# Analyse socio-économique des communes des Pyrénées-Orientales

Projet réalisé avec R

## 1 Objectif du projet

L'objectif principal de ce projet est d'utiliser une méthode de classification non supervisée, ici les k-means, afin de regrouper les communes des Pyrénées-Orientales qui présentent des caractéristiques socio-économiques proches.

La problématique retenue est la suivante :

> *Peut-on regrouper les communes des Pyrénées-Orientales selon leurs caractéristiques socio-économiques afin d'identifier différents profils territoriaux ?*

Six variables correspondant à l'année 2023, dernier millésime disponible au moment de la réalisation du projet, ont été sélectionnées pour mener l'étude :

- Le taux de chômage
- Le revenu médian
- La part des familles monoparentales
- Le nombre d'allocataires du RSA pour 1000 habitants de 15 à 64 ans
- La part des actifs ayant uniquement le baccalauréat
- La part des actifs ayant un diplôme de niveau bac+5 ou plus

Ces variables ne résument pas à elles seules la situation sociale d'une commune, mais elles donnent une vision assez large de l'emploi, des revenus, de la structure familiale et du niveau de qualification.

## 2 Données et outils utilisés

Le projet a été réalisé avec R (version 4.6.1) à l'aide de RStudio. R a été choisi car il est couramment utilisé pour l'analyse statistique et permet de produire assez facilement des graphiques et des cartes. Les principaux packages utilisés sont :

- `readxl` pour lire les fichiers Excel
- `dplyr`, `tidyr` et `stringr` pour préparer les données
- `ggplot2` et `factoextra` pour les graphiques statiques
- `cluster` pour le calcul de la silhouette

— `plotly` et `htmlwidgets` pour les graphiques interactifs
— `jsonlite` pour lire les contours géographiques des communes

Les données proviennent principalement de fichiers communaux de l'Insee et de la CNAF. Elles concernent l'emploi, les diplômes, les revenus, les familles, la population et le RSA. Les différentes tables sont rapprochées grâce au code Insee de chaque commune, qui est plus fiable que le nom de la commune pour effectuer les jointures.

Les sources des différentes tables sont les suivantes :

— Taux de chômage et niveau de diplôme des actifs : Insee, Emploi et population active en 2023
— Revenu médian : Insee, Filosofi 2023
— Familles monoparentales : Insee, Couples, familles et ménages en 2023
— Population communale utilisée pour estimer les revenus manquants : Insee, Population en 2023
— Allocataires du RSA pour 1 000 habitants : Observatoire des territoires, ANCT

Une graine de génération égale à 66 est utilisée. Elle permet de retrouver les mêmes résultats à chaque exécution du script, notamment lors de la validation du modèle de revenu et de l'initialisation des k-means.

## 3 Préparation des données

Les fichiers d'origine contiennent des données pour toute la France. Le script sélectionne uniquement les communes dont le code commence par 66. Les indicateurs sont ensuite calculés sous forme de taux afin de pouvoir comparer des communes de tailles différentes.

Le revenu médian n'est pas disponible pour 56 petites communes. Supprimer ces communes aurait créé un biais important dans l'étude. Les revenus manquants sont donc estimés avec une régression linéaire utilisant le chômage, la monoparentalité, le niveau de diplôme et la population de la commune. La population est transformée avec un logarithme pour limiter l'écart entre les petites communes et les villes les plus peuplées, notamment Perpignan, qui est de loin la commune la plus peuplée du département.

Le modèle est évalué par validation croisée en cinq groupes. Il obtient une erreur absolue moyenne d'environ 1423 € et un coefficient $R^2$ de 0,36. La qualité reste moyenne : les estimations sont suffisantes pour conserver les communes dans une analyse exploratoire, mais elles ne peuvent pas remplacer les données officielles dans le cadre d'une analyse plus approfondie.

