

Helyn - Étude statistique

Prédire les résultats des élections présidentielles françaises d'un bureau de vote.

D'après Cyril Deschamps, Etienne Ducros, Hugo Brun et Lennon Herrmann
Application du projet : <https://helyn.cyrildeschamps.fr>

2025-01-15

Contents

1	Introduction	2
1.1	Problématique	2
1.2	Cadre théorique	2
1.3	Modèle de l'étude	3
1.4	Objectifs de l'étude	3
1.5	Livrable	3
2	Identification et récupération des données	3
2.1	Élections par bureau de vote	3
2.2	Données socio-démographiques	3
3	Nettoyage et fusion des données	4
3.1	Traitemennt des données électORALES	5
3.2	Traitemennt des données socio-démographiques	5
4	Visualisation par la carte	5
5	Data Mining Descriptif	6
5.1	Recherche de corrélation globale	7
5.2	Richesse	8
5.3	Khi-2	9
5.4	Impact du nombre d'individu	10
5.5	Recherche globale par visualisation poussée	10
6	Recherche de corrélations poussées	11
6.1	Régression linéaire	11
6.2	Régression multiple	12
6.3	Random forest	13
6.3.1	Entraînement	13
6.3.2	Etude de la random forest	13
7	Synthèse	18
7.1	Résultats principaux	19
7.2	Profils types	19
7.3	Modélisation	19
7.4	Limites et biais	19
7.5	Conclusion	19
7.6	Pour aller plus loin	20

8 Annexe	20
8.1 Script de régression linéaire et résultat	20
8.2 Script d'Entraînement random forest (via ranger, sur plusieurs coeurs)	21
8.3 Script de merge spatiale (simplifié)	22
8.4 Code initialisation carte React avec MapLibre	22

1 Introduction

1.1 Problématique

Prédire les résultats d'élections présidentielles française d'un bureau de vote, dans l'idéal. Mais en réalité le but était surtout de trouver des corrélations et dépendances entre les résultats d'élections présidentielles française d'un bureau de vote, et des données démographiques.

1.2 Cadre théorique

Les préférences politiques des individus peuvent être façonnées par des variables socio-démographiques telles que l'âge, le niveau d'éducation, la situation économique ou encore le lieu de résidence. Nous voulions essayer d'observer des tendances que créent ces cadres théoriques. Plusieurs cadres théoriques permettent de formuler des hypothèses sur ces influences potentielles :

La **théorie de la socialisation politique** suppose que des facteurs comme l'âge, le genre et la composition familiale jouent un rôle dans la formation des opinions politiques. Par exemple, on pourrait s'attendre à ce que les jeunes adultes soient plus attirés par des idées progressistes, tandis que les générations plus âgées privilégient des positions conservatrices, influencées par leurs expériences passées.

Les théories liées au **clivage de classe sociale**, inspirées par les travaux de Karl Marx et Pierre Bourdieu, postulent que la position des individus dans la hiérarchie socio-professionnelle peut orienter leurs préférences politiques. Les individus occupant des emplois manuels ou précaires pourraient être davantage sensibles aux discours en faveur de la justice sociale, tandis que les catégories socio-professionnelles supérieures seraient plus réceptives aux politiques favorisant l'investissement et la stabilité économique.

Le **capital culturel**, tel que défini par Pierre Bourdieu, pourrait également jouer un rôle dans le comportement électoral. Les niveaux de qualification peuvent influencer la perception des enjeux politiques : les personnes disposant d'un diplôme supérieur pourraient privilégier des valeurs universalistes et libérales, alors que les moins diplômés pourraient être plus sensibles aux discours protectionnistes ou nationalistes.

La **théorie du contexte géographique** met en avant l'importance du lieu de vie et des interactions sociales locales sur les comportements politiques. Les caractéristiques urbaines et rurales, la proportion de logements sociaux ou encore l'ancienneté des constructions peuvent être des indicateurs pertinents pour comprendre les dynamiques électorales locales.

La **théorie de la privation relative** suggère que les inégalités perçues entre les attentes des individus et leur situation réelle peuvent les amener à soutenir des mouvements politiques contestataires. Des indicateurs comme le taux de pauvreté, le niveau de vie ou la précarité de l'emploi pourraient refléter ce sentiment de frustration.

Enfin, la **théorie du choix rationnel** propose que les électeurs adoptent des comportements stratégiques, en cherchant à maximiser leurs bénéfices individuels à travers leurs choix politiques. Par exemple, les propriétaires immobiliers ou les ménages vivant seuls pourraient orienter leurs votes en fonction des politiques perçues comme avantageuses ou défavorables à leurs intérêts.

1.3 Modèle de l'étude

Quel lien existe entre les différents facteurs socio-démographiques et influencent-ils les choix politiques des citoyens ?

1.4 Objectifs de l'étude

L'objectif est d'explorer ces hypothèses théoriques en s'appuyant sur des données socio-démographiques et électorales de plusieurs années. Cette étape théorique permet de poser les bases d'une analyse empirique qui pourra confirmer ou infirmer ces hypothèses en fonction des observations. Nous étudierons les liens au travers de différentes techniques d'analyse statistiques.

1.5 Livrable

L'application sous forme d'une carte interactive, ainsi que des diagrammes de notre étude, est accessible via l'URL : <https://helyn.cyrildeschamps.fr>

2 Identification et récupération des données

2.1 Élections par bureau de vote

D'après le cadre théorique, la première étape consiste à collecter les résultats des élections présidentielles. Pour enrichir notre jeu de données, nous avons choisi de les détailler au niveau des bureaux de vote. La plateforme *data.gouv.fr* nous donne accès à ce type de données. Lors de chaque élection, environ 70 000 bureaux de vote partagent leurs résultats. Nous limiterons toutefois notre étude aux trois dernières élections présidentielles (2012, 2017 et 2022), ce qui sera largement suffisant pour répondre à notre problématique.

De plus, nous nous concentrerons uniquement sur les bureaux de vote situés en France métropolitaine afin de simplifier la visualisation des données sur une carte. Cette restriction n'a qu'un faible impact sur la représentativité globale du jeu de données.

Enfin, nous n'utiliserons que les données des premiers tours. Elles sont plus représentatives du choix de chacun, car non-influencé par l'obligation de voter pour le "moins pire".

2.2 Données socio-démographiques

À partir du cadre théorique, nous avons identifié plusieurs groupes de facteurs socio-démographiques susceptibles d'influencer les orientations politiques des individus :

- **Facteurs démographique** : ce groupe comprend des variables telles que l'âge, le genre et la composition familiale. Ces éléments permettent d'évaluer l'impact de la socialisation politique, notamment l'influence des générations et des expériences de vie sur le comportement électoral.
- **Facteurs travail et revenus** : ces variables incluent la catégorie socio-professionnelle, le type d'emploi (salarié ou indépendant), le taux d'activité et le niveau de vie. Elles permettent d'explorer le rôle des disparités économiques et de la hiérarchie sociale sur les préférences politiques, en lien avec la théorie du clivage de classe.
- **Facteurs éducation** : le niveau de diplôme et les qualifications constituent des indicateurs clés du capital culturel. Ces données sont utiles pour comprendre la manière dont l'accès au savoir et la formation influencent la perception des enjeux politiques.
- **Facteurs lieu de vie** : ce groupe comprend des variables liées au lieu de résidence, telles que la répartition urbaine ou rurale, la proportion de logements sociaux, et l'ancienneté des constructions. Ces éléments permettent d'examiner l'impact du contexte spatial et des interactions sociales locales sur le vote.

- **Facteurs conditions de vie** : le taux de pauvreté, la précarité de l'emploi et la composition des ménages permettent de mesurer le sentiment de privation relative et son effet potentiel sur l'adhésion à des mouvements politiques contestataires.
- **Facteurs mode de vie** : certaines variables, comme la propriété immobilière ou la structure des ménages (par exemple, vivre seul ou en famille), peuvent refléter des comportements électoraux stratégiques, en lien avec des choix rationnels visant à préserver ou à optimiser une situation personnelle.

Certaines de ces données ont été directement récupérées sur le site de l'INSEE, via des études ou des cartes interactives avec export de données (par exemple, statistiques-locales.insee.fr). Nous avons rencontré des difficultés pour obtenir des données complètes et précises, par exemple pour l'âge moyen des habitants. En effet, la donnée était séparée en tranches d'âge, ce qui ne correspondait pas exactement à notre besoin. De plus, ces tranches d'âge étaient réduites à moins de 25 ans, 25-64 ans et plus de 65 ans. Nous avons donc dû nous contenter de ces données, en espérant qu'elles soient suffisantes pour notre étude.

Nous avons aussi rencontré des problèmes pour obtenir des données sur le niveau d'études, qui n'étaient pas disponibles pour toutes les années. Pour ces paramètres, nous avons choisi des laisser ces valeurs manquantes plutôt que de les calculer avec des valeurs approximatives. Cela nous permet de conserver une certaine rigueur dans notre analyse, nous laissant toujours la possibilité de revenir sur ce choix et d'imputer les données si besoin.

Les données socio-démographiques concernant les communes d'outre-mer n'ont pas été incluses dans notre étude, car il manquait trop de données pour être comparables avec celles de la France métropolitaine. Cela aurait introduit un biais dans notre analyse, en mélangeant des données provenant de contextes socio-économiques différents.

La grande majorité de ces données étant regroupées par commune, nous avons décidé de les répartir proportionnellement entre les bureaux de vote en fonction du nombre de votants.

Nous avons aussi voulu utiliser d'autres données, ou des jeux plus précis pour notre analyse, mais ces données étaient rapportés aux départements ou région et non par commune. Cela aurait nécessité un travail de fusion et de répartition des données plus complexe, que nous n'avons pas jugé utile pour notre étude.

En revanche, la collecte des données concernant les revenus, le lieu de vie et les conditions de vie a été plus complexe. Ces données sont fournies sous forme de « carreaux » géographiques de 200 mètres de côté. Il a donc été nécessaire de récupérer les contours géographiques des bureaux de vote, de convertir et harmoniser les formats de coordonnées (CRS), puis de fusionner ces ensembles de données. La librairie R sf a été utilisée pour effectuer ces opérations, notamment en croisant les carreaux de 200 mètres avec les contours des bureaux de vote. Cette tâche s'est avérée chronophage en raison du coût élevé des calculs nécessaires pour chaque opération. Certaines données ont dû être additionnées, d'autres moyennées. Une vérification finale a montré une perte de seulement 0,01 % des données, ce qui constitue un excellent résultat pour ce type de fusion géographique.

