IAE Nantes
Économie & Management
Pôle Sociétés
Nantes Université

Nantes Université

Master 1 - Econométrie et Statistiques, parcours Econométrie Appliquée

Analyse de données et descriptive

Dossier réalisé par :
QUINTIN DE KERCADIO Pierre
CROCHET Florian

Année universitaire 2024 - 2025

Résumé

Le revenu moyen annuel représente un indicateur clé des disparités économiques et sociales entre les départements français. Cette étude a pour objectif de mieux comprendre les facteurs économiques, sociaux et environnementaux qui influencent les variations de ce revenu. Dans cette perspective, nous avons analysé une base de données comprenant 11 variables quantitatives et 3 qualitatives, couvrant des dimensions démographiques, économiques et territoriales. L'Analyse en Composantes Principales a permis de synthétiser ces informations en quatre dimensions principales, et un modèle économétrique a ensuite été utilisé pour tester leur influence. En conclusion, notre étude met en évidence l'impact de variables comme le PIB, la population et la surface verte par habitant sur le revenu moyen annuel. Elle révèle également que des facteurs comme la mortalité ou la présence d'une métropole jouent un rôle clé dans la compréhension des disparités entre départements. Enfin, nos analyses montrent que le taux d'éducation, bien qu'essentiel, peut être en tension avec d'autres indicateurs socio-économiques, ce qui complexifie son interprétation. En réponse, nous proposons des solutions telles le soutien économique des zones rurales et littorales, et l'augmentation des initiatives éducatives pour réduire les inégalités territoriales.

Sommaire

I - Introduction

L'étude des disparités économiques et sociales entre les départements français vise à comprendre les facteurs qui influencent le revenu annuel moyen par territoire. En s'appuyant sur une base de données composée de 13 variables explicatives, réparties entre 10 variables quantitatives et 3 variables qualitatives, cette analyse mobilise des indicateurs démographiques, économiques et environnementaux pour identifier les principales relations entre les variables explicatives et le revenu moyen.

Parmi les variables quantitatives, on retrouve la population en millions, qui représente la taille démographique du département, le taux de chômage exprimé en pourcentage, qui mesure l'accès à l'emploi, et le PIB départemental en milliards d'euros, qui reflète l'activité économique locale. L'espérance de vie en années témoigne du niveau global de bien-être, tandis que le taux de natalité et le taux de mortalité pour 1000 habitants illustrent la dynamique démographique. D'autres indicateurs, tels que le taux d'éducation en pourcentage, la part des ménages avec voiture, la part des logements sociaux et la surface verte par habitant en mètres carrés, permettent d'appréhender des dimensions variées du cadre de vie et du niveau de développement.

Les variables qualitatives comprennent la proximité d'un littoral, indiquée par 1 pour oui et 0 pour non, la distinction entre zone rurale et zone urbaine, avec une catégorisation similaire, ainsi que la présence d'une grande métropole, qui permet d'évaluer l'impact des pôles économiques et urbains sur les niveaux de revenus.

En combinant ces variables explicatives, cette étude cherche à déterminer l'influence relative de chaque facteur sur le revenu annuel moyen par département. Après avoir présenté nos données, nous réaliserons une analyse descriptive, puis une analyse des composantes principales , et une régression linéaire multiple, avant de conclure.

II - Présentation des données

Tableau 1 : Récapitulatif du modèle

Variables	Correspondance sur R
Revenu moyen annuel (€)	revenu
Nombre d'habitants (millions)	population
Taux de chômage (%)	chomage
PIB départemental (milliards €)	pib
Espérance de vie (ans)	esperance
Taux de natalité (pour 1000)	natalite
Taux de mortalité (pour 1000)	mortalite
Taux d'éducation (%)	education
Part des ménages avec voiture (%)	voiture
Part des logements sociaux (%)	social
Surface verte par habitant (m²)	surface
Proche littoral	littoral
Zone rurale	rurale
Accès métropole	metropole

La taille de la population d'un département peut influencer le revenu moyen. Les départements densément peuplés, notamment urbains, bénéficient souvent d'un accès accru aux opportunités économiques, ce qui peut entraîner des revenus plus élevés.

Un taux de chômage élevé est souvent associé à une diminution des revenus moyens, car il reflète des difficultés économiques locales et un accès limité à des emplois stables et bien rémunérés.

Un PIB par habitant élevé indique une économie locale dynamique, favorisant des revenus moyens plus élevés grâce à une activité économique prospère et diversifiée.

Une espérance de vie élevée peut refléter des conditions de vie favorables, telles qu'un meilleur accès aux soins, une alimentation équilibrée et des infrastructures de qualité, ce qui est souvent corrélé à des revenus moyens plus élevés.

