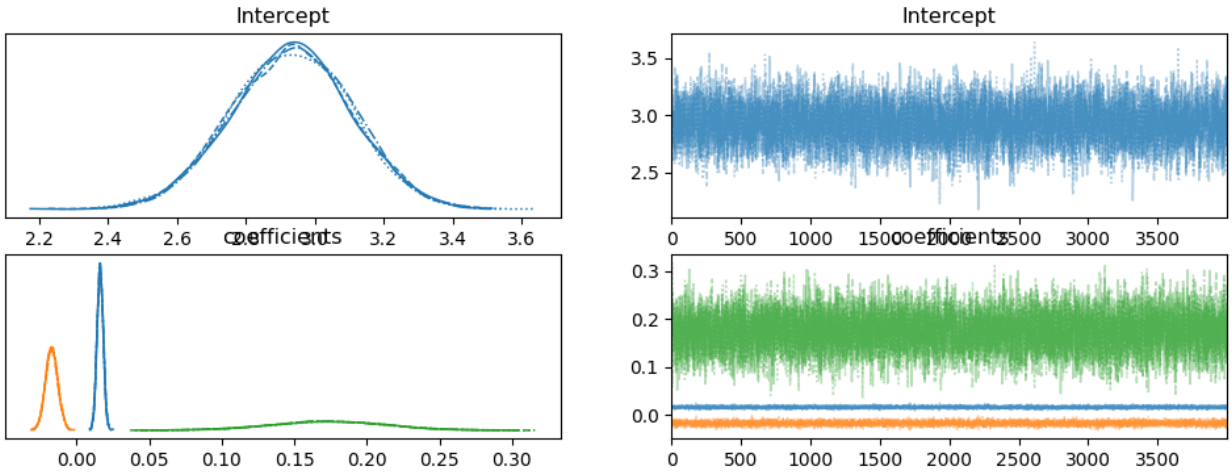

FIGURE 15 – Graphique des traces pour la régression de Poisson bayésienne.

Pour le modèle Poisson bayésien, les graphiques de trace (voir Figure 15) montrent également des coefficients suivant une distribution normale, suggérant des estimations stables et fiables des paramètres. Voici l'analyse détaillée :

- **Distribution des coefficients** : Les coefficients suivent une distribution normale, suggérant des estimations stables et fiables des paramètres.
- **Variabilité des résidus (sigma)** : La variabilité des résidus reste faible dans l'ensemble, bien que certaines fluctuations puissent indiquer des périodes d'incertitude ou des points aberrants.
- **Distribution des résidus** : Les résidus sont globalement distribués autour de zéro, ce qui montre que le modèle capture bien la structure des données sans introduire de biais.

En conclusion, le modèle Poisson est bien ajusté, avec des coefficients stables et des résidus bien répartis. Les valeurs de sigma restent généralement faibles, ce qui suggère une bonne prédiction des données, bien qu'une attention particulière soit nécessaire pour les périodes de variabilité accrue, qui pourraient nécessiter un ajustement ou une exploration plus approfondie.

7.3 Performances de la Régression Bayésienne avec Prior Laplace

Tout comme pour le modèle linéaire classique, on obtient ici encore des résultats très prometteurs. La convergence du modèle est confirmée par les valeurs de R-hat, qui sont toutes égaux à 1.0 pour tous les paramètres.

Trace des échantillons de données factices pour l'intercept et les coefficients du modèle bayésien avec Prior Laplace