3 Nettoyage et fusion des données

Concernant le nettoyage des données, plusieurs étapes ont été nécessaires pour garantir la qualité et la cohérence des informations. Les données électorales et socio-démographiques ont été collectées auprès de différentes sources, nécessitant une harmonisation des formats et des variables. Nous avons d'abord sélectionné les colonnes pertinentes pour notre étude, en éliminant les doublons et les variables inutiles. Un travail de normalisation a été effectué pour uniformiser les noms des colonnes et faciliter les opérations de fusion et de jointure. Les données ont été vérifiées afin d'éviter les valeurs aberrantes et garantir leur cohérence. Certaines transformations ont également été effectuées pour faciliter la visualisation et les rendre compatibles avec notre modèle d'analyse.

Le point de centralisation des données finales est le **code bureau de vote**, un identifiant unique attribué à chaque bureau de vote.

3.1 Traitement des données électorales

Pour les résultats des élections, il a d'abord été nécessaire de transformer la liste des votes par candidat dans chaque bureau de vote en catégories agrégées correspondant aux principaux courants politiques : Extrême Gauche, Gauche, Centre, Droite et Extrême Droite. Cette étape permet de simplifier l'analyse des tendances politiques tout en conservant une granularité suffisante.

Une colonne “Code” a été ajoutée en combinant les codes du département, de la commune et du bureau de vote, ce qui facilite la fusion avec d'autres jeux de données. Certaines colonnes, telles que les noms des départements ou les pourcentages inutiles (ex. : % Abstentions/Inscrits), ont été supprimées pour alléger le jeu de données. Une colonne “nonExp” a été créée pour regrouper les bulletins blancs et nuls. Cette information nous aide à mieux comprendre les niveaux de participation et de contestation.

Les données finales ont été organisées de manière à regrouper les colonnes liées à chaque courant politique, en mettant en avant les variables principales (comme les pourcentages par votant, inscrits et suffrages exprimés).

Un extrait des données nettoyées et organisées est présenté ci-dessous :

Année	Code	Libellé de la commune	Votants	nonExp	VoixEG	% Voix/VotantEG	VoixG	% Voix/VotantG
2022	010010001	L'Abergement-Clémenciat	537	17	6	1.117318	107	19.9%
2022	010020001	L'Abergement-de-Varey	175	4	5	2.857143	61	34.8%
2022	010040001	Ambérieu-en-Bugey	863	23	12	1.390498	286	33.3%

3.2 Traitement des données socio-démographiques

Concernant les facteurs socio-démographiques, certaines données étaient déjà disponibles au niveau des bureaux de vote, tandis que d'autres étaient regroupées au niveau communal. Ces dernières ont été réparties entre les bureaux de vote en suivant une répartition proportionnelle basée sur le nombre de votants, comme décrit précédemment. Les données ont ensuite été allégé et les colonnes renommées pour améliorer la lisibilité globale des données. Les colonnes inutiles ont été supprimées, pour éviter de garder des informations superflues.

Afin de préparer la visualisation de nos données sur des cartes, nous avons conservé un fichier spatial au format GeoJSON contenant les contours des bureaux de vote, identifiés par l'ID code_bureau_vote. Ce fichier, d'une taille de 600 Mo, a été isolé pour éviter de ralentir le traitement des données lors des analyses statistiques. En cas de besoin, il peut être facilement fusionné avec notre jeu de données pour générer des visualisations cartographiques. Le script est accessible en annexe.

Ces étapes ont permis de disposer d'un jeu de données homogène et prêt pour l'analyse statistique des corrélations entre les variables socio-démographiques et les résultats électoraux. Voici un échantillon des 72 colonnes finales :

code_bureau_vote	votants	non_exp	voix_eg	pourcentage_voix_votant_eg	voix_g	pourcentage_voix_votant_g
010010001	508	9	29	5.708661	125	24.6%
010020001	179	5	17	9.497207	49	27.3%
010040001	800	19	118	14.750000	227	28.3%

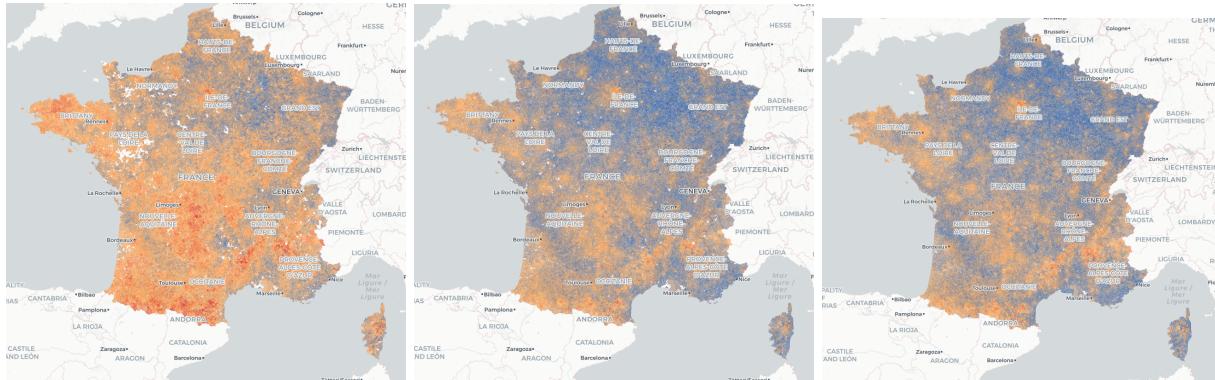
4 Visualisation par la carte

Après avoir fusionné toutes les données issues des élections présidentielles de 2012, 2017 et 2022, nous avons construit des cartes géographiques illustrant l'évolution de l'intention de vote des Français au fil du temps. Ces cartes mettent en évidence les changements significatifs dans les orientations politiques de l'électorat, avec une représentation des zones dominées par les votes à droite (teintes bleues) et à gauche (teintes rouges).

En analysant les cartes de 2012, 2017 et 2022, nous pouvons observer un basculement marqué vers la droite dans de nombreuses régions, particulièrement dans le Nord-Est de la France. Toutefois, certaines régions, comme la Bretagne, conservent une stabilité de vote à gauche.

Cette transition dans les résultats électoraux nous invite à explorer plus en détail les facteurs sous-jacents à ces changements, notamment en cherchant des corrélations entre différentes variables socio-économiques et les tendances politiques observées.

Ces dernières sont disponibles directement sur <https://helyn.cyrildeschamps.fr>, elles sont interactives et des données sont accessibles en cliquant sur la commune de votre choix.



5 Data Mining Descriptif

L'analyse des données doit être la plus précise possible. Pour cela, nous avons choisi de créer un jeu de données intermédiaire qui élimine la duplication des bureaux de vote sur plusieurs années. Nous sélectionnerons aléatoirement une seule observation par bureau de vote parmi différentes années afin d'éviter un biais de similitude. En effet, un bureau ayant historiquement voté à droite a de fortes chances de continuer à voter de la même manière, ce qui pourrait fausser l'analyse si nous incluons plusieurs observations temporelles sans ajustement.

Nous ajoutons également une variable score comprise entre 0 et 1, qui indique l'orientation politique moyenne :

- 0 correspond à l'extrême gauche,
- 0.5 représente le centre,
- 1 correspond à l'extrême droite.

```
data <- data %>%
  mutate(
    score_orientation = round(
      (voix_ed * 1 + voix_d * 0.75 + voix_c * 0.5 + voix_g * 0.25 + voix_eg * 0) /
      (voix_eg + voix_g + voix_c + voix_d + voix_ed),
      2
    )
  ) %>%
  filter(!is.na(score_orientation))

score_moyen <- mean(data$score_orientation)

# Intervalle de confiance
n <- length(data$score_orientation)
score_var <- var(data$score_orientation)
alpha <- 0.05
quant_stud <- qt(1-alpha/2, df=n-1)
borne_inf <- score_moyen - quant_stud * sqrt(score_var/n)
```

```

borne_sup <- score_moyen + quant_stud * sqrt(score_var/n)
paste(borne_inf, "<", score_moyen, "<", borne_sup)

## [1] "0.58203329077629 < 0.582397657293831 < 0.582762023811372"

```

Cela indique que l'orientation politique française actuelle se situe plutôt au centre-droit.

Notre objectif initial est d'explorer la corrélation entre le niveau de richesse et l'orientation politique. Les politiques de droite étant souvent perçues comme plus favorables aux personnes disposant de revenus élevés, il est raisonnable d'anticiper une corrélation positive entre un haut niveau de richesse et un soutien aux partis de droite.

5.1 Recherche de corrélation globale

Ensuite, nous avons cherché à analyser les corrélations entre nos données de vote et nos données socio-démographiques. Dans un premier temps, nous avons tenté de relier le nombre de votes aux autres variables. Cependant, nous nous sommes rapidement rendu compte que cette approche introduisait un biais : en utilisant le nombre de votes, nous observions systématiquement une corrélation avec les autres données. Cela s'explique par le fait que l'augmentation de la taille d'une commune entraîne mécaniquement une augmentation du nombre de ménages, d'habitants, et donc du nombre de votes.

