Gentrification et Vote électoral

February 8, 2026

1 Rapport d'Analyse : Déterminants du Vote et Gentrification

Auteurs : Isaline JOUVE, Romain RATAJCZYK, Vincent VASYLCHENKO

1.1 1. Contexte et Objectifs

Les élections municipales de 2020 ont été marquées par une poussée écologiste dans les grandes villes et une abstention record. Ce projet vise à comprendre si ce vote est corrélé à des dynamiques de **gentrification** (évolution des revenus, des cadres, de l'éducation).

Nous croisons ici les données électorales (Ministère de l'Intérieur) et socio-économiques (INSEE) pour répondre à la problématique : > *Les dynamiques socio-démographiques locales (gentrification) permettent-elles de prédire l'évolution du vote de gauche et écologiste ?*

Dans toute cette étude, nous nous intéressons aux résultats obtenus par le "bloc de gauche" aux élections municipales de 2014 et 2020. Nous considérons ici que ce bloc correspond à au groupe Europe Ecologie-Les Verts (LVEC), au groupe Ecologiste (LECO), au groupe Union de la Gauche (LUG), au groupe Divers Gauche (LDVG), au groupe (LSOC) et au groupe Liste du Parti de Gauche (LPG).

1.2 Exécution du code

Installation automatique des dépendances si manquantes

[1]:
```
# Installation automatique des dépendances si manquantes
! pip install -r "required_libraries.txt"
```

```
Requirement already satisfied: pandas in /opt/python/lib/python3.13/site-
packages (from -r required_libraries.txt (line 1)) (2.3.3)
Requirement already satisfied: numpy in /opt/python/lib/python3.13/site-packages
(from -r required_libraries.txt (line 2)) (2.4.0)
Requirement already satisfied: matplotlib in /opt/python/lib/python3.13/site-
packages (from -r required_libraries.txt (line 3)) (3.10.8)
Requirement already satisfied: scikit-learn in /opt/python/lib/python3.13/site-
packages (from -r required_libraries.txt (line 4)) (1.8.0)
Requirement already satisfied: statsmodels in /opt/python/lib/python3.13/site-
packages (from -r required_libraries.txt (line 5)) (0.14.6)
Requirement already satisfied: geopandas in /opt/python/lib/python3.13/site-
packages (from -r required_libraries.txt (line 6)) (1.1.2)
Requirement already satisfied: shapely in /opt/python/lib/python3.13/site-
packages (from -r required_libraries.txt (line 7)) (2.1.2)
```

```
Requirement already satisfied: openpyxl in /opt/python/lib/python3.13/site-
packages (from -r required_libraries.txt (line 8)) (3.1.5)
Requirement already satisfied: xlrd in /opt/python/lib/python3.13/site-packages
(from -r required_libraries.txt (line 9)) (2.0.2)
Requirement already satisfied: mapclassify in /opt/python/lib/python3.13/site-
packages (from -r required_libraries.txt (line 10)) (2.10.0)
Requirement already satisfied: python-dateutil>=2.8.2 in
/opt/python/lib/python3.13/site-packages (from pandas->-r required_libraries.txt
(line 1)) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /opt/python/lib/python3.13/site-
packages (from pandas->-r required_libraries.txt (line 1)) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in
/opt/python/lib/python3.13/site-packages (from pandas->-r required_libraries.txt
(line 1)) (2025.3)
Requirement already satisfied: contourpy>=1.0.1 in
/opt/python/lib/python3.13/site-packages (from matplotlib->-r
required_libraries.txt (line 3)) (1.3.3)
Requirement already satisfied: cycler>=0.10 in /opt/python/lib/python3.13/site-
packages (from matplotlib->-r required_libraries.txt (line 3)) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/opt/python/lib/python3.13/site-packages (from matplotlib->-r
required_libraries.txt (line 3)) (4.61.1)
Requirement already satisfied: kiwisolver>=1.3.1 in
/opt/python/lib/python3.13/site-packages (from matplotlib->-r
required_libraries.txt (line 3)) (1.4.9)
Requirement already satisfied: packaging>=20.0 in
/opt/python/lib/python3.13/site-packages (from matplotlib->-r
required_libraries.txt (line 3)) (25.0)
Requirement already satisfied: pillow>=8 in /opt/python/lib/python3.13/site-
packages (from matplotlib->-r required_libraries.txt (line 3)) (12.0.0)
Requirement already satisfied: pyparsing>=3 in /opt/python/lib/python3.13/site-
packages (from matplotlib->-r required_libraries.txt (line 3)) (3.3.1)
Requirement already satisfied: scipy>=1.10.0 in /opt/python/lib/python3.13/site-
packages (from scikit-learn->-r required_libraries.txt (line 4)) (1.16.3)
Requirement already satisfied: joblib>=1.3.0 in /opt/python/lib/python3.13/site-
packages (from scikit-learn->-r required_libraries.txt (line 4)) (1.5.3)
Requirement already satisfied: threadpoolctl>=3.2.0 in
/opt/python/lib/python3.13/site-packages (from scikit-learn->-r
required_libraries.txt (line 4)) (3.6.0)
Requirement already satisfied: patsy>=0.5.6 in /opt/python/lib/python3.13/site-
packages (from statsmodels->-r required_libraries.txt (line 5)) (1.0.2)
Requirement already satisfied: pyogrio>=0.7.2 in
/opt/python/lib/python3.13/site-packages (from geopandas->-r
required_libraries.txt (line 6)) (0.12.1)
Requirement already satisfied: pyproj>=3.5.0 in /opt/python/lib/python3.13/site-
packages (from geopandas->-r required_libraries.txt (line 6)) (3.7.2)
Requirement already satisfied: et-xmlfile in /opt/python/lib/python3.13/site-
packages (from openpyxl->-r required_libraries.txt (line 8)) (2.0.0)
```

```
Requirement already satisfied: networkx>=3.2 in /opt/python/lib/python3.13/site-
packages (from mapclassify->-r required_libraries.txt (line 10)) (3.6.1)
Requirement already satisfied: certifi in /opt/python/lib/python3.13/site-
packages (from pyogrio>=0.7.2->geopandas->-r required_libraries.txt (line 6))
(2025.11.12)
Requirement already satisfied: six>=1.5 in /opt/python/lib/python3.13/site-
packages (from python-dateutil>=2.8.2->pandas->-r required_libraries.txt (line
1)) (1.17.0)
```