Un taux de natalité élevé peut indiquer une population jeune et dynamique, mais il peut aussi engendrer des charges économiques accrues, ce qui peut influencer le revenu moyen différemment selon les départements.

Les taux de mortalité élevés peuvent indiquer des problèmes socio-économiques sous-jacents, tels qu'un accès limité aux soins de santé, qui peuvent affecter négativement les revenus moyens.

Le niveau d'éducation joue un rôle clé dans le revenu moyen, car une meilleure éducation favorise l'accès à des emplois qualifiés et bien rémunérés, renforçant ainsi le potentiel économique d'un département.

Un taux élevé de possession de voitures peut refléter un niveau de vie plus élevé et une capacité financière accrue des ménages, influençant positivement le revenu moyen.

Des dépenses sociales importantes peuvent indiquer une population en situation économique précaire, ce qui pourrait correspondre à des revenus moyens plus faibles dans ces zones.

Les départements plus vastes, souvent ruraux, peuvent avoir des revenus moyens inférieurs en raison d'une faible densité de population et d'un accès limité aux opportunités économiques. »

Les départements côtiers, grâce au tourisme et à l'activité économique maritime, ont souvent des revenus moyens plus élevés par rapport aux départements non côtiers. »

Les départements ruraux peuvent présenter des revenus moyens plus faibles, car ils dépendent davantage des secteurs primaires comme l'agriculture, souvent moins rémunérateurs que les secteurs industriels ou tertiaires.

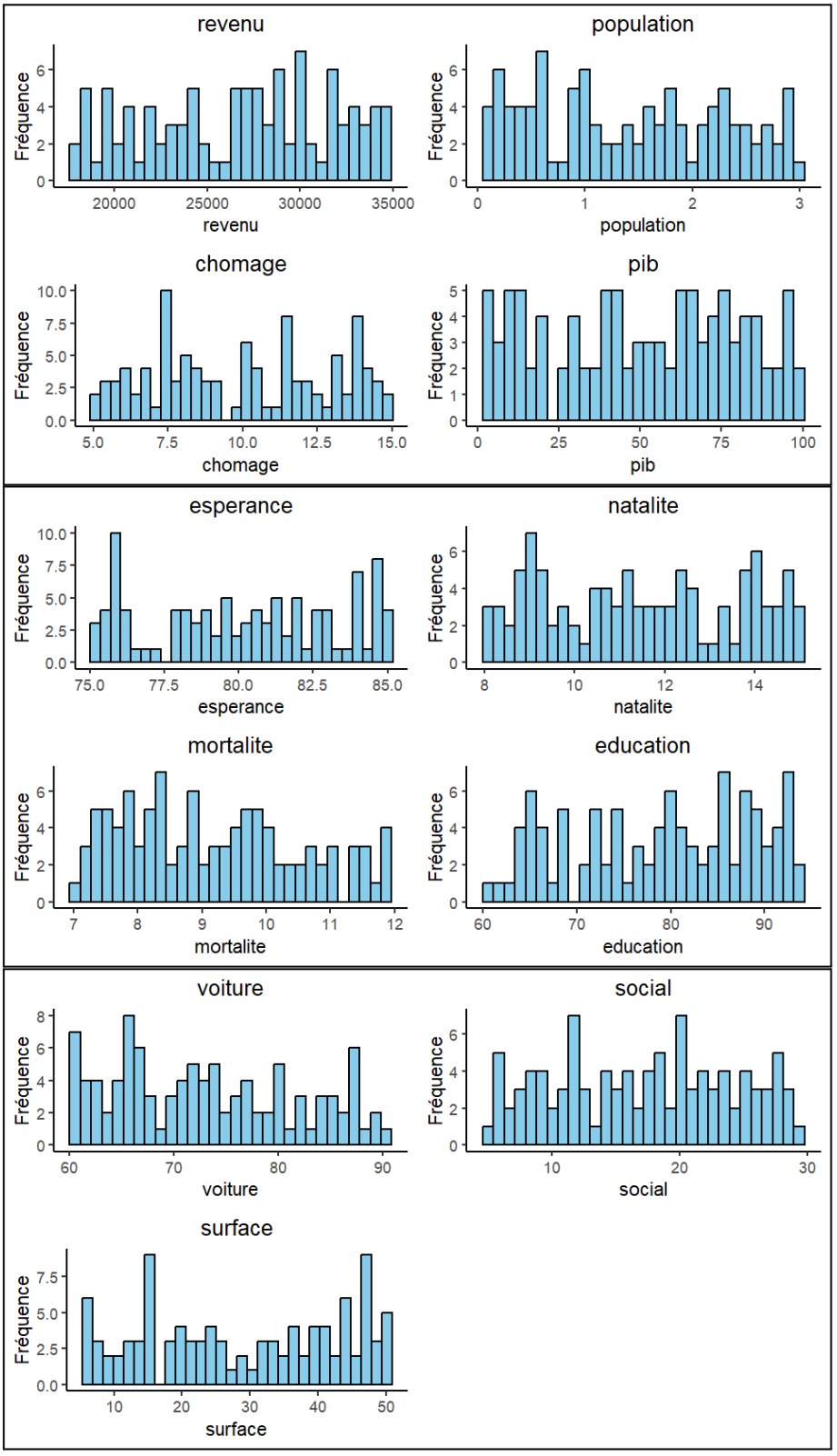
Les départements métropolitains concentrent généralement les infrastructures, les industries et les emplois bien rémunérés, ce qui entraîne des revenus moyens plus élevés par rapport aux zones rurales ou peu peuplées.

