Sports et société(s): prédire les pratiques sportives par les caractéristiques socioéconomiques.

Une analyse spatiale.

Charlotte Combier, Léopold Maurice, Guilhem Sirot

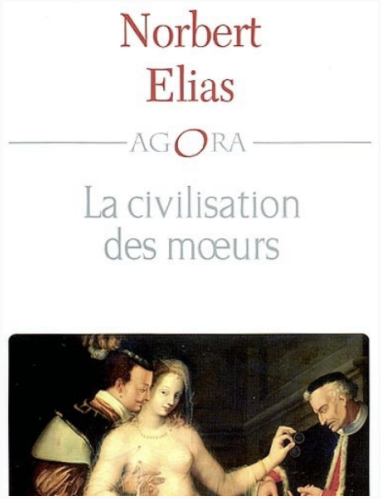
6 janvier 2023

ENSAE

Introduction

Prédire les pratiques sportives : un angle mort de la sociologie du sport ?

- Goûts socialement situés dans l'espace (social) - *La Distinction*, Bourdieu
- Les pratiques sportives n'échappent pas à ces lois sociales - *La civilisation des moeurs*, Elias
- Pourtant, il existe un manque de travaux sur le sujet (spatialisés)



- **Question de recherche : Peut-on prédire les pratiques sportives de manière spatialisée à partir de données socioéconomiques ?**

Données

Données au niveau communal

- Données socioéconomiques : fusion de plusieurs bases complémentaires
  1. Filosofi (Insee)
  2. Données de chômage des jeunes (data.gouv)
  3. Données de demandeurs d'emploi (DARES)
  4. Populations légales (Insee)

- Données sportives
  1. Enquête sur les pratiques physiques et sportives (l'ENPPS) de l'INJEP (SSM Jeunesse et Sports)
     - Données sur les licences sportives
     - Données sur les clubs sportifs
  2. Base permanente des équipements sportifs (Ministère des Sports)

Nettoyage et fusion des bases

- Principales étapes de nettoyage
  - **Transformation de la structure des données** : une ligne par commune, une colonne par fédération/nombre d'équipements etc...
  - **Sélection des fédérations** : football, basketball, tennis, équitation, judo, handball, pétanque (choix 1. "représentatif" fédérations dont le nombre total de communes avec des clubs dépasse 0,2 après normalisation + 2. choix "sociologiques")
  - **Nettoyage de routine** : doublons, valeurs aberrantes (IQR)

- Fusion
  - Fusion par code géographique Insee
  - Conservation des variables d'intérêts :
    - **Socioéconomiques** : taux de pauvreté, revenu médian, populations légales, demandeurs d'emploi, chômage des jeunes et ratio interdécile (non utilisé *in fine*)
    - **Sportives** : infrastructures, clubs, licenciés

