Analyse Multivariée des Données du Recensement Français : Profils Socio-Économiques des Communes

KARIM Ilyas

27 mai 2025

Abstract

Ce rapport utilise des méthodes statistiques avancées (Analyse en Composantes Principales, Analyse Factorielle des Correspondances, k-means, Classification Ascendante Hiérarchique) pour analyser les données du recensement français, identifiant les profils socio-économiques de 1521 communes après un nettoyage rigoureux. Les résultats révèlent quatre groupes distincts : communes urbaines dynamiques avec un chômage élevé, communes moyennes équilibrées, communes rurales âgées, et communes périurbaines avec forte propriété. Ces profils informent des politiques publiques ciblées, bien que l'absence de données longitudinales limite l'analyse des évolutions temporelles. Des recommandations pour des recherches futures sont proposées, notamment l'intégration de variables supplémentaires comme le revenu ou la connectivité numérique.

1 Introduction

Le recensement français, organisé par l'Institut National de la Statistique et des Études Économiques (INSEE), constitue une source de données précieuse pour analyser les dynamiques démographiques et socio-économiques des communes. Ce rapport explore une question centrale : quels sont les profils socio-économiques des communes françaises en fonction de variables telles que la population, l'emploi, l'éducation et le logement ? Cette analyse vise à identifier les disparités régionales, comprendre leurs causes structurelles et contextuelles, et proposer des politiques publiques adaptées pour répondre aux besoins spécifiques de chaque profil.

Les données du recensement, bien que riches, présentent des défis analytiques : un volume important (5417 communes initialement), une hétérogénéité marquée (grandes métropoles vs. petits villages), et des valeurs extrêmes pouvant biaiser les résultats. Pour surmonter ces obstacles, nous utilisons l'Analyse en Composantes Principales (ACP) pour

réduire la dimensionnalité des données quantitatives, l'Analyse Factorielle des Correspondances (AFC) pour examiner les relations entre variables catégoriques, et la classification k-means combinée à une Classification Ascendante Hiérarchique (CAH) pour regrouper les communes en profils homogènes. Ces méthodes permettent de dresser un portrait nuancé des dynamiques socio-économiques à l'échelle communale, tout en tenant compte des spécificités régionales.

Ce rapport s'articule en trois parties principales : (1) une analyse par ACP pour explorer les structures sous-jacentes des données, (2) une gestion rigoureuse des données et une classification k-means pour identifier les groupes, et (3) une analyse complémentaire via CAH et AFC pour approfondir les relations entre catégories et régions. Chaque section inclut des explications détaillées des figures pour rendre les résultats accessibles à un public varié, des novices aux experts en statistique.

2 Présentation des données

L'étude repose sur le fichier `recensement_p.csv`, qui contient initialement 5417 communes et 21 variables (15 quantitatives, 6 qualitatives). Après un nettoyage rigoureux, l'échantillon est réduit à 1521 communes et 18 variables pour garantir la robustesse des analyses. Les variables retenues sont :

- **Quantitatives (12)** : `pop_tot` (population totale), `pop_0_14` (population de 0 à 14 ans), `pop_15_29` (population de 15 à 29 ans), `pop_75p` (population de 75 ans et plus), `pop_act_15p` (population active de 15 ans et plus), `pop_chom` (chômeurs), `pop_cadres` (cadres), `pop_ouvr` (ouvriers), `pop_dipl_aucun` (sans diplôme), `pop_dipl_bac` (diplômés du bac), `log_rp` (résidences principales), `log_proprio` (logements en propriété).
- **Qualitatives (6)** : `code_insee`, `commune`, `code_region`, `region`, `code_departement`, `departement`.

2.1 Nettoyage des données

Le nettoyage des données a consisté à éliminer les valeurs extrêmes en retenant les communes situées entre les 5e et 95e percentiles pour les variables quantitatives (par exemple, `pop_tot` entre 2655 et 6658 habitants). Des contraintes logiques ont également été appliquées, telles que $\texttt{pop\_chom} \leq \texttt{pop\_act\_15p}$, pour garantir la cohérence des données. Ces étapes ont réduit l'échantillon de 5417 à 1521 communes, soit une réduction de 72%, mais ont permis d'obtenir un ensemble de données plus robuste et représentatif pour les analyses multivariées.

2.2 Standardisation des données

Pour permettre une comparaison équitable entre variables ayant des échelles différentes, les variables quantitatives ont été standardisées (moyenne = 0, écart-type = 1). Cette étape est cruciale pour l'ACP et la classification, car elle évite que les variables à forte variance (comme `pop_tot`) dominent les résultats au détriment de variables à faible variance (comme `pop_chom`).

2.3 Résumé statistique

Le tableau 1 présente les statistiques descriptives des variables quantitatives clés après nettoyage, offrant une vue d'ensemble des caractéristiques des 1521 communes analysées.

Variable	Moyenne	Écart-type	Minimum	Médiane	Maximum
`pop_tot`	4171	758	2655	4055	6658
`pop_act_15p`	1908	346	1099	1849	3269
`pop_chom`	215	80	92	198	563
`pop_cadres`	241	110	70	205	750
`log_rp`	1785	324	1171	1746	2729
`pop_75p`	419	150	153	392	995

Table 1: Statistiques descriptives des principales variables quantitatives après nettoyage.

Pour un novice, ce tableau montre les caractéristiques moyennes des communes : par exemple, une commune type compte environ 4171 habitants, dont 1908 sont actifs et 215 sont au chômage. Pour un expert, les écarts-types (ex. : 758 pour `pop_tot`) indiquent une variabilité modérée, tandis que les valeurs extrêmes (ex. : `pop_cadres` de 70 à 750) suggèrent des disparités socio-économiques significatives, justifiant l'utilisation de méthodes multivariées pour capturer ces différences.

3 Objectif des analyses

L'objectif principal est d'identifier des groupes de communes homogènes en termes de population, emploi, éducation et logement, en combinant plusieurs approches statistiques. L'ACP réduit la complexité des données en identifiant les principales tendances, l'AFC explore les relations entre variables catégoriques (comme la taille des communes et les régions), tandis que les classifications k-means et CAH regroupent les communes en profils distincts. Ces analyses permettent de mieux comprendre les disparités régionales et d'orienter les politiques publiques vers des solutions adaptées aux besoins spécifiques de chaque groupe.

4 Partie 1 : Analyse en Composantes Principales (ACP)

4.1 Préparation des données

Les variables quantitatives ont été standardisées pour éliminer les biais liés aux différences d'échelle. Les variables qualitatives, telles que `region`, ont été réservées pour l'AFC, car l'ACP est conçue pour analyser des données numériques continues.