III - Analyse descriptive

A - Analyse univariée

1. Variables quantitatives

Figure 1 : Distribution des valeurs des variables quantitatives



Avec ces graphiques, nous pouvons dans la mesure du possible identifier la tendance de la distribution des variables quantitatives. Cependant, ces graphiques ne sont pas précis, nous avons donc besoin d'un tableau avec des indicateurs statistiques (moyenne, médiane, minimum…).

Tableau 2 : Statistiques descriptives des variables quantitatives

	Min	1Q	Médiane	Moyenne	3Q	Sd
revenu	18156.34	19725.01	20747.68	24042.748	28309.29	6472.938795498533
population	0.51	0.51	1.15	1.366	1.81	0.9888781522513277
chomage	7.49	8.14	10.08	10.228	11.36	2.6420011355031625
pib	33.64	52.59	64.47	64.296	73.15	23.803360687096266
esperance	75.1	78.2	78.89	80.038	83.95	3.888440561459055
natalite	9.13	13.08	13.75	12.994	14.37	2.2423269163973387
mortalite	7.19	8.03	8.82	8.8	9.51	1.266096362841312
education	68.91	80.84	83.85	81.708	86.6	7.698140684606901
voiture	63.39	63.73	73.52	75.068	85.16	12.027151366803364
sociaux	11.7	14.49	15.23	16.354	19.75	3.740284748518487
surface	6.77	14.95	36.75	31.07	47.48	19.296058405798835

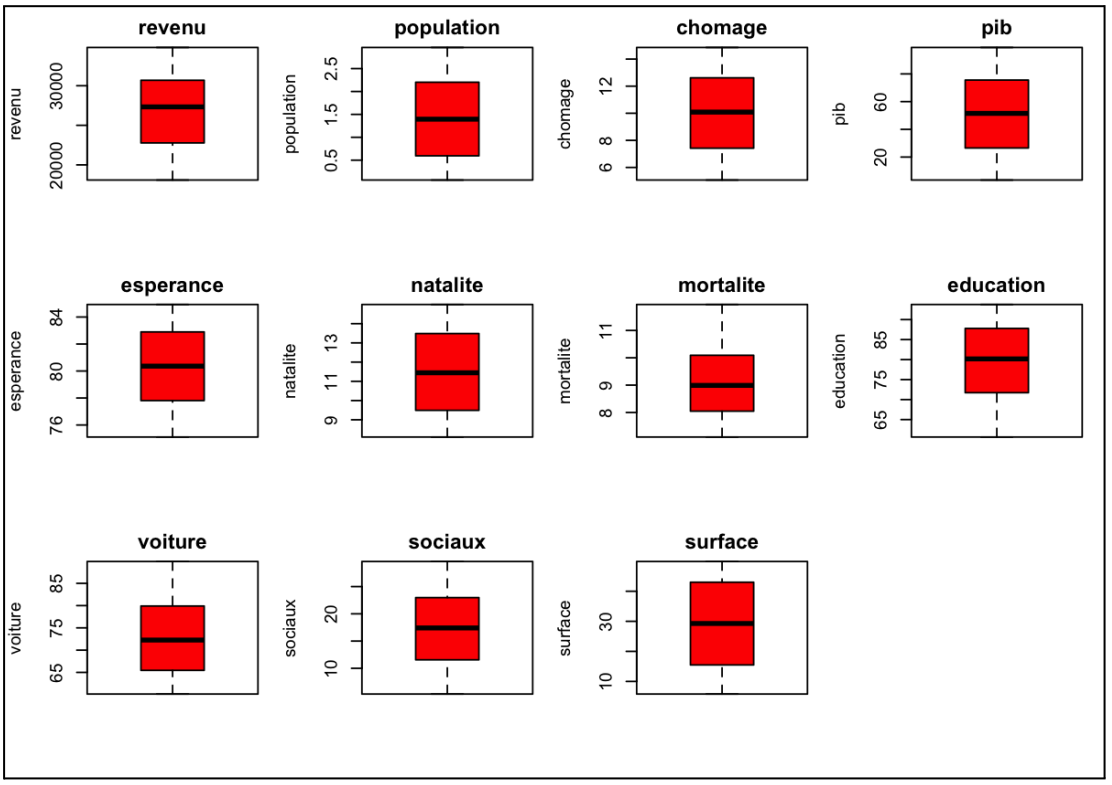
Les résultats des statistiques descriptives montrent que les médianes et les moyennes des variables quantitatives telles que le revenu, la population, le chômage, le PIB, et l'espérance de vie sont globalement proches, suggérant une distribution symétrique des données avec un impact limité des valeurs atypiques.