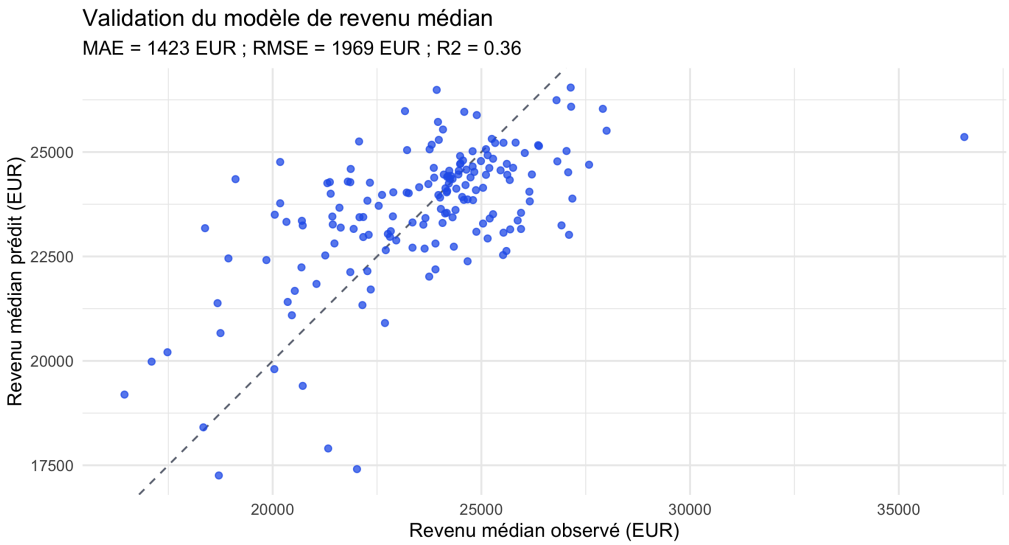



FIGURE 1 – Comparaison entre les revenus observés et les revenus prédits.

La droite en pointillés représente une prédiction parfaite. La majorité des communes se situe autour de cette droite, mais certains écarts restent importants, surtout pour les revenus les plus faibles ou les plus élevés. Le modèle a tendance à rapprocher les prédictions de la moyenne départementale.

### 3.1 Construction du score global

En complément des six indicateurs, un score global est calculé pour proposer une lecture synthétique de la vulnérabilité relative de chaque commune. Pour chaque variable, les communes sont d'abord classées les unes par rapport aux autres avec la fonction `percent_rank()`. Le rang obtenu est ensuite ramené sur une échelle comprise entre 0 et 100.

Le sens de certains indicateurs est inversé afin qu'un score élevé corresponde toujours à une situation considérée comme plus fragile. Ainsi, un chômage, une monoparentalité, un recours au RSA ou une part d'actifs ayant uniquement le bac élevés augmentent le score. À l'inverse, un revenu médian ou une part de diplômés bac+5 faibles augmentent également le score.

Le score global correspond à la moyenne simple des six scores :

$$\text{Score global} = \frac{S_{\text{chômage}} + S_{\text{revenu}} + S_{\text{monoparentalité}} + S_{\text{RSA}} + S_{\text{bac}} + S_{\text{bac+5}}}{6}$$

Une valeur proche de 100 indique donc qu'une commune se situe parmi les communes les plus fragiles du département sur plusieurs indicateurs. Une valeur proche de 0 correspond à une situation globalement plus favorable. Par exemple, un score de 67 ne signifie pas que la commune est

vulnérable à 67 %. Il s'agit d'un indice relatif, qui dépend des communes et des variables présentes dans l'étude.

