



UNIVERSITÉ CLAUDE BERNARD LYON 1

DATA MINING
PROJET
RAPPORT

Clustering des admissions aux CPGE en 2024

Élèves :
Thomas BARAND
Bastien MORON

Enseignant :
Rémy CAZABET

Code disponible sur : https://github.com/thomas0barand/parcoursup_clustering.git

Table des matières

1 Problématique	2
2 EDA	2
2.1 Vue d'ensemble du dataset	2
2.2 Variables d'intérêt	2
2.3 Statistiques descriptives clés	3
2.4 Distribution géographique	3
3 Traitement des données	3
3.1 Construction des caractéristiques	3
3.2 Nettoyage et normalisation des données	3
3.3 Analyse en composantes principales (PCA)	5
3.3.1 Investigation plus approfondie	6
4 Clustering	7
4.1 K-Nearest Neighbors (KNN)	7
4.2 DBSCAN	7
4.3 Hierarchical	9
4.3.1 Contexte et objectif	9
4.3.2 Principe de la méthode	9
4.3.3 Vérification de cohérence : corrélation cophénique et analyse d'inconsistency	9
4.3.4 Choix du nombre de clusters et extraction des labels	10
4.3.5 Interprétations	11
5 Visualisations des groupes	11
5.1 Facteurs discriminants	11
5.2 Autres caractéristiques	13
6 Conclusion	14
7 Répartition du travail	15

1 Problématique

Dans quelle mesure les écoles d'ingénieurs et les Classes Préparatoires aux Grandes Écoles (CPGE) scientifiques forment-elles des groupes distincts selon leurs caractéristiques d'élitisme ? Cette question, au cœur de notre étude, vise à décrypter la structuration de ces formations d'excellence de l'enseignement supérieur français. À travers l'analyse des données Parcoursup 2024, nous explorons comment ces établissements se différencient selon plusieurs dimensions de l'élitisme : sélectivité académique, diversité sociale, géographique et de genre. Notre démarche mobilise des techniques d'analyse exploratoire et de clustering pour dépasser les classifications traditionnelles (public/privé, Paris/province) en combinant différents indicateurs : taux d'accès, proportion de boursiers, diversité des mentions au bac, mixité et origines géographiques des admis. L'identification de "profils types" émergents pourrait non seulement éclairer les mécanismes de reproduction sociale à l'œuvre dans ces filières, mais aussi contribuer à la réflexion sur des politiques éducatives plus inclusives.

La question est maintenant placée en ouverture, annonçant clairement l'objectif de l'étude et la démarche qui sera suivie pour y répondre. Le reste du paragraphe détaille la méthodologie et les enjeux, tout en conservant le fil conducteur de cette problématique centrale.

2 EDA

2.1 Vue d'ensemble du dataset

Le jeu de données Parcoursup 2024 contient les informations sur les candidatures et admissions pour l'ensemble des formations post-bac en France. Notre analyse se concentre sur les écoles d'ingénieurs et les CPGE.

2.2 Variables d'intérêt

Les principales variables analysées sont :

- Sélectivité académique :
 - Ratio de mentions Très Bien et TB avec félicitations
 - Taux d'accès
 - Rang du dernier appelé
- Diversité sociale :
 - Pourcentage de boursiers (candidats et admis)
 - Ratio de sélectivité des boursiers
 - Origine géographique des admis
- Attractivité :
 - Ratio candidats/places
 - Capacité d'accueil
 - Taux de remplissage
- Diversité des profils :
 - Distribution des types de bac
 - Ratio filles/garçons
 - Origine académique

Session	Statut de l'établissement de la filière de formation (public, privé...)	Code UAI de l'établissement	Établissement	Code départemental de l'établissement	Département de l'établissement	Région de l'établissement	Académie de l'établissement	Commune de l'établissement
231 2024	Public	0310036W	Lycée Pierre De Fermat	31.0	Haute-Garonne	Occitanie	Toulouse	Toulouse
4914 2024	Public	0310036W	Lycée Pierre De Fermat	31.0	Haute-Garonne	Occitanie	Toulouse	Toulouse
7836 2024	Public	0310036W	Lycée Pierre De Fermat	31.0	Haute-Garonne	Occitanie	Toulouse	Toulouse
13378 2024	Public	0310036W	Lycée Pierre De Fermat	31.0	Haute-Garonne	Occitanie	Toulouse	Toulouse

FIGURE 1 – Extrait du tableau des données brutes

2.3 Statistiques descriptives clés

Pour les écoles d'ingénieurs et CPGE analysées :

- Nombre total de formations : 984
- Nombre de formations de type Ecole d'Ingénieur : 523
- Nombre de formations de type CPGE scientifique : 487
- Nombre de statistiques connues pour chaque établissement : 121
- Capacité d'accueil moyenne : 48 places
- Ratio moyen candidats/places : 15.3
- Taux moyen de boursiers admis : 14%
- Pourcentage moyen de mentions TB : 32%

2.4 Distribution géographique

La répartition géographique montre une forte concentration dans les régions :

- Île-de-France (31% des places)
- Auvergne-Rhône-Alpes (15%)
- Occitanie (11%)

3 Traitement des données

Cette partie s'attarde sur le choix des caractéristiques et de leur traitement. L'objectif ici est de préparer le bon jeu de caractéristiques pour un clustering efficace. Pour cela, il est nécessaire de s'assurer que les caractéristiques sont adaptées et discriminantes vis à vis de la problématique.

De plus, ces caractéristiques se doivent d'être nettoyées et nettoyées pour éviter les valeurs aberrantes et les erreurs de calculs.

3.1 Construction des caractéristiques

La construction des caractéristiques se fait en : - gardant des caractéristiques déjà présentes dans le dataset, - construisant de nouvelles caractéristiques depuis les données

On peut donner quelques exemples de caractéristiques récupérées et créées. Ces caractéristiques visent à avoir un lien avec la *sélectivité* de la formation.

Les caractéristiques sont à diviser en différents groupes, elles portent sur les potentiels biais dans la sélection des candidats : -La formation en elle même (Nombre total de places, candidats, ...) -Le genre -La population de boursiers -L'origine du bachelier(quel type de bac avant d'entrer / de candidater) -mention au bac -L'origine géographique (est ce que le candidat provient de la même académie, même établissement) -Des facteurs de pressions sur la formation (taux d'accès, taux de refus, ...)

Ainsi on peu donner une liste non exhaustive des caractéristiques calculées ou extraites :

À ces caractéristiques nous ajoutons pour chaque formation des clés :

Clés et identifiants (non utilisés dans la distance)

- Code UAI de l'établissement,
- Nom de l'établissement.
- ...

Sur le plan catégoriel, nous retenons différentes caractéristiques qui semblent adéquates pour contextualiser l'interprétation des clusters. Les autres libellés administratifs et identifiants servent de clés de suivi, sans intervenir dans la distance.

Caractéristiques catégorielles considérées

- Code UAI de l'établissement ; Établissement (identifiants et libellés).
- Département, Région, Commune (localisation administrative).
- Filière détaillée bis (descripteurs) (MPSI, PCSI, ECG, ...).