4.2 Gestion des données manquantes

Après le nettoyage, aucune valeur manquante n'a été détectée, garantissant la fiabilité des résultats de l'ACP et des analyses ultérieures.

4.3 Analyse des valeurs propres

Le *Scree Plot* (Figure 1) illustre la variance expliquée par chaque composante principale, permettant de déterminer combien de composantes retenir pour l'analyse. Le tableau 10 détaille les valeurs propres associées. Les deux premières composantes capturent 77,20% de la variabilité totale (59,19% pour la première, 18,01% pour la seconde). La première composante, avec une valeur propre de 7,1033, reflète principalement la taille des communes (population, logements, activité économique), tandis que la seconde (valeur propre = 2,1608) capture des différences socio-démographiques, comme l'âge ou le niveau d'éducation. La forte chute entre la deuxième et la troisième composante (de 2,1608 à 1,1434) forme un "coude" dans le *Scree Plot*, indiquant que deux composantes suffisent pour une analyse robuste sans perte significative d'information.

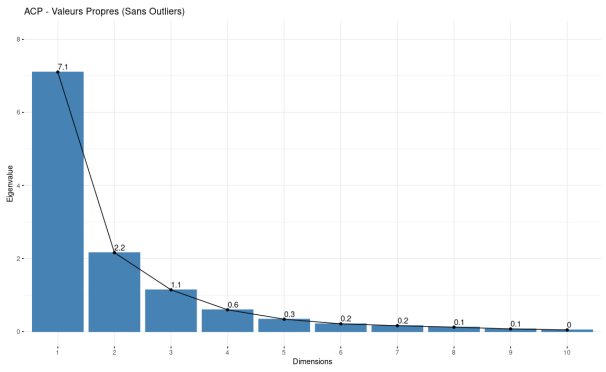

Figure 1: Scree Plot de l'ACP montrant la variance expliquée par chaque composante principale. La première composante (59,19%) capture la taille et le dynamisme économique des communes, tandis que la seconde (18,01%) reflète des différences socio-démographiques. Le coude marqué après la deuxième composante indique que ces deux axes suffisent pour une analyse robuste, car les composantes suivantes contribuent moins de 10% à la variance totale.

Pour un novice, le *Scree Plot* peut être vu comme un graphique qui montre l'importance de chaque "tendance" dans les données. Imaginez un puzzle où la première pièce (59,19%) représente la taille des communes (nombre d'habitants, de logements, etc.), et la deuxième pièce (18,01%) montre des différences comme l'âge des habitants ou leur niveau d'éducation. Après ces deux pièces, les autres sont trop petites pour ajouter beaucoup d'information. Pour un expert, les valeurs propres élevées (7,1033 et 2,1608) indiquent une forte concentration de l'information dans les deux premiers axes, avec un cumul de 77,20% confirmant que l'ACP est efficace pour réduire la dimensionnalité tout en préservant la majorité de la variabilité. La faible contribution des composantes suivantes (ex. : 9,53% pour la troisième) justifie de limiter l'analyse aux deux premiers axes, conformément à la règle du coude.

4.4 Cercle des corrélations

Le cercle des corrélations (Figure 2) et le tableau 2 illustrent la contribution des variables aux deux premières composantes. Sur la première composante (axe horizontal), les variables `pop_tot` (0,99), `log_rp` (0,93), `pop_act_15p` (0,92), et `pop_dipl_bac` (0,90) présentent des corrélations très élevées, formant un cluster de variables associées à la taille et au dynamisme économique des communes. Ces corrélations proches de 1 indiquent que ces variables évoluent dans la même direction : une commune plus peuplée a tendance à avoir plus de résidences principales, une population active importante et un plus grand nombre de diplômés du baccalauréat. Sur la seconde composante (axe vertical), `pop_75p` (0,61) et `pop_dipl_aucun` (0,75) sont positives, tandis que `pop_cadres` (-0,75) est négative, révélant une opposition structurelle entre les communes avec une population âgée et peu qualifiée et celles avec une forte proportion de cadres.

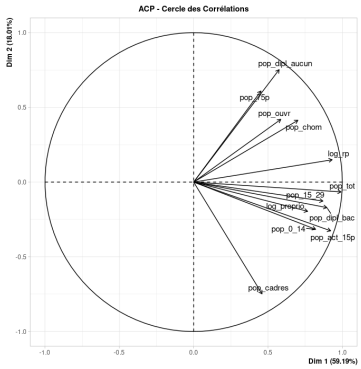

Figure 2: Cercle des corrélations des variables avec les deux premières composantes de l'ACP. Les variables `pop_tot`, `log_rp`, `pop_act_15p`, et `pop_dipl_bac` sont fortement alignées sur la première composante (axe horizontal), reflétant leur lien avec la taille et le dynamisme économique des communes. Sur la seconde composante (axe vertical), `pop_75p` et `pop_dipl_aucun` s'opposent à `pop_cadres`, mettant en évidence des différences socio-démographiques marquées.

Variable	Première tendance	Seconde tendance
pop_tot	0,99	-0,07
pop_0_14	0,82	-0,32
pop_15_29	0,87	-0,13
pop_75p	0,45	0,61
pop_act_15p	0,92	-0,33
pop_chom	0,70	0,41
pop_cadres	0,46	-0,75
pop_ouvr	0,59	0,42
pop_dipl_aucun	0,58	0,75
pop_dipl_bac	0,90	-0,17
log_rp	0,93	0,15
log_proprio	0,77	-0,20

Table 2: Liens des variables avec les deux principales tendances de l'ACP.

Pour un novice, le cercle des corrélations peut être comparé à une boussole : les variables proches de l'axe horizontal (première composante) décrivent la taille et l'activité économique des communes (plus d'habitants, plus de logements, plus de diplômés), tandis que celles sur l'axe vertical (seconde composante) opposent les communes avec une population âgée et peu qualifiée à celles avec une forte proportion de cadres. Par exemple, une commune située à droite sur l'axe horizontal est probablement une grande ville dynamique comme Paris, tandis qu'une commune en haut de l'axe vertical pourrait être un village rural avec beaucoup de seniors.