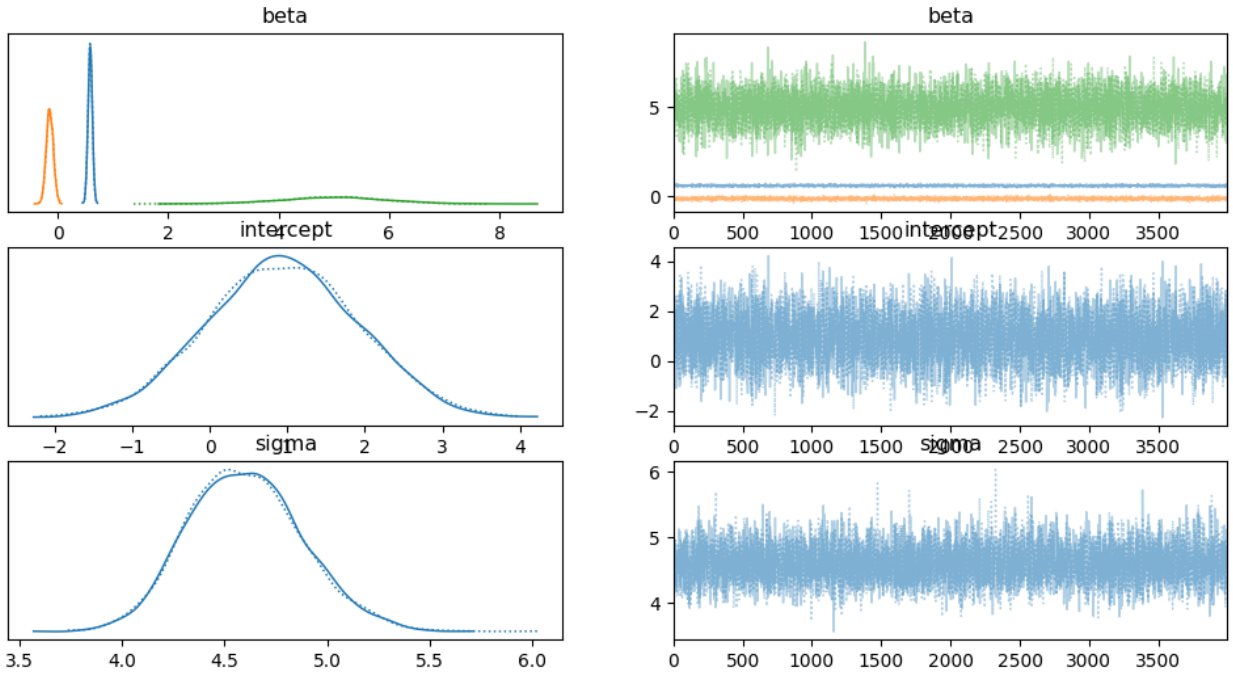

FIGURE 16 – Graphique des traces pour la régression de Poisson avec Prior Laplace.

Le graphique 16 montre que le modèle statistique analysé est robuste et stable :

— **Intercept** : Très stable, centré autour de zéro, indiquant une bonne définition et stabilité.

— **Sigma (écart-type)** : Distribution centrée autour de 4.5 à 5.0, montrant une stabilité autour de cette valeur centrale.

— **Beta (coefficients)** : Les coefficients suivent une distribution normale bien centrée autour de la moyenne, ce qui indique que les estimations des paramètres sont stables et fiables.

— **Évolution temporelle** : Les paramètres (intercept, sigma, beta) restent relativement stables sur les observations, avec des fluctuations mineures, ce qui est un bon signe pour la robustesse du modèle.

En conclusion, le modèle semble fiable pour des prédictions, avec des paramètres bien définis et une variabilité contrôlée.

7.4 Analyse comparative des trois modèles

Les trois modèles bayésiens analysés — régression linéaire, régression de Poisson, et régression avec prior Laplace — ont montré des résultats solides :

— **Identification des effets** : Tous les modèles ont bien capturé les relations entre les prédicteurs et la variable cible, notamment l'impact du traitement (variable z).
— **Précision et variabilité** : Les coefficients sont bien estimés avec une faible variabilité des résidus. La convergence des chaînes MCMC, validée par les valeurs de $\hat{R}$, assure la fiabilité des résultats.
— **Biais potentiel** : Aucun biais majeur n'a été observé, mais l'effet du traitement semble fortement marqué, suggérant une attention particulière à la structure des données et à l'attribution du traitement.

En somme, ces modèles sont fiables pour l'analyse des données factices, avec une bonne identification des effets et une faible variabilité des résultats. Une vigilance particulière est nécessaire concernant l'attribution du traitement afin d'éviter tout biais.

8 Performances sur les Données Réelles

8.1 Régression Linéaire Bayésienne

— **R-hat** : La majorité des R-hat pour les variables sont proches de 1, ce qui indique que le modèle a correctement convergé. Les estimations des coefficients sont donc stables et fiables.
— **HDI (Intervalles de Crédibilité)** : Certaines variables, telles que `wagesal` avec un intervalle [-6.587, 2.930], montrent des intervalles de crédibilité chevauchant 0, suggérant une incertitude importante concernant l'effet de ces variables sur `wastecents`.

Trace des Échantillons pour l'Intercept et les Coefficients

Les graphiques ci-dessous (Figure 19) montrent les traces des échantillons pour l'intercept et les coefficients du modèle linéaire bayésien, ainsi que les densités postérieures correspondantes. Les courbes de densité montrent une concentration autour des valeurs moyennes, avec des intervalles de crédibilité à 95% illustrant l'incertitude des estimations.