3.2 Nettoyage et normalisation des données

Les données brutes contiennent des valeurs problématiques (`inf`, `-inf`, `NaN`) issues de divisions par zéro ou de champs manquants. Nous remplaçons `inf/-inf` par les correspondantes de la caractéristique

TABLE 1 – Description des caractéristiques des formations

Caractéristique	Type	Description
Capacité de l'établissement par formation	Extraite	Capacité totale de la formation.
f_ratio_candidats	Calculée	Taux de filles parmi les candidats à cette formation.
f_ratio_admis	Extraite	Taux de filles parmi les admis à cette formation.
f_selectivity_candidats	Calculée	Ratio entre les taux d'admisses et de candidates.
assez_bien_mention_ratio	Extraite	Taux de mention « assez bien » parmi les admis.
meme_academie_ratio	Extraite	Taux d'admis issus de la même académie.
last_call_rank_ratio	Calculée	Rang du dernier appelé comparativement aux places disponibles et au nombre de candidats.
pressure_ratio	Calculée	Taux de pression (places / candidats).
taux_acces_ratio	Extraite	Rapport entre le nombre de candidats dont le rang de classement est inférieur ou égal au rang du dernier appelé de leur groupe et le nombre de candidats ayant validé un vœu pour la formation étudiée en phase principale.
longitude	Extraite	Longitude de la formation.

(bornes sup/borne inf) et imputons `NaN` par la médiane de la colonne, choix robuste aux outliers et cohérent avec une mise à l'échelle ultérieure.

Pour la normalisation, nous utilisons **RobustScaler** (centrage par la médiane, échelle par l'IQR). Cette transformation atténue l'influence des valeurs extrêmes et place toutes les variables numériques sur une échelle comparable, condition nécessaire à l'usage de distances pour le clustering. Les colonnes catégorielles et les identifiants sont explicitement exclues de la normalisation.

Les variables catégorielles seront, si leur cardinalité n'est pas trop élevée, one-hot encodée avant clustering.

Les variables géographiques non catégorielles (longitude et latitude) sont traitées afin de limiter l'effet des valeurs extrêmes (dû au formation hors métropole) tout en gardant de l'information, nous appliquons un *clipping* de type Winsorization contrôlé par un paramètre α (typiquement 0,90–0,98), qui conserve la majorité des données tout en bornant les extrêmes. Le but étant de garder l'information "formation proche des autres sans fausser les résultats par des formation plus lointaines".

Après clipping, nous appliquons une normalisation MinMax sur $[0, 1]$ à *longitude* et *latitude*. Cette mise à l'échelle assure que le signal spatial contribue de manière contrôlée aux distances, sans dominer les autres familles de variables. Le paramétrage d' α permet d'ajuster le poids effectif de la dimension géographique selon les besoins analytiques.

Enfin, nous vérifions la pertinence des variables via leur variance post-normalisation et des visualisations de distributions (voir Fig. 2a et Fig. 2b).

Ainsi, certaines variables ont été écartées pour plusieurs raisons principales : leur redondance avec d'autres (corrélations trop fortes), une distribution insuffisamment dispersée ou encore la présence de valeurs aberrantes par rapport aux autres variables.

L'efficacité de ces écartements a aussi été vérifié empiriquement avec une projection PCA pour différentes combinaisons de variables. La vérification consistait à tester différentes combinaisons de groupes de variables pour voir elles étaient significativement importantes par les distances entre points.

Variables candidates écartées

- Variables numériques :
 - Les variables caractérisant les origines en termes de bac (prof, tech, gen * _ratio) ne sont pas dispersés, ces variables sont trop discriminantes pour notre problématique.
 - Certaines variables de type *_selectivity* sont écartées car elles comportent trop de valeurs aberrantes.

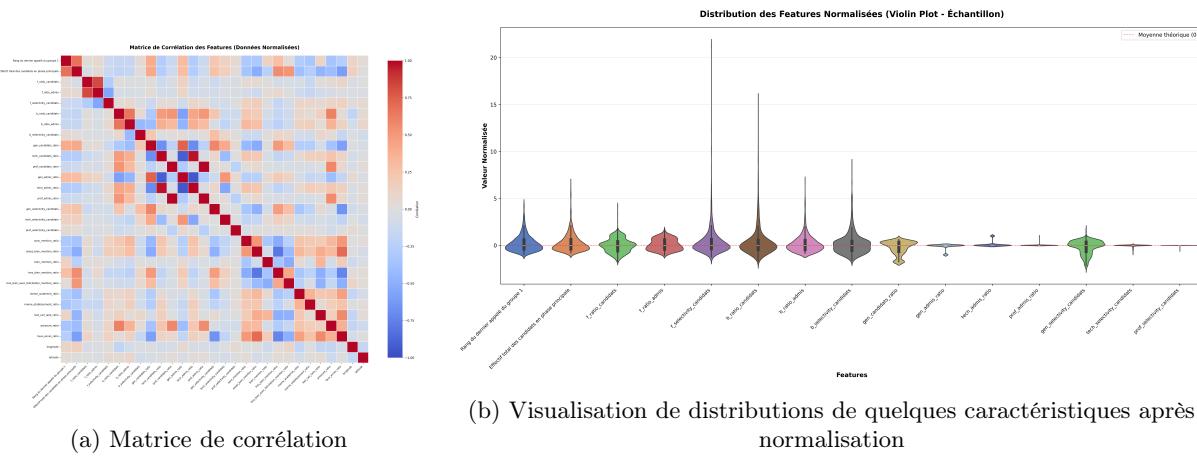


FIGURE 2 – Analyse exploratoire des caractéristiques : corrélations et distributions normalisées.

— *Variables catégorielles*

- Le département et la région de l'établissement sont retiré, car leur cardinalités étaient trop élevées pour permettre un one-hot encodage efficace, et ils n'apportaient pas d'information supplémentaires pertinentes.

Conclusion Le jeu de données comporte ainsi 22 caractéristiques distinctes, dont deux de nature catégorielle : *Filière de formation détaillée* et *Statut de l'établissement* qui seront les variables du clustering.

3.3 Analyse en composantes principales (PCA)

La représentation des données à l'aide d'une Analyse en Composantes Principales (PCA) permet de réduire la dimensionnalité du jeu de données tout en conservant la majeure partie de la variance. Cette approche facilite la visualisation des relations entre les différentes formations et met en évidence les caractéristiques les plus discriminantes, c'est-à-dire celles qui expliquent le mieux les variations observées.

Afin de prendre en compte la dimension géographique, un paramètre d'ajustement, noté α , a été introduit. Celui-ci contrôle l'influence des variables géographiques (longitude et latitude) dans la projection. En fixant α autour de 0.98, on s'assure que les informations spatiales contribuent au positionnement des points dans l'espace réduit sans dominer totalement les autres variables explicatives.