Nous avons donc décidé de passer à l'analyse des proportions de voix afin d'éliminer ce biais lié à la taille des communes. En utilisant les proportions de votes exprimées pour chaque catégorie, nous avons pu observer les corrélations entre ces proportions et nos données socio-démographiques. Certaines variables se sont révélées particulièrement significatives, comme la proportion d'employés (part_employé) et celle d'ouvriers (part_ouvrier), qui présentaient des relations notables avec les résultats des votes.

```

##          VarVote           VarDemo   corrélation
## 176 pourcentage_voix_votant_eg    part_employes -0.8246447
## 282 pourcentage_voix_votant_g    score_orientation -0.8228443
## 285 pourcentage_voix_votant_ed   score_orientation  0.7716417
## 178 pourcentage_voix_votant_c    part_employes  0.7450642
## 181 pourcentage_voix_votant_eg    part_ouvriers  0.7430580
## 183 pourcentage_voix_votant_c    part_ouvriers -0.5946924
## 280 pourcentage_voix_votant_ed   part_bac5_plus -0.5599796
## 53  pourcentage_voix_votant_c part_non_peu_diplomes -0.5566079
## 281 pourcentage_voix_votant_eg    score_orientation -0.5517057
## 275 pourcentage_voix_votant_ed   part_bac3_4    -0.5310849

## # A tibble: 285 x 3
## # Groups:   VarDemo [57]
##      VarVote           VarDemo   corrélation
##      <fct>           <fct>     <dbl>
## 1 pourcentage_voix_votant_g population_ratio -0.365
## 2 pourcentage_voix_votant_ed population_ratio  0.287
## 3 pourcentage_voix_votant_d population_ratio  0.0790
## 4 pourcentage_voix_votant_c population_ratio -0.0749
## 5 pourcentage_voix_votant_eg population_ratio  0.0231
## 6 pourcentage_voix_votant_g nb_25            0.400
## 7 pourcentage_voix_votant_ed nb_25           -0.313
## 8 pourcentage_voix_votant_d nb_25           -0.0572
## 9 pourcentage_voix_votant_eg nb_25            0.0287
## 10 pourcentage_voix_votant_c nb_25           0.0142
## # i 275 more rows

```

Nous avons également entrepris une analyse ciblée des variables démographiques en regroupant les données par type de variable. Cela nous a permis d'examiner les corrélations spécifiques entre chaque variable

démographique et les proportions de votes. Pour ce faire, nous avons trié les corrélations au sein de chaque groupe démographique afin de mettre en évidence les relations les plus fortes et les plus significatives.

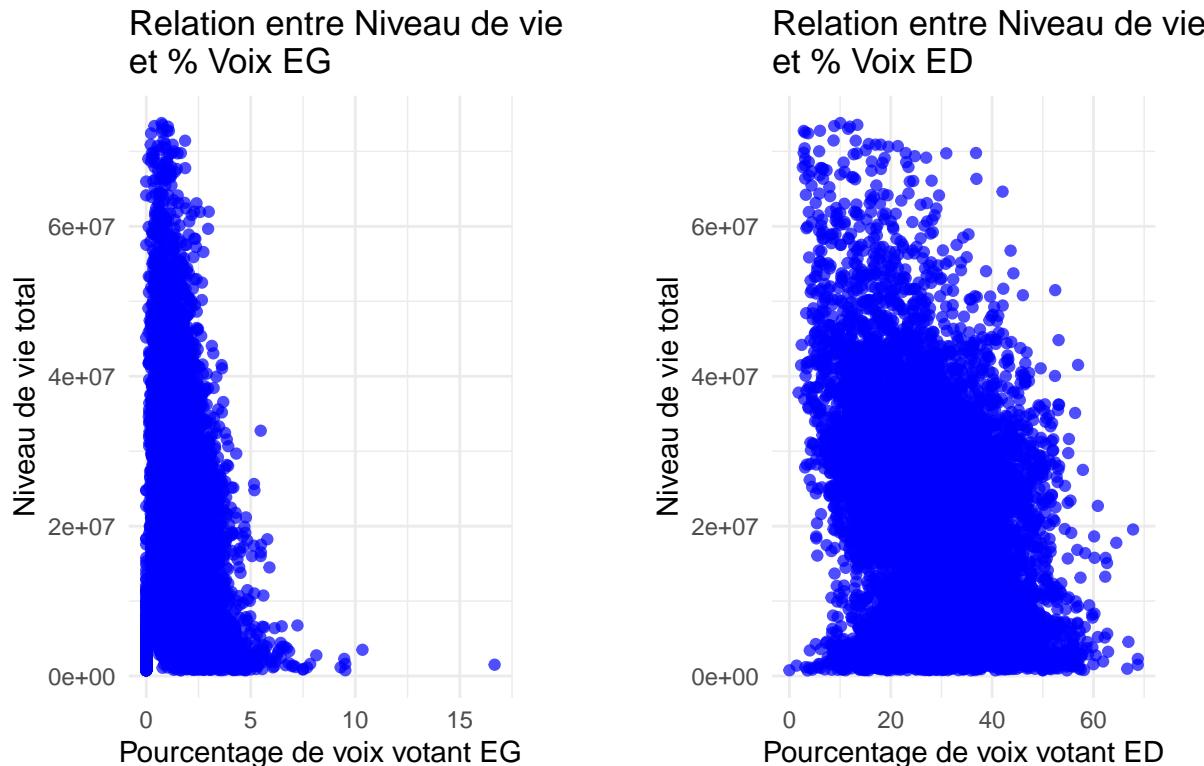
À ce stade, nous nous sommes interrogés sur la possibilité de modéliser nos données en utilisant un simple modèle de régression linéaire. L'objectif était de déterminer si une telle approche pouvait capturer les tendances observées et expliquer les variations des proportions de voix en fonction des différentes variables socio-démographiques.

5.2 Richesse

N'ayant pas de données sur le niveau de vie en 2012, nous ne sélectionnerons pas de données sur cette année. Voici le dataset que nous utiliserons :

code_bureau_vote	votants	non_exp	voix_eg	pourcentage_voix_votant_eg	voix_g	pourcentage_voix_votant_g
920230021	1048	26	12	1.1450382	367	35.019
840920005	631	9	4	0.6339144	161	25.515
544490001	640	17	2	0.3125000	118	18.437

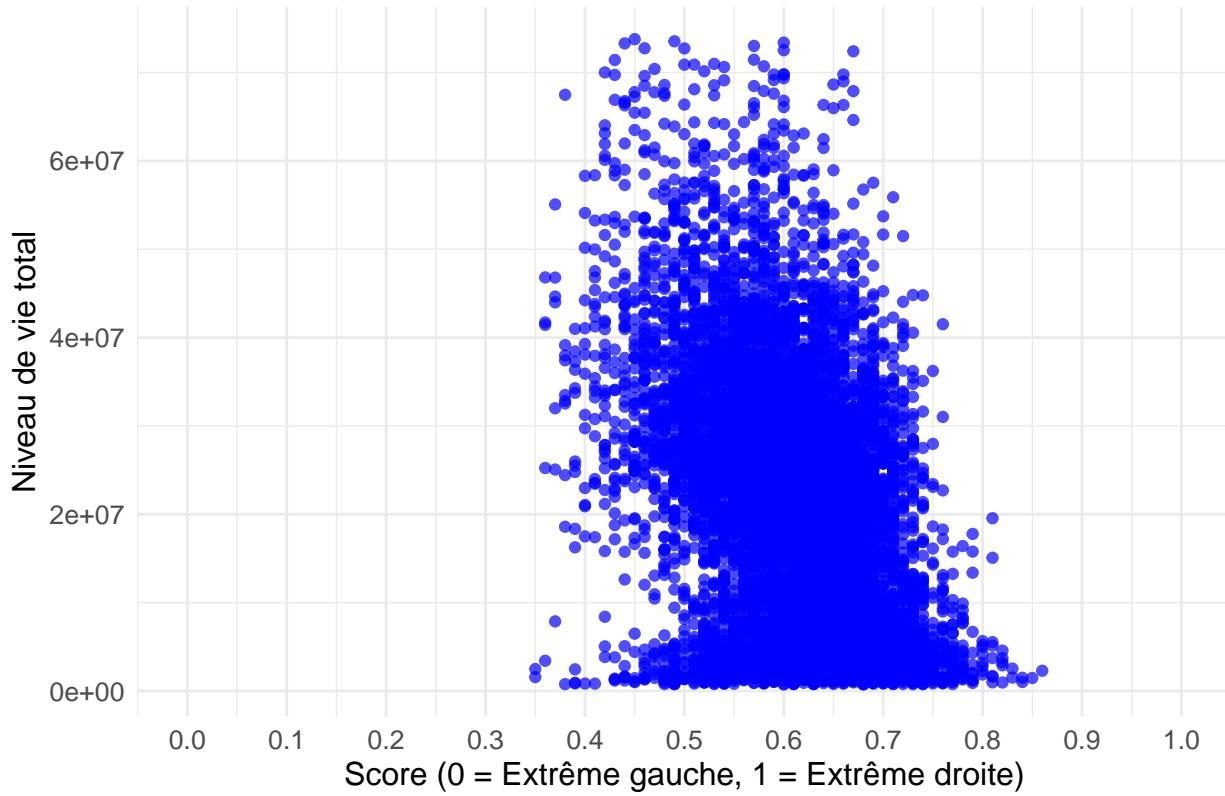
Observons rapidement la répartition vote droite/gauche en fonction de la richesse des habitants du secteur du bureau de vote :



On constate qu'effectivement le niveau de vie semble impacter le pourcentage de voix d'extrême gauche mais n'explique pas tout. Pour ce qui est de l'extrême droite, on observe une grande répartition donc le niveau de vie ne semble pas un facteur premier. Mais il y a tout de même cette tendance, ce qui laisse penser que la pauvreté pousse à voter dans les extrêmes.

Comparons maintenant avec notre score :

Relation entre le niveau de vie et l'orientation de vote



Nous cherchons une diagonale allant de l'angle inférieur gauche vers l'angle supérieur droit. Ce n'est pas le cas ici, ce qui indique que l'influence de la richesse sur les votes n'est pas significative lorsqu'elle est observée isolément. En revanche, notre théorie semble se confirmer : plus le niveau de richesse est faible, plus les votes tendent vers les extrêmes.

En conclusion, la richesse permet simplement d'expliquer qu'un niveau de vie bas est associé à une probabilité plus élevée de vote pour les extrêmes.

5.3 Khi-2

Nous avons ensuite tenté de réaliser des tests du Khi-2 afin d'examiner les relations entre nos données catégoriques et les proportions de votes. Pour cela, nous avons décidé de segmenter les données en quatre catégories, représentées par 4 colonnes et 4 lignes, pour simplifier l'analyse et réduire la complexité des relations à tester.

Malheureusement, cette approche s'est avérée encore moins concluante. Les résultats obtenus étaient chaotiques et incohérents, avec des p-valeurs élevées suggérant une absence de relation significative entre les variables étudiées.

L'un des principaux problèmes rencontrés résidait dans la distribution déséquilibrée des données au sein des catégories. Certaines cellules du tableau de contingence contenaient des effectifs trop faibles, voire nuls, ce qui a réduit considérablement la validité statistique du test du Khi-2. Cela a également généré des résultats peu fiables et difficilement généralisables.