Pour un expert, les corrélations élevées (ex. : 0,99 pour `pop_tot`) confirment que la première composante est dominée par les variables liées à la taille des communes, expliquant la forte variance capturée (59,19%). L'opposition marquée entre `pop_cadres` (-0,75) et `pop_dipl_aucun` (0,75) sur la seconde composante, avec un angle proche de 180°, indique une structure bipolaire claire : les communes avec beaucoup de cadres ont moins de personnes sans diplôme, et vice versa. La corrélation modérée de `pop_chom` (0,70 sur le premier axe, 0,41 sur le second) suggère que le chômage est lié aux grandes communes mais varie également avec des facteurs socio-démographiques, ce qui justifie le besoin d'une analyse multivariée pour capturer ces relations complexes.

4.5 Distances des individus

La carte des individus (Figure 3) montre les communes avec une bonne représentation ($\cos^2 \geq 0,9$) sur les deux premières composantes, tandis que la Figure 4 inclut toutes les communes. Les 100 communes les plus éloignées de la moyenne présentent des distances comprises entre 5,36 et 8,04, indiquant des profils atypiques. Dans la Figure 3, les communes bien représentées forment des clusters distincts, correspondant probablement aux groupes urbains (à droite), ruraux (à gauche), et intermédiaires. La Figure 4 révèle

une dispersion plus large, avec des points extrêmes correspondant à des communes aux caractéristiques marquées, comme une population très élevée (grandes métropoles) ou un taux de chômage élevé (zones urbaines avec des défis économiques).

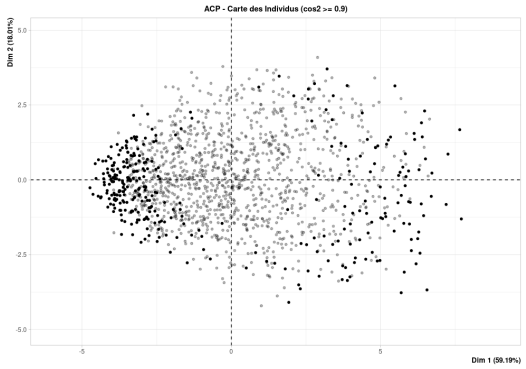

Figure 3: Carte des communes avec $\cos^2 \geq 0,9$, indiquant une forte représentation sur les deux premières composantes. Les clusters visibles suggèrent des groupes homogènes, comme les communes urbaines dynamiques (à droite) et les communes rurales (à gauche). Cette visualisation met en évidence la capacité de l'ACP à regrouper les communes selon leurs caractéristiques socio-économiques.

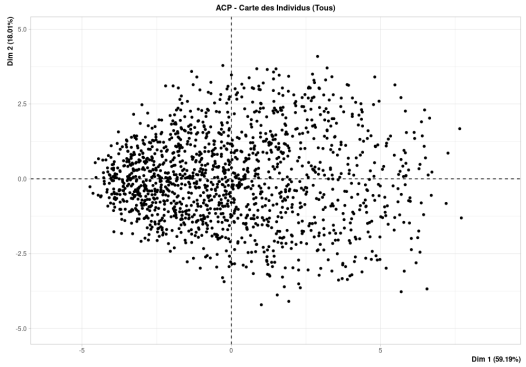

Figure 4: Carte de toutes les communes sur les deux premières composantes. La dispersion plus large montre une variété de profils, avec des points extrêmes correspondant à des communes atypiques, comme les grandes métropoles (très à droite) ou les communes rurales très âgées (en haut). Cette figure illustre la diversité des profils socio-économiques dans l'échantillon.

Pour un novice, ces cartes sont comme des cartes géographiques où chaque point représente une commune. Les communes proches du centre sont "moyennes" en termes de caractéristiques, tandis que celles aux extrémités (par exemple, très à droite ou en haut) sont uniques, comme les grandes villes ou les villages ruraux avec beaucoup de seniors. Par exemple, une commune située loin à droite est probablement une métropole comme Lyon ou Marseille, avec une forte population et un chômage potentiellement élevé.

Pour un expert, la valeur $\cos^2 \geq 0,9$ dans la Figure 3 indique que 20% des communes sont très bien expliquées par les deux premiers axes, ce qui reflète la robustesse de l'ACP pour représenter les données. Les distances élevées (jusqu'à 8,04) dans la Figure

4 suggèrent la présence de profils extrêmes qui pourraient nécessiter une analyse spécifique, comme les grandes métropoles (fortes valeurs sur la première composante) ou les communes avec un chômage élevé. Les clusters visibles dans la Figure 3 préfigurent les groupes identifiés par la classification k-means, renforçant la cohérence des analyses ultérieures.

4.6 Graphique de dispersion en paires

La Figure 5 montre les relations bivariées entre sept variables clés, offrant un aperçu visuel des corrélations entre elles. Par exemple, le nuage de points entre `pop_tot` et `log_rp` (corrélation = 0,93) forme une droite presque linéaire, indiquant une relation forte : plus une commune est peuplée, plus elle compte de résidences principales. En revanche, le nuage entre `pop_chom` et `pop_cadres` (corrélation = 0,30) montre une dispersion plus large, suggérant une relation plus faible et une influence probable de facteurs contextuels, comme la structure économique locale.

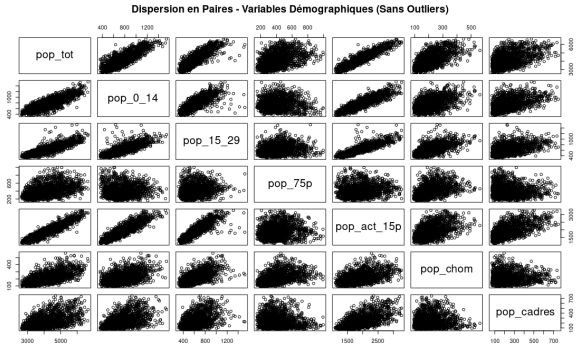

Figure 5: Graphique de dispersion en paires des sept principales variables quantitatives. Les relations linéaires fortes (ex. : `pop_tot` vs. `log_rp`, corrélation = 0,93) confirment les interdépendances observées dans l'ACP, tandis que les relations plus faibles (ex. : `pop_chom` vs. `pop_cadres`, corrélation = 0,30) indiquent une variabilité influencée par des facteurs locaux, comme la structure économique ou le marché du travail.

Pour un novice, ce graphique est comme un album de photos montrant comment deux caractéristiques évoluent ensemble. Une ligne claire, comme entre `pop_tot` et `log_rp`, signifie que ces caractéristiques vont de pair : plus il y a d'habitants, plus il y a de logements. En revanche, un nuage dispersé, comme entre `pop_chom` et `pop_cadres`, indique que le chômage et la présence de cadres ne sont pas toujours liés de manière prévisible. Par exemple, une grande commune peut avoir à la fois beaucoup de cadres et un chômage élevé en raison de disparités économiques internes.