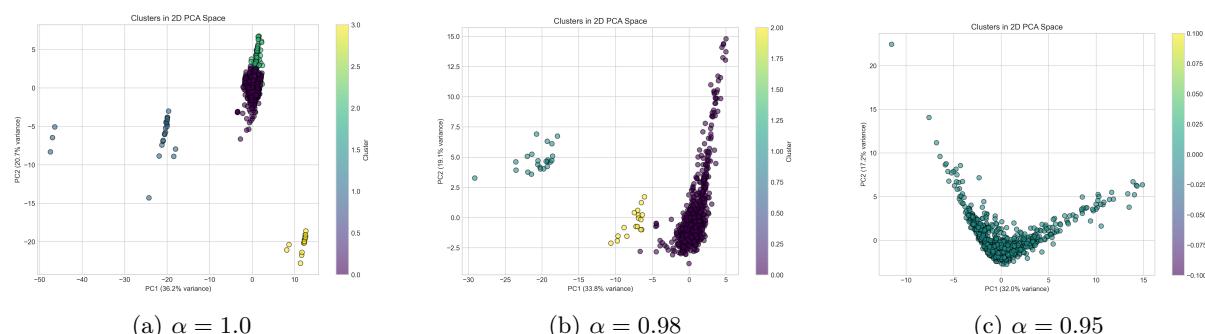


FIGURE 3 – Visualisation de la PCA pour différentes valeurs de α .

Pour la suite nous choisirons $\alpha = 0.965$ qui est un bon compromis entre influence du paramètre et non-sur-représentation de celui-ci.

On peut étudier plus en profondeur cette représentation PCA en s'intéressant aux taux de mentions des admis dans chaque formations, cela nous donnera une bonne idée de la "qualité"

L'analyse PCA indique que les caractéristiques sélectionnées permettent de disperser correctement les formations, avec des variances expliquées de 19 % pour la première composante principale (PC1) et de 30 % pour la deuxième composante (PC2). Cette représentation suggère également que la dispersion des données dans l'espace PCA reflète de manière pertinente la sélectivité des candidats, au vue de leur résultats au baccalauréat. ("Les bons élèves ont tendances à aller dans les formation à droite" dans la dimension PC1)

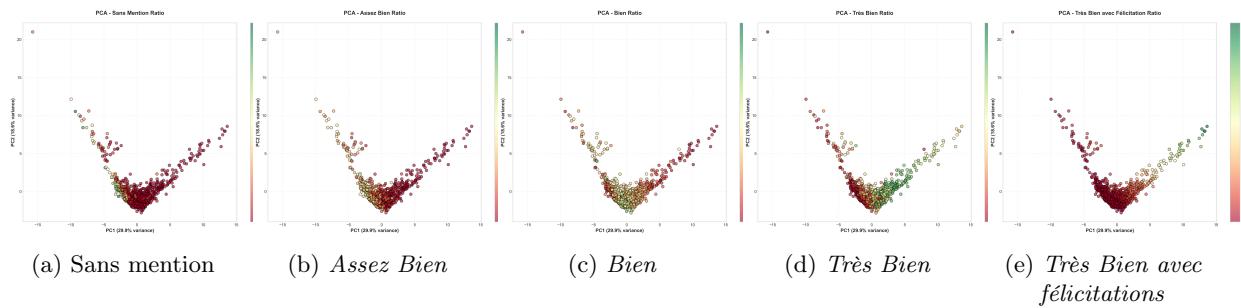


FIGURE 4 – Représentation des formations dans l'espace PCA, colorée selon le taux d'admis par type de mention au baccalauréat, de la moins à la plus élevée.

3.3.1 Investigation plus approfondie

Afin de mieux interpréter la représentation obtenue par l'ACP, les différentes caractéristiques ont été visualisées en coloriant les points de l'espace PCA en fonction de leur valeur, qu'elles soient numériques ou catégorielles. Cette approche permet d'examiner de manière plus détaillée l'influence de chaque variable sur la projection et d'évaluer comment les distances observées dans l'espace réduit reflètent les différences entre formations.

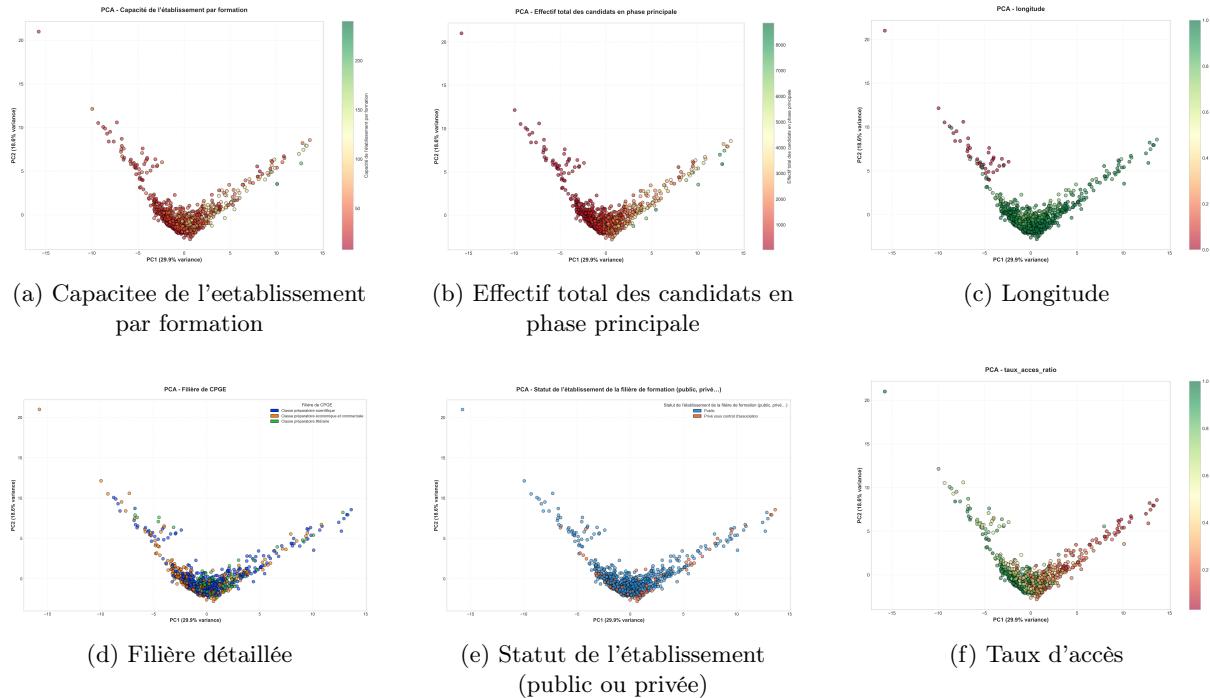


FIGURE 5 – Visualisation de différentes caractéristiques dans l'espace PCA, permettant d'interpréter la projection et la signification des distances entre formations.

Ces visualisations suggèrent que certaines caractéristiques contribuent de manière plus significative à la dispersion des formations dans l'espace PCA, et, par conséquent, aux distances observées entre celles-ci.

Il apparaît que la dispersion des formations est principalement influencée par des facteurs de sélectivité, tels que la mention au baccalauréat et le taux d'accès. Les différentes filières et le statut des établissements au sein des CPGE semblent relativement mélangés, à l'exception d'un groupe de CPGE situées hors de la métropole (voir Fig. 5c).