[ ]:
```
# Étape 1 : Préparation des données et statistiques descriptives
%run "./notebooks/data_preparation.ipynb"
```

```
/opt/python/lib/python3.13/site-packages/nbformat/__init__.py:96:
MissingIDFieldWarning: Cell is missing an id field, this will become a hard
error in future nbformat versions. You may want to use `normalize()` on your
notebooks before validations (available since nbformat 5.1.4). Previous versions
of nbformat are fixing this issue transparently, and will stop doing so in the
future.
  validate(nb)
/opt/python/lib/python3.13/site-packages/nbformat/__init__.py:96:
DuplicateCellId: Non-unique cell id '5d2f7dc3' detected. Corrected to
'a305122b'.
  validate(nb)
/opt/python/lib/python3.13/site-packages/nbformat/__init__.py:96:
DuplicateCellId: Non-unique cell id '6ddd9105' detected. Corrected to
'3594556a'.
  validate(nb)
/opt/python/lib/python3.13/site-packages/nbformat/__init__.py:96:
DuplicateCellId: Non-unique cell id 'd3a2c9f7' detected. Corrected to
'7d4a478d'.
  validate(nb)
/opt/python/lib/python3.13/site-packages/nbformat/__init__.py:96:
DuplicateCellId: Non-unique cell id '40fd2b19' detected. Corrected to
'63775aa6'.
  validate(nb)
```

1.3 Objectif

Ce code effectue le premier traitement des bases de données qui seront utilisées pour ce projet. Il agrège dans un premier temps les bases de données produites par l'INSEE : * Revenus et pauvreté des ménages * Revenus localisés sociaux et fiscaux * Diplômes-Formation

et dans un deuxième temps ajoute les résultats des élections présidentielles et communales de 2014, 2017, 2020 et 2022 qui ont été mis en forme dans le code "Data formatting".