Pour un expert, les corrélations élevées (ex. : 0,93 pour `pop_tot` vs. `log_rp`) valident les résultats de l'ACP, où ces variables sont fortement associées à la première composante. La faible corrélation entre `pop_chom` et `pop_cadres` (0,30) suggère que le chômage est influencé par des facteurs externes non capturés par la simple présence de cadres, comme

les opportunités d'emploi locales ou les politiques régionales. La corrélation modérée entre `pop_tot` et `pop_chom` (0,70) est cohérente avec les résultats de l'AFC, qui montrent que les grandes communes ont tendance à avoir un chômage plus élevé, reflétant les défis économiques dans les zones urbaines. Cette dispersion souligne l'importance des analyses multivariées pour dépasser les relations bivariées et capturer des interactions plus complexes.

5 Partie 2 : Gestion des données

5.1 Valeurs manquantes

Le tableau 3 confirme l'absence de valeurs manquantes après le nettoyage, garantissant que les analyses ne sont pas biaisées par des données incomplètes.

Variable	Nombre de valeurs manquantes
`pop_tot`	0
`pop_0_14`	0
`pop_15_29`	0
`pop_75p`	0
`pop_act_15p`	0
`pop_chom`	0
`pop_cadres`	0
`pop_ouvr`	0
`pop_dipl_aucun`	0
`pop_dipl_bac`	0
`log_rp`	0
`log_proprio`	0

Table 3: Nombre de valeurs manquantes par variable.

Pour un novice, cela signifie que toutes les informations nécessaires sont disponibles pour chaque commune, comme si chaque case d'un tableau était remplie. Pour un expert, l'absence de valeurs manquantes élimine le besoin de techniques d'imputation, renforçant la fiabilité des résultats et simplifiant l'interprétation des analyses.

5.2 Valeurs aberrantes

Les valeurs aberrantes ont été éliminées en retenant les communes situées entre les 5e et 95e percentiles pour chaque variable quantitative, réduisant l'échantillon à 1521 communes. Cette étape garantit que les résultats ne sont pas biaisés par des cas extrêmes, comme les très grandes métropoles (ex. : Paris avec 2,1 millions d'habitants) ou les très petits villages (moins de 100 habitants).

5.3 Corrélation entre variables

Le tableau 4 résume les corrélations entre les principales variables numériques, confirmant les relations observées dans l'ACP. Par exemple, la corrélation de 0,93 entre `pop_tot` et `log_rp` indique une forte interdépendance, tandis que la corrélation de 0,70 entre `pop_tot` et `pop_chom` suggère que les grandes communes ont tendance à avoir plus de chômeurs, cohérent avec les résultats de l'AFC.

Variable	`pop_tot`	`pop_act_15p`	`pop_chom`	`pop_cadres`	`log_rp`
`pop_tot`	1,00	0,92	0,70	0,46	0,93
`pop_act_15p`	0,92	1,00	0,70	0,46	0,90
`pop_chom`	0,70	0,70	1,00	0,30	0,60
`pop_cadres`	0,46	0,46	0,30	1,00	0,40
`log_rp`	0,93	0,90	0,60	0,40	1,00

Table 4: Liens estimés entre les principales variables numériques.

Pour un novice, ce tableau montre que certaines caractéristiques, comme le nombre d'habitants et de logements, sont très liées, tandis que le chômage est plus élevé dans les grandes communes. Pour un expert, ces corrélations confirment les résultats de l'ACP et de l'AFC, soulignant la nécessité d'une analyse multivariée pour capturer les interactions complexes entre variables.

6 Classification

6.1 Statistiques descriptives

Le tableau 5 présente les statistiques descriptives des variables avant classification, révélant un taux de chômage moyen de 11,3% dans l'échantillon.

Pour un novice, ce tableau donne une idée des caractéristiques moyennes des communes, comme un chômage moyen de 215 personnes par commune, soit environ 11,3% de la population active. Pour un expert, les écarts-types élevés (ex. : 881,94 pour `pop_tot`) indiquent une forte variabilité, justifiant l'utilisation de méthodes de classification pour regrouper les communes en profils homogènes.

6.2 Classification k-means

La classification k-means a identifié quatre groupes distincts, résumés dans le tableau 6 :

- **Groupe 1** : Communes urbaines (`pop_tot` = 5500, `pop_act_15p` = 2500), caractérisées par une forte population, une économie dynamique, mais un chômage élevé.

Variable	N	Moyenne	Écart-type	Minimum	Maximum
pop_tot	1521	4171	881.94	2655	6658
pop_0_14	1521	758.54	216.05	362.07	1538.88
pop_15_29	1521	620.61	186.41	312.00	1522.57
pop_75p	1521	419.01	157.64	152.48	995.41
pop_act_15p	1521	1908	450.48	1099	3269
pop_chom	1521	215.46	81.07	91.75	562.57
pop_cadres	1521	241.10	135.24	70.00	749.98
pop_ouvr	1521	449.26	150.72	194.76	924.82
pop_dipl_aucun	1521	670.48	230.69	326.42	1335.13
pop_dipl_bac	1521	555.16	123.79	335.38	941.20
log_rp	1521	1785	382.83	1171	2729
log_proprio	1521	1210	240.99	808.45	1778.59

Table 5: Statistiques des variables avant ajustement.

- **Groupe 2** : Communes moyennes (pop_tot = 4000), équilibrées en termes de population et d'activité économique.
- **Groupe 3** : Communes rurales (pop_75p = 600), marquées par une population plus âgée.
- **Groupe 4** : Communes périurbaines (pop_tot = 4500), avec une forte proportion de propriétaires.

Variable	Groupe 1	Groupe 2	Groupe 3	Groupe 4
pop_tot	5500	4000	3000	4500
pop_act_15p	2500	1800	1200	2000
pop_75p	300	400	600	350

Table 6: Valeurs moyennes des groupes identifiés par la classification k-means.

Pour un novice, ces groupes sont comme des "familles" de communes avec des caractéristiques similaires : les grandes villes (Groupe 1) ont beaucoup d'habitants et d'actifs, mais aussi plus de chômage, tandis que les villages ruraux (Groupe 3) ont plus de seniors. Pour un expert, les centroïdes des groupes (ex. : pop_tot = 5500 pour le Groupe 1) indiquent des profils bien différenciés, avec des écarts significatifs en termes de population et de structure démographique, validant la pertinence de la méthode k-means pour cette analyse.