Cette analyse indique que les points sont suffisamment dispersés pour envisager une approche de clustering, suggérant que des résultats pertinents peuvent être obtenus à partir de ces données. Néanmoins les formations semblent se répartir sur un spectre continu ce qui limite l'applicabilité du clustering.

4 Clustering

Dans cette section, les techniques de clustering sont appliquées au jeu de données précédemment constitué afin d'explorer la structure sous-jacente et d'identifier d'éventuels regroupements naturels.

4.1 K-Nearest Neighbors (KNN)

La méthode de regroupement KNN a été appliquée pour tenter de segmenter les formations. Compte tenu de la distribution des données, il est raisonnable de s'attendre à une distinction principale entre les formations situées en métropole et hors métropole, avec peu de chances d'obtenir plus de deux clusters cohérents.

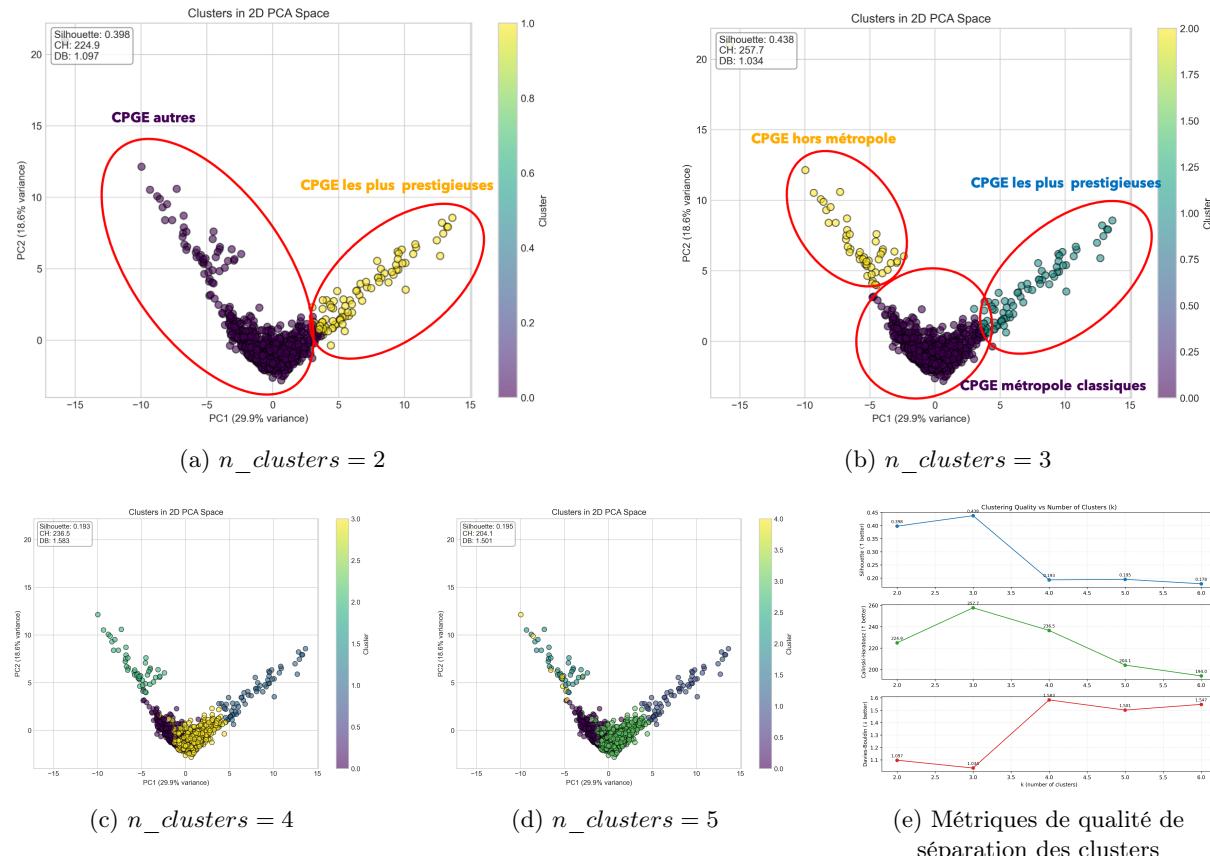


FIGURE 6 – Résultats du clustering KNN pour différentes valeurs de $n_clusters$, ainsi que les métriques de qualité de séparation des clusters.

Analyse et interprétation À partir de l'analyse croisée qualitative et quantitative, le nombre optimal de clusters semble se situer entre 2 et 3. Pour $n_clusters = 2$, les formations d'élite telles que Louis-le-Grand et Henri-IV se distinguent clairement des autres, situées globalement dans un second groupe.

Pour $n_clusters = 3$, il est possible de séparer un cluster correspondant aux CPGE situées hors métropole. Au-delà de trois clusters, la segmentation tend à subdiviser les formations selon leur performance perçue (très bonnes, bonnes, moyennes) ou leur localisation outre-mer, ce qui semble moins pertinent.

Conclusion L'analyse indique que, bien que les données soient correctement dispersées, la formation de clusters cohérents reste limitée. Les regroupements identifiés par KNN ne révèlent pas de clusters distincts au-delà de la distinction principale entre formations d'élite, les formations hors métropoles et autres CPGE.

4.2 DBSCAN

Afin d'améliorer les résultats obtenus avec la méthode KNN, l'algorithme DBSCAN a été appliqué. Ce dernier permet un raffinement du clustering en autorisant la détection de formes de clusters plus

complexes et non linéaires, sans nécessiter de spécifier à l'avance le nombre de groupes.

Les paramètres principaux de DBSCAN, à savoir ε (epsilon) et min_samples, ont été systématiquement variés afin d'identifier la combinaison offrant le meilleur score de silhouette. Cette analyse quantitative est complétée par une évaluation qualitative des clusters obtenus, permettant d'apprécier la cohérence des regroupements au regard des caractéristiques des formations.

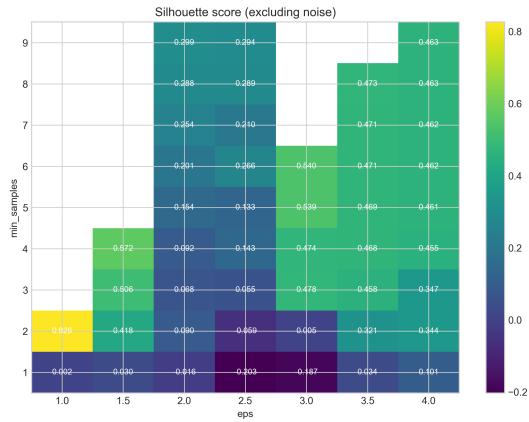


FIGURE 7 – Score de silhouette moyen en fonction des valeurs de ε et de min_samples, utilisé pour guider le choix des paramètres de DBSCAN.

Afin d'optimiser les performances du clustering avec l'algorithme DBSCAN, plusieurs combinaisons de paramètres ε et min_samples ont été testées. La figure 7 présente la carte des scores de silhouette obtenus, permettant d'identifier les zones du paramètre espace menant à des regroupements plus cohérents.