Le code totalement reproductible est particulièrement lourd à exécuter. Aussi, le script vérifie si les données sont déjà chargées en mémoire (session active) ou si le fichier final est déjà présent (disponible sur le dépôt Git). Si ces conditions sont remplies, l'étape de préparation est ignorée pour fluidifier la lecture du notebook. Mais ce code le génère si besoin (pour vérifier la réplicabilité

par exemple si vous le souhaitez !)

Chargement/Rechargement des données brutes...

```
- Chargement Population…
```

```
/tmp/ipykernel_9431/86269345.py:56: DtypeWarning: Columns (0,3) have mixed
types. Specify dtype option on import or set low_memory=False.
  df_pop_2020 = pd.read_csv(os.path.join(RAW_INSEE_DIR,
files_insee['pop_2020']), sep=';', dtype={'COM': str})
```

```
- Chargement Diplômes…
```

```
/tmp/ipykernel_9431/86269345.py:63: DtypeWarning: Columns (0,3) have mixed
types. Specify dtype option on import or set low_memory=False.
  df_diplo_2020 = pd.read_csv(os.path.join(RAW_INSEE_DIR,
files_insee['diplo_2020']), sep=';', header=0, dtype={'COM': str})
```

```
- Chargement Revenus…
```

[ ]:
```
# Étape 2 : Géoreprésentation des données des élections municipales
%run "./notebooks/Geo representation.ipynb"
```

[ ]:
```
# Étape 3 : Régressions et Statistiques
%run "./notebooks/regression_statsmodel.ipynb"
```

[1]:
```
# Étape 4 : Exploration Machine Learning
%run "./notebooks/ML_exploration.ipynb"
```

```
/opt/anaconda3/envs/datasci/lib/python3.11/site-
packages/nbformat/__init__.py:96: MissingIDFieldWarning: Cell is missing an id
field, this will become a hard error in future nbformat versions. You may want
to use `normalize()` on your notebooks before validations (available since
nbformat 5.1.4). Previous versions of nbformat are fixing this issue
transparently, and will stop doing so in the future.
  validate(nb)
```

2 Exploration Machine Learning et ACP

Cette section complète l'approche économétrique classique par des méthodes de Machine Learning et d'analyse multivariée, afin d'exploiter leurs complémentarités interprétatives et prédictives.

La **régression linéaire (OLS)** est mobilisée principalement pour l'analyse du sens des relations entre variables explicatives et vote, à travers le signe et l'ordre de grandeur des coefficients estimés. Elle fournit un cadre interprétable permettant d'identifier les effets marginaux moyens.

Le **Random Forest**, fondé sur l'agrégation d'arbres de décision entraînés sur des sous-échantillons aléatoires, est utilisé pour établir une hiérarchie des déterminants du vote, à partir de leur contribution à la réduction de l'erreur de variance. Il permet de capter des non-linéarités, interactions et effets de seuil.

La comparaison des coefficients de détermination (R^2) constitue un élément central de l'analyse. Un R^2 plus élevé pour le Random Forest par rapport à l'OLS indique que la relation entre caractéristiques socio-économiques et comportement électoral ne se limite pas à des effets linéaires, mais repose sur des structures non linéaires indispensables à une prédiction correcte du vote.

Enfin, l'**Analyse en Composantes Principales (ACP)** est utilisée en complément afin de visualiser l'espace sociologique des communes et de synthétiser la structure des corrélations entre les variables explicatives.

2.0.1 Application du Random Forest

```
R² Random Forest (entraînement): 0.8075

Importance des variables :
score_gauche_pres_2017    0.856400
pct_sup_2022              0.081408
log_pop_2022              0.042223
log_med_19                0.019969
dtype: float64
R² moyen (prédiction) : 0.7303 (+/- 0.0272)
```

2.0.2 Apports et limites du Random Forest

La standardisation des variables ne constitue pas un enjeu pour le Random Forest, celui-ci étant invariant au changement d'échelle. Le réglage conjoint de la profondeur des arbres et du nombre d'estimateurs permet d'atteindre un compromis satisfaisant biais-variance, garantissant la stabilité des résultats.