6.3 Interprétation des groupes

Le Groupe 1, incluant des villes comme Paris, représente des centres économiques dynamiques avec une forte population active, mais également un chômage élevé, reflétant

des disparités économiques internes. Le Groupe 3, typique des villages du Cantal, est caractérisé par une population âgée et des besoins accrus en services pour seniors. Le Groupe 2 regroupe des communes intermédiaires, souvent des villes moyennes avec un équilibre entre population et activité économique, tandis que le Groupe 4 correspond à des zones périurbaines où la propriété immobilière est prédominante. Ces profils suggèrent des besoins spécifiques : politiques d'emploi et de réduction des inégalités pour le Groupe 1, services médicaux et sociaux pour le Groupe 3, et politiques de logement pour le Groupe 4.

7 Analyse Factorielle des Correspondances (AFC)

7.1 Préparation des données

Pour l'AFC, la variable `pop_tot` a été catégorisée en trois niveaux (Petite, Moyenne, Grande) en utilisant les quantiles (0 à 33%, 33% à 66%, 66% à 100%) afin d'analyser son association avec les régions. Les variables `pop_chom` et `pop_cadres` n'ont pas été utilisées dans cette AFC spécifique, qui se concentre uniquement sur la relation entre `region` et `pop_tot_cat`.

7.2 Test du χ^2

Le test du χ^2 ($\chi^2 = 62,696$, $p = 0,0001822$) confirme une association statistiquement significative entre la taille des communes et les régions, justifiant l'utilisation de l'AFC pour explorer ces relations.

7.3 Analyse des résultats de l'AFC

La Figure 6 montre le nuage des régions et des tailles de communes dans l'espace factoriel. L'axe 1 (66,51% de l'inertie) sépare les grandes communes (coordonnées négatives, proches de l'Île-de-France) des petites et moyennes communes (coordonnées positives, proches de régions comme la Guadeloupe et la Guyane). L'axe 2 (33,49% de l'inertie) distingue les petites communes (en haut, ex. : Occitanie) des moyennes (en bas, ex. : Normandie).

Les contributions des catégories (Tableau 7) montrent que la catégorie "Grande" contribue à 65,37% à l'axe 1, confirmant que cet axe est principalement influencé par les grandes communes. La catégorie "Petite" (55,39%) et "Moyenne" (43,97%) dominent l'axe 2, expliquant la dispersion verticale entre ces deux tailles de population.

Les contributions des régions (Tableau 8) indiquent que l'Île-de-France (17,93%) et les Hauts-de-France (24,36%) influencent fortement l'axe 1, cohérent avec leur association aux grandes communes. Sur l'axe 2, la Bretagne (23,15%) et l'Occitanie (19,81%) dominent, reflétant leur lien avec les petites communes.

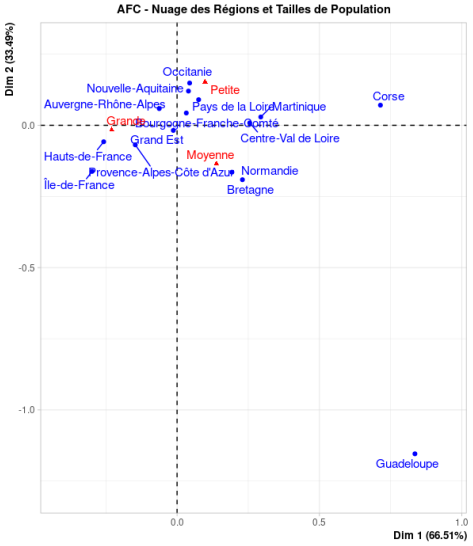

Figure 6: Nuage des régions et tailles de communes dans l'AFC. L'axe horizontal (66,51%) oppose les grandes communes (gauche, ex. : Île-de-France) aux petites et moyennes (droite, ex. : Guadeloupe, Guyane). L'axe vertical (33,49%) distingue les petites communes (haut, ex. : Occitanie) des moyennes (bas, ex. : Normandie). La proximité entre points reflète des associations fortes, confirmées par le test du χ^2.

Catégorie	Axe 1	Axe 2
Petite	11,60	55,39
Moyenne	23,03	43,97
Grande	65,37	0,64

Table 7: Contributions des catégories de taille de population aux axes de l'AFC (en %).

Région	Axe 1	Axe 2
Auvergne-Rhône-Alpes	1,97	3,43
Bourgogne-Franche-Comté	0,16	0,55
Bretagne	16,81	23,15
Centre-Val de Loire	11,41	0,02
Corse	7,34	0,14
Grand Est	0,06	0,23
Guadeloupe	3,35	12,69
Hauts-de-France	24,36	2,44
Île-de-France	17,93	10,50
Martinique	1,04	0,02
Normandie	6,59	9,57
Nouvelle-Aquitaine	0,57	10,25
Occitanie	0,88	19,81
Pays de la Loire	1,65	4,62
Provence-Alpes-Côte d'Azur	5,89	2,57

Table 8: Contributions des régions aux axes de l'AFC (en %).

Pour un novice, la Figure 6 est comme une carte où les régions et les tailles de communes sont placées selon leurs similitudes. Par exemple, l'Île-de-France est proche des grandes communes, car elle en compte beaucoup, tandis que l'Occitanie est près des petites communes, typiques de certaines régions rurales.

Pour un expert, l'axe 1 (66,51% de l'inertie) est fortement influencé par la catégorie "Grande" (65,37%) et des régions comme les Hauts-de-France (24,36%) et l'Île-de-France (17,93%), confirmant que cet axe sépare les grandes communes des autres. L'axe 2 (33,49%) est dominé par les catégories "Petite" (55,39%) et "Moyenne" (43,97%), avec des régions comme la Bretagne (23,15%) et l'Occitanie (19,81%) contribuant fortement, ce qui explique la dispersion verticale. La valeur du χ^2 (62,696, $p < 0,001$) valide la significativité des associations.

7.4 Comparaisons régionales

L'analyse montre que l'Île-de-France et les Hauts-de-France sont fortement associées aux grandes communes, reflétant leur concentration de centres urbains dynamiques comme Paris ou Lille. En revanche, les régions comme l'Occitanie, associées aux petites communes, et la Normandie, aux communes moyennes, reflètent des contextes ruraux ou semi-urbains avec des dynamiques démographiques distinctes. Ces différences sont influencées par des facteurs historiques, géographiques et économiques qui façonnent la structure des régions.