Dataviz et premières statistiques descriptives

Dataviz spatiale : cartes socio-économiques

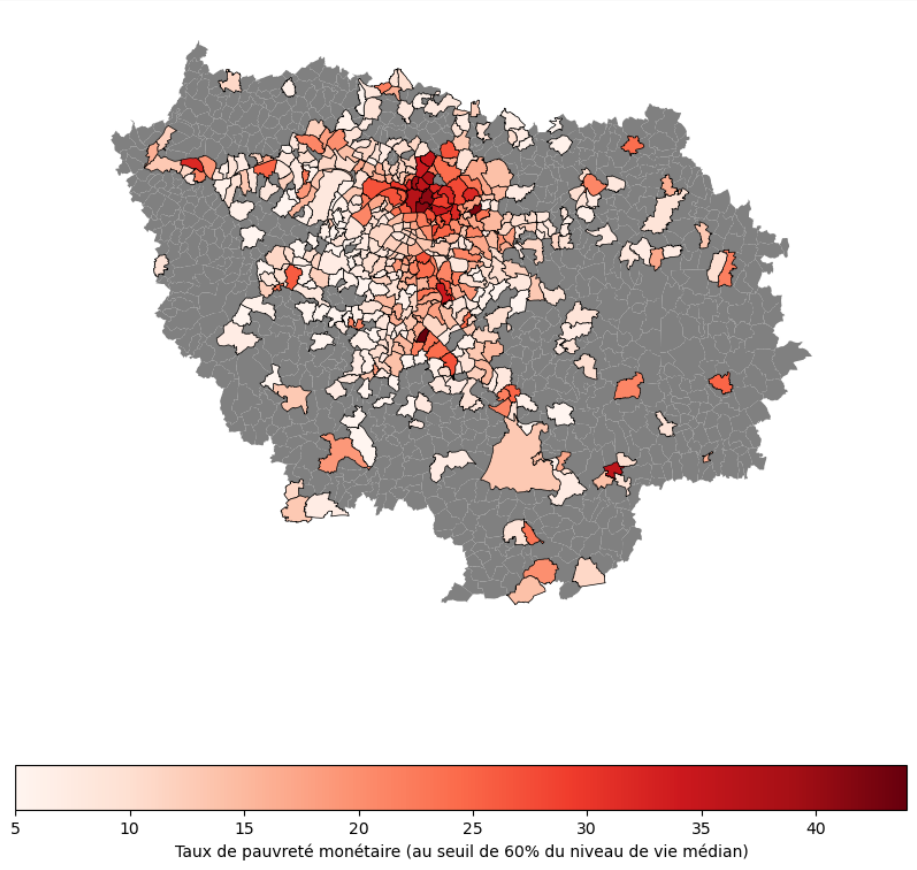


Figure 1 – Taux de pauvreté en IdF

On voit les difficultés économiques de la Seine-Saint-Denis. On note aussi que le taux de pauvreté n'est disponible que dans les villes moyennes, ce qui va orienter notre analyse future.