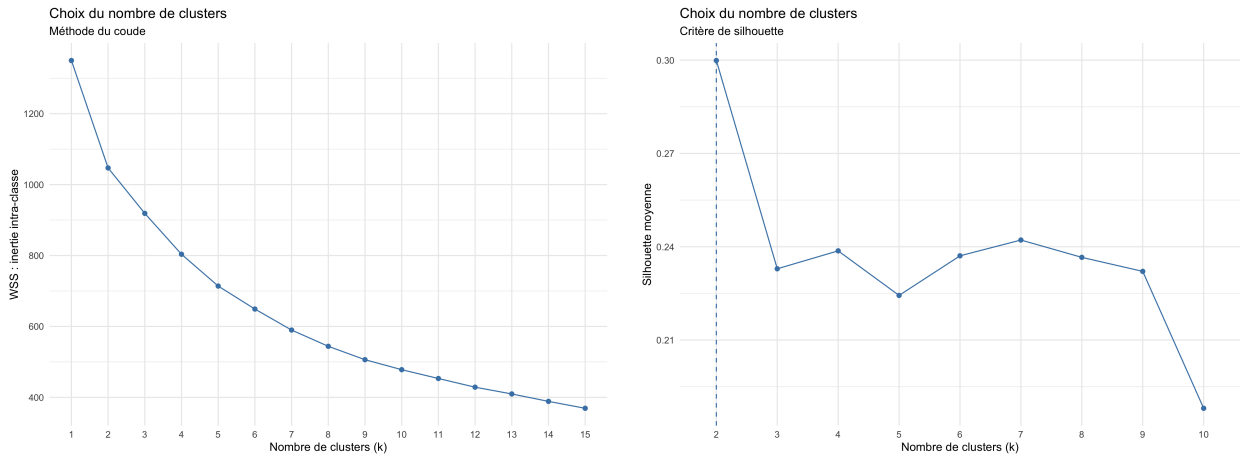
Ce score n'est pas utilisé pour construire les clusters, car cela reviendrait à intégrer une seconde fois les six variables dans le k-means. Il sert uniquement à résumer les profils obtenus et à faciliter leur comparaison. Tous les indicateurs ont le même poids dans son calcul, ce qui constitue un choix méthodologique simple mais discutable selon les objectifs d'une politique publique.

# 4 Méthode de classification

Les variables n'utilisent pas les mêmes unités. Le revenu est exprimé en euros, les autres variables en pourcentages ou pour 1000 habitants. Elles sont donc centrées et réduites avant le clustering (standardisation). Cette standardisation évite que le revenu prenne trop de poids uniquement parce que ses valeurs numériques sont plus grandes.

La méthode des k-means regroupe ensuite les communes en minimisant la distance entre les communes d'un même groupe. Le modèle est relancé 100 fois, grâce au paramètre `nstart`, avec des centres de départ différents, puis la meilleure solution est conservée.

Deux méthodes sont utilisées pour choisir le nombre de groupes : le critère du coude et la silhouette.




(a) Méthode du coude.

(b) Critère de silhouette.

FIGURE 2 – Choix du nombre de clusters.

La courbe du coude diminue fortement pour les premières valeurs de $k$, puis la baisse devient plus régulière. Le coude n'est pas parfaitement net, mais il se situe dans les premières valeurs. La silhouette apporte une réponse plus claire : sa valeur maximale est obtenue pour $k = 2$, avec une silhouette moyenne de 0,30. Cette valeur indique que la séparation existe, mais qu'elle reste assez modérée. Certaines communes ont donc un profil intermédiaire entre plusieurs groupes.

La remontée ponctuelle de la silhouette pour $k = 5$ ne remet pas en cause ce choix. La silhouette n'est pas obligatoirement décroissante. Une nouvelle séparation peut parfois isoler un sous-groupe assez homogène. Toutefois, la valeur obtenue pour $k = 5$ reste inférieure à celle de $k = 2$ et créer cinq profils rendrait l'interprétation beaucoup plus complexe. Pour le projet, $k = 2$ constitue donc

la lecture principale. La solution avec $k = 3$ est conservée comme lecture complémentaire, car elle permet d'affiner les profils tout en restant facilement interprétable.

# 5 Analyse des résultats

## 5.1 Solution avec deux clusters

La solution retenue automatiquement sépare les communes en deux grands profils.

| | Cluster 1 | Cluster 2 |
|---|---|---|
| Nombre de communes | 68 | 158 |
| Taux de chômage moyen | 21,6 % | 11,7 % |
| Revenu médian moyen | 19 823 € | 24 289 € |
| Familles monoparentales | 24,7 % | 14,8 % |
| RSA pour 1000 habitants | 103,0 | 46,1 |
| Actifs ayant uniquement le bac | 25,1 % | 25,0 % |
| Actifs bac+5 ou plus | 8,3 % | 11,1 % |
| Score global moyen | 67,0 | 42,1 |

Table 1 – Profil moyen des deux clusters.