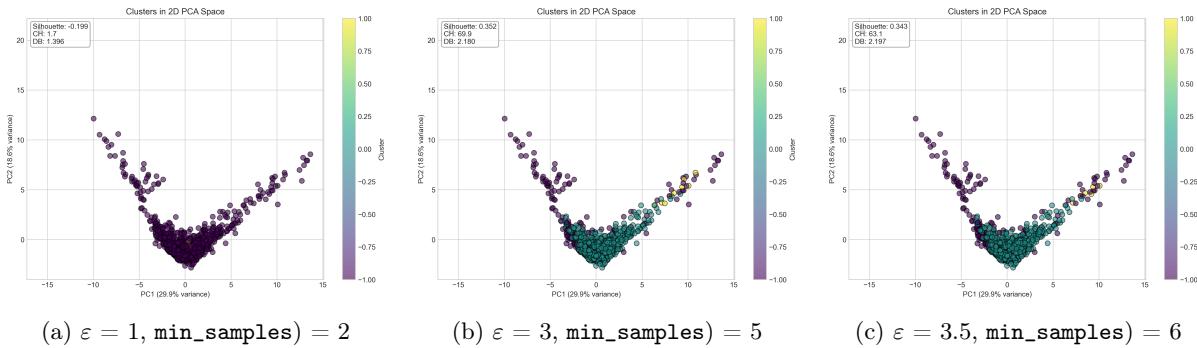


FIGURE 8 – Résultats du clustering DBSCAN pour différents couples de paramètres ($\varepsilon, \text{min_samples}$).

Dans une première configuration (7), les clusters identifiés apparaissent aberrants : leur distribution est très déséquilibrée et le score de silhouette ne traduit pas une séparation pertinente entre les groupes. Les regroupements sont trop singuliers et ne reflètent pas de véritables structures dans les données.

En revanche, pour les deux dernières configurations testées, l'algorithme DBSCAN parvient à isoler les formations dites « atypiques » ou extrêmes (par exemple certaines CPGE très sélectives ou situées hors métropole) dans le bruit, noté cluster = -1. Les formations plus communes sont quant à elles regroupées dans un ou deux grands clusters relativement homogènes.

Toutefois, cette approche ne surpassait pas les résultats obtenus avec la méthode KNN : elle ne permet pas de distinguer clairement les formations selon des critères cohérents. Cette limitation semble liée à la continuité du spectre des distances entre points dans le jeu de données, rendant difficile la séparation naturelle en groupes distincts.

Ainsi, une méthode de clustering hiérarchique pourrait constituer une alternative plus pertinente. Ce type d'approche offre la possibilité d'analyser la structure des regroupements à différents niveaux de granularité et pourrait aider à mieux représenter les relations de similarité entre formations. La section suivante explore cette méthode et ses résultats sur le même jeu de données.

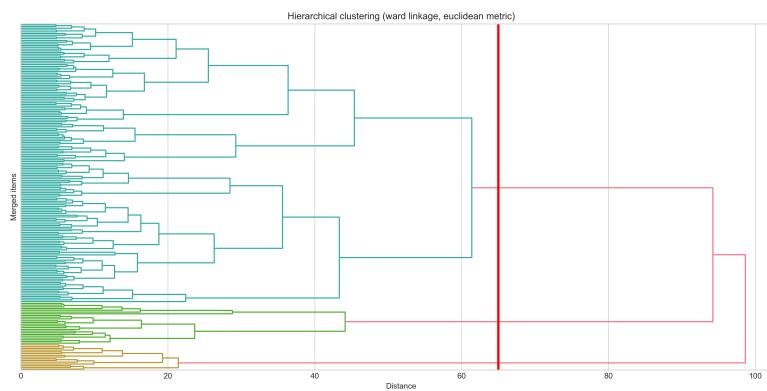


FIGURE 9 – Dendrogramme hiérarchique (Ward + Euclidean). La ligne rouge matérialise le seuil de coupe.

4.3 Hierarchical

4.3.1 Contexte et objectif

Cette section complète l'analyse après KMeans et DBSCAN. L'objectif est d'explorer une méthode non paramétrique en nombre de clusters et lisible visuellement : le clustering hiérarchique agglomératif (HAC). Il produit un dendrogramme qui révèle des découpages « naturels » à différentes granularités et permet de justifier un choix de nombre de groupes par inspection visuelle et par mesures de cohérence.

4.3.2 Principe de la méthode

- Au départ, chaque formation est un cluster isolé ; on fusionne itérativement les clusters les plus proches selon :
 - une métrique de dissimilarité (distance) entre objets ;
 - une règle de liaison (linkage) qui définit la distance entre clusters (single, complete, average, ward, ...).
- Le résultat est un arbre (dendrogramme) où la hauteur de chaque lien correspond à la distance de fusion.

Choix de la métrique et du linkage Pour garantir que les hauteurs du dendrogramme reflètent bien les distances d'origine, nous évaluons la corrélation pour plusieurs métriques (euclidean, cosine, correlation, seuclidean, mahalanobis, ...). Nous retenons :

- Évaluation « dissimilarité » : corrélation élevée = distances bien restituées par l'arbre.
- Rappel : le linkage ‘ward’ exige la distance euclidienne.

Dans nos essais :

- Sur linkage *average*, la métrique *mahalanobis* obtenait la meilleure corrélation cophénique (0.95) et *euclidean* arrivait ensuite.
- Pour construire un dendrogramme « compact » par variance intra-cluster, nous avons utilisé ‘ward + euclidean’ (conforme aux contraintes de Ward).

4.3.3 Vérification de cohérence : corrélation cophénique et analyse d'inconsistency

L'évaluation de la cohérence de la hiérarchie issue du clustering a été réalisée à l'aide de deux indicateurs principaux : la corrélation cophénique et les coefficients d'*inconsistency*.

Corrélation cophénique. La corrélation cophénique mesure la fidélité du dendrogramme par rapport aux distances originales entre les points (`pdist(X)`) Dans notre cas, la corrélation obtenue est satisfaisante, traduisant une bonne adéquation entre les distances du dendrogramme et les distances initiales.

Analyse d'inconsistency. L'analyse de l'*inconsistency* permet d'identifier les niveaux du dendrogramme où se trouvent des discontinuités marquées. Ces sauts révèlent souvent des « coupures naturelles » dans la hiérarchie et orientent le choix du nombre optimal de clusters.

Les coefficients les plus élevés apparaissent pour des regroupements correspondant à un nombre de clusters de **24, 56, 142, 282, 294, 296, 369 et 395**. L'analyse de la partie terminale du dendrogramme (*tail analysis*) met également en évidence des ratios élevés pour **3, 5, 8 et 9 clusters**, traduisant des ruptures de haut niveau menant des clusters de grandes tailles.