Dataviz spatiale : cartes sportives

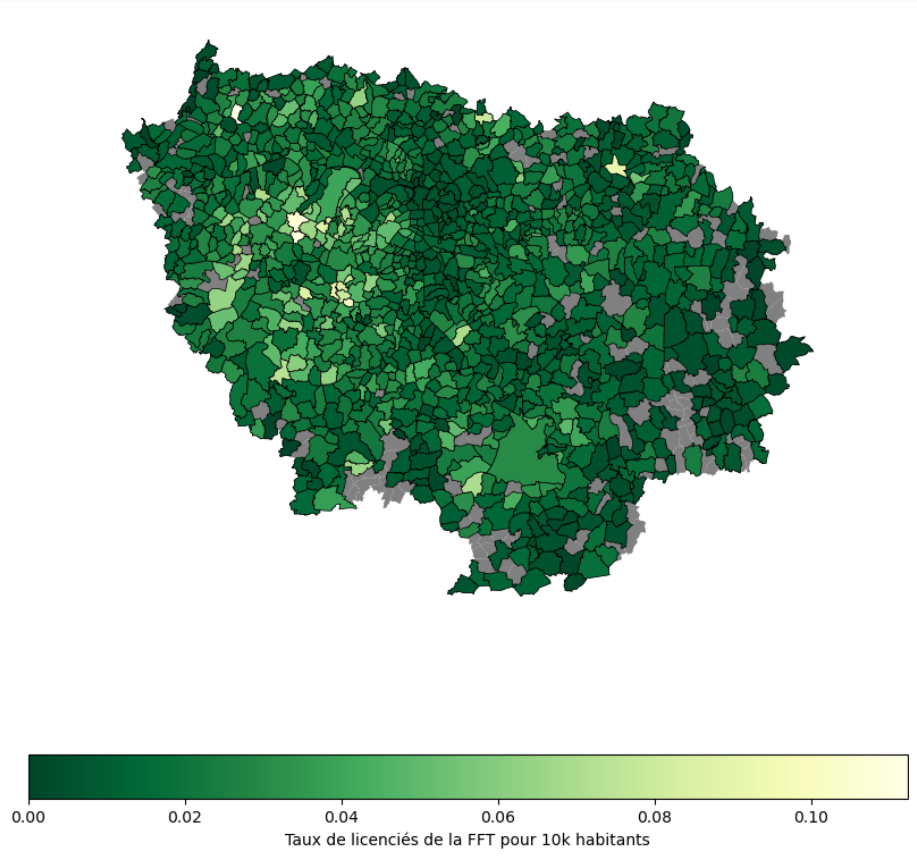


Figure 2 – Carte des taux de licenciés de tennis en IdF

On pressent déjà ce qu'il y a bien homogamie entre dimension sportive et dimension socio-économique et ces dernières trouvent sens au niveau communal : les licenciés de tennis sont plus nombreux dans le sud-ouest parisien.

Dataviz des corrélations

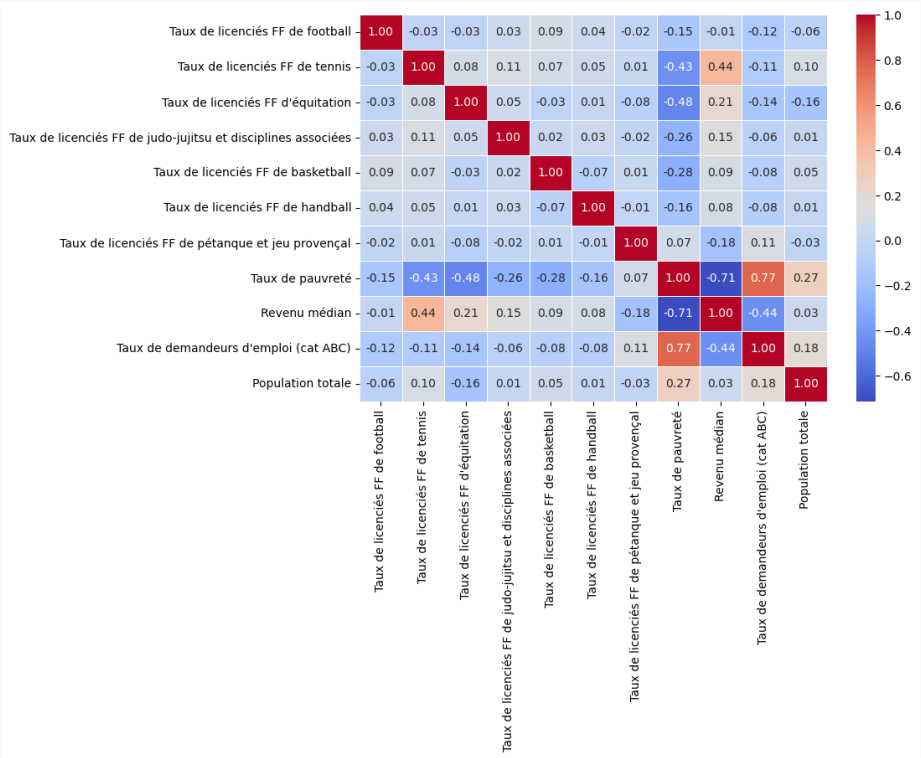


Figure 3 – Carte de chaleur des corrélations entre pratiques sportives et variables socio-économiques, licenciés en relatif à la population légale

On trouve bien des corrélations pertinentes : le taux de licenciés de tennis est corrélé positivement (0.44) au revenu médian et corrélé négativement (-0.41) au taux de pauvreté.