Le cluster 1 regroupe 68 communes présentant globalement davantage de fragilités : le chômage, la monoparentalité et le recours au RSA y sont plus élevés, tandis que le revenu médian et la part de bac+5 sont plus faibles. Le cluster 2, qui contient 158 communes, présente en moyenne une situation plus favorable.

Les écarts sont importants. Le chômage moyen du cluster 1 dépasse celui du cluster 2 d'environ 10 points. Le nombre d'allocataires du RSA y est également plus de deux fois supérieur et le revenu médian est inférieur d'environ 4 500 €. La classification ne repose donc pas sur une seule variable ; plusieurs indicateurs de fragilité évoluent dans la même direction.

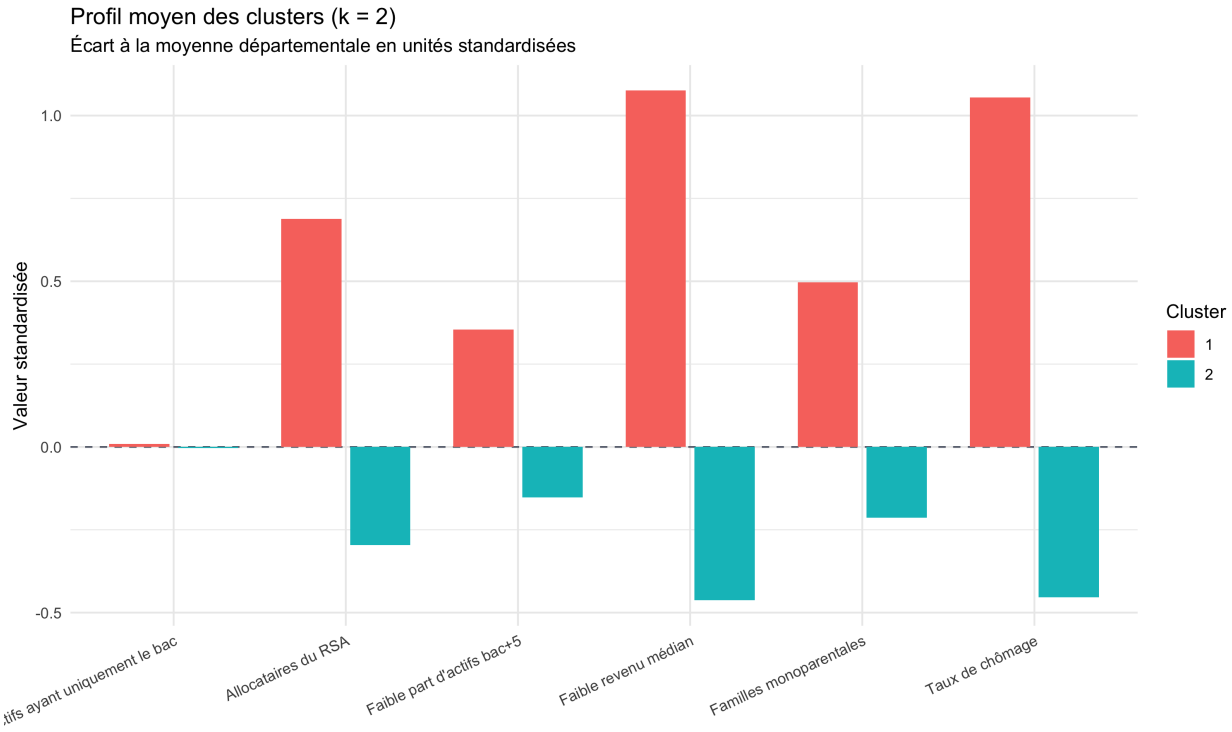



FIGURE 3 – Écart des deux clusters à la moyenne départementale.

Le graphique confirme que les écarts les plus importants concernent le chômage, le revenu médian et le RSA. En revanche, la part des actifs ayant uniquement le bac est presque identique dans les deux groupes. Cette variable participe donc peu à la séparation obtenue avec $k = 2$.

Cette première solution a l'avantage d'être facile à lire, mais elle oppose assez fortement deux ensembles. Elle masque notamment les différences entre les communes intermédiaires et celles qui disposent de revenus ou de niveaux de diplôme plus élevés. C'est pour cette raison qu'une seconde lecture avec trois clusters est utile.