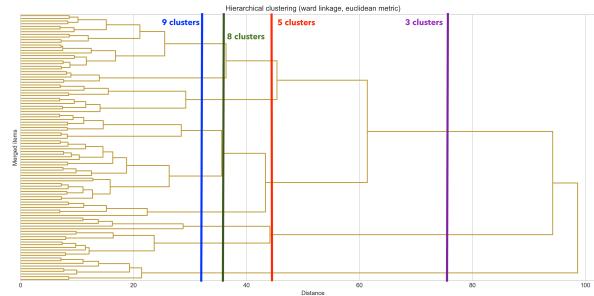


FIGURE 10 – Séparation des clusters pour les valeurs choisies 3, 5, 8 et 9.

Interprétation. Ces résultats suggèrent plusieurs niveaux possibles de segmentation :

- À faible granularité (3 à 9 clusters), la hiérarchie regroupe les formations selon des différences marquées de profil global (par exemple, distinction entre CPGE très sélectives, moyennement sélectives et régionales ou outre-mer comme observée précédemment) ;
- À granularité plus fine (24 à 395 clusters), elle permettrait de détailler davantage les sous-groupes internes au sein de ces catégories principales.

La sélection d'un seuil de coupe se repose donc sur une base solide donnée par l'analyse de l'inconsistency.

4.3.4 Choix du nombre de clusters et extraction des labels

Deux stratégies sont mises en place :

- Le choix de la distance de coupe : tracer une ligne verticale à la distance choisie (p.ex. 50) et compter les branches coupées (C'est ce qui a été fait dans l'exemple 9). Ici on dénombre 3 clusters.
- Imposer un nombre maximum de clusters. On peut arriver au même résultats en utilisant l'une ou l'autre méthode.

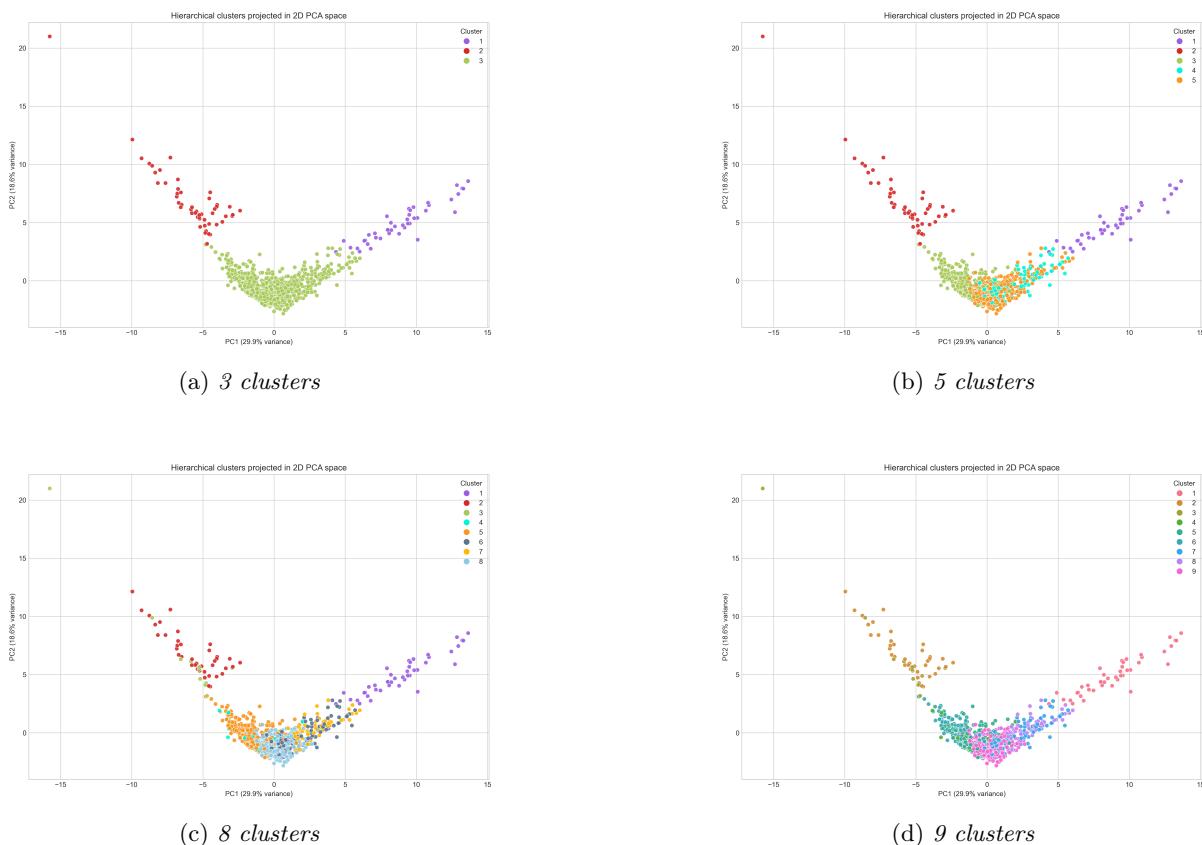


FIGURE 11 – Clusters hiérarchiques en fonction du nombre de clusters

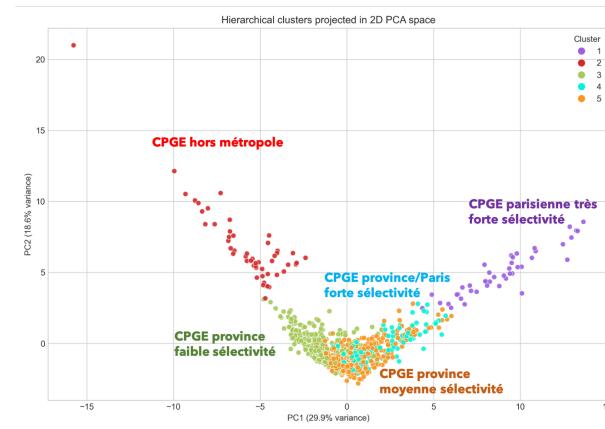


FIGURE 12 – Interprétation des différents clusters hiérarchiques

4.3.5 Interprétations

On choisit de se concentrer sur les clusters de grande taille, plus facilement interprétables. Les clusters de plus petite taille pourraient également faire l'objet d'une analyse approfondie, mais une telle étude nécessiterait un temps supplémentaire.

La démarche de *clustering hiérarchique* s'avère particulièrement adaptée au traitement des données considérées. Elle permet de mettre en évidence trois groupes distincts, interprétables de manière cohérente avec les observations précédentes : *CPGE prestigieuses*, *CPGE hors métropole* et *CPGE autres*.

Lorsque le nombre de clusters est plus élevé, l'interprétation devient plus complexe. Il est alors possible de supposer que les *CPGE classiques* se subdivisent en sous-groupes selon leur niveau de *sélectivité*. On observe également l'émergence de structures différenciant les établissements situés en *province* de ceux situés à *Paris*, au sein du groupe des *CPGE classiques*.

Cette interprétation est qualitative et une exploration des données groupes par groupe est nécessaire pour plus de précision.