Dataviz des corrélations : effet mécanique de la population

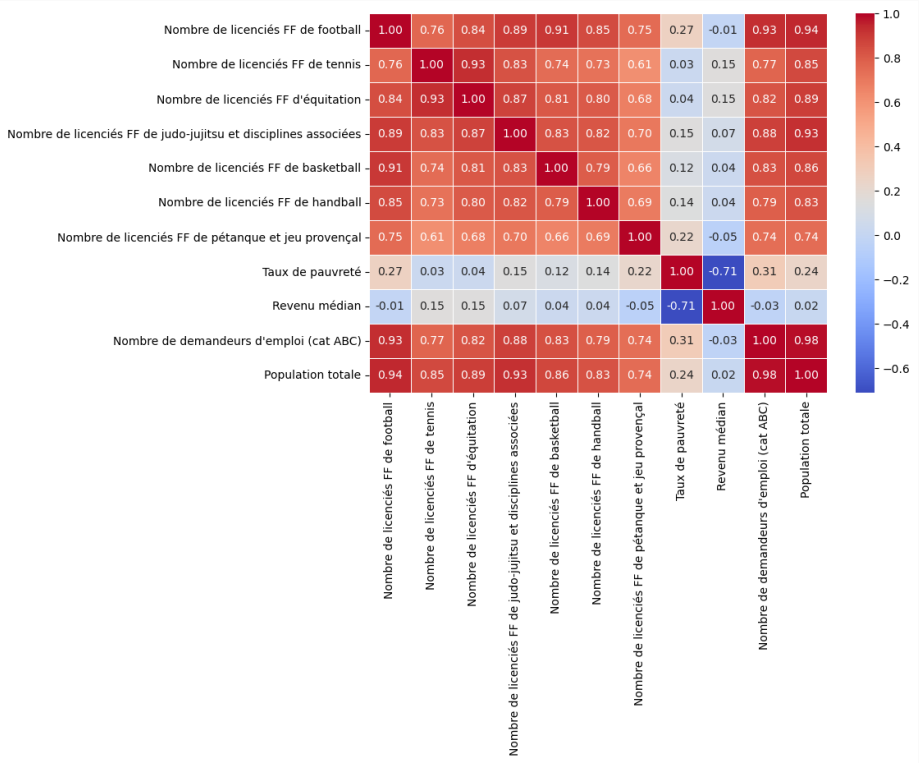
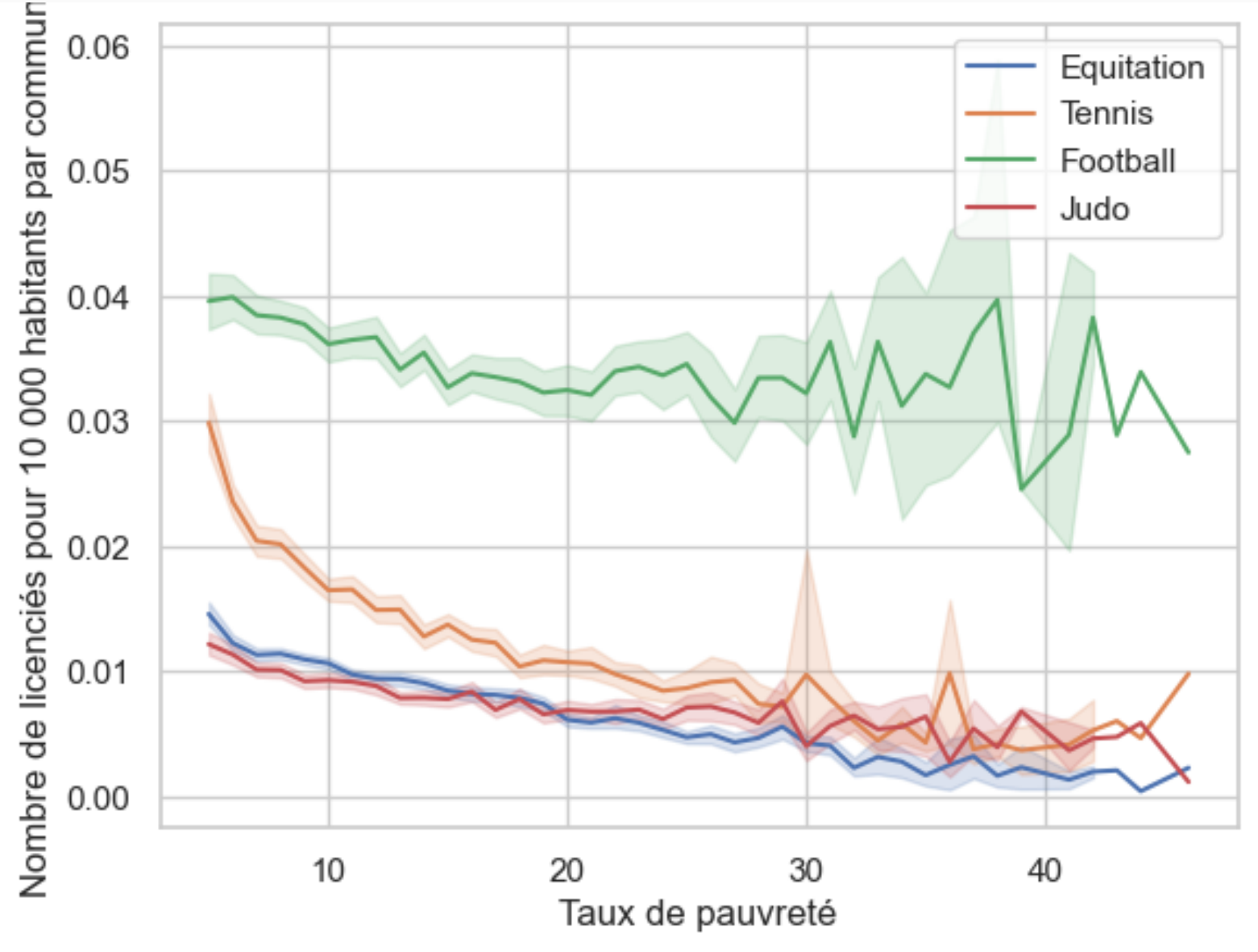