### 5.2 Lecture complémentaire avec trois clusters

Même si $k = 2$ est retenu par la silhouette, une solution avec trois groupes est également produite afin d'obtenir une lecture plus détaillée :

- Le cluster 1 rassemble 57 communes plus fragiles, avec un chômage moyen de 22,8 %, un revenu médian proche de 19 568 € et environ 113 allocataires du RSA pour 1 000 habitants
- Le cluster 2 comprend 39 communes plus favorables, avec le revenu moyen le plus élevé et une part importante d'actifs ayant un bac+5 ou plus
- Le cluster 3 regroupe 130 communes intermédiaires, proches de la moyenne sur plusieurs indicateurs

| | Communes | Chômage | Revenu | Monoparentalité | RSA/1000 | Bac | Bac+5 ou plus | Score |
|---|---|---|---|---|---|---|---|---|
| Cluster 1 | 57 | 22,8 % | 19 568 € | 23,4 % | 113,5 | 24,9 % | 8,1 % | 68,3 |
| Cluster 2 | 39 | 12,9 % | 24 857 € | 17,3 % | 46,3 | 20,6 % | 18,9 % | 32,4 |
| Cluster 3 | 130 | 11,7 % | 23 852 € | 15,4 % | 46,2 | 26,3 % | 8,6 % | 46,6 |

TABLE 2 – Valeurs moyennes des trois profils de communes.

Le cluster 1 est clairement le profil le plus fragile. Il cumule le chômage et le recours au RSA les plus élevés, le revenu le plus faible et davantage de familles monoparentales. Son score global moyen atteint 68,3, contre 46,6 pour le cluster 3 et 32,4 pour le cluster 2. Ces valeurs résument la position relative des communes sur les six indicateurs, mais ne constituent pas un classement officiel des territoires.

Le cluster 2 correspond à un profil plus favorisé et plus diplômé. Son revenu médian moyen est le plus élevé et près de 19 % de ses actifs possèdent un diplôme bac+5 ou plus, soit plus du double des deux autres groupes. Le cluster 3 présente lui aussi un chômage et un RSA relativement faibles, mais se distingue du cluster 2 par une part plus forte d'actifs ayant uniquement le bac et une part beaucoup plus faible de bac+5. Il représente donc un profil intermédiaire plutôt qu'un simple prolongement du groupe favorisé.