L'analyse des écarts-types révèle que certaines variables, comme le revenu et l'espérance de vie, sont homogènes, avec une faible dispersion autour de la moyenne, tandis que d'autres, comme la population et la surface, affichent une grande variabilité, indiquant

une hétérogénéité importante. Ces écarts peuvent s'expliquer par des différences significatives entre les départements en termes de densité ou de taille géographique.

Par ailleurs, des valeurs maximales élevées, notamment pour la population et le PIB, pourraient signaler la présence de valeurs atypiques, bien que leur impact sur la distribution globale semble limité. En somme, la plupart des variables sont homogènes, mais une attention particulière doit être portée aux variables hétérogènes, comme la population et la surface, pour comprendre les facteurs sous-jacents à leur variabilité.

Figure 2: Boxplot des variables explicatives quantitatives et de la variable expliquée



Comme nous pouvons l'observer et après le test de rosner, il n'y a pas de valeur atypique dans notre base de données

2. Variables qualitatives

Tableau 3 : Résumé des résultats des variables qualitatives

Zone	0	1
littoral	72	29
rurale	64	37
metropole	83	18

Les résultats montrent une prédominance de la catégorie "0" dans toutes les zones géographiques étudiées. Le variable “littoral” compte 72 observations dans la catégorie "0" contre 29 dans la catégorie "1", ce qui indique que la plupart des départements ne sont pas situés à proximité du littoral. Ensuite, la variable “rurale” présente 64 observations pour "0" et 37 pour "1", ce qui montre que la majorité des départements sont davantage urbains. De même, “metropole” affiche 83 observations pour "0" et seulement 18 pour "1", indiquant que peu de départements comprennent une grande métropole. Ces données indiquent une tendance générale où la catégorie "0" domine, bien que les proportions varient légèrement selon les zones.