Le gain d'environ 7 points de R^2 par rapport à la régression linéaire indique que le Random Forest parvient à capter des **effets non linéaires et des effets de seuil** absents du cadre OLS. Par exemple, l'impact du niveau de diplôme sur le vote apparaît conditionné à des seuils démographiques ou contextuels, effets que l'OLS tend à lisser en les ramenant à un effet moyen.

Toutefois, malgré cette amélioration explicative, le Random Forest n'apporte pas de gain prédictif décisif par rapport à l'OLS, les niveaux de R^2 sur la prédiction restant proches (R^2 moyen calculé avec cross-validation sur 5 blocs, donc 20% des données cachées pour tester le modèle). Ce constat suggère que le comportement électoral étudié demeure largement structuré par une inertie linéaire dominante, rendant toute complexification excessive du modèle peu justifiée voire contre-productive (rasoir d'Ockham).

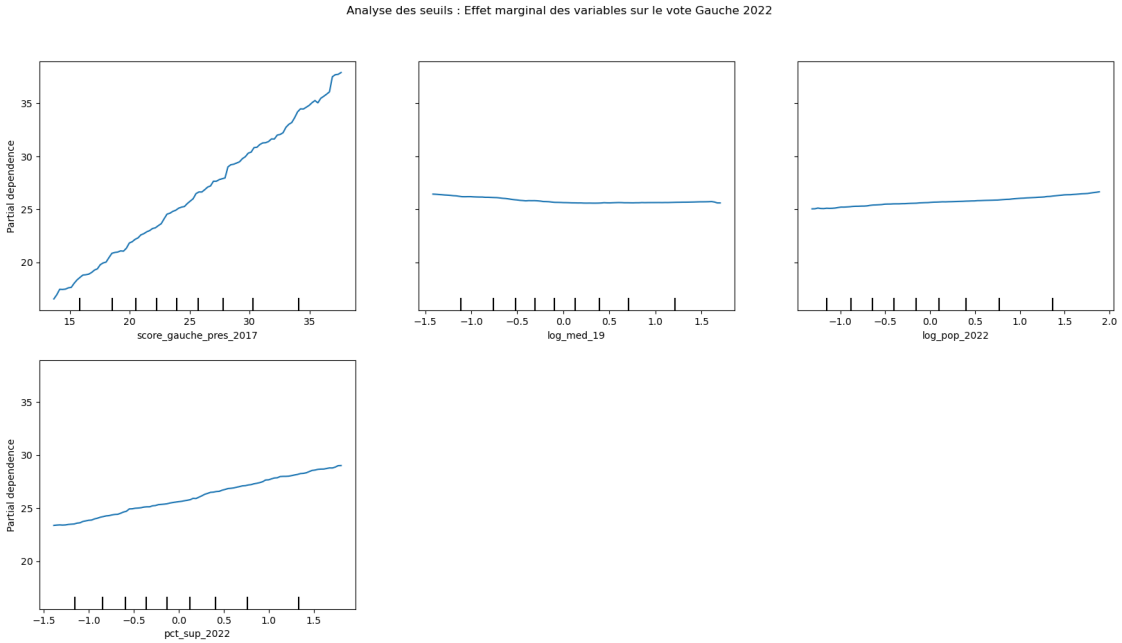
L'analyse de l'importance des variables révèle enfin une domination écrasante du **score électoral passé (score_2017)**, qui concentre environ 85,6 % de la contribution explicative, confirmant que le vote observé relève avant tout d'une **reproduction géographique et historique**. Parmi les variables socio-économiques, le **niveau de diplôme** apparaît comme un déterminant plus puissant du vote de gauche que le revenu, avec une contribution relative nettement supérieure (8,1 % contre 1,9 %).

2.1 Analyse des effets non linéaires : Partial Dependence Plots (PDP)

Pour mettre en évidence les **effets non linéaires et les seuils** identifiés par le Random Forest, nous utilisons les *Partial Dependence Plots (PDP)*.

Ces graphiques permettent de visualiser l'effet marginal moyen d'une variable explicative sur la prédiction du modèle, en isolant son impact des autres variables. Cela est particulièrement utile pour détecter : - des seuils, par exemple lorsque l'effet du niveau de diplôme sur le vote n'apparaît qu'à partir d'une certaine densité de population, - des zones de plateau ou de saturation, - des interactions implicites, où la pente change selon la valeur d'une autre variable.