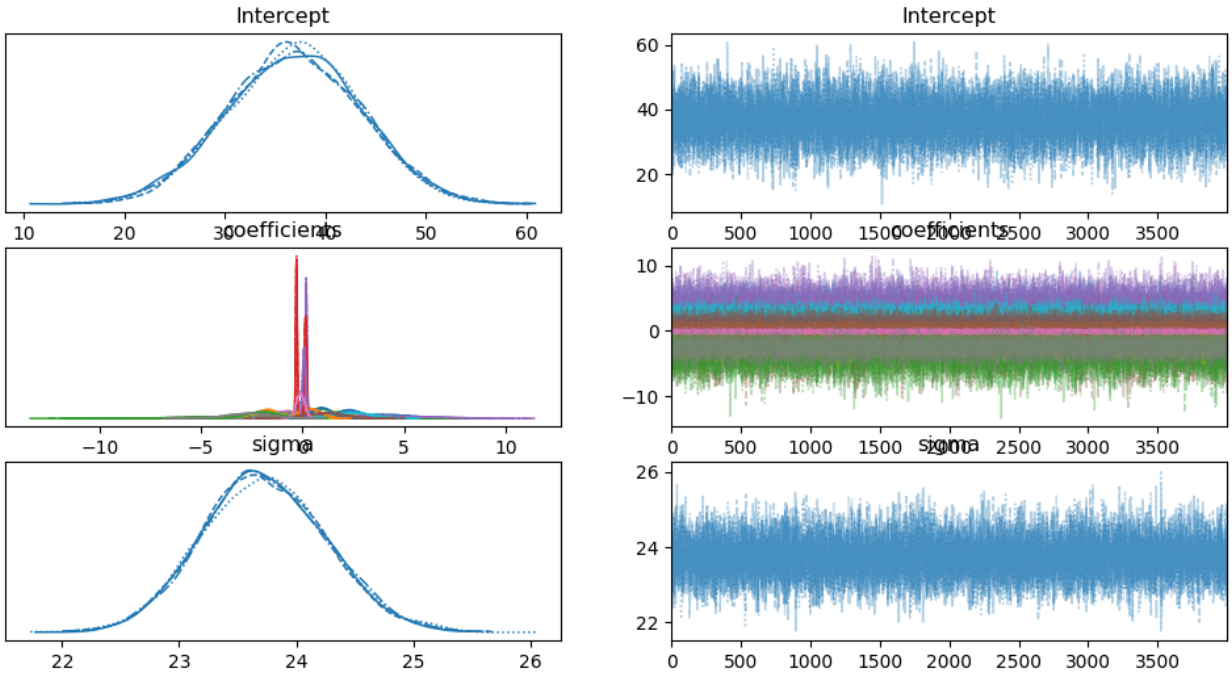

FIGURE 17 – Trace des échantillons pour l'intercept et les coefficients du modèle bayésien Linéaire

Les traces montrent des évolutions stables autour des valeurs moyennes, ce qui indique que les chaînes de Markov ont exploré correctement l'espace des paramètres. De plus, l'absence de tendances dans les traces et l'absence de chevauchement avec zéro pour certains coefficients suggèrent que ces variables ont un effet statistiquement significatif.

Variables Influentes

Les variables suivantes ont un effet statistiquement significatif sur la prédiction de `wastecents`, basées sur leurs intervalles de crédibilité (HDI) qui ne contiennent pas 0 et les valeurs de R-hat proches de 1 :

— `gender`, `educ`, `percenttp`, `child`, `eitcother`, `polideo`, `polknow2`, `polkno3`, `stateresid`, `raceeth`

Variables Moins Influentes

Les variables suivantes montrent un effet faible ou non significatif, avec des coefficients proches de 0, des HDIs chevauchant 0 et des écarts-types élevés :

— `partyid`, `taxpayer`, `wagesal`, `eitcself`, `polinffreq`, `regvote`, `discusspol`, `poleffic`, `polvol`, `polknow1`, `hhinc`

8.2 Régression Bayésienne Poisson

— **R-hat** : Les R-hat pour la majorité des variables sont proches de 1, suggérant une bonne convergence du modèle. Cependant, certaines variables, comme l'intercept (R-hat = 1.20) et `regvote` et `voted` (R-hat = 1.59), montrent des problèmes de convergence, indiquant que ces paramètres nécessitent peut-être plus d'itérations ou un ajustement des paramètres de convergence.

— **HDI (Intervalles de Crédibilité)** : Certaines variables, telles que `regvote`, `voted`, et `raceeth`, ont des intervalles de crédibilité contenant 0, ce qui suggère que ces variables n'ont probablement pas d'effet significatif. D'autres variables, telles que `gender`, `educ`, `wagesal`, et `eitcother`, ont des intervalles de crédibilité qui ne contiennent pas 0, indiquant un effet statistiquement significatif.

Trace des Échantillons pour l'Intercept et les Coefficients

Les graphiques ci-dessous montrent les traces des échantillons pour l'intercept et les coefficients du modèle Poisson bayésien. En comparaison avec le modèle linéaire, certaines traces des coefficients montrent des fluctuations plus importantes, indiquant des problèmes de convergence pour certains paramètres.

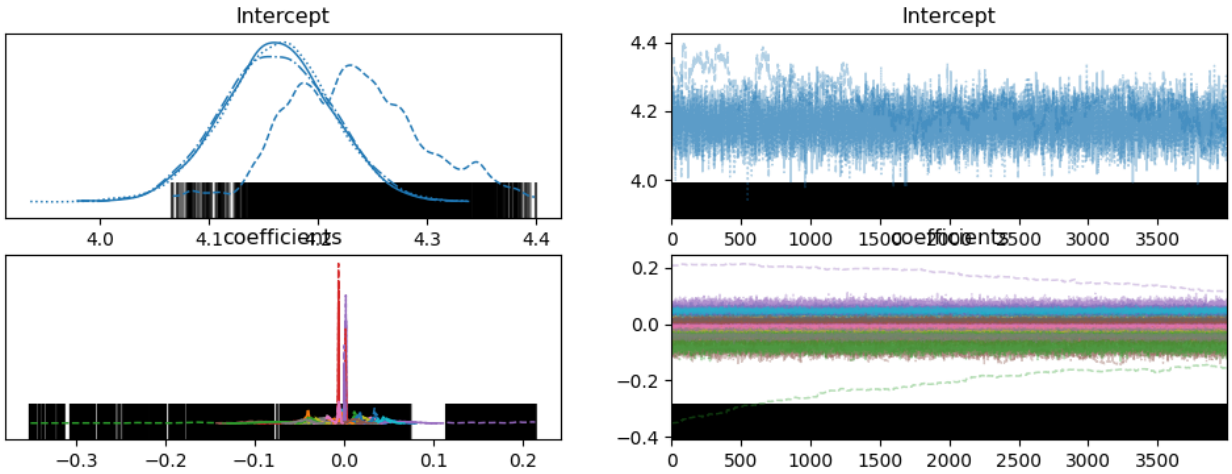

FIGURE 18 – Trace des échantillons pour l'intercept et les coefficients du modèle bayésien Poisson

Bien que la plupart des paramètres montrent des traces stables, des fluctuations pour `regvote`, `voted`, et `raceeth` suggèrent des problèmes de convergence. Un nombre plus élevé d'itérations ou un ajustement des paramètres de convergence pourrait améliorer la stabilité. De plus, les intervalles de crédibilité pour certaines variables, telles que `regvote`, montrent un effet non significatif.