B - Analyse bivariée

1. Deux variables quantitatives

Figure 3: Analyse des corrélations entre les variables quantitatives

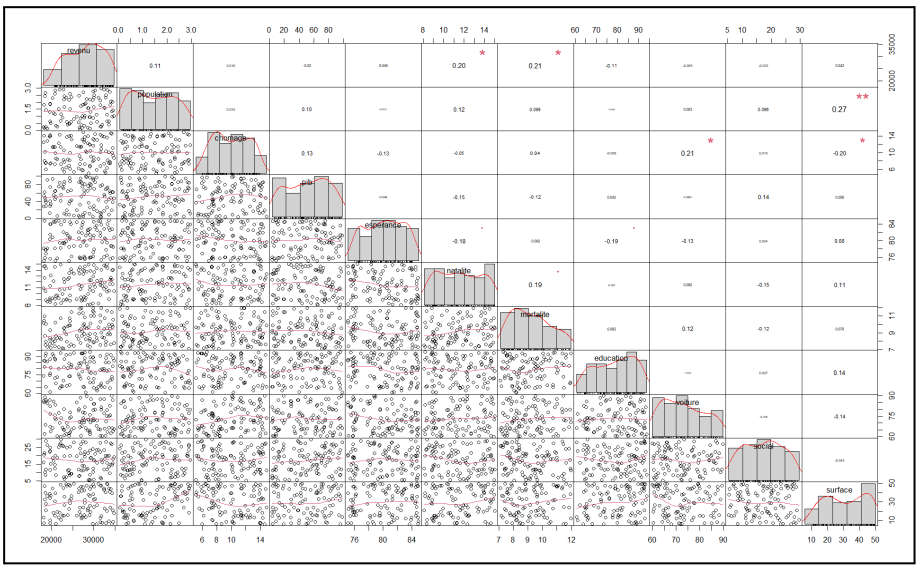
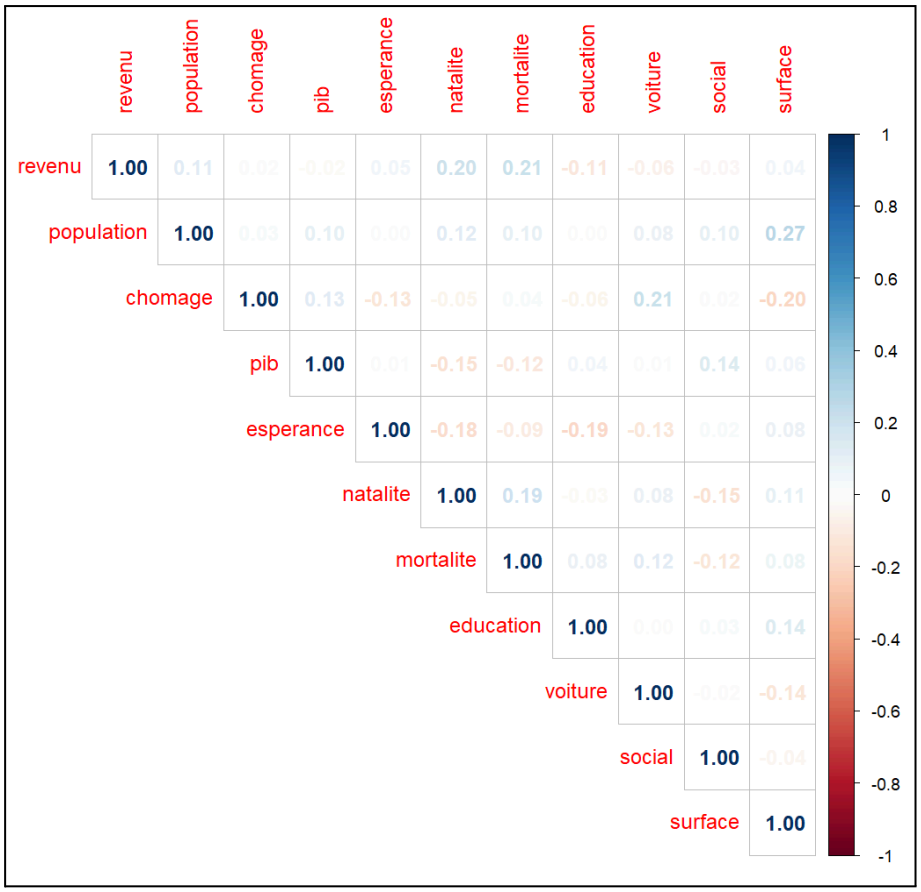
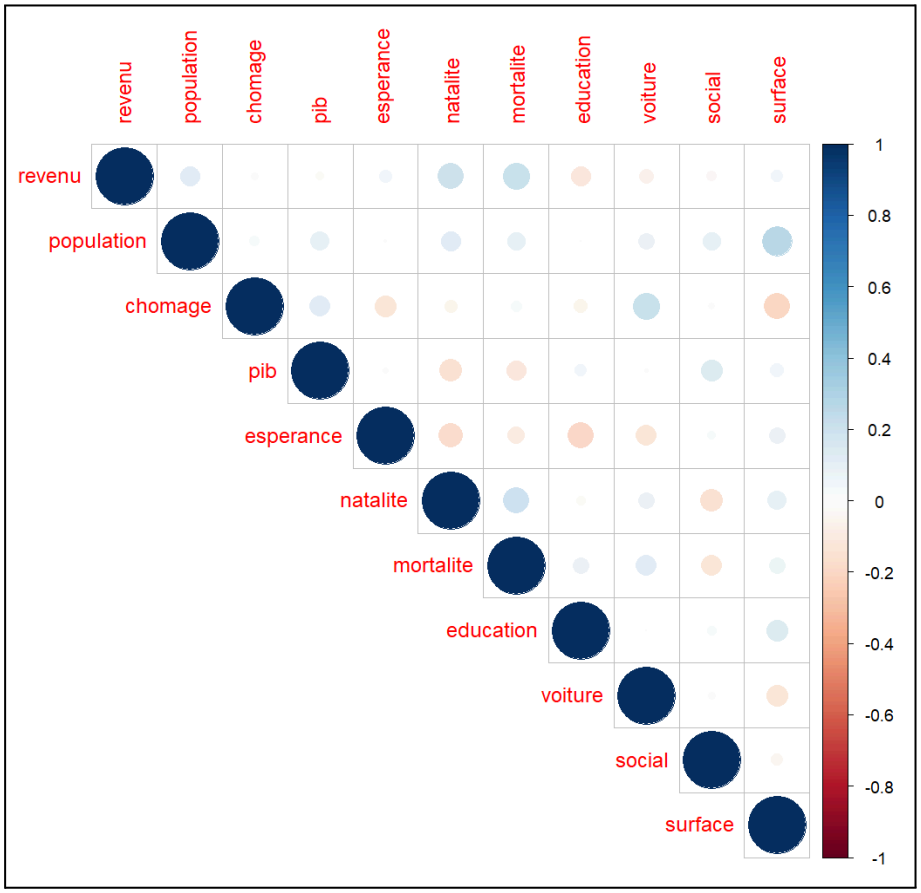