L'interprétation des PDP complète ainsi l'importance des variables, en montrant non seulement quelles variables comptent, mais comment elles influencent la prédiction à différents niveaux.



2.1.1 Analyse des Graphiques de Dépendance Partielle (PDP)

Les *Partial Dependence Plots* permettent d'interpréter le modèle du Random Forest en isolant l'effet marginal de chaque variable. Ils révèlent des dynamiques que la régression linéaire lisse ou ignore.

- Inertie du vote (Score 2017) : La relation est fortement linéaire et positive. Cela confirme une dépendance au sentier : l'historique électoral reste le déterminant principal, témoignant de la stabilité de l'ancrage territorial des partis. C'est la confirmation visuelle du R^2 élevé, le vote de 2022 est une réplication presque linéaire de celui de 2017.

Effet de taille (Log Population) : Contrairement à l'intuition d'un "gradient urbain", la courbe est quasi-plate (légèrement croissante). Une fois le niveau de diplôme contrôlé, la taille de la commune n'a pas d'effet intrinsèque sur le vote (effet de composition). On peut interpréter sa légère croissance comme un "petit bonus métropolitain". Mais si les grandes villes votes à gauche, c'est avant tout parce qu'elles sont peuplées de diplômés, pas parce qu'elles sont grandes.

Niveau de vie (Revenu médian) : La relation est quasi-neutre (plate), signe que le revenu n'apporte plus d'information prédictive supplémentaire une fois le diplôme et le vote passé pris en compte. C'est une dynamique que l'OLS ne pouvait détecter, suggérant un effet de saturation.

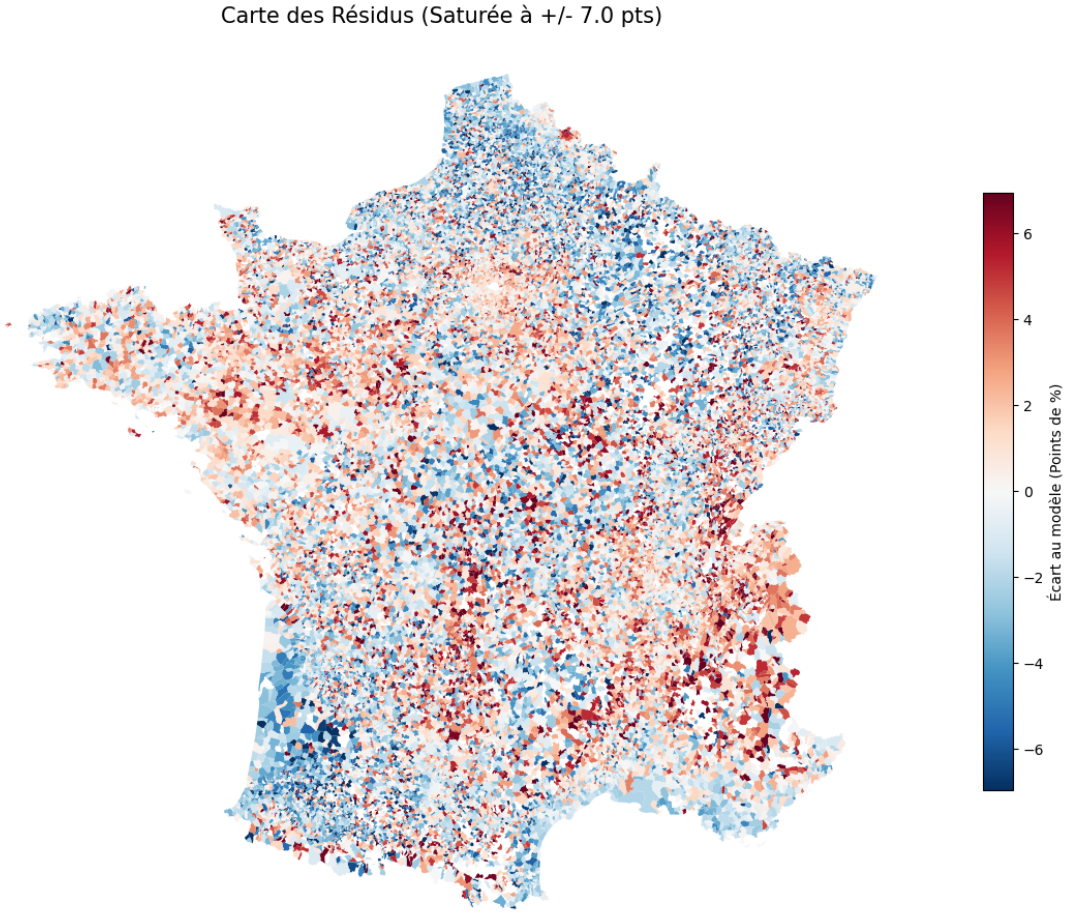
Capital culturel (Diplômés du supérieur) : C'est la seule variable sociologique conservant une dynamique croissante et linéaire. Elle confirme que le clivage éducatif reste le principal moteur actif du vote de gauche, au-delà de la simple inertie historique.