Variables avec un Effet Statistiquement Significatif

Les variables suivantes montrent un effet significatif sur `wastecents`, car leurs intervalles de crédibilité ne contiennent pas 0 :

— `gender`, `educ`, `percenttp`, `eitcother`, `polideo`, `polvol`, `polknow2`

Variables avec un Effet Incertain ou Faible

Les variables suivantes montrent un effet faible ou non significatif, avec des intervalles de crédibilité chevauchant 0 :

— `wagesal`, `regvote`, `voted`, `raceeth`

8.3 Régression Bayésienne avec un Prior Laplace

— **R-hat** : Comme pour le modèle linéaire classique, les R-hat de la grande majorité des variables sont proches de 1, ce qui indique que le modèle a correctement convergé. Les estimations des coefficients sont donc stables et fiables.

— **HDI (Intervalles de Crédibilité)** : Certaines variables, comme gender, educ et taxpayer, ont cette fois des intervalles de crédibilité contenant 0, ce qui rend l'interprétation de leur effet sur `wastecents` incertain. En revanche, d'autres variables présentent un effet statistiquement notable. C'est le cas de feelfedgov_1, qui montre une relation négative, indiquant que la méfiance envers le gouvernement est associée à une perception réduite des dépenses. De même, *weightvec* affiche un effet négatif modéré, tandis que *wagesal_1.0* a un impact marqué, soulignant le rôle des revenus dans la perception des dépenses publiques.

Trace des Échantillons pour l'Intercept et les Coefficients

Les traces des échantillons pour l'intercept et les coefficients du modèle bayésien Laplace sont présentées dans la figure suivante (Figure 19).

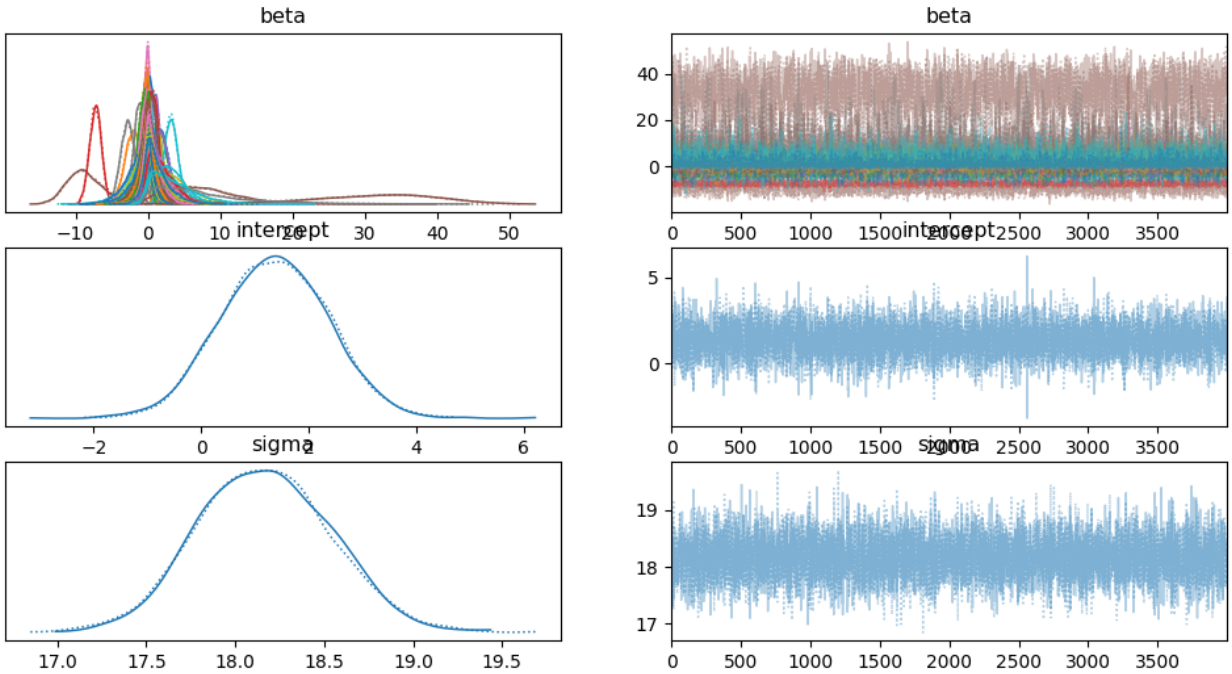

FIGURE 19 – Trace des échantillons pour l'intercept et les coefficients du modèle bayésien Laplace

Le résultat est encore une fois très proche de celui obtenu pour la régression bayésienne linéaire classique. Avec les mêmes observations que pour le modèle linéaire, on peut dire que les Chaînes de Markov ont correctement exploré l'espace des paramètres et que certains de nos paramètres reflètent bel et bien un effet statistiquement significatif.