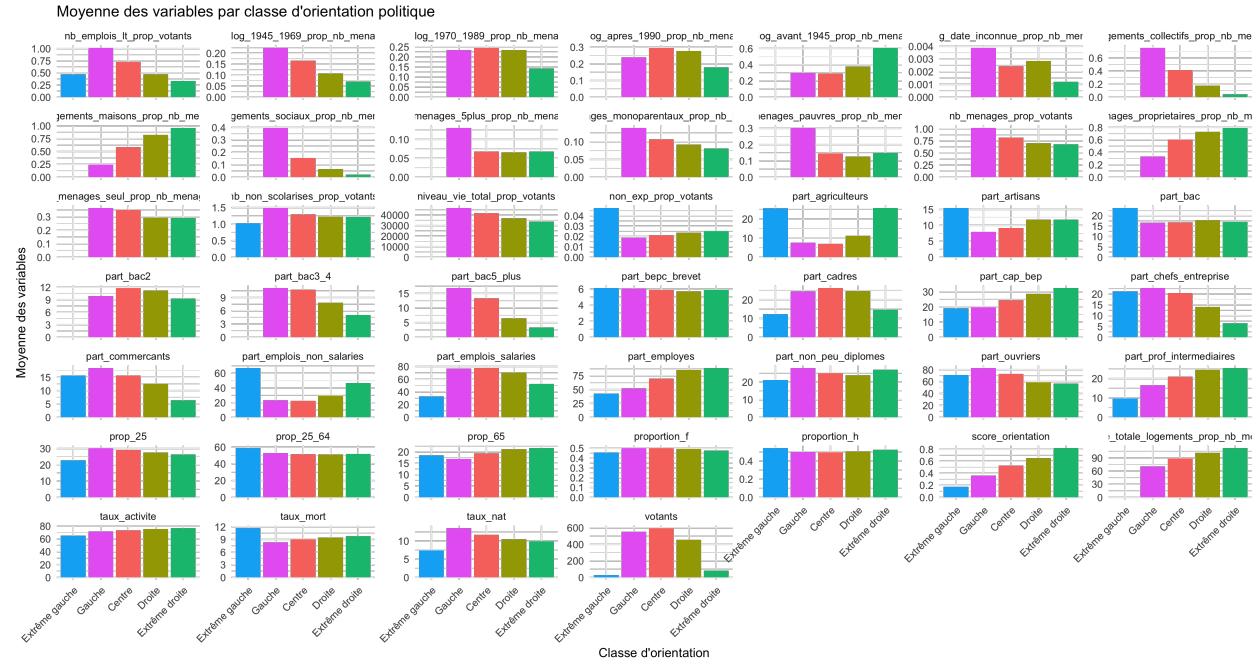
À la lumière de ces difficultés, nous avons conclu que le test du Khi-2 n'était pas adapté à nos données dans leur état actuel.

5.4 Impact du nombre d'individu

Beaucoup de variable étant des quantités, il est intéressant de plutôt utiliser des proportions. Il faut également supprimer les colonnes inutiles car déjà représenté par d'autres (ex: Taux homme et Nb d'homme).

5.5 Recherche globale par visualisation poussée

Comme évoqué précédemment, il est pertinent de visualiser notre score en fonction de nos différentes variables explicatives. Nous allons donc les tester rapidement et observer les tendances qui en ressortent.



Cette matrice de graphiques met en évidence les relations complexes entre diverses caractéristiques socio-démographiques et les choix électoraux. Un premier constat notable est que la proportion de ménages propriétaires tend à être un indicateur clé d'un vote orienté vers la droite. Cette observation s'explique probablement par une volonté de préserver un certain cadre fiscal et des politiques favorisant le patrimoine immobilier. À l'inverse, le niveau d'études apparaît comme un facteur corrélé à un vote en faveur des partis de gauche.

Les catégories socio-professionnelles montrent également des tendances intéressantes. Les ouvriers et employés semblent privilégier des partis plus populaires, souvent porteurs de mesures sociales fortes. Les cadres et professions intermédiaires affichent une propension plus marquée pour des votes modérés ou progressistes, reflétant probablement des préoccupations liées à la stabilité économique.

La structure familiale influence également les comportements électoraux. Les ménages monoparentaux apparaissent légèrement plus enclins à voter pour des partis de gauche, sans doute en raison de politiques davantage orientées vers le soutien familial. Par ailleurs, les zones où les ménages sont majoritairement composés de couples sans enfants montrent des tendances plus diverses, probablement influencées par d'autres facteurs comme le revenu et le cadre de vie.

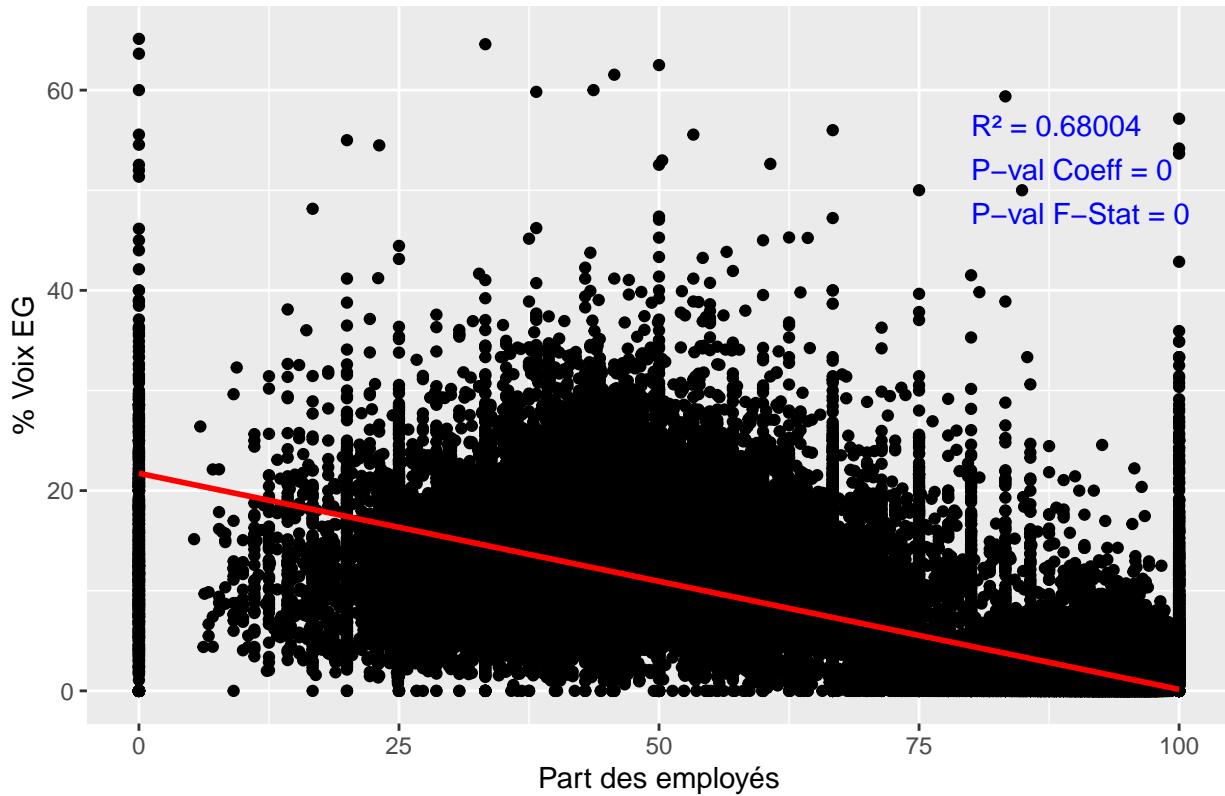
L'âge des populations étudiées révèle également de légères tendances. Les tranches d'âge les plus élevées tendent à privilégier des votes conservateurs, peut-être en raison d'une volonté de stabilité face aux changements sociétaux. À l'inverse, les jeunes adultes semblent davantage orientés vers des partis écologistes ou progressistes, reflétant une sensibilité accrue aux enjeux environnementaux et aux politiques de justice sociale.

6 Recherche de corrélations poussées

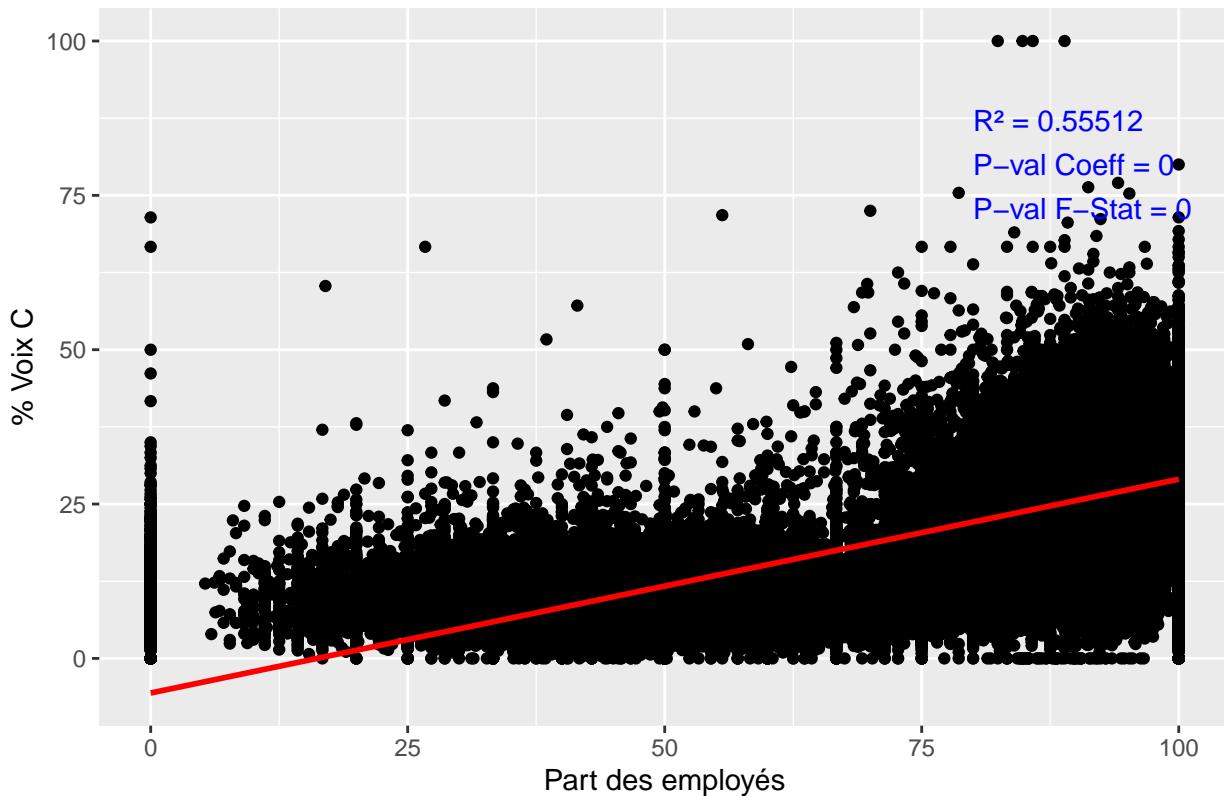
6.1 Régression linéaire

Le script est accessible en annexe.