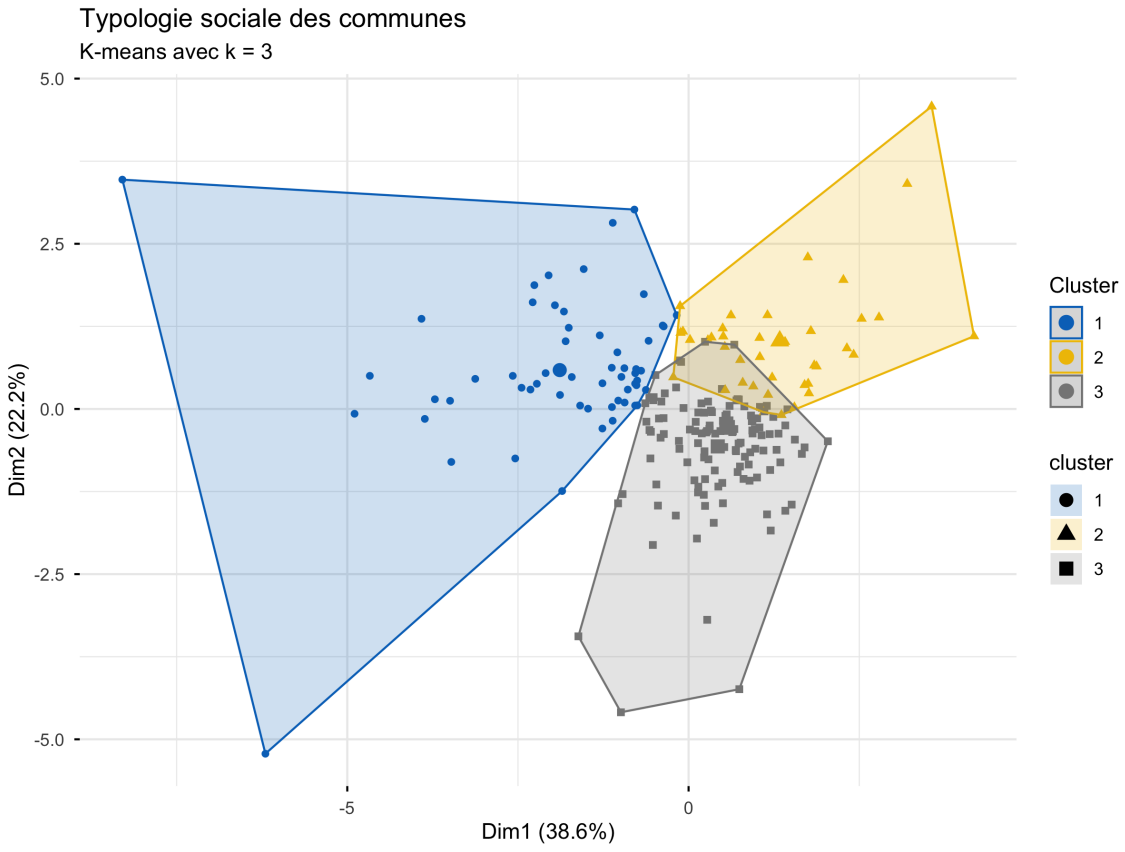



FIGURE 4 – Projection des communes et des trois clusters sur les deux premières dimensions.

Ce graphique représente les communes sur deux dimensions de synthèse. La première explique 38,6 % de l'information et la deuxième 22,2 %, soit 60,8 % au total. Il s'agit donc d'une représentation utile, mais partielle, des six variables initiales. La position des points sert surtout à visualiser les proximités entre communes ; elle ne correspond pas à leur position géographique.

Le cluster 1, en bleu, occupe principalement la partie gauche du graphique et présente une dispersion importante. Cette diversité indique que toutes les communes fragiles ne le sont pas exactement pour les mêmes raisons. Le cluster 2, en jaune, se situe plutôt en haut à droite et se sépare notamment grâce au revenu et à la forte présence de diplômés bac+5. Le cluster 3, en gris, est plus concentré dans la partie basse et centrale. Le chevauchement visible au centre explique pourquoi la silhouette reste modérée, certaines communes se trouvent à la frontière entre deux profils et pourraient changer de groupe si les variables ou l'année étudiée étaient modifiées.