Figure 4 – Carte de chaleur des corrélations entre pratiques sportives et variables socio-économiques, licenciés en nombre total par communes

Il y a un effet mécanique entre le nombre de licenciés et la population, que l'on vient amoindrir en prenant le taux de licenciés dans le reste de l'étude.

Dataviz des statistiques descriptives : importance du taux de pauvreté



Modélisation : clustering et prédiction (ML)

Clustering : méthode

- Culstering des communes sur les données sportives
- **Limitation aux villes moyennes et grandes**
  - cohérence du lien pratique, licenciés, clubs, équipements
  - base de 470 communes
  - arrondissements de Paris traités comme des communes)
- **Méthode KMeans** : choix de 6 clusters (méthode du coude et interprétation sociologique des clusters)

Clustering : représentation géographique I

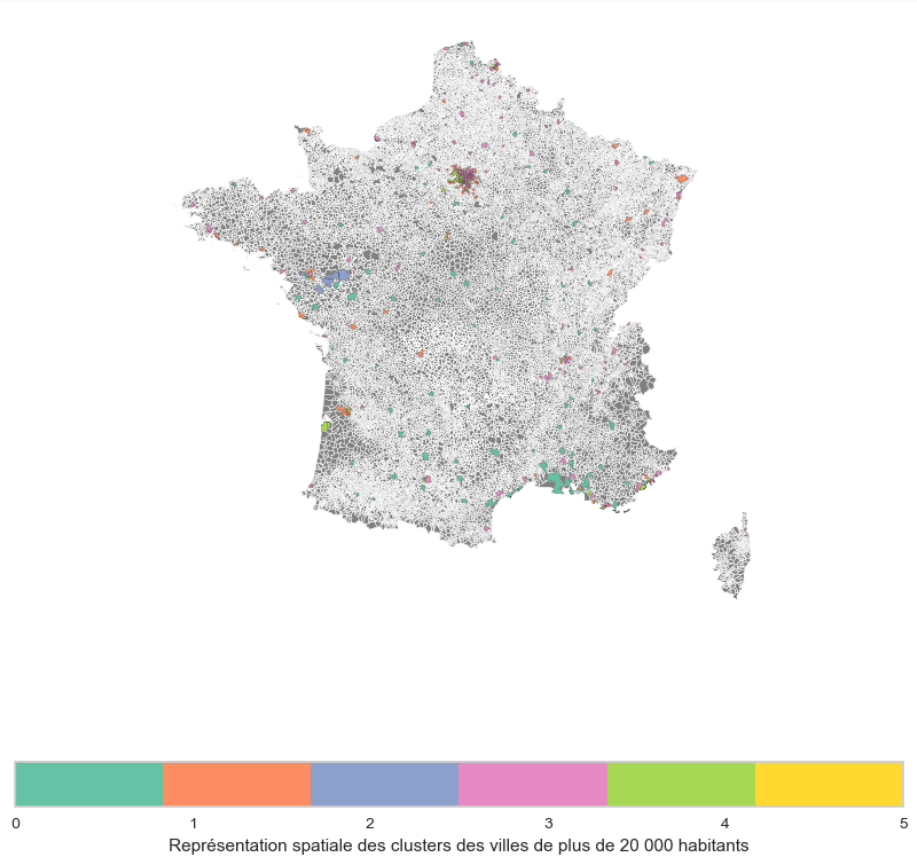


Figure 6 – Carte des clusters en France

On exclut du reste de l'analyse les clusters 2 qui regroupe les banlieues de Nantes (5) communes. On peut déjà voir que le cluster 0 et 1 contient surtout des petites villes, et une tendance pour le Sud pour le cluster 0.

Clustering : représentation géographique I

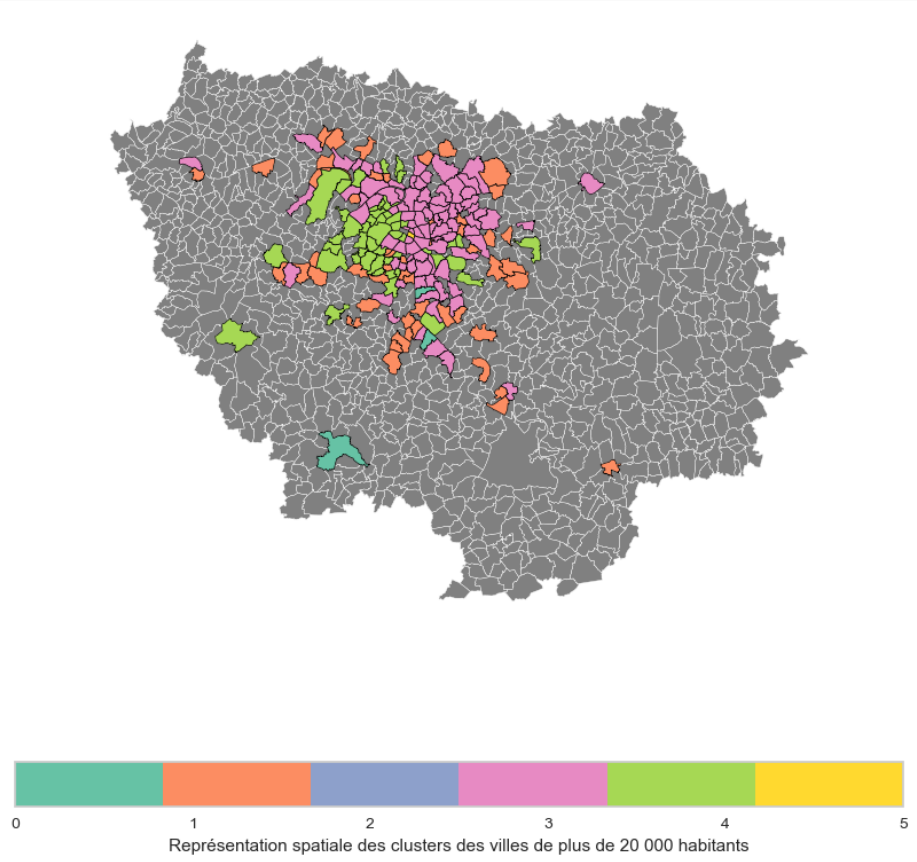


Figure 7 – Carte des clusters en France

On exclut du reste de l'analyse les clusters 5 qui ne contient que le 14ème arrondissement de Paris.