Régression linéaire : EG ~ part_employes



Régression linéaire : C ~ part_employes



Nous avons donc entrepris de modéliser les relations entre les variables les plus prometteuses à l'aide de régressions linéaires. Dans un premier temps, nous avons effectué des analyses unitaires et observé des résultats plutôt encourageants. Par exemple, la proportion de votes d'un groupe spécifique montrait une forte corrélation avec la part d'employés (part_employé), indiquant un lien significatif entre ces deux variables (schéma ci dessus).

Nous avons ensuite étendu cette approche à d'autres données, notamment celles relatives au travail plus généralement et à l'âge. Cependant, les résultats se sont avérés bien moins concluants. Les modèles de régression linéaire ne parvenaient pas à capturer efficacement les variations des proportions de votes en fonction de ces variables. Nous avons constaté que ces relations semblaient plus complexes et non linéaires, ce qui suggérait que des modèles plus sophistiqués pourraient être nécessaires pour mieux représenter ces interactions.

En outre, l'analyse visuelle des données a révélé une limitation supplémentaire : les points sur les graphiques présentaient une distribution resserrée autour du centre, avec peu de variabilité dans les extrêmes. Cette forme compressée rendait difficile l'identification de tendances claires et limitait la capacité de la régression linéaire à fournir des prédictions robustes.

Face à ces constats, nous avons conclu que la régression linéaire simple, bien qu'utile pour une première exploration, était insuffisante pour modéliser les liens complexes entre les variables socio-démographiques et les proportions de votes. Ces résultats ont souligné la nécessité d'envisager des approches plus avancées, telles que les régressions polynomiales, les modèles non linéaires ou encore des algorithmes d'apprentissage automatique, pour capturer des relations plus subtiles et non linéaires présentes dans les données.

6.2 Régression multiple

En ce qui concerne la régression multiple, nous avons consacré un certain temps pour tester différentes combinaisons de variables afin de trouver des modèles offrant des résultats satisfaisants. Nous avons exploré plusieurs approches, notamment les sélections backward et forward, en essayant différentes configurations

et en ajustant les paramètres. Malgré nos efforts, les résultats obtenus n'ont pas été convaincants : aucun modèle ne semblait réellement capable de capturer de manière significative les liens complexes entre les variables démographiques et les proportions de votes.

6.3 Random forest

6.3.1 Entraînement

Le choix du modèle s'est orienté vers une random forest pour sa capacité à modéliser des interactions complexes entre les variables explicatives. L'implémentation ranger a été retenue en raison de ses performances optimisées, tant en termes de rapidité d'entraînement que de gestion de grandes quantités de données. L'objectif est de détecter des corrélations non linéaires qui pourraient influencer le vote de manière subtile et difficilement perceptible avec des modèles plus simples. Cette approche permet d'obtenir un bon équilibre entre le temps de calcul, la robustesse des prédictions et l'interprétabilité des résultats. Le script détaillé, incluant les paramètres d'entraînement et les étapes de prétraitement, est disponible en annexe pour une analyse complète et reproductible.

Le script est accessible en annexe.

Notre modèle affiche un R-squared de 0.6196, un RMSE de 0.0514 et un MAE de 0.0385, sur 71 013 observations et 43 variables explicatives. Cela signifie que les facteurs socio-démographiques expliquent environ 61 % de la variabilité du vote, ce qui traduit une bonne capacité à identifier des corrélations pertinentes entre les variables et le comportement électoral.

Cependant, cette performance doit être nuancée par la faible variance du score cible, ce qui limite l'amplitude des variations à expliquer. Les erreurs faibles (MAE et RMSE) indiquent une précision notable, mais environ 40 % de la variabilité reste inexpliquée, probablement en raison de facteurs non modélisés, comme des préférences individuelles ou des contextes locaux spécifiques.

6.3.2 Etude de la random forest

Nous commençons par identifier les variables significatives de notre modèle en utilisant le score d'importance basé sur l'impureté. Ce score correspond à la somme des réductions d'impureté réalisées par l'ensemble des arbres lorsqu'une variable est utilisée pour diviser un noeud. Bien que cette notion d'impureté puisse nous être peu familière, elle reste un indicateur fiable pour évaluer l'importance des variables. Nous obtenons ainsi :

```
##                                     Overall
## nb_logements_collectifs_prop_nb_menages 27.01208
## part_ouvriers                           26.95906
## part_employes                            25.93545
## nb_logements_maisons_prop_nb_menages    22.48439
## part_cap_bep                             20.89493
## surface_totale_logements_prop_nb_menages 19.49009
## prop_25                                  15.98144
## votants                                   15.40846
## nb_emplois_lt_prop_votants                14.87260
## part_emplois_non_salaries                 14.10320
## part_emplois_salaries                     13.65021
##                                         variable
## nb_logements_collectifs_prop_nb_menages nb_logements_collectifs_prop_nb_menages
## part_ouvriers                           part_ouvriers
## part_employes                            part_employes
## nb_logements_maisons_prop_nb_menages    nb_logements_maisons_prop_nb_menages
## part_cap_bep                             part_cap_bep
## surface_totale_logements_prop_nb_menages surface_totale_logements_prop_nb_menages
## prop_25                                 prop_25
```

```

## votants
## nb_emplois_lt_prop_votants
## part_emplois_non_salaries
## part_emplois_salaries

```

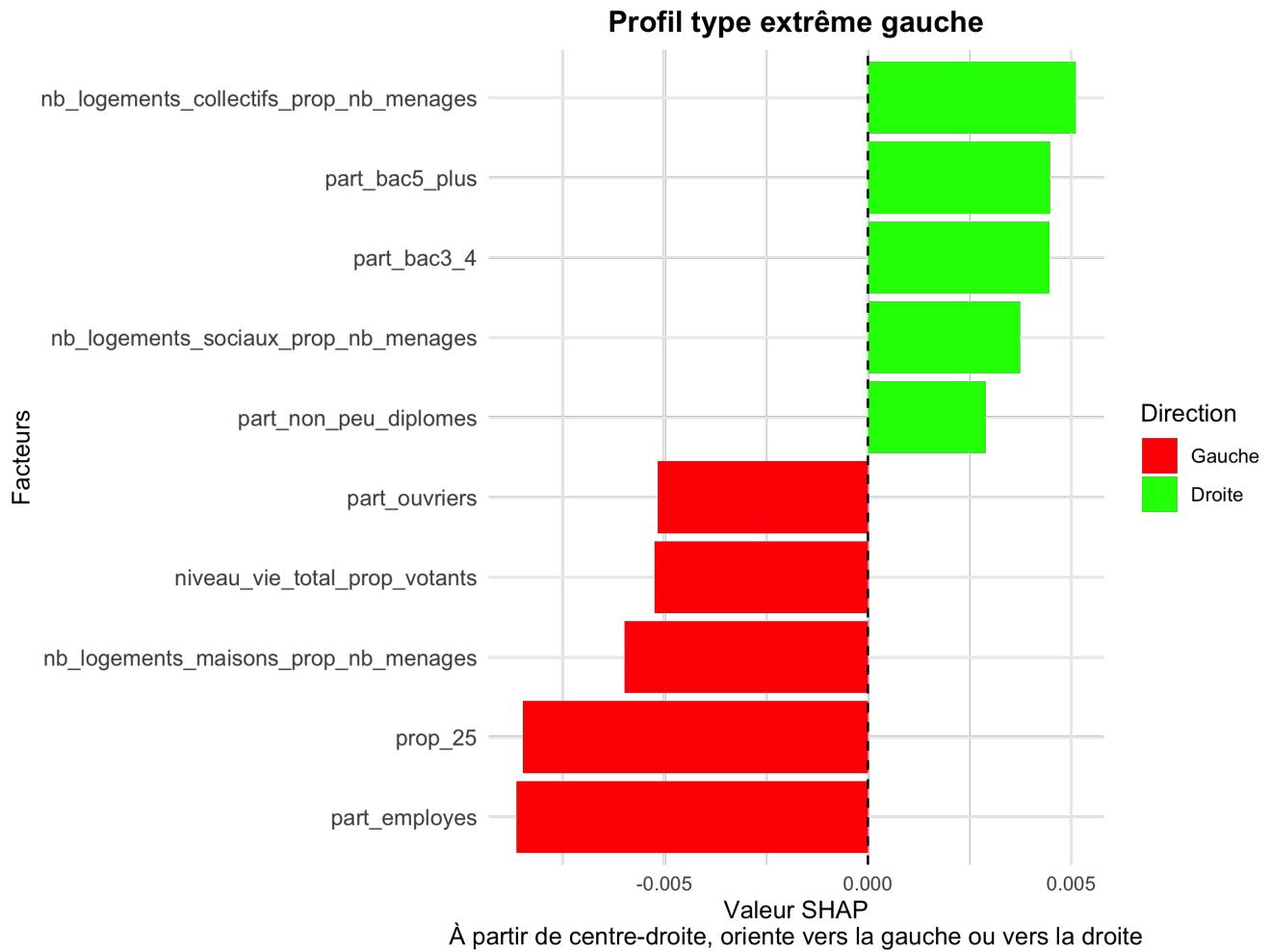
```

votants
nb_emplois_lt_prop_votants
part_emplois_non_salaries
part_emplois_salaries

```

On peut en conclure que des facteurs tels que la proportion de logements collectifs, de maisons et leur taille influencent l'orientation des votes au sein d'un bureau de vote. On observe également l'influence de facteurs liés à l'emploi, comme la proportion d'ouvriers et d'employés. Ces éléments seront décrits plus en détail lors de la synthèse.