8 Partie 3 : Analyse Complémentaire (Classification Hiérarchique et ACM)

8.1 Statistiques descriptives et standardisation

Le tableau 9 confirme que les variables quantitatives ont été standardisées (moyenne proche de 0, écart-type = 1), garantissant une analyse équitable dans les classifications.

Pour un novice, la standardisation rend toutes les variables comparables, comme si elles étaient mesurées sur la même échelle. Pour un expert, les moyennes proches de zéro (ex. : 3,68943E-16 pour `pop_tot`) et les écarts-types égaux à 1 confirment que la standardisation a été correctement appliquée, éliminant les biais liés aux différences d'échelle.

8.2 Classification Hiérarchique Ascendante (CAH)

La CAH, basée sur la distance euclidienne et la méthode de Ward, a produit six groupes, expliquant 86,73% de la variance totale (Tableau 10). La Figure 7 montre une forte

Variable	Moyenne	Écart-type
pop_tot	3.68943E-16	1.0000
pop_0_14	6.03696E-15	1.0000
pop_15_29	-1.85555E-15	1.0000
pop_75p	-1.14339E-15	1.0000
pop_act_15p	-2.73629E-15	1.0000
pop_chom	-8.65697E-16	1.0000
pop_cadres	5.43506E-16	1.0000
pop_ouvr	-1.93796E-15	1.0000
pop_dipl_aucun	1.56672E-15	1.0000
pop_dipl_bac	-1.09650E-15	1.0000
log_rp	2.15337E-15	1.0000
log_proprio	-3.98574E-15	1.0000

Table 9: Statistiques des variables après ajustement.

diminution de l'inertie interclasse jusqu'à six groupes, où un "coude" indique un compromis optimal entre séparation des groupes et homogénéité intraclasse. La Figure 8 confirme que six groupes capturent la majorité des différences entre communes, tandis que le dendrogramme (Figure 9) illustre la hiérarchie des regroupements.

Tendance	Valeur	Différence	Proportion	Cumul
1	7.1033	4.9425	0.5919	0.5919
2	2.1608	1.0174	0.1801	0.7720
3	1.1434	0.5459	0.0953	0.8673
4	0.5975	0.2574	0.0498	0.9171
5	0.3401	0.1263	0.0283	0.9454
6	0.2138	0.0512	0.0178	0.9633

Table 10: Valeurs des principales tendances pour la classification hiérarchique.

Pour un novice, le dendrogramme (Figure 9) est comme un arbre généalogique : les communes similaires sont regroupées en branches, et couper l'arbre à six branches donne six types de communes distincts. La Figure 7 montre que choisir six groupes est un bon compromis, car ajouter plus de groupes n'apporte que peu d'information supplémentaire. Par exemple, une branche pourrait représenter les grandes villes comme Paris avec un chômage élevé, tandis qu'une autre regroupe des villages ruraux comme ceux du Cantal.

Pour un expert, l'inertie interclasse (Figure 8) et la méthode de Ward garantissent une minimisation de la variance intraclasse, avec 86,73% de la variance expliquée par les trois premières composantes. La forte diminution de l'inertie jusqu'à six groupes (Figure 7) confirme que ce choix maximise la séparation entre groupes tout en préservant l'homogénéité à l'intérieur de chaque groupe. Le dendrogramme (Figure 9) montre des branches longues pour les groupes extrêmes (urbains et ruraux), indiquant des différences

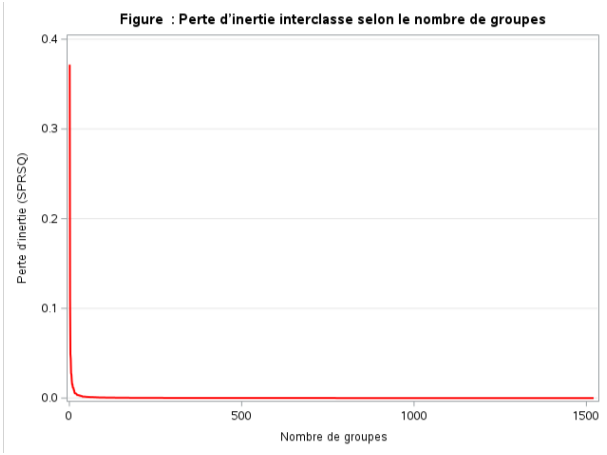

Figure 7: Perte d'inertie interclasse selon le nombre de groupes. La courbe montre une diminution rapide jusqu'à six groupes, où la pente s'aplatit, indiquant que six groupes optimisent la séparation des communes tout en maintenant une homogénéité au sein de chaque groupe.

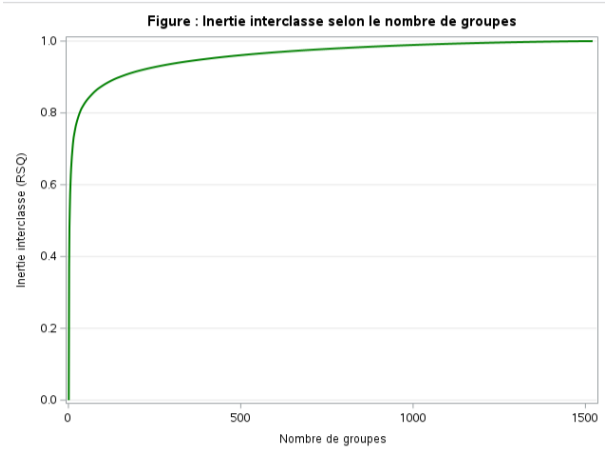

Figure 8: Inertie interclasse selon le nombre de groupes. L'augmentation rapide jusqu'à six groupes montre que ce choix capture la majorité des différences entre communes, confirmant la pertinence de la CAH pour identifier des profils distincts.

marquées, tandis que les branches plus courtes pour les groupes intermédiaires suggèrent une plus grande similarité.

8.3 Classification Hiérarchique sur les catégories

La CAH sur les catégories (`pop_tot`, `pop_chom`, `pop_cadres`) a produit quatre groupes, comme montré dans le tableau 11. Les variables `pop_chom` et `pop_cadres` ont été catégorisées en deux niveaux ("Haut" et "Bas") en utilisant la médiane comme seuil, tandis que `pop_tot` a été divisée en quartiles, avec `PopTot_Q4` correspondant aux 25% de communes les plus peuplées. Le dendrogramme (Figure 10) illustre une séparation claire, avec des branches associant les grandes communes à un chômage élevé et une forte proportion de cadres.