Clustering : description des clusters

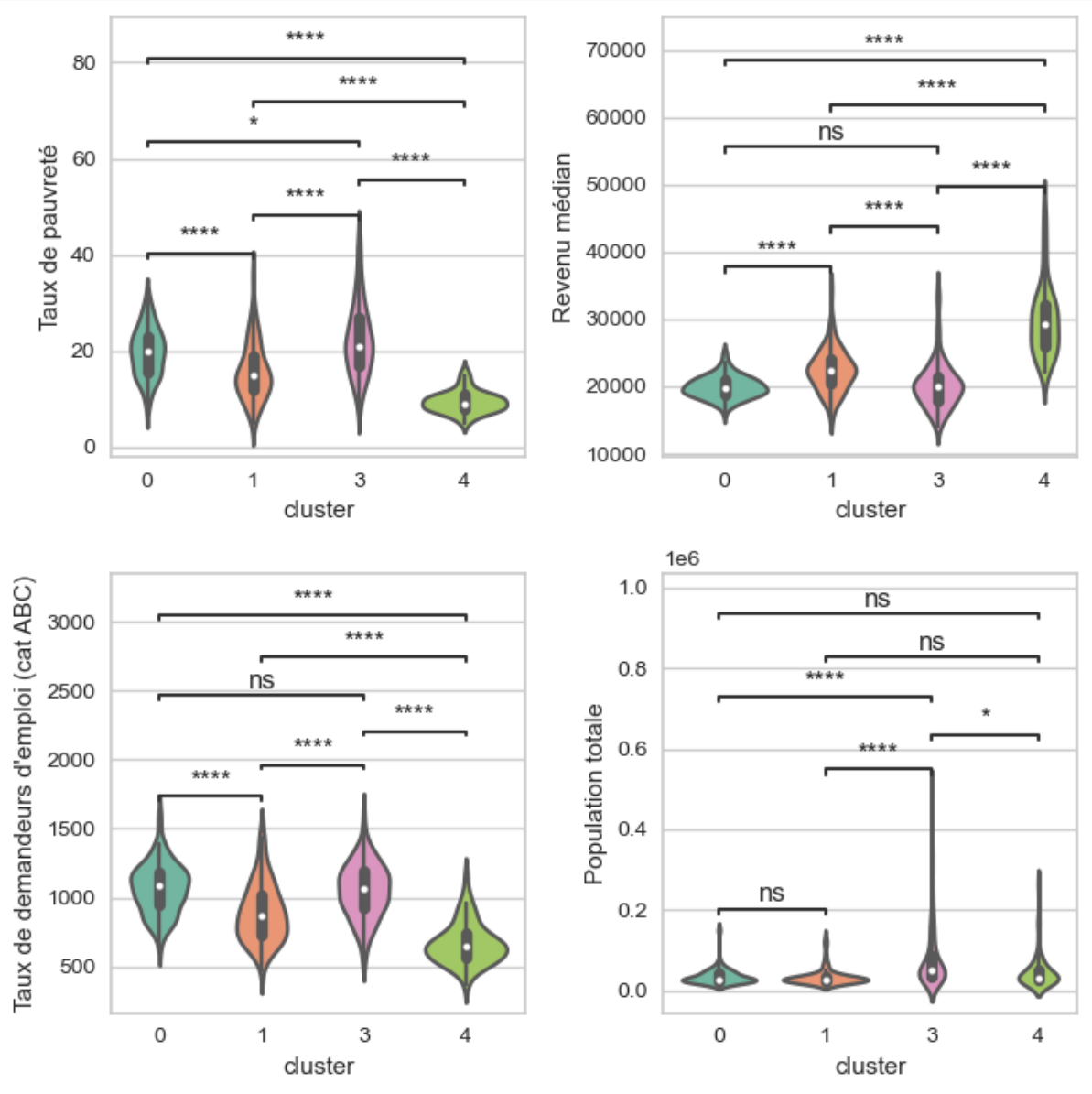


Figure 8 – Violinplot des clusters 0, 1, 3, 4 en fonction des variables

Clustering : description des clusters II

- Cluster 0 :
  - **Villes de petite taille, rurales et localisées dans le Sud**
  - **Spécificités** : pétanque et nombre d'infrastructures
  - Assez homogène et moyen en terme socio-économiques
- Cluster 1 :
  - Assez proche du cluster 1, à part pétanque.
- Cluster 3 :
  - Contient les **territoires les plus pauvres**, mais pas seulement
  - Territoires les plus peuplés
  - Sous-dotés en infrastructures
- Cluster 4 :
  - **Communes aisées**