Figure 4: Corrélation des variables quantitatives



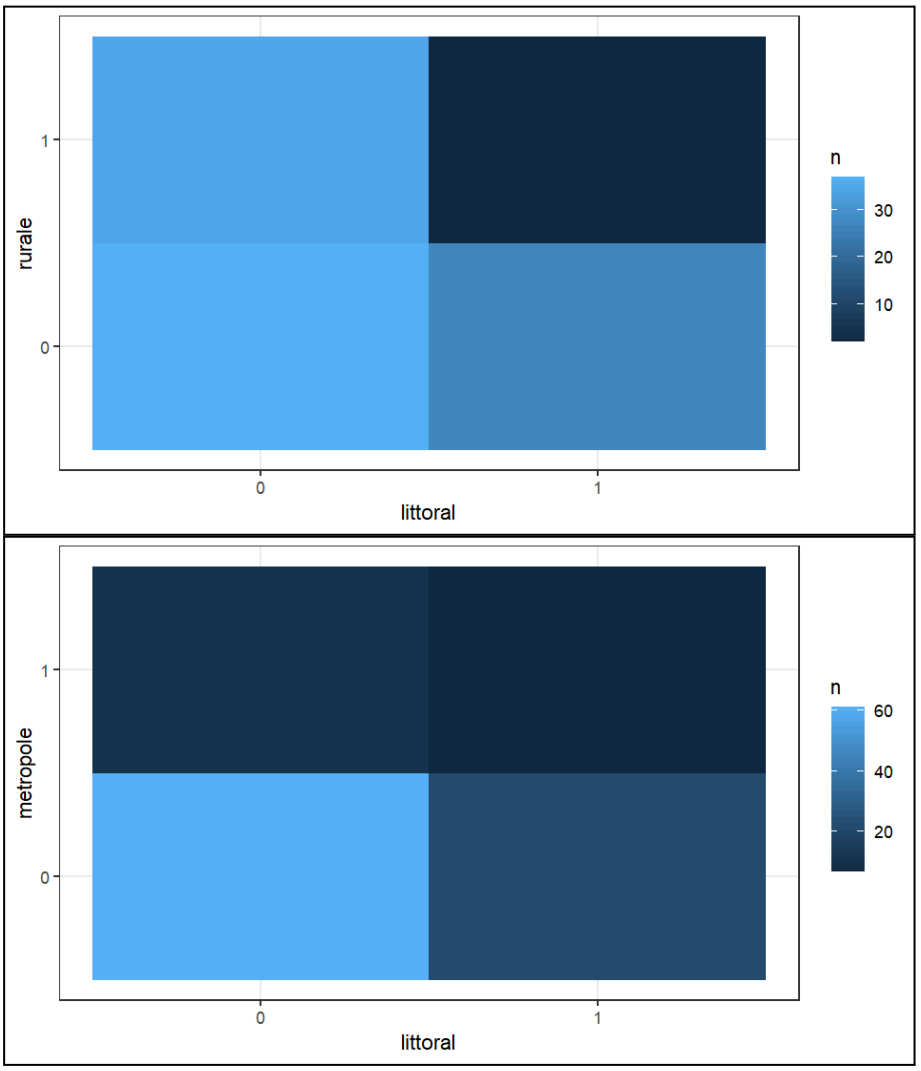


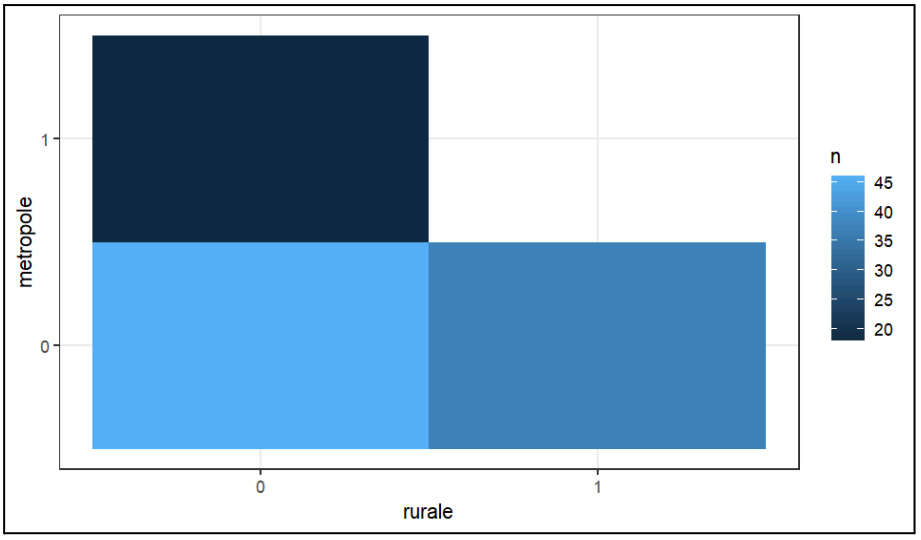
Les coefficients de corrélation oscillent autour de 0 ou restent relativement faibles, bien en dessous de la limite de 0,6 (positive ou négative). Cela indique qu'il n'existe pas de relation linéaire significative entre les différentes variables explicatives.

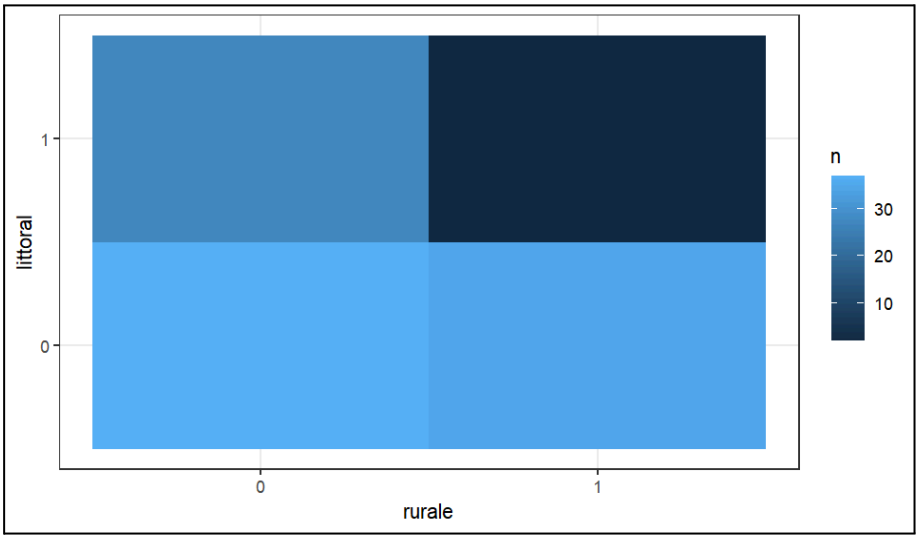
Cette absence de corrélations fortes indique que les variables explicatives sont relativement indépendantes les unes des autres, ce qui peut être positif dans un contexte de modélisation. En effet, une faible corrélation entre les variables explicatives réduit les risques de multicolinéarité, ce qui pourrait biaiser les estimations dans un modèle de régression.