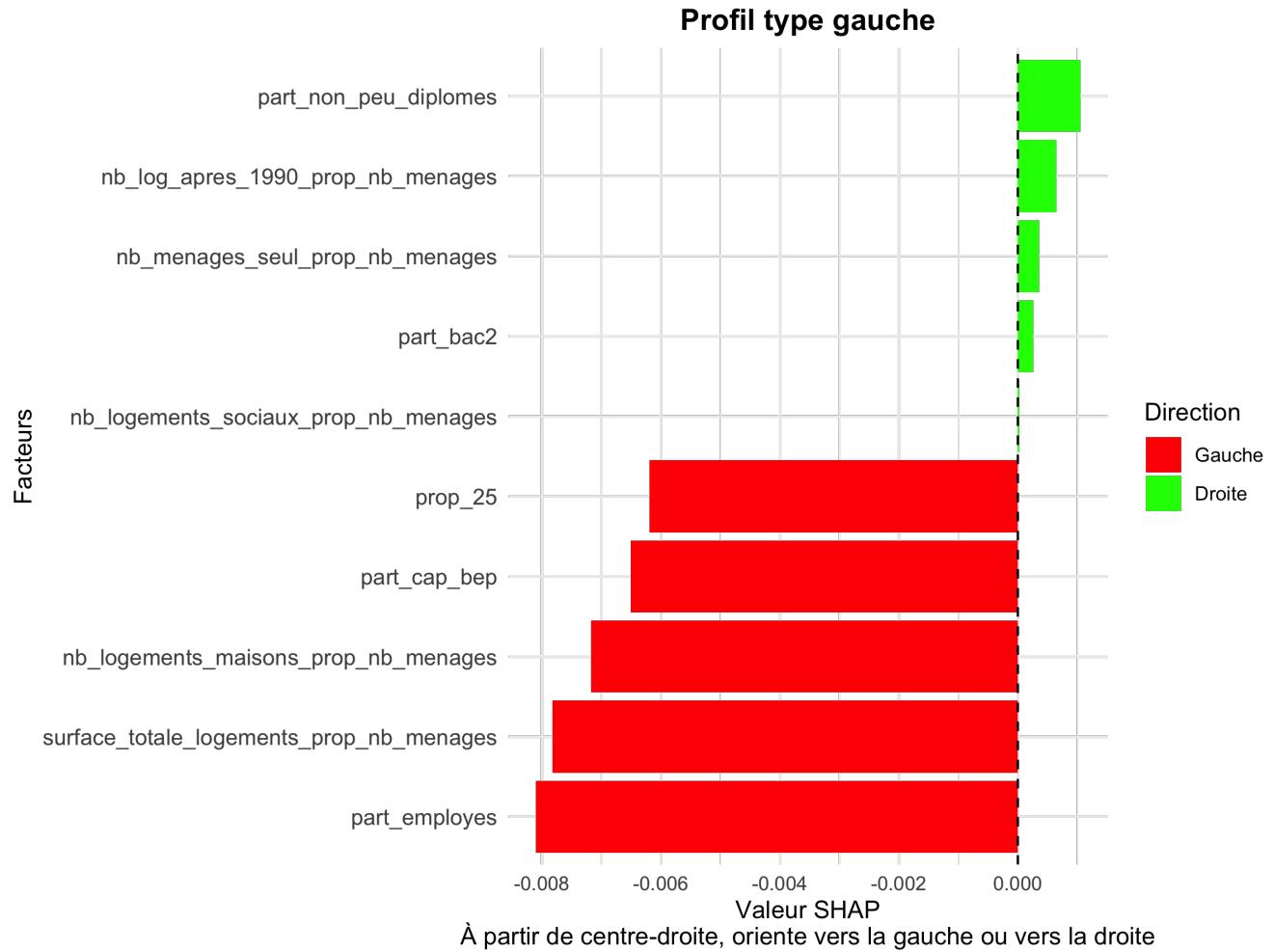
Enfin, nous pouvons analyser des profils types et identifier les variables qui influencent ces profils. Pour cela, nous effectuons des prédictions à partir de notre modèle tout en catégorisant les observations afin d'examiner, pour chaque catégorie, les principales variables influentes. Nous utilisons la méthode de Shapley et les valeurs SHAP pour déterminer ces contributions. Ces résultats doivent être interprétés avec prudence, car la random forest établit des liens entre les données, ce qui peut introduire un biais et ne pas refléter des causalités directes. Voici les diagrammes correspondants :



Le diagramme des profils types met en évidence que les individus ayant une tendance à voter pour l'extrême gauche sont généralement de jeunes personnes de moins de 25 ans, appartenant aux catégories socioprofessionnelles des employés ou des ouvriers. Fait intéressant, ce profil type inclut également des résidents de maisons individuelles avec un niveau de vie modéré à élevé.

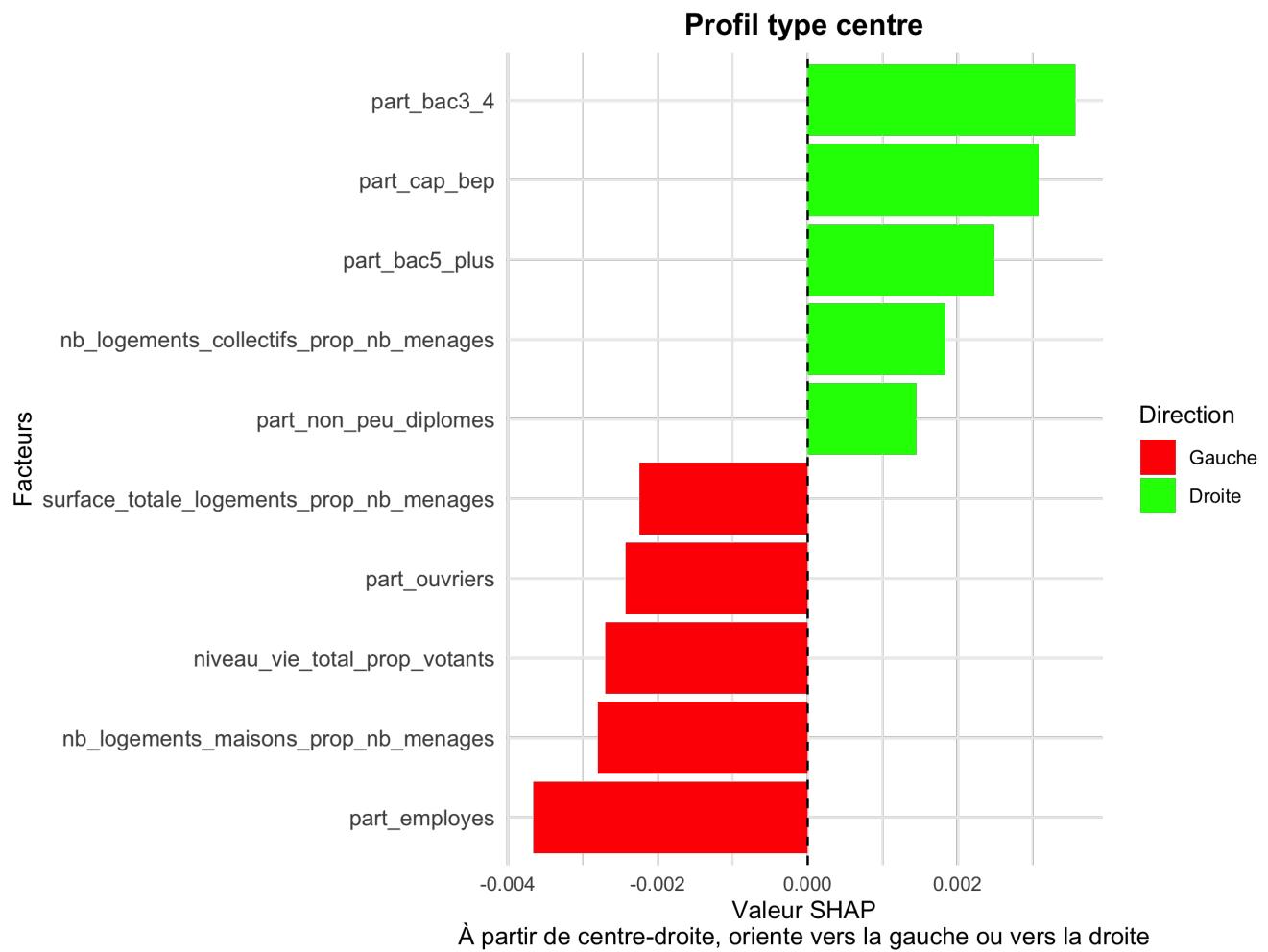
À l'inverse, les personnes les moins enclines à voter pour l'extrême gauche se distinguent par des caractéristiques telles qu'un niveau d'éducation élevé (bac+3, bac+4, bac+5 et plus) ou, à l'opposé, un faible niveau de