Les variables les plus influentes, avec un effet significatif sur la prédiction de `wastecents`, incluent :

— `wagesal_1.0`, `partyid`, `taxpayer_1`, `feelfedgov_1`

Les variables moins influentes, dont les coefficients sont proches de 0 ou ont des écarts-types élevés, incluent :

— `polinffreq`, `gender_3`, `labforce_4.0`, `hhinc_3.0`, `labforce_8.0`

On aurait pu s'attendre à ce que les variables significatives soient les mêmes ici que pour le modèle linéaire classique, cependant on observe malgré tout une différence notable entre nos deux résultats. En effet, le modèle linéaire classique semblait porter peu d'importance à certaines variables comme `partyid` ou encore `wagesal` alors même qu'elles sont considérés comme étant très importantes ici. Cela va dans le sens de l'analyse exploratoire effectué plus tôt qui semblait souligner des différences notables dans la perception du gachis des taxes en fonction de ces dernières.

8.4 Comparaison entre les modèles

Les résultats des critères **WAIC** et **LOO** pour les trois modèles sont présentés dans le tableau ci-dessous.

Critère	Modèle Linéaire	Modèle Poisson	Modèle avec Prior Laplace
elpd_waic	-4603.00 ± 19.72	-8742.07 ± 247.09	-4657.02 ± 28.77
p_waic	30.01	307.89	33.95
elpd_loo	-4603.07 ± 19.72	-8742.58 ± 247.10	-4657.09 ± 28.78
p_loo	30.07	308.40	34.01
Pareto k (good)	100%	99.7%	100%
Pareto k (bad)	0%	0.3%	0%
Pareto k (very bad)	0%	0%	0%

TABLE 8 – Comparaison des critères WAIC et LOO entre le modèle linéaire et le modèle de Poisson.

Analyse des Résultats

Les résultats indiquent que le **modèle linéaire** est le plus performant parmi les trois modèles analysés :

- **Meilleur ajustement aux données** : Le **modèle linéaire** présente les valeurs les plus élevées pour **elpd_waic** et **elpd_loo**, ce qui signifie qu'il fournit la meilleure généralisation aux données.
- **Modèle avec prior Laplace** :
  - Le **modèle avec prior Laplace** obtient des performances légèrement inférieures au modèle linéaire, mais reste nettement meilleur que le modèle de Poisson.
  - Cette amélioration par rapport au modèle de Poisson suggère que l'utilisation d'un prior Laplace peut aider à mieux capturer certaines structures des données.
- **Modèle de Poisson** :
  - Le **modèle de Poisson** est le moins performant, avec des scores **elpd_waic** et **elpd_loo** nettement plus faibles.
  - De plus, il est beaucoup plus complexe que les autres modèles, comme l'indiquent les valeurs élevées de **p_waic** et **p_loo**, suggérant un risque de surajustement.
- **Stabilité des modèles** :
  - L'analyse des valeurs de **Pareto k** montre que le modèle linéaire et le modèle avec prior Laplace ont une bonne stabilité (**0% de valeurs problématiques**).
  - En revanche, le modèle de Poisson présente **quelques valeurs aberrantes**, ce qui peut affecter sa fiabilité.

Graphiques des Log-Vraisemblances

Afin de visualiser la différence de log-vraisemblance entre les modèles, nous avons tracé les courbes de densité des log-vraisemblances (Figure 20) pour les deux modèles.

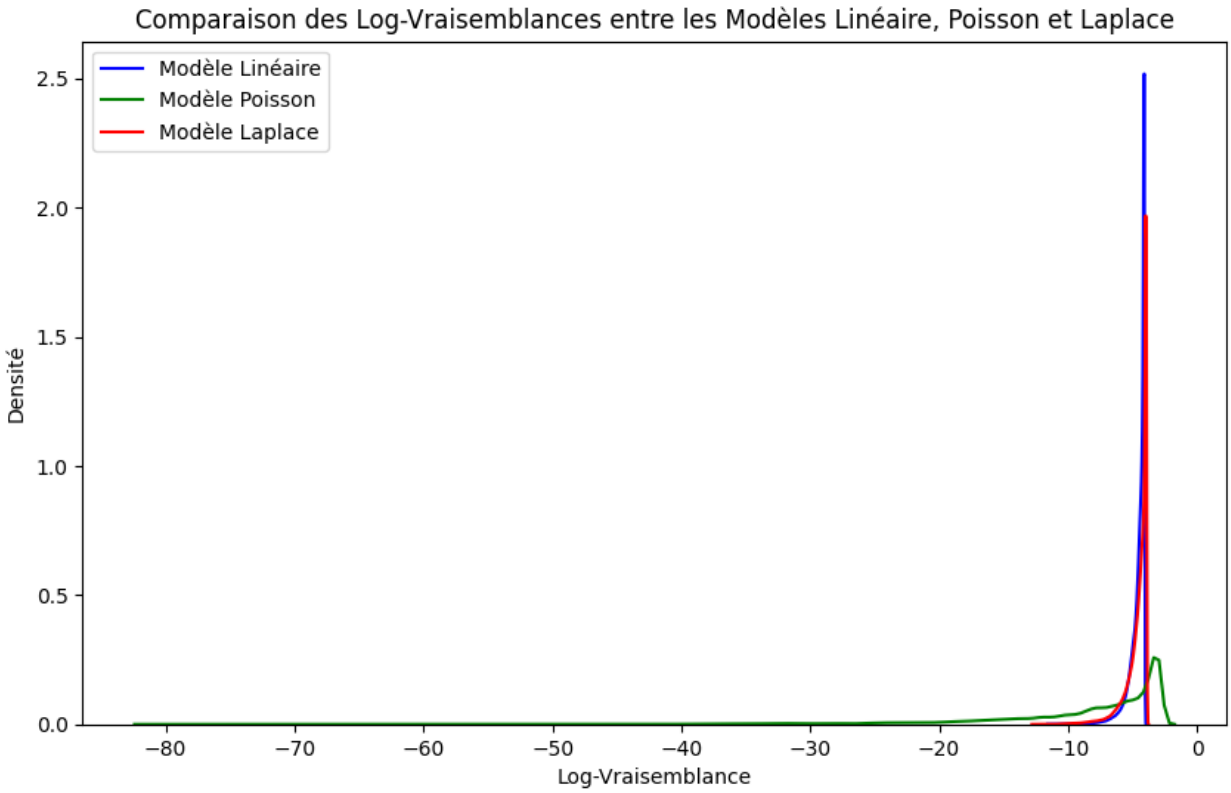

FIGURE 20 – Comparaison des Log-Vraisemblances entre les Modèles Linéaire et Poisson