L'absence de non-linéarités marquées dans les profils PDP démontre que, une fois l'inertie historique contrôlée, les déterminants socio-économiques agissent de manière stable et linéaire (diplôme) ou deviennent neutres (revenu, taille), contredisant l'hypothèse d'effets de seuil complexes.

2.1.2 Cartographie des Résidus : Analyse Spatiale de l'Erreur

La cartographie des résidus consiste à projeter spatialement l'écart entre le vote réel et le vote prédit par notre modèle.

L'objectif est de détecter une éventuelle "autocorrélation spatiale" des erreurs : * Si les résidus sont répartis de manière aléatoire sur le territoire, cela valide la robustesse du modèle : les variables socio-économiques introduites (revenus, densité, diplômes) suffisent à expliquer la géographie du vote. * Si les résidus apparaissent groupés géographiquement (des "tâches" de couleurs uniformes sur des régions entières), cela indique que le modèle souffre d'un biais de variable omise. Cela signifie qu'une dimension non prise en compte par nos données (facteur culturel, historique, ou présence d'un leader local) joue un rôle déterminant dans ces territoires.



2.1.3 Interprétation des disparités territoriales

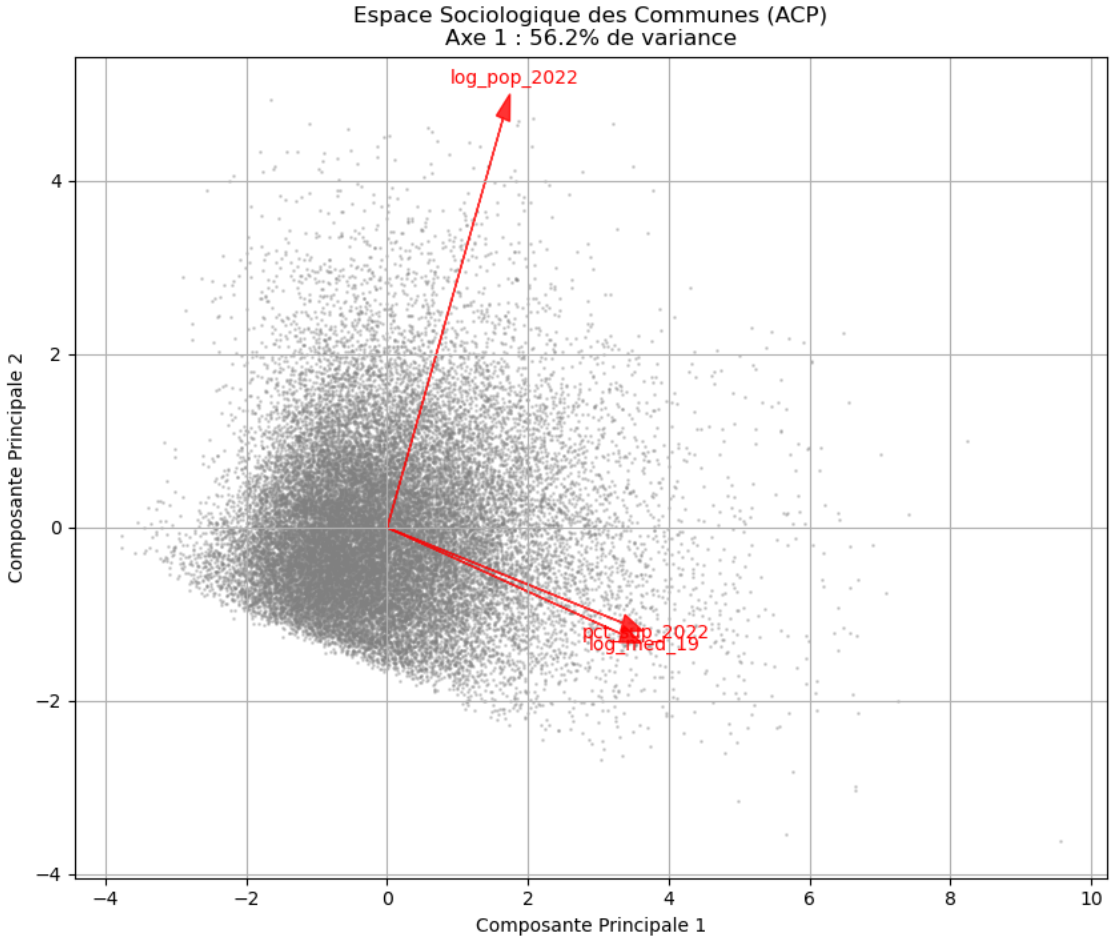
La cartographie des résidus met en évidence les territoires où les variables socio-économiques (revenus, diplômes, densité) ne suffisent pas à expliquer le comportement électoral. Ces écarts révèlent l'influence de dynamiques politiques et culturelles locales.