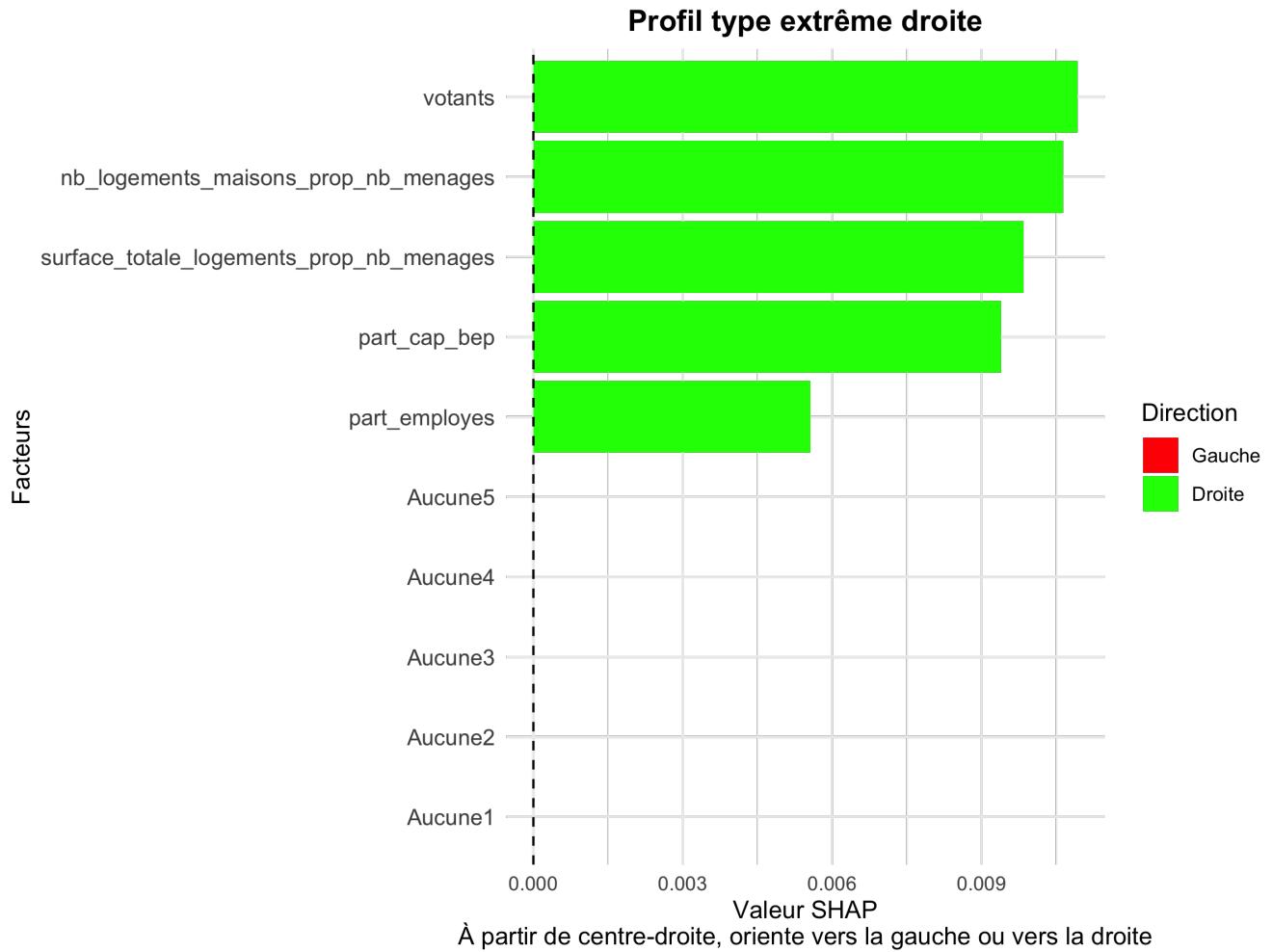
qualification. Ce profil inclut également des résidents de logements collectifs ou sociaux, reflétant des conditions de vie plus modestes ou une dynamique sociale différente.



Tout d'abord, nous n'avons pas identifié d'éléments spécifiques qui les inciteraient à ne pas voter pour ce courant politique. Cependant, des similarités avec les facteurs liés à l'extrême gauche sont apparentes, notamment le fait d'être un jeune employé de moins de 25 ans.

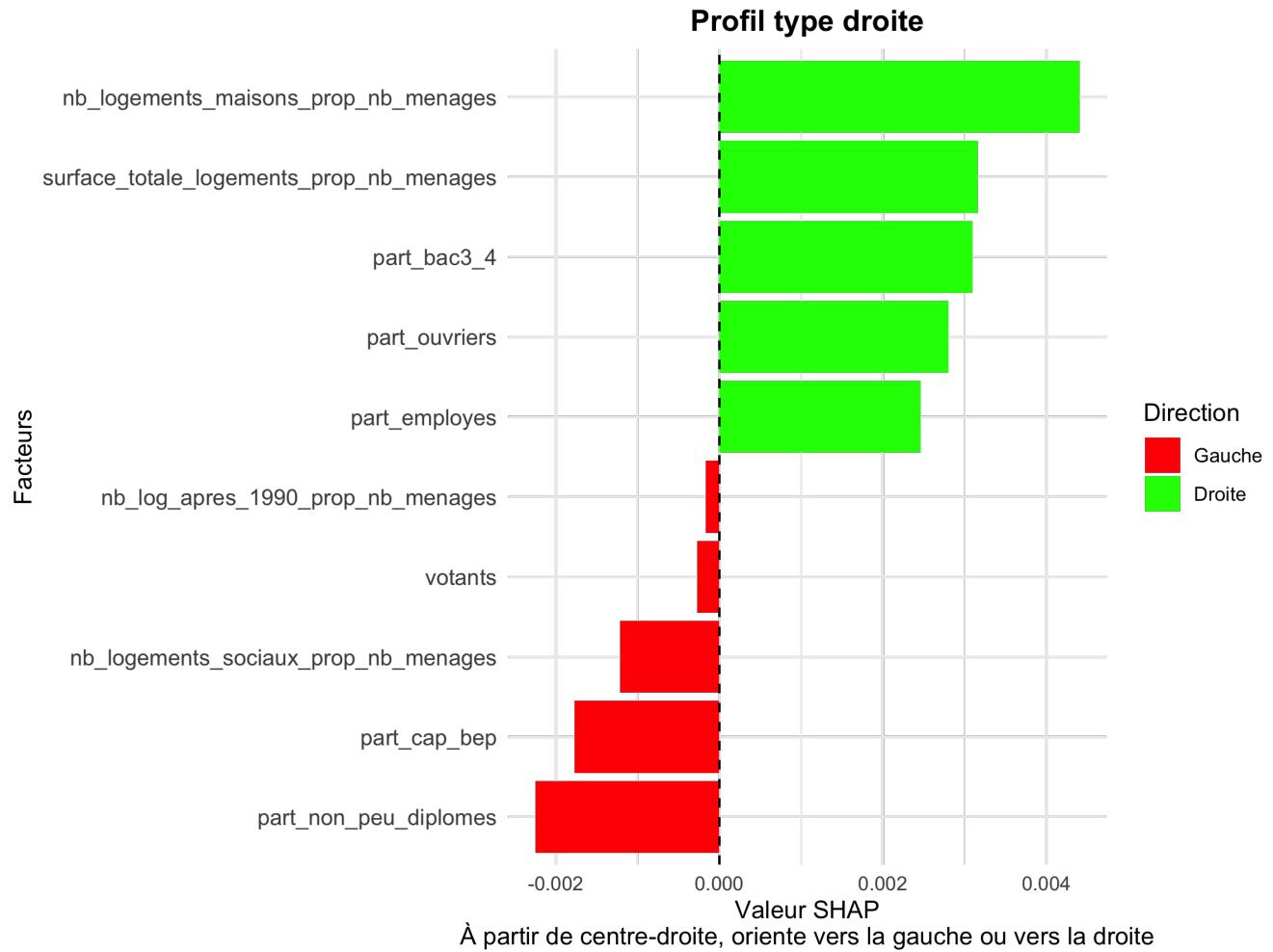
Ensuite, d'autres facteurs spécifiques favorisent un vote à gauche. Parmi ceux-ci, un niveau d'éducation correspondant à un CAP ou un BEP se démarque. On observe également une corrélation positive avec la surface totale des logements proportionnellement au nombre de ménages, ce qui peut être interprété comme un indicateur de résidence dans des petites communes ou des zones plus rurales, où les logements tendent à être plus spacieux. Enfin, vivre dans une maison apparaît également comme un facteur influençant positivement le choix de voter à gauche, reflétant un certain mode de vie propre à ces zones géographiques.





On remarque qu'il y a 5 variables majeures qui expliquent un vote à l'extrême droite. Par exemple, lorsque la proportion de logements individuels augmente dans une ville, celle-ci aura tendance à voter plus à l'extrême droite. De même pour la surface moyenne des habitations. Les personnes vivant dans des maisons individuelles valorisent souvent la stabilité, la sécurité et les traditions, ce qui peut les rendre plus réceptives aux discours conservateurs (la préservation des valeurs "traditionnelles") et nationalistes de l'extrême droite. C'est l'un des résultats qui confirme une hypothèse nous avions initialement formée.

On remarque aussi que cette intention de vote est présente chez les personnes ayant un CAP ou BEP. On l'expliquerait intuitivement par le fait que les détenteurs de CAP ou BEP occupent souvent des emplois manuels ou techniques, qui peuvent être vulnérables face à la délocalisation, l'automatisation ou la concurrence étrangère. Les partis d'extrême droite adoptant un discours protectionniste (problème de l'immigration) ce qui peut convaincre cette catégorie en promettant une sauvegarde de leurs emplois.



Sur les variables qui influent le vote à droite, on en repère deux plutôt intéressantes : la proportion d'employés et la part de Bac+3 ou Bac+4. Les employés, appartenant à des catégories modestes, peuvent être sensibles aux politiques économiques de droite, qui mettent souvent en avant des réformes favorisant l'emploi, la compétitivité des entreprises et une fiscalité plutôt modérée. La droite traditionnelle peut être perçue comme offrant un équilibre entre la valorisation du travail et la stabilité économique, sans les excès perçus de l'extrême droite ou des politiques redistributives plus marquées de la gauche. Les Bac+3 ou Bac+4 peuvent avoir une position relativement confortable et privilégier des politiques qui protègent leurs acquis économiques et sociaux, sans forcément soutenir les discours plus extrêmes. Ils souvent issus de la classe moyenne, et peuvent rechercher des politiques stables et modérées, ce que la droite traditionnelle incarne.

On retrouve aussi les variable désignant les logements individuels et leur surface, qui, comme pour l'extrême droite, peuvent très bien expliquer un vote à droite. Les autres variables sont un peu plus inattendues et nous n'avons pas trouvé de lien logique. Cela peut être un facteur qui ne s'explique pas ou être un biais

7 Synthèse

À travers cette étude, nous avons démontré que les résultats électoraux d'un bureau de vote peuvent être influencés par une multitude de variables socio-démographiques, notamment la catégorie socioprofessionnelle, le niveau d'éducation, les conditions de vie et les caractéristiques géographiques. Les corrélations observées, bien que significatives, soulignent la complexité de ces interactions, et mettent en évidence qu'aucun facteur

isolé ne permet d'expliquer pleinement les orientations politiques. Les modèles utilisés, notamment la régression multiple et la forêt aléatoire (Random Forest), ont révélé la nécessité de considérer ces facteurs comme un ensemble interconnecté.