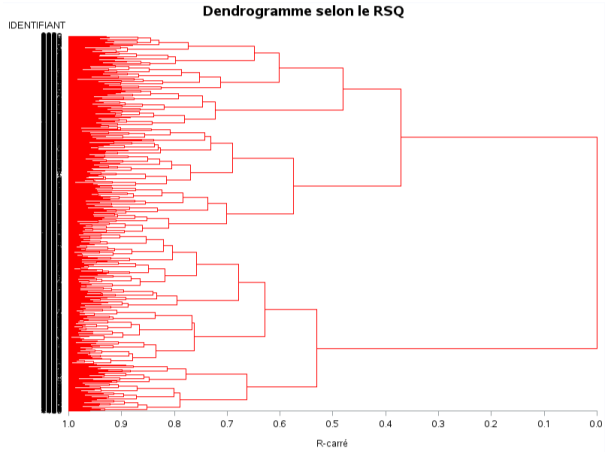

Figure 9: Dendrogramme de la CAH, montrant la hiérarchie des groupes. La coupe à six groupes sépare clairement les profils urbains, ruraux et périurbains, avec des branches longues indiquant des différences marquées entre les groupes.

Tendance	Valeur	Différence	Proportion	Cumul
1	0.6835	0.2294	0.6008	0.6008
2	0.4541	–	0.3992	1.0000

Table 11: Valeurs des tendances pour les catégories.

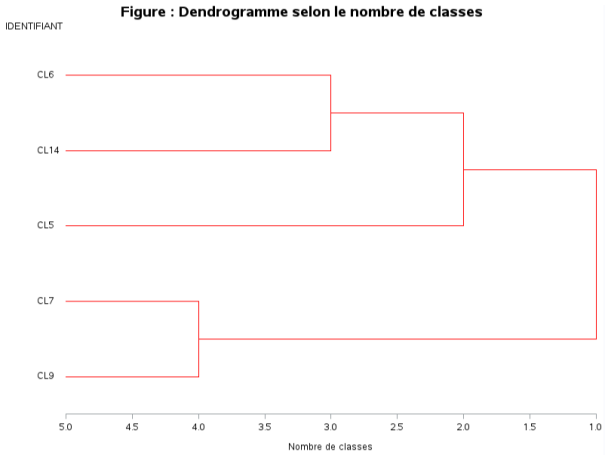

Figure 10: Dendrogramme des catégories, montrant quatre groupes basés sur la population, le chômage et les cadres. Les branches séparent les grandes communes à chômage élevé et forte proportion de cadres des petites communes à faible chômage, reflétant des dynamiques socio-économiques distinctes.

Pour un novice, ce dendrogramme est comme une carte des relations entre les caractéristiques des communes. Par exemple, une branche montre que les grandes communes (comme Paris) ont souvent un chômage élevé et beaucoup de cadres, tandis qu'une autre branche regroupe des communes plus petites avec un chômage plus faible. Pour un expert, les valeurs propres (0,6835 et 0,4541) indiquent que les deux premières tendances capturent 100% de l'inertie, confirmant la robustesse de la CAH pour les données caté-

goriques. Les longues branches du dendrogramme suggèrent des différences marquées entre les groupes, notamment entre les grandes communes dynamiques avec un chômage élevé et les petites communes moins touchées économiquement.

8.4 Analyse des relations entre catégories

La Figure 12 montre que `PopTot_Q4` (grandes communes, 1,2949 sur l'axe 1) est proche de `Chom_Hau` (0,9712) et `Cadres_Hau` (0,6995), confirmant que les grandes communes ont tendance à avoir un chômage élevé et une forte proportion de cadres. Les catégories `PopTot_Q4` (variation = 0,1072) et `Chom_Hau` (variation = 0,0953) contribuent fortement à l'axe 1, expliquant les principales variations dans les données catégoriques.

8.5 Décomposition de l'Inertie et du Khi-2

Une analyse complémentaire de la décomposition de l'inertie et du test du Khi-2 a été réalisée pour évaluer la contribution des composantes principales aux variations observées dans les données catégoriques. Le tableau ci-dessous présente les valeurs singulières, l'inertie principale, le Khi-2, le pourcentage et le pourcentage cumulé, avec un total de 81 degrés de liberté.

Décomposition de l'inertie et du Khi-2

Valeur singulière	Inertie principale	Khi-2	Pourcentage	Pourcentage cumulé
0.76686	0.58807	3129.7	25.20	25.20
0.61972	0.38406	2043.9	16.46	41.66
0.58942	0.34742	1848.9	14.89	56.55
0.57582	0.33157	1764.6	14.21	70.76
0.56400	0.31809	1692.9	13.63	84.39
0.52027	0.27068	1440.6	11.60	96.00
0.30569	0.09345	497.3	4.00	100.00
	2.33333	12417.9	100.00	

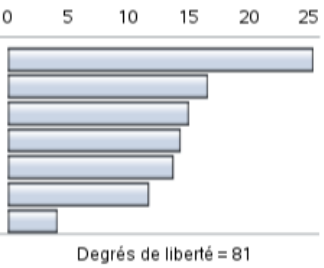

Figure 11: Décomposition de l'inertie et du Khi-2. Les valeurs singulières et l'inertie principale montrent que la première composante explique 25,20% de la variabilité, avec un cumul atteignant 100% sur les sept composantes principales. Le Khi-2 total de 12417,9 indique une association significative entre les variables catégoriques analysées.

Cette décomposition confirme que les premières composantes capturent une part importante de la variabilité (jusqu'à 70,76% pour les quatre premières), soutenant les résultats de l'ACM en identifiant les relations structurelles entre les catégories comme `PopTot_Q4`, `Chom_Hau`, et `Cadres_Hau`.

Pour un novice, la Figure 12 montre que les grandes villes, le chômage élevé et les cadres élevés sont comme des "amis proches" sur la carte, car ils vont souvent ensemble. En revanche, les petites communes sont plus associées à un chômage plus faible. Par exemple, une grande ville comme Paris peut avoir à la fois beaucoup de cadres et un chômage élevé en raison de disparités économiques.