Modèle Linéaire (bleu)

- La courbe bleue montre la distribution des log-vraisemblances pour le modèle linéaire.
- Elle présente un pic très prononcé autour de -5, indiquant que la plupart des log-vraisemblances pour ce modèle sont concentrées autour de cette valeur.
- Cependant, cela peut aussi signifier que le modèle est très sensible aux données et pourrait surajuster.

Modèle de Poisson (vert)

- La courbe verte montre la distribution des log-vraisemblances pour le modèle de Poisson.
- Elle est plus étalée que celle du modèle linéaire, avec une densité plus faible et plus uniforme sur une large gamme de valeurs de log-vraisemblance.
- Cela peut suggérer que le modèle est plus robuste à différentes configurations de données, mais il pourrait aussi indiquer une moins bonne adéquation globale par rapport au modèle linéaire.

Modèle de Laplace (rouge)

- La courbe rouge montre la distribution des log-vraisemblances pour le modèle de Laplace.
- Elle suit de près la distribution du modèle linéaire, avec un pic également autour de -5, mais légèrement moins prononcé.
- La distribution est similaire à celle du modèle linéaire, mais légèrement moins concentrée.
- Cela pourrait indiquer que le modèle de Laplace offre un compromis entre la précision du modèle linéaire et la robustesse du modèle de Poisson.

8.5 Choix du Modèle

En conclusion, bien que les trois modèles aient leurs avantages et inconvénients, le **modèle linéaire** s'avère être le meilleur choix pour cette analyse en raison de sa capacité à ajuster les données de manière optimale tout en maintenant une faible complexité. Le modèle avec prior Laplace constitue un bon compromis pour des données présentant des structures plus complexes, offrant une performance légèrement inférieure à celle du modèle linéaire mais toujours meilleure que celle du modèle de Poisson.

Le **modèle de Poisson**, en revanche, est moins performant, notamment en raison de sa complexité plus élevée et de son risque accru de surajustement. En conséquence, le modèle linéaire est recommandé si l'objectif est d'obtenir des prédictions précises et stables, tandis que le modèle avec prior Laplace pourrait être envisagé dans des situations où une certaine robustesse supplémentaire est requise, notamment pour des jeux de données plus hétérogènes ou bruités.

9 Conclusion du Projet

Ce projet a exploré la perception du gaspillage des fonds publics en analysant deux variables clés : `wastecent` et `wastethink`. Les résultats montrent que plusieurs facteurs influencent cette perception. Les modèles bayésiens utilisés ont permis de quantifier ces influences, avec une préférence pour le modèle linéaire en raison de sa capacité à ajuster les données de manière optimale tout en maintenant une faible complexité.

Les principales conclusions sont les suivantes :

- **Variables les Plus Influentes** : Les variables les plus influentes dans la perception du gaspillage des fonds publics sont :
  - `gender` : Le genre a un impact notable, avec des différences dans la perception du gaspillage en fonction de l'identité de genre.
  - `educ` : Le niveau d'éducation est fortement lié à la perception du gaspillage des fonds publics, les personnes ayant un niveau d'éducation plus élevé ayant tendance à percevoir moins de gaspillage.

- `percenttp` : La perception des contribuables aux États-Unis influence la manière dont les individus évaluent l'efficacité des dépenses publiques.
- `polknow2` : La connaissance des partis politiques, en particulier ceux présents au Sénat, montre un lien entre la familiarité avec la politique et la perception du gaspillage gouvernemental.
- `eitcother` : La connaissance de l'EITC en dehors du foyer, qui reflète l'exposition à des expériences fiscales personnelles, modère la perception du gaspillage des fonds publics.

— **Performance des Modèles Bayésiens** : Le modèle linéaire bayésien s'est avéré être le plus performant, offrant un bon équilibre entre ajustement des données et complexité. Le modèle avec prior Laplace offre une alternative robuste, tandis que le modèle de Poisson est moins adapté en raison de sa complexité élevée et de son risque de surajustement.

Ces résultats soulignent l'importance de prendre en compte divers facteurs socio-économiques et politiques lors de l'analyse des perceptions publiques concernant l'utilisation des fonds publics. Ils mettent également en évidence la nécessité pour les gouvernements de communiquer de manière transparente sur l'utilisation des fonds publics afin de rétablir la confiance des citoyens.