5 Visualisations des groupes

5.1 Facteurs discriminants

Les caractéristiques les plus déterminantes dans la formation des clusters sont globalement similaires, quel que soit l'algorithme de clustering employé. Trois grands types de critères se distinguent :

Le type de baccalauréat obtenu par les candidats à l'établissement. Les variables associées, telles que *prof_candidats_ratio* ou *tech_candidats_ratio*, représentent respectivement la proportion de postulants issus d'un cursus professionnel ou technologique. Ces indicateurs se révèlent être les plus discriminants, car ils traduisent directement le profil de recrutement des établissements. Ils ne mettent pas en évidence un élitisme, mais reflètent plutôt la séparation opérée dès le lycée entre les filières générales, technologiques et professionnelles.

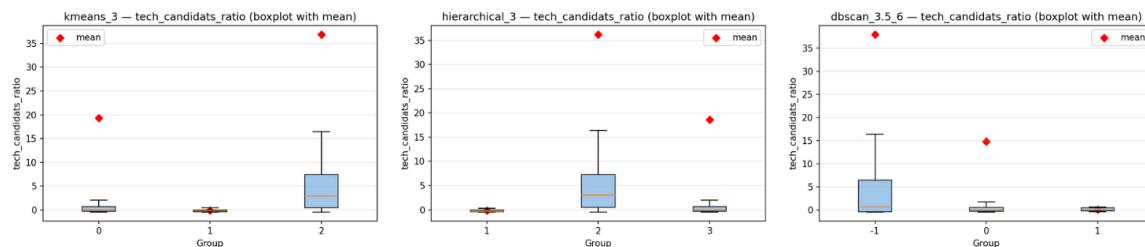


FIGURE 13 – Visualisation par cluster de la caractéristique *tech_candidats_ratio*

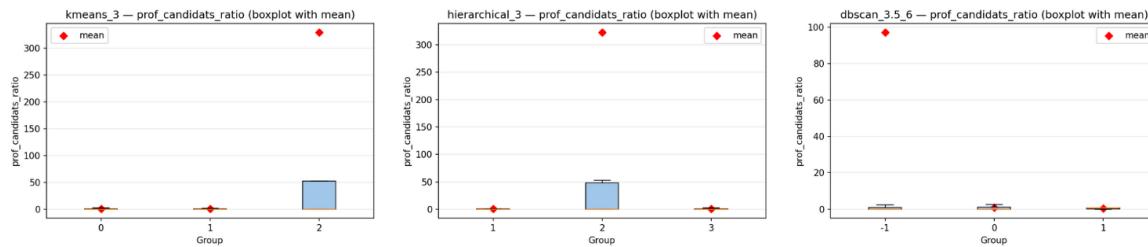


FIGURE 14 – Visualisation par cluster de la caractéristique pro_candidats_ratio

Pour les trois méthodes de clustering, un groupe d'établissements se distingue nettement selon ce critère : celui accueillant majoritairement des élèves issus de ces filières. Cette séparation apparaît toutefois moins marquée avec *DBSCAN*, où ces établissements sont en partie classés dans le cluster « bruit ».

Le type de mention obtenue au baccalauréat par les élèves admis dans l'établissement. La variable la plus pertinente dans cette catégorie est *tres_bien_avec_felicitation_mention*, qui met en évidence une forme d'élitisme, le critère étant directement lié au niveau académique des élèves plutôt qu'à leur origine de filière. Ce facteur permet d'identifier les établissements qui recrutent les candidats les plus performants scolairement.

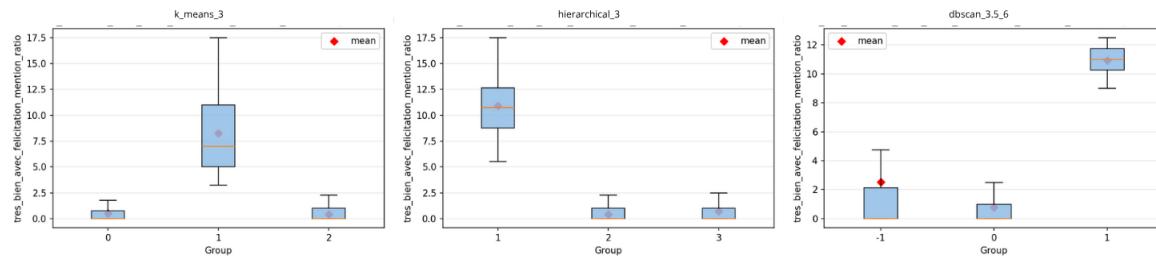


FIGURE 15 – Visualisation par cluster de la caractéristique tres_bien_avec_felicitation_mention_ratio

Les établissements élitistes se démarquent ainsi comme les seuls à admettre une proportion notable d'élèves ayant obtenu cette mention, et ce dans les trois méthodes de clustering. Le cluster représentant les établissements d'élite apparaît donc clairement dans chaque cas, même lorsque le nombre de clusters est limité à trois.

Les informations géographiques des établissements. Nous nous intéressons ici plus particulièrement à la variable *longitude*, qui permet notamment de distinguer les établissements de France métropolitaine de ceux situés en Outre-Mer.

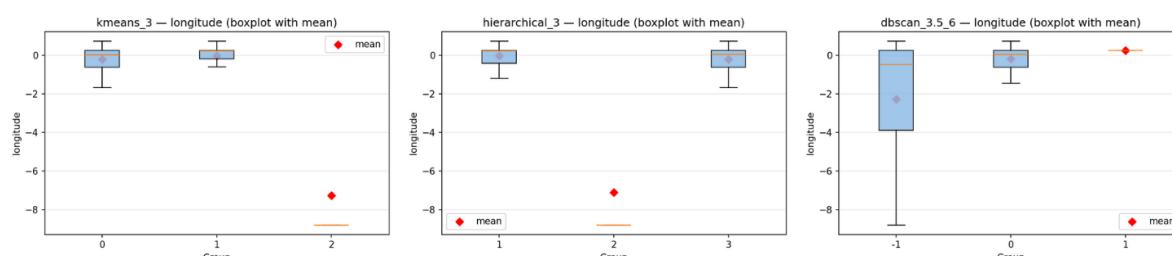


FIGURE 16 – Visualisation par cluster de la caractéristique longitude

Cette variable apporte un éclairage complémentaire sur les regroupements identifiés via le type de baccalauréat. En effet, les établissements accueillant majoritairement des élèves issus de cursus techniques et professionnels appartiennent au même cluster que ceux présentant une longitude négative, c'est-à-dire localisés en Outre-Mer. Pour vérifier cette corrélation, on observe le graphe reliant la longitude au ratio d'élèves admis avec mention « très bien avec félicitations » :

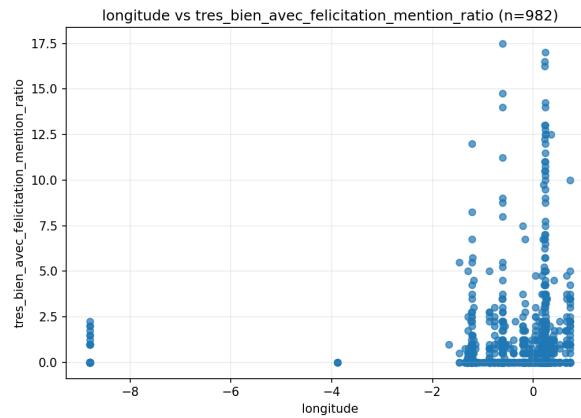


FIGURE 17 – Corrélation entre la longitude et la mention au bac des entrants

Bien que le faible nombre d'établissements ultramarins limite la portée des conclusions, il apparaît qu'aucun d'entre eux n'attire autant d'élèves d'élite que les établissements situés en métropole.