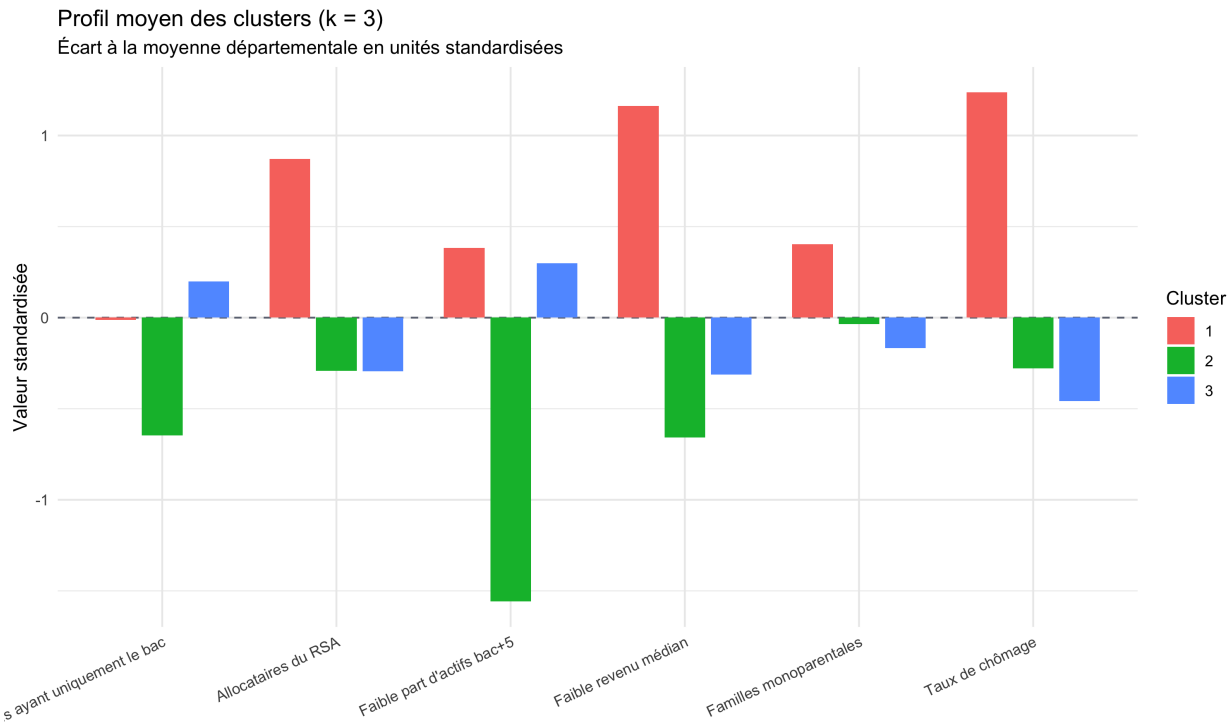



FIGURE 5 – Profil moyen des clusters pour $k = 3$.

La solution avec trois clusters distingue mieux les communes disposant d'une forte proportion de diplômés bac+5. Elle est plus précise, mais aussi un peu moins stable statistiquement que la solution en deux groupes. Elle doit donc être considérée comme une analyse complémentaire.

Le graphique des profils standardisés facilite la comparaison d'indicateurs qui n'ont pas la même unité. Une barre proche de zéro correspond à la moyenne départementale. Le cluster 1 s'écarte nettement de cette moyenne pour le chômage, le faible revenu, le RSA et les familles monoparentales. Le cluster 2 se distingue surtout par sa faible valeur sur l'indicateur "faible part de bac+5", ce qui signifie au contraire que la part réelle de bac+5 y est élevée. Le cluster 3 reste plus proche de la moyenne, avec une part d'actifs ayant uniquement le bac légèrement supérieure.