Annexes

Description du Jeu de Données

Nom	Description	Format
gender	Quel genre vous identifiez-vous ?	Entier
educ	Quel est le plus haut niveau d'éducation que vous avez atteint ?	Entier
partyid	Comment vous identifiez-vous politiquement ?	Entier
firstthought	Quelle est votre première pensée lorsque vous entendez le mot "impôts" ?	Texte
taxpayer	Êtes-vous un contribuable ?	Booleen
percenttp	Quel pourcentage des adultes aux États-Unis sont des contribuables selon vous ?	Décimal
recent	Décrivez la dernière fois que vous avez payé des impôts.	Texte
tpfeel	Comment vous sentez-vous d'être un contribuable ?	Texte
biggest	Quel type d'impôt représente la plus grande part du budget familial ?	Texte
wagesal	Avez-vous déjà travaillé pour un salaire ou un salaire fixe ?	Booleen
paystub	Quels impôts apparaissent sur votre fiche de paie ?	Texte
child	Combien d'enfants de moins de 17 ans vivent dans votre foyer ?	Entier
depend	Avez-vous d'autres personnes à charge sur vos impôts ?	Booleen
glad	Pour quoi êtes-vous content que vos impôts financent ?	Texte
upset	Pour quoi êtes-vous contrarié que vos impôts financent ?	Texte
benefit	Avez-vous personnellement bénéficié des dépenses fiscales ? Si oui, comment ?	Texte
wastecents	Combien de centimes de chaque dollar d'impôt pensez-vous que le gouvernement gaspille ?	Décimal
wastethink	Que pensez-vous du gaspillage gouvernemental ?	Texte
eitcself	Avez-vous déjà bénéficié du crédit d'impôt sur le revenu gagné (EITC) ou du crédit d'impôt pour enfants ?	Entier
eitcexp	Comment ce crédit d'impôt affecte-t-il vos impôts et votre famille ?	Texte
eitcother	Connaissez-vous quelqu'un en dehors de votre foyer ayant reçu l'EITC ou le crédit d'impôt pour enfants ?	Entier
eitcthink	Que pensez-vous du crédit d'impôt sur le revenu gagné et/ou du crédit d'impôt pour enfants ?	Texte
labforce	Quel était votre statut la semaine dernière ?	Entier
polideo	Comment vous situez-vous politiquement sur une échelle de 1 à 7 ?	Entier
polinffreq	À quelle fréquence avez-vous cherché des informations sur un candidat ou des questions politiques ?	Entier
regvote	Étiez-vous inscrit pour voter lors de l'élection du 4 novembre 2014 ?	Entier
feelfedgov_1	Quelle est votre opinion sur le gouvernement fédéral ?	Décimal

Nom	Description	Format
voted	Avez-vous voté lors de l'élection du 4 novembre 2014 ?	Entier
discusspol	À quelle fréquence discutez-vous de politique avec vos amis ou famille ?	Entier
poleffic	Pensez-vous que les gens comme vous ont leur mot à dire sur le gouvernement ?	Entier
polvol	Avez-vous travaillé pour un parti ou un candidat lors des élections ?	Entier
polknow1	Quel parti avait le plus grand nombre de membres à la Chambre des représentants avant les élections ?	Entier
polknow2	Quel parti avait le plus grand nombre de membres au Sénat avant les élections ?	Entier
polkno3	Quelle fonction occupe John Roberts ?	Entier
marital	Statut marital actuel.	Entier
ownhome	Propriété du logement.	Entier
stateresid	État de résidence.	Entier
yearbirth	Année de naissance.	Entier
raceeth	Race ou ethnie.	Entier
hhinc	Revenu familial total de l'année précédente (avant impôts).	Entier

Code

1. Simulation des Données Factices

```
# Importer les librairies necessaires
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import pymc as pm
import numpy as np

# Simulation de donnees factices
np.random.seed(42)

n_samples = 200  # Nombre d'observations
n_predictors = 5  # Nombre de predicteurs

# Generer des donnees
X_simulated = np.random.randn(n_samples, n_predictors)

coefficients_true = np.random.randn(n_predictors)
intercept_true = np.random.randn(1)
wastecents_simulated = intercept_true + np.dot(X_simulated,
    coefficients_true) + np.random.randn(n_samples) * 0.5  # ajouter du
    bruit

# Creer un DataFrame avec les donnees simulees
df_simulated = pd.DataFrame(X_simulated, columns=[f'X{i+1}' for i in range
    (n_predictors)])
df_simulated['wastecents'] = wastecents_simulated

print(df_simulated.head())
```

2. Traitement des Données Factices

```
df_numeric = df_simulated.select_dtypes(include=['number']).copy()

# Remplacer les valeurs manquantes par la mediane des colonnes numeriques
df_numeric = df_numeric.apply(lambda col: col.fillna(col.median()) if pd.
    api.types.is_numeric_dtype(col) else col)

# Verification
print("Valeurs manquantes apres imputation :\n", df_numeric.isnull().sum()
    )

# Selectionner les variables explicatives (toutes sauf wastecent)
predictors = df_numeric.drop(columns=['wastecents']).columns.tolist()

# Verification des types de donnees des predicteurs
print("Types de donnees des predicteurs :\n", df_numeric[predictors].
    dtypes)
```

3. Modèles Bayésiens sur les Données Factices

3.1. Modèle Bayésien Linéaire (Données Factices)

```
with pm.Model() as linear_model:
    # Prior
    intercept = pm.Normal('Intercept', mu=0, sigma=10)
    coefficients = pm.Normal('coefficients', mu=0, sigma=10, shape=len(predictors))

    # Lien lineaire
    X = df_numeric[predictors].values
    mu = intercept + pm.math.dot(X, coefficients)

    # Likelihood
    sigma = pm.HalfNormal('sigma', sigma=10)
    wastecents_obs = pm.Normal('wastecents_obs', mu=mu, sigma=sigma, observed=df_numeric['wastecents'])

    # Echantillonnage
    trace_linear = pm.sample(4000, tune=1000, cores=4, return_inferencedata=True)

    # Calculer la log-vraisemblance
    log_likelihood_linear = pm.compute_log_likelihood(trace_linear)

log_likelihood_group = log_likelihood_linear['log_likelihood']
log_likelihood_values = log_likelihood_group['wastecents_obs'].values

# Calculer WAIC et LOO pour le modele
import arviz as az
WAIC_linear = az.waic(log_likelihood_linear)
LOO_linear = az.loo(log_likelihood_linear)
```