  - Fort taux de licenciés en tennis et équitation
  - Faible taux de pauvreté et de demandeurs d'emploi

Prédiction : méthode

- Modèle de ML pour prédire les clusters de pratiques sportives à partir de données socio-écnomiques
- **Modèle linéaire** (meilleur *a*ccuracy score)
- **Echantillon entraînement** : 376 communes tirées aléatoirement
- **Echantillon test** : 94 communes tirées aléatoirement

Prédiction : Résultats

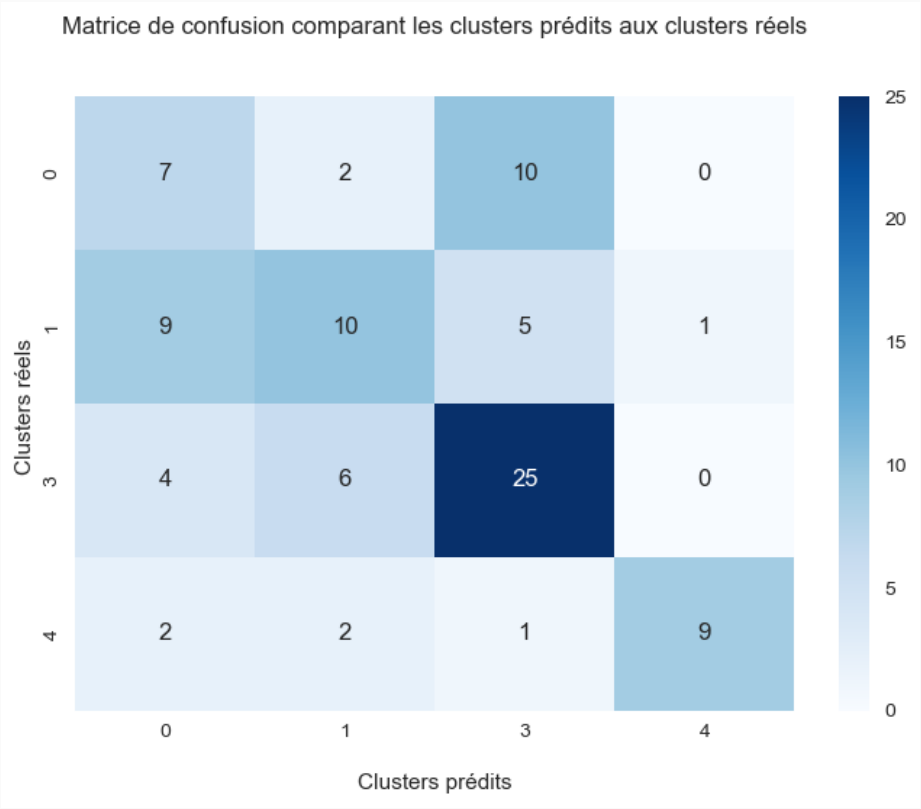


Figure 9 – Matrice de confusion

- Prédiction correcte par les données socio-économiques de la pratique sportive
- « Il y a homologie entre l'espace des sports et l'espace des positions sociales » - Pociello