7.1 Résultats principaux

- **Les proportions de voix exprimées pour les extrêmes (gauche et droite)** augmentent à mesure que le niveau de vie diminue, reflétant une tendance à voter pour des partis plus radicaux dans des contextes de précarité.
- **À l'inverse, des niveaux de vie élevés** favorisent un vote davantage centré.

7.2 Profils types

- **Extrême gauche :**
 - Jeunes (moins de 25 ans), employés ou ouvriers.
 - Résidant dans des maisons, avec un niveau de vie modéré à élevé.
- **Gauche :**
 - Facteurs similaires à l'extrême gauche.
 - Niveau d'éducation intermédiaire (CAP, BEP), résidents de zones rurales ou de petites communes.
- **Centre :** Résultats moins distincts, reflétant une orientation modérée et des caractéristiques socio-démographiques diversifiées.
- **Droite et extrême droite :**
 - Population plus âgée, souvent propriétaire, résidant dans des maisons individuelles.
 - Corrélation avec des niveaux d'éducation plus élevés et des surfaces de logement importantes.

7.3 Modélisation

- **Régressions linéaires simples et multiples :** Les résultats sont mitigés, notamment à cause de la distribution resserrée des données autour de moyennes. Ces modèles se révèlent insuffisants pour capter les interactions complexes.
- **Forêt aléatoire :** R^2 ajusté de 0.6196, confirmant une capacité notable à expliquer les comportements électoraux, bien que 40 % de la variabilité reste inexpliquée.

7.4 Limites et biais

- **Biais des données :** Les données collectées ne couvrent pas toutes les dimensions possibles (ex. : absence de données individuelles, limitations géographiques).
- **Complexité des interactions :** Bien que la forêt aléatoire soit performante, elle reste limitée pour identifier des causalités directes.
- **Effet contextuel :** Certaines tendances locales ou régionales ne sont pas complètement expliquées par les variables utilisées.

7.5 Conclusion

Cette étude met en lumière la complexité des comportements électoraux en France. Bien que des corrélations significatives aient été établies entre certaines variables socio-démographiques et les tendances de vote, elles ne permettent pas de prédire entièrement les résultats d'un bureau de vote. Les comportements électoraux sont influencés par des facteurs multiples, interconnectés et parfois contextuels.

7.6 Pour aller plus loin

Des modèles plus sophistiqués, intégrant des données qualitatives et des dynamiques temporelles, pourraient fournir une vision encore plus fine de ces interactions. Cela pourrait également ouvrir la voie à des applications pratiques, telles que l'analyse prédictive ou l'optimisation des stratégies électorales.

8 Annexe

8.1 Script de régression linéaire et résultat

```
# Fonction pour générer un graphique avec p-values et R-squared
generate_plot <- function(data, x_var, y_var, x_label, y_label, title) {
  # Modèle de régression
  model <- lm(as.formula(paste0(y_var, " ~ ", x_var)), data = data)
  model_summary <- summary(model)

  # Extraction des p-values, R-squared
  coef_pvalue <- signif(model_summary$coefficients[2, 4], 5) # P-value du coefficient
  fstat_pvalue <- signif(pf(
    model_summary$fstatistic[1],
    model_summary$fstatistic[2],
    model_summary$fstatistic[3],
    lower.tail = FALSE
  ), 5) # P-value de la F-statistic
  r_squared <- signif(model_summary$r.squared, 5) # R-squared

  # Génération du graphique
  ggplot(data, aes_string(x = x_var, y = y_var)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(
      title = title,
      x = x_label,
      y = y_label
    ) +
    annotate(
      "text",
      x = max(data[[x_var]], na.rm = TRUE) * 0.8,
      y = max(data[[y_var]], na.rm = TRUE) * 0.8,
      label = paste0(
        "R2 = ", r_squared,
        "\nP-val Coeff = ", coef_pvalue,
        "\nP-val F-Stat = ", fstat_pvalue
      ),
      hjust = 0,
      size = 4,
      color = "blue"
    )
}

# Liste des régressions à effectuer
regressions <- list(
```

```

list(x = "part_employes", y = "pourcentage_voix_votant_eg", x_label = "Part des employés", y_label = "Pourcentage de voix votant pour l'option Egalité")
list(x = "part_employes", y = "pourcentage_voix_votant_c", x_label = "Part des employés", y_label = "Pourcentage de voix votant pour l'option Citoyenneté")

# Générer et afficher les graphiques
plots <- lapply(regressions, function(reg) {
  generate_plot(
    data = data,
    x_var = reg$x,
    y_var = reg$y,
    x_label = reg$x_label,
    y_label = reg$y_label,
    title = reg$title
  )
})

# Afficher les graphiques
for (p in plots) print(p)

```

8.2 Script d'Entraînement random forest (via ranger, sur plusieurs coeurs)

```

# Préparation des données
df_model <- clean_data %>%
  group_by(code_bureau_vote) %>%
  slice_sample(n = 1) %>%
  ungroup() %>%
  select(where(is.numeric)) %>%
  select(-contains("voix")) %>%
  filter(!is.na(score_orientation))

write.csv(df_model, file = "big_score_training_data.csv", row.names = FALSE)

nzv <- nearZeroVar(df_model, saveMetrics = TRUE)
df_model <- df_model[, !nzv$nzv]

preProcValues <- preProcess(df_model, method = "medianImpute")
df_model <- predict(preProcValues, df_model)

num_cores <- parallel::detectCores() - 1
cl <- makeCluster(num_cores)
registerDoParallel(cl)

train_ctrl <- trainControl(
  method      = "repeatedcv",
  number      = 5,
  repeats     = 3,
  verboseIter = FALSE,
  allowParallel = TRUE
)

tune_grid <- expand.grid(
  mtry       = c(2, 4, 6, 8),
  splitrule  = c("variance", "extratrees"),

```

```

min.node.size = c(5, 10, 15)
)

df_model <- as.data.frame(df_model)
X <- df_model[, setdiff(names(df_model), "score_orientation")]
Y <- df_model$score_orientation

set.seed(123)
rf_model <- train(
  x           = X,
  y           = Y,
  method      = "ranger",
  trControl   = train_ctrl,
  tuneGrid    = tune_grid,
  num.trees   = 500,
  importance  = "impurity"
)

stopCluster(cl)
saveRDS(rf_model, file = "big_score_model.rds")

```

8.3 Script de merge spatiale (simplifié)

```

# Rendre les géométries valides avant l'intersection
bureau_vote_data <- st_make_valid(bureau_vote_data)
revenu_data <- st_make_valid(revenu_data)

# Intersection géographique entre les carreaux de 200m et les bureaux de vote
intersections <- st_intersection(revenu_data, bureau_vote_data)

# Calcul des aires et des proportions, run plusieurs heures (préparation MERGE)
intersections$area_intersection <- st_area(intersections)
intersections$area_carre <- st_area(revenu_data[match(intersections$id, revenu_data$id), ])
intersections$proportion <- as.numeric(intersections$area_intersection / intersections$area_carre)

# Calcul des indicateurs pondérés par les proportions d'intersection
resultats_final <- intersections[, .(
  nb_menages = round(sum(men * proportion, na.rm = TRUE), 3),
  nb_menages_pauvres = round(sum(men_pauv * proportion, na.rm = TRUE), 3),
  nb_logements_maisons = round(sum(men_mais * proportion, na.rm = TRUE), 3),
  niveau_vie_total = round(sum(ind_snv * proportion, na.rm = TRUE), 3)
), by = codebureauvote]

# Ajout de la géométrie correspondante au bureau de vote
resultats_final$geom <- bureau_vote_data$geom[match(resultats_final$codebureauvote, bureau_vote_data$co
resultats_final <- st_as_sf(resultats_final, crs = 2154)

```

8.4 Code initialisation carte React avec MapLibre

```

// Initialisation de la carte
const map = new maplibregl.Map({
  container: mapContainerRef.current,

```

```

style: {
  version: 8,
  sources: {
    "osm-tiles": {
      type: "raster",
      tiles: [
        "https://a.basemaps.cartocdn.com/light_nolabels/{z}/{x}/{y}.png",
        "https://b.basemaps.cartocdn.com/light_nolabels/{z}/{x}/{y}.png",
        "https://c.basemaps.cartocdn.com/light_nolabels/{z}/{x}/{y}.png",
      ],
      tileSize: 256,
    },
    // On ajuste ici la source de la couche vectorielle en fonction de l'année
    political_map: {
      type: "vector",
      tiles: [
        `${tilesHostBaseURL}/data/${politicalMapSourceName}/{z}/{x}/{y}.pbf`,
      ],
    },
    "osm-tiles-labels": {
      type: "raster",
      tiles: [
        "https://a.basemaps.cartocdn.com/light_only_labels/{z}/{x}/{y}.png",
        "https://b.basemaps.cartocdn.com/light_only_labels/{z}/{x}/{y}.png",
        "https://c.basemaps.cartocdn.com/light_only_labels/{z}/{x}/{y}.png",
      ],
      tileSize: 256,
    },
  },
  layers: [
    {
      id: "osm-background",
      type: "raster",
      source: "osm-tiles",
    },
    {
      id: "political_map",
      type: "line",
      source: "political_map",
      "source-layer": "political_map",
      paint: {
        "line-color": "#e3cfad",
        "line-width": [
          "interpolate",
          ["exponential", 1.5],
          ["zoom"],
          0,
          0.1,
          10,
          3,
        ],
      },
    },
  ],
}

```

```
{
  id: "political_map_fill",
  type: "fill",
  source: "political_map",
  "source-layer": "political_map",
  paint: {
    "fill-color": [
      "interpolate",
      ["linear"],
      ["get", "score_orientation"],
      0,
      "#313695",
      0.25,
      "#4575b4",
      0.5,
      "#fd6e61",
      0.75,
      "#d73027",
      1,
      "#a50026",
    ],
    "fill-opacity": 0.7,
    "fill-outline-color": [
      "interpolate",
      ["linear"],
      ["get", "score_orientation"],
      0,
      "#313695B3",
      0.25,
      "#4575b4B3",
      0.5,
      "#fd6e61B3",
      0.75,
      "#d73027B3",
      1,
      "#a50026B3",
    ],
    "fill-antialias": true,
  },
},
{
  id: "osm-labels",
  type: "raster",
  source: "osm-tiles-labels",
},
],
},
},
center: [2.2137, 46.2276],
zoom: 5,
maxZoom: 13,
minZoom: 5,
});
```