1. Zones de Sur-estimation (Résidus négatifs) : le modèle prédit un score élevé pour le bloc gauche/écologiste, mais le vote réel est plus faible. C'est par exemple le cas pour les Hauts-de-France, le Grand Est et le littoral de Nouvelle_Aquitaine.
    - La concurrence du Rassemblement National (Hauts-de-France / Grand Est) : Dans ces régions industrielles ou post-industrielles, les caractéristiques socio-économiques (précarité, ouvriers) pourraient théoriquement favoriser un vote de gauche traditionnelle. Le modèle surestime ce vote car il ne capte pas le basculement d'une partie de l'électorat populaire vers le RN.
    - La variable omise de l'âge (Nouvelle-Aquitaine) : Sur la façade atlantique, la sur-estimation signale probablement l'absence de la variable démographique. Ces zones attirent des retraités aisés. Bien que dotés de capitaux économiques ou culturels parfois associés au vote progressiste, ces populations conservent un comportement électoral plus conservateur.
2. Zones de Sous-estimation (Résidus positifs) : le vote réel pour la gauche est nettement supérieur à la prédiction du modèle. C'est le cas pour la Bretagne et le Sud-Ouest. On peut supposer un effet culturel.
    - En Bretagne, la tradition démocrate-chrétienne et un tissu associatif dense favorisent structurellement le vote centre-gauche et écologiste, au-delà de ce que le niveau de diplôme seul pourrait prédire.
    - Dans le Sud-Ouest, l'héritage du Radical-Socialisme maintient un ancrage à gauche très fort, créant une norme sociale de vote qui résiste aux évolutions économiques ou à la gentrification.

2.1.4 Analyse en Composantes Principales (ACP) : L'espace sociologique

Notre modèle économétrique étant déjà parcimonieux et efficace, l'ACP n'est pas ici mobilisée dans une optique de réduction de dimension. Elle est utilisée à des fins exploratoires et heuristiques.

L'objectif est de visualiser la structure des corrélations entre les variables explicatives (revenus, diplômes, densité). En projetant ces dimensions sur un plan factoriel, nous cherchons à faire émerger "l'espace sociologique" des communes françaises et à comprendre comment les inégalités territoriales se superposent ou s'opposent.



Interprétation : La polarisation du vote L'ACP révèle que l'espace sociologique est structuré par deux dimensions distinctes (Axe 1 : 56.2% de variance) :

Une colinéarité quasi-parfaite entre Capital Économique et Culturel : Les vecteurs log_med_19 et pct_sup_2022 sont confondus. En France, les territoires riches sont les territoires diplômés. Cela valide la redondance d'information gérée par le Random Forest.