3.2. Modèle Bayésien Poisson (Données Factices)

```
with pm.Model() as poisson_model:
    # Prior
    intercept = pm.Normal('Intercept', mu=0, sigma=10)
    coefficients = pm.Normal('coefficients', mu=0, sigma=10, shape=len(predictors))

    # Lien log-lineaire
    X = df_numeric[predictors].values
    log_lambda = intercept + pm.math.dot(X, coefficients)

    # Likelihood - Distribution de Poisson
    wastecents_obs = pm.Poisson('wastecents_obs', mu=pm.math.exp(log_lambda), observed=df_numeric['wastecents'])

```

```
    # Echantillonnage
    trace_poisson = pm.sample(4000, tune=1000, cores=4,
   return_inferencedata=True)

    # Calculer la log-vraisemblance
    log_likelihood_poisson = pm.compute_log_likelihood(trace_poisson)

log_likelihood_group_poisson = log_likelihood_poisson['log_likelihood']
log_likelihood_poisson_values = log_likelihood_group_poisson['
   wastecents_obs'].values

# Calculer WAIC et LOO pour le modele
WAIC_poisson = az.waic(log_likelihood_poisson)
LOO_poisson = az.loo(log_likelihood_poisson)
```

3.3. Modèle Bayésien avec Prior Laplace (Données Factices)

```
with pm.Model() as laplace_model:
    # Prior sur les coefficients (Laplace)
    beta = pm.Laplace("beta", mu=0, b=1, shape=X.shape[1])

    # Prior sur l'intercept
    intercept = pm.Normal("intercept", mu=0, sigma=1)

    # Fonction de lien lineaire
    mu = intercept + pm.math.dot(X, beta)

    # Prior sur l'ecart-type (sigma) de la vraisemblance
    sigma = pm.HalfNormal("sigma", sigma=1)

    # Likelihood - Distribution normale
    y_obs = pm.Normal("y_obs", mu=mu, sigma=sigma, observed=df_numeric['
   wastecents'])

    # Echantillonnage
    trace_laplace = pm.sample(4000, tune=1000, return_inferencedata=True,
   cores=2, target_accept=0.95)

# Calcul manuel de la log-vraisemblance
beta_samples = trace_laplace.posterior['beta'].values
intercept_samples = trace_laplace.posterior['intercept'].values
sigma_samples = trace_laplace.posterior['sigma'].values

log_likelihood_values = -0.5 * np.sum(
    np.log(2 * np.pi * sigma_samples**2) + ((df_numeric['wastecents'] -
   intercept_samples - np.dot(X, beta_samples.T))**2) / (sigma_samples**2)
   ,
    axis=1
)

```

```
# Calcul de WAIC et LOO
WAIC_laplace = az.waic(log_likelihood_values)
LOO_laplace = az.loo(log_likelihood_values)
```

5. Données Réelles et Traitement

```
# Charger les donnees reelles
df_real = pd.read_csv("Q14 survey_for_dataverse.csv", encoding="ISO-8859-1")
print(df_real.head())

# Traitement similaire aux donnees factices
df_numeric_real = df_real.select_dtypes(include=['number']).copy()
df_numeric_real = df_numeric_real.apply(lambda col: col.fillna(col.median()) if pd.api.types.is_numeric_dtype(col) else col)

# Verification
print("Valeurs manquantes apres imputation :\n", df_numeric_real.isnull().sum())

# Variables explicatives pour les donnees reelles
predictors_real = df_numeric_real.drop(columns=['wastecents']).columns.tolist()
```

6. Modèles Bayésiens sur les Données Réelles

```
# Application des modeles bayesiens sur les donnees reelles (le meme processus que pour les donnees factices)
```

Références

[1] Guillaume KON KAM KING, *Description of the projects*. Université Paris-Saclay, INRAE, MaIAGE. https://drive.google.com/file/d/1vREoVvRL9wP_eppeXMcJem0dhlENKW34/view

[2] V. Williamson, (2017), *Read My Lips : Why Americans Are Proud to Pay Taxes, Introduction*, Princeton University Press. https://assets.press.princeton.edu/chapters/i10977.pdf

[3] Paul Gonzalez, (2014), *Poll : Americans estimate federal government wastes 51 cents on the dollar*. https://www.washingtonexaminer.com/opinion/698845/poll-americans-estimate-federal-government-wastes-51-cents-on-the-dollar/

[4] Gelman, Hill, Vehtari (2020), *Regression and Other Stories*. https://users.aalto.fi/~ave/ROS.pdf

[5] Jianming Ma, Kara M. Kockelman (2006), *Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity*. https://www.caee.utexas.edu/prof/kockelman/public_html/TRB06MVPBayesian.pdf

[6] Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016), *Probabilistic programming in Python using PyMC3*, PeerJ Computer Science, 2, e55. https://doi.org/10.7717/peerj-cs.55