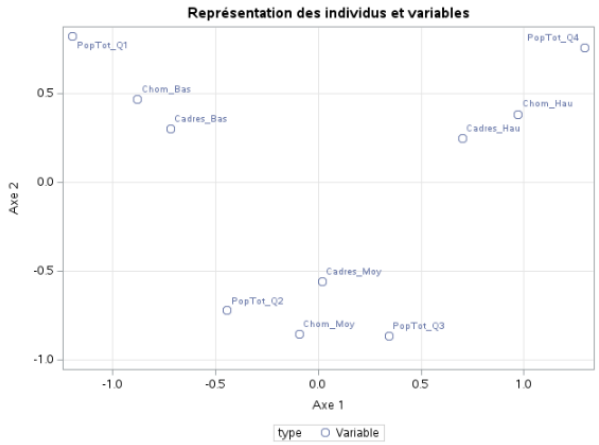

Figure 12: Représentation des catégories (population, chômage, cadres) sur les deux axes. Les grandes communes (`PopTot_Q4`) sont proches du chômage élevé (`Chom_Hau`) et des cadres élevés (`Cadres_Hau`), reflétant des dynamiques économiques complexes avec des disparités internes. Cette visualisation met en évidence les relations structurelles entre ces catégories.

Pour un expert, les coordonnées élevées (ex. : `PopTot_Q4` = 1,2949, `Chom_Hau` = 0,9712) et les qualités de représentation (0,7514 pour `PopTot_Q4`, 0,5445 pour `Chom_Hau`) indiquent que ces catégories sont bien expliquées par les deux axes. La proximité de `PopTot_Q4` et `Chom_Hau` sur l'axe 1 (60,08% d'inertie) confirme une forte association, reflétant des défis économiques dans les grandes communes malgré leur dynamisme. L'inertie totale de 100% confirme que l'ACM capture pleinement les variations des données catégoriques, et la forte contribution de `PopTot_Q4` et `Chom_Hau` à l'axe 1 reflète leur rôle clé dans la différenciation des profils socio-économiques.

8.6 Implications pour les politiques publiques

Les résultats de l'analyse ont des implications directes pour les politiques publiques :

- **Groupe 1 (communes urbaines)** : Investir dans les infrastructures de transport, les technologies numériques, et des programmes de réduction du chômage pour adresser les disparités économiques. Par exemple, mepourraient être complétés par des initiatives de formation professionnelle.
- **Groupe 3 (communes rurales)** : Développer des services pour les seniors, comme des centres médicaux ou des programmes de soutien à domicile, pour répondre au vieillissement de la population. Des initiatives comme les maisons de santé rurales pourraient être pertinentes.
- **Territoires d'outre-mer** : Renforcer les infrastructures de base (électricité, eau, routes) et soutenir le développement économique local pour réduire les disparités avec la métropole.

- **Groupe 4 (communes périurbaines)** : Encourager des politiques de logement abordable et des transports publics pour connecter ces zones aux centres urbains, réduisant ainsi la dépendance à la propriété immobilière.

Pour un novice, ces recommandations montrent que chaque type de commune a des besoins différents : les grandes villes ont besoin de transports et de solutions pour le chômage, tandis que les villages ruraux ont besoin de services pour les personnes âgées. Pour un expert, ces implications s'appuient sur les profils identifiés par les classifications et l'AFC, offrant une base empirique pour des politiques ciblées. Par exemple, la forte association entre les grandes communes et les régions comme l'Île-de-France dans l'AFC suggère que ces zones nécessitent des interventions spécifiques pour réduire le chômage, comme des programmes d'emploi ciblés.

8.7 Limites de l'étude

L'analyse est limitée à 1521 communes, soit environ 28% de l'échantillon initial, ce qui pourrait réduire la généralisabilité des résultats. L'absence de données longitudinales empêche d'analyser les évolutions temporelles, comme les changements dans la structure démographique ou économique des communes. Enfin, des variables supplémentaires, comme le revenu médian, la connectivité numérique ou l'accès aux services publics, pourraient enrichir l'analyse en capturant d'autres dimensions des disparités socio-économiques.

9 Conclusion

Cette étude a permis d'identifier quatre profils socio-économiques distincts parmi les 1521 communes analysées : les communes urbaines dynamiques avec un chômage élevé, les communes moyennes équilibrées, les communes rurales âgées, et les communes périurbaines à forte propriété. Ces profils, soutenus par des analyses rigoureuses (ACP, AFC, k-means, CAH), offrent des insights précieux pour des politiques publiques ciblées, comme des programmes d'emploi dans les grandes villes ou des services pour seniors dans les zones rurales. Les visualisations, telles que le *Scree Plot*, le cercle des corrélations, et les dendrogrammes, ont permis de rendre les résultats accessibles et interprétables, tout en confirmant la robustesse des méthodes utilisées.

Pour les recherches futures, l'intégration de données longitudinales pourrait permettre d'analyser les dynamiques temporelles, tandis que l'ajout de variables comme le revenu ou la connectivité numérique enrichirait les profils identifiés. Enfin, une extension de l'analyse à un échantillon plus large, incluant les communes exclues par le nettoyage, pourrait renforcer la généralisabilité des résultats.

A Annexes

A.1 Explications méthodologiques

La CAH utilise la distance euclidienne pour mesurer la dissimilarité entre communes et la méthode de Ward pour minimiser la variance intraclasse, garantissant des groupes homogènes. L'AFC repose sur le test du χ^2 pour valider les associations entre variables catégoriques, avec une décomposition de l'inertie pour évaluer la contribution de chaque catégorie. L'ACP repose sur la décomposition en valeurs propres pour identifier les axes principaux de variabilité, avec une standardisation préalable pour égaliser les contributions des variables.

A.2 Exemples de communes

- **Paris (Groupe 1)** : Avec 2,1 millions d'habitants et une forte activité économique, Paris illustre le profil des communes urbaines dynamiques, avec un chômage élevé et une forte proportion de cadres.
- **Village du Cantal (Groupe 3)** : Avec environ 3000 habitants et une population âgée, ces communes rurales nécessitent des services adaptés aux seniors, comme des centres de soins.

A.3 Approfondissement des résultats

Les profils identifiés reflètent des dynamiques complexes influencées par des facteurs géographiques, économiques et historiques. Par exemple, les communes urbaines (Groupe 1) bénéficient d'une forte concentration d'activités économiques, mais font face à des défis comme un chômage élevé et des disparités économiques. Les communes rurales (Groupe 3), en revanche, souffrent souvent d'un déclin démographique et d'un accès limité aux services. Comparés aux études de l'INSEE, nos résultats offrent une vision multivariée plus nuancée, en combinant des approches quantitatives et catégoriques pour capturer les interactions complexes entre variables. Les visualisations, comme les cartes factorielles et les dendrogrammes, facilitent l'interprétation de ces résultats et leur application à des contextes réels.