L'orthogonalité partielle de la taille urbaine : Le vecteur log_pop_2022 diverge des deux précédents. Cela indique que la densité urbaine constitue une dimension sociologique indépendante du statut social.

Conclusion : Le vote de gauche ne répond pas à un simple gradient 'Centre-Périphérie' uniforme, mais à une interaction entre ces deux dimensions : la densité (taille de la ville) et la composition sociale (richesse/éducation).

Note importante par rapport au lien entre l'ACP et les PDP.

Loin de se contredire, les deux analyses confirment la robustesse du modèle :

Sur la population : L'ACP prouve que la densité est une dimension géographique distincte. Le

PDP confirme qu'elle est politiquement neutre (courbe plate). Le modèle distingue donc bien la taille de la ville de la sociologie électorale.

Sur le couple revenu/diplôme : L'ACP révèle leur colinéarité quasi-parfaite (redondance). Face à ce doublon, le Random Forest a logiquement "arbitré" en faveur du diplôme, rendant la variable revenu muette (saturation) sans perdre d'information.

3 Conclusion Générale

Ce projet a pris racine dans la volonté d'expliquer ressorts de la "vague verte" et plus largement de la dynamique du bloc de gauche lors des élections municipales de 2020. En croisant des données électorales et socio-économiques à l'échelle communale, notre analyse a cherché à dépasser le simple constat électoral pour interroger les mutations structurelles du vote en France, notamment au prisme de la gentrification.

1. La validation des déterminants sociologiques

L'approche économétrique a confirmé que la recomposition politique actuelle est fortement corrélée aux variables de la "nouvelle sociologie urbaine". La densité de population et la part des diplômés du supérieur apparaissent comme les vecteurs les plus puissants du vote écologiste et de gauche. L'analyse en première différence a permis de valider une dynamique causale : ce n'est pas seulement le niveau de richesse qui détermine le vote, mais bien la transformation sociologique d'un territoire (arrivée de cadres, élévation du niveau de diplôme) qui favorise la bascule politique.

2. Au-delà des modèles linéaires : Une validation par la parcimonie

L'apport du Random Forest est ici critique plus que prédictif. En affichant une performance équivalente au modèle linéaire (OLS) et des courbes de dépendance partielle (PDP) majoritairement plates ou linéaires, l'algorithme invalide l'hypothèse d'effets de seuil complexes.

Il permet surtout de simplifier les effets de composition : une fois l'inertie et le diplôme contrôlés, la densité urbaine et le revenu perdent leur pouvoir explicatif (courbes neutres). Le Machine Learning confirme ainsi que le vote de gauche ne dépend pas d'une "masse critique" géographique, mais reste structuré par une dynamique linéaire fondamentale : l'inertie historique et le capital culturel.

3. Les limites du déterminisme social

Cependant, l'enseignement majeur de ce travail réside peut-être dans l'analyse des résidus. La cartographie des erreurs du modèle démontre que la sociologie ne fait pas l'élection. La persistance de clusters géographiques de sous-estimation (notamment en Bretagne et dans le Sud-Ouest) met en lumière une "dépendance au sentier". Dans ces territoires, l'héritage historique et culturel maintient le vote à gauche à des niveaux bien supérieurs à ce que la simple réalité économique prédirait.

4. Vers une nouvelle géographie électorale

En conclusion, si la gentrification et l'élévation du niveau d'éducation sont des moteurs indéniables du vote EELV/Gauche, elles ne peuvent occulter une fracture territoriale plus profonde. Nos résultats illustrent la polarisation croissante entre des centres urbains intégrés, où le vote progressiste se consolide sur des bases culturelles, et des périphéries ou des zones industrielles (Hauts-de-France, Grand Est) où le modèle peine à prédire le recul de la gauche au profit du Rassemblement National. L'analyse suggère ainsi que le clivage gauche-droite traditionnel s'efface progressivement au profit

d'une opposition entre lieux, structurée par le rapport à la métropolisation et l'ancrage historique local.