5.2 Autres caractéristiques

Même si la séparation n'est pas toujours parfaitement nette, les trois méthodes de clustering révèlent une distinction claire entre établissements élitistes, établissements spécialisés et autres établissements. Il est donc pertinent d'examiner si certains critères supplémentaires apportent des informations sur la composition de ces trois groupes.

Commençons par étudier l'influence du genre sur cette répartition.

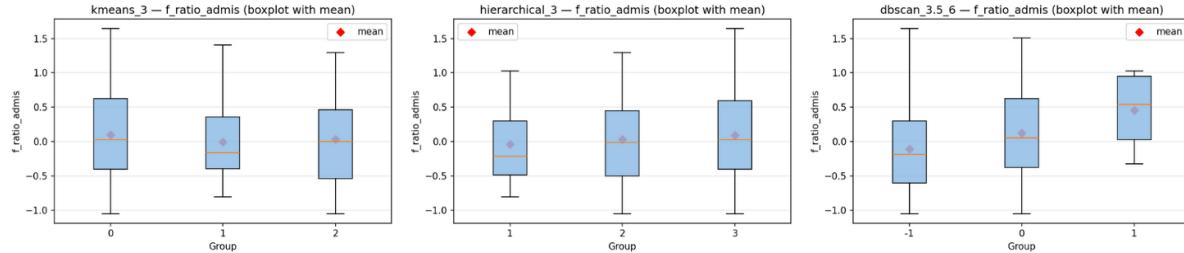


FIGURE 18 – Visualisation par cluster de la caractéristique f_ratio_admis

Aucune corrélation nette n'émerge. Les méthodes *k-means* et hiérarchique suggèrent une légère sous-représentation des femmes dans les établissements élitistes et une sur-représentation dans les établissements spécialisés, mais ces tendances demeurent insuffisantes pour conclure à un phénomène sociétal avéré.

Examinons ensuite la situation des élèves boursiers dans l'enseignement supérieur, à travers les variables représentant le taux de candidats boursiers et celui des élèves boursiers effectivement admis.

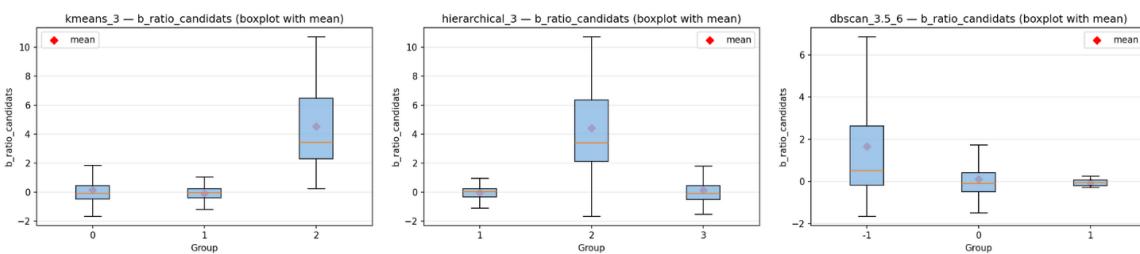


FIGURE 19 – Visualisation par cluster de la caractéristique *b_ratio_candidats*

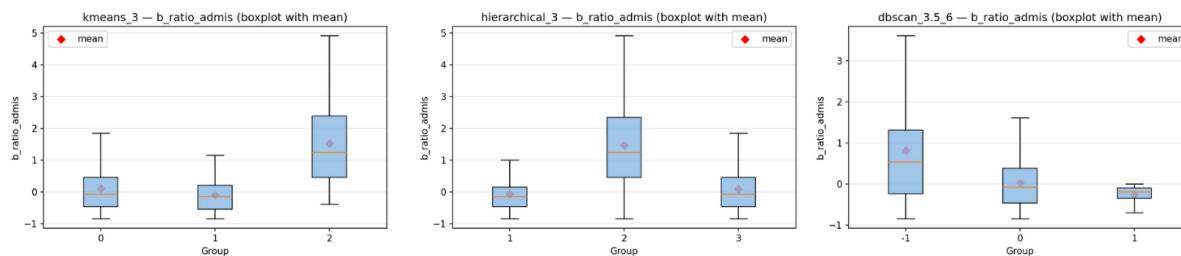


FIGURE 20 – Visualisation par cluster de la caractéristique *b_ratio_admis*

Les élèves boursiers apparaissent clairement regroupés dans le cluster des établissements spécialisés (issus des filières techniques ou professionnelles) et sont sous-représentés dans les établissements élitistes. Les causes de cette répartition sont probablement multiples, mais le constat demeure sans ambiguïté.

Les trois méthodes de clustering produisent ainsi des résultats globalement similaires, mais la méthode *DBSCAN* permet d'expliquer un nombre plus limité de phénomènes. Cette différence s'explique par la présence d'un cluster de bruit dominant, qui regroupe en réalité un mélange de plusieurs sous-groupes. Ce résultat peut être expliqué par l'absence de structures bien définies au sein du jeu de données.

6 Conclusion

Cette analyse de clustering a permis de mettre en évidence trois groupes distincts au sein des établissements d'enseignement supérieur français. On distingue ainsi : un cluster d'établissements spécialisés (techniques et professionnels), incluant notamment ceux situés en Outre-Mer; un cluster général plus hétérogène, regroupant des établissements aux profils variés; et enfin, un cluster clairement identifié d'établissements élitistes. En revanche, la structure interne du cluster général reste plus difficile à caractériser. Ce découpage met également en lumière le déterminisme social présent dans le système éducatif français, qui limite l'accès des élèves boursiers aux établissements les mieux classés. Une tendance similaire, bien que moins marquée, peut être observée concernant la représentation des femmes.

Pour renforcer ces conclusions, il serait pertinent de comparer nos résultats avec des études analogues menées dans d'autres pays d'Europe et à l'international. Une telle comparaison permettrait de déterminer si le système spécifique des CPGE en France contribue à la formation de cette élite ou si des processus similaires existent ailleurs. Par ailleurs, afin d'affiner et de valider statistiquement nos observations, une analyse statistique approfondie serait nécessaire : tests appropriés et estimation d'intervalles de confiance permettraient d'obtenir des mesures d'incertitude plus précises et de quantifier la robustesse des effets observés.

7 Répartition du travail

Étudiant	Code	Rédaction de rapport
Bastien Moron	EDA, Data visualisation, Visualisations de groupes	Problématique, EDA, Visualisations des groupes, Conclusion
Thomas Barand	Prétraitement des données : Nettoyage, normalisation, création de features, Visualisation PCA, Clustering : KNN, DBSCAN, Hierarchical	Traitement des données, Clustering : KNN, DBSCAN, Hierarchical

TABLE 2 – Répartition des tâches entre les étudiants