## 5.3 Analyse de la carte communale

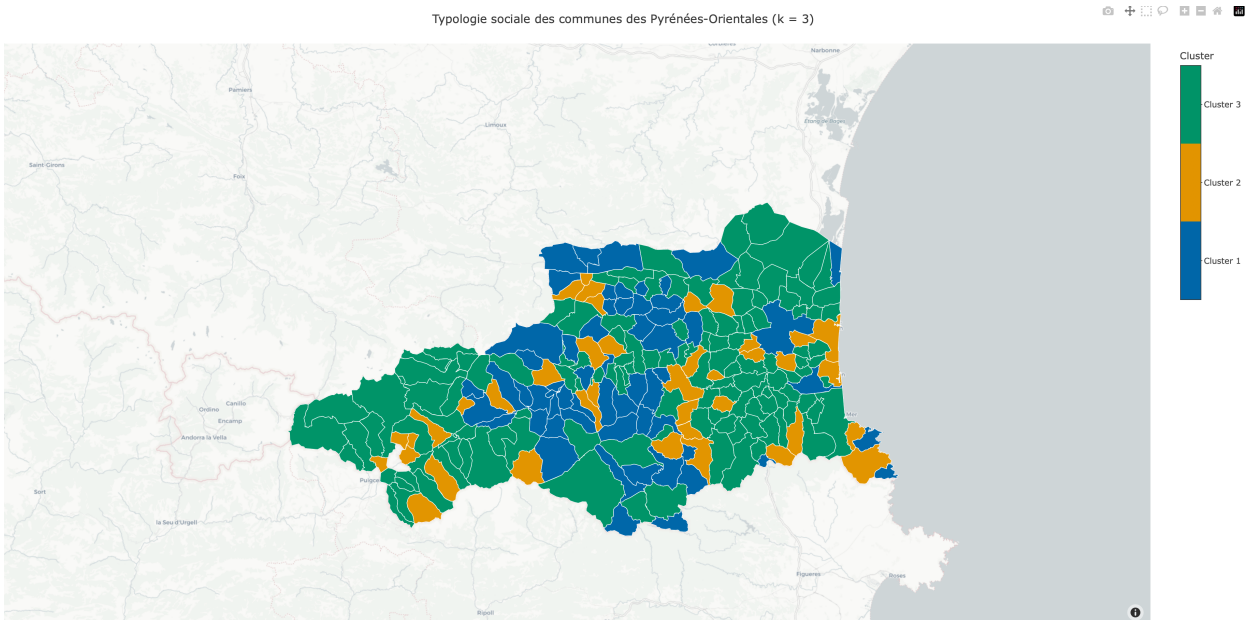



FIGURE 6 – Répartition géographique des communes selon la solution à trois clusters.

Une carte interactive a été réalisée pour la solution avec $k = 3$ ; elle est disponible dans le dossier des résultats graphiques. La carte montre que les clusters ne forment pas trois zones parfaitement continues. Des communes appartenant à des profils différents sont souvent voisines. Cela confirme que le modèle regroupe les territoires selon leurs caractéristiques sociales et économiques, et non selon leur proximité géographique.

Le cluster 1 comprend notamment Perpignan, Prades et Elne. Perpignan présente, dans les données utilisées, un chômage de 22,5 %, un revenu médian de 19 110 € et environ 158 allocataires du RSA pour 1000 habitants de 15 à 64 ans. Prades possède également un chômage élevé et un recours important au RSA. Ce groupe contient aussi de petites communes rurales comme Prugnanes ou Le Tech. Leur présence rappelle qu'un même profil statistique peut concerner aussi bien un pôle urbain que des territoires peu peuplés.

Le cluster 2 rassemble plusieurs communes au profil plus favorisé ou plus diplômé, parmi lesquelles Cabestany, Collioure, Canet-en-Roussillon et Saint-Cyprien. Cabestany possède par exemple un revenu médian proche de 26 360 €, une part de bac+5 de 14,8 % et un niveau de RSA plus faible. Collioure présente également un revenu élevé. Canet-en-Roussillon et Saint-Cyprien ont toutefois un chômage supérieur à la moyenne de leur cluster. Leur appartenance s'explique par la combinaison de l'ensemble des variables et non par un seul indicateur.

Le cluster 3, le plus nombreux, est largement réparti dans le département. Il comprend notamment Rivesaltes, Thuir, Céret, Argelès-sur-Mer, Bompas ou encore Font-Romeu-Odeillo-Via. Ces communes ont des situations différentes mais restent globalement intermédiaires dans l'espace formé par les six variables. Rivesaltes, par exemple, présente un chômage de 17 % et un revenu médian de 22 080 €, tandis que Céret dispose d'un revenu plus élevé. La carte interactive complète cette lecture en permettant d'afficher les indicateurs de chaque commune au survol.

Cette répartition peut servir d'outil d'aide au diagnostic. Elle permet de repérer des communes

confrontées à des enjeux comparables, même lorsqu'elles sont éloignées, et d'imaginer des analyses ou des échanges entre territoires de même profil. Elle ne permet cependant pas, à elle seule, de définir une priorité d'intervention publique.

## 6 Limites et conclusion

Cette étude reste exploratoire. Le choix des six variables influence directement les groupes obtenus et les indicateurs ont tous le même poids. Par ailleurs, les revenus de 56 communes sont estimés, et les taux des très petites communes peuvent varier fortement à cause de leurs faibles effectifs. Cette instabilité est visible pour certaines petites communes affichant des taux très élevés ou très faibles. Enfin, les clusters décrivent des ressemblances statistiques : ils ne constituent ni un classement officiel, ni une mesure directe des besoins sociaux.

Malgré ces limites, le projet fournit une méthode reproductible pour comparer les communes du département des Pyrénées-Orientales. La solution en deux clusters donne une lecture simple et statistiquement plus nette. La solution en trois clusters apporte une information supplémentaire en isolant un profil plus diplômé et un profil intermédiaire. L'association des tableaux, des graphiques et de la carte permet ainsi de passer d'un résultat statistique à une lecture territoriale plus concrète.