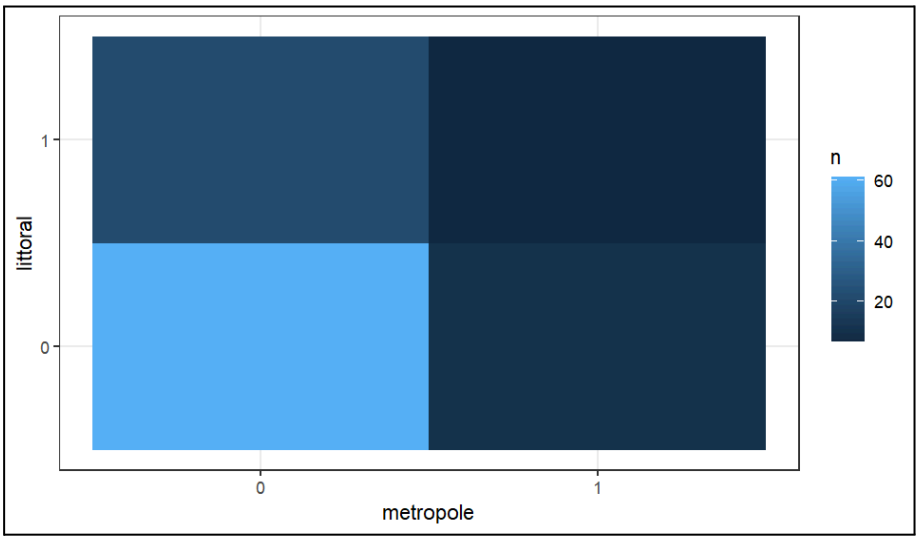
2. Deux variables qualitatives

Figure 5: Cartes des points chauds




metropole
1
0
0
1
rurale
n
45
40
35
30
25
20


littoral
1
0
0
1
rurale
n
30
20
10


littoral
1
0
0
1
metropole
n
60
40
20

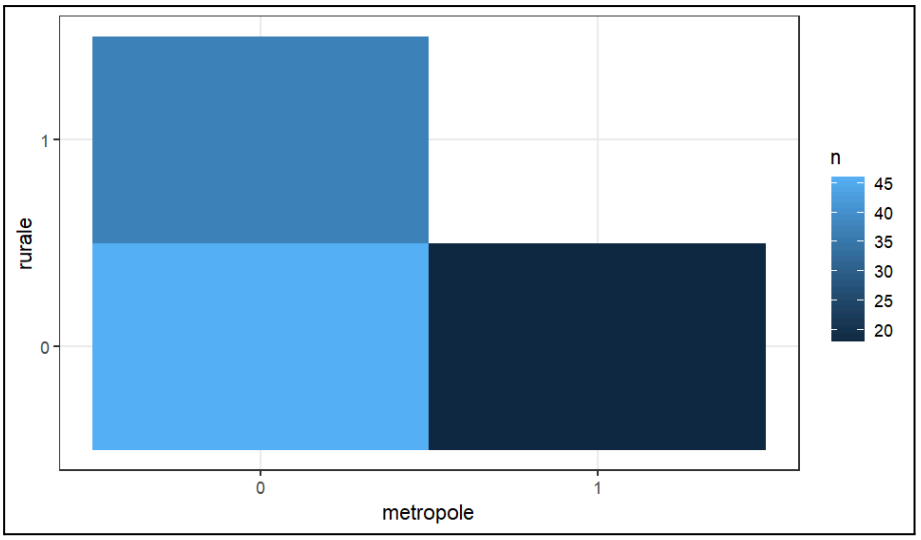
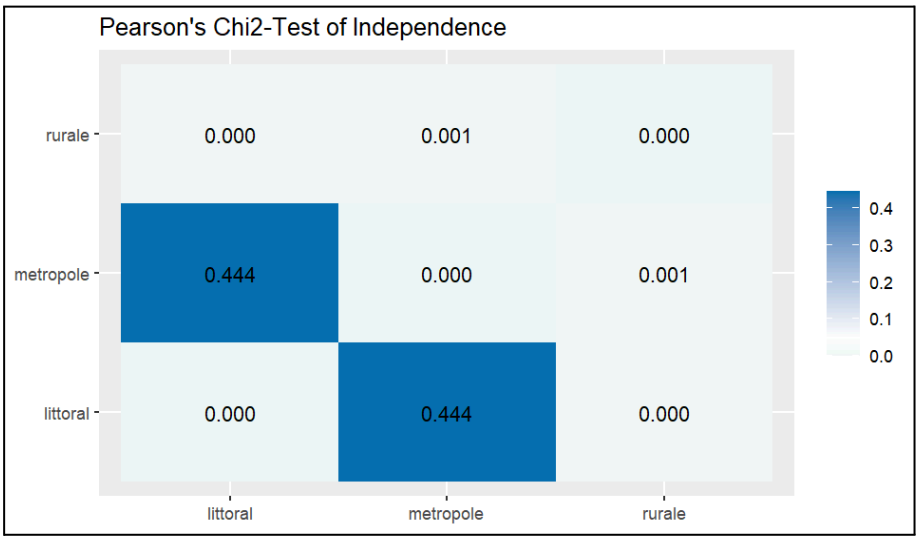


Figure 6: matrice d'indépendance des variables qualitatives



La matrice d'indépendance des variables qualitatives permet d'évaluer les relations statistiques entre les variables Littoral, Métropole et Rural à l'aide du test du Chi-2. Les p-valeurs affichées indiquent si les variables sont indépendantes ou liées. Une p-valeur inférieure à 0,05 (ex. Littoral vs Rural) indique une relation significative, suggérant une forte dépendance entre ces variables. Une p-valeur supérieure à 0,05 (ex. Littoral vs Métropole, p = 0,444) suggère que ces variables sont indépendantes et qu'il n'y a pas de lien statistique notable. Ainsi, cette matrice met en évidence une segmentation claire entre les zones rurales et littorales ou métropolitaines, tandis que Littoral et Métropole semblent indépendantes.

3. Une variable quantitative et une qualitative

Avant d'effectuer un test de comparaison de moyennes, nous avons vérifié la normalité des variables quantitatives pour chaque modalité des variables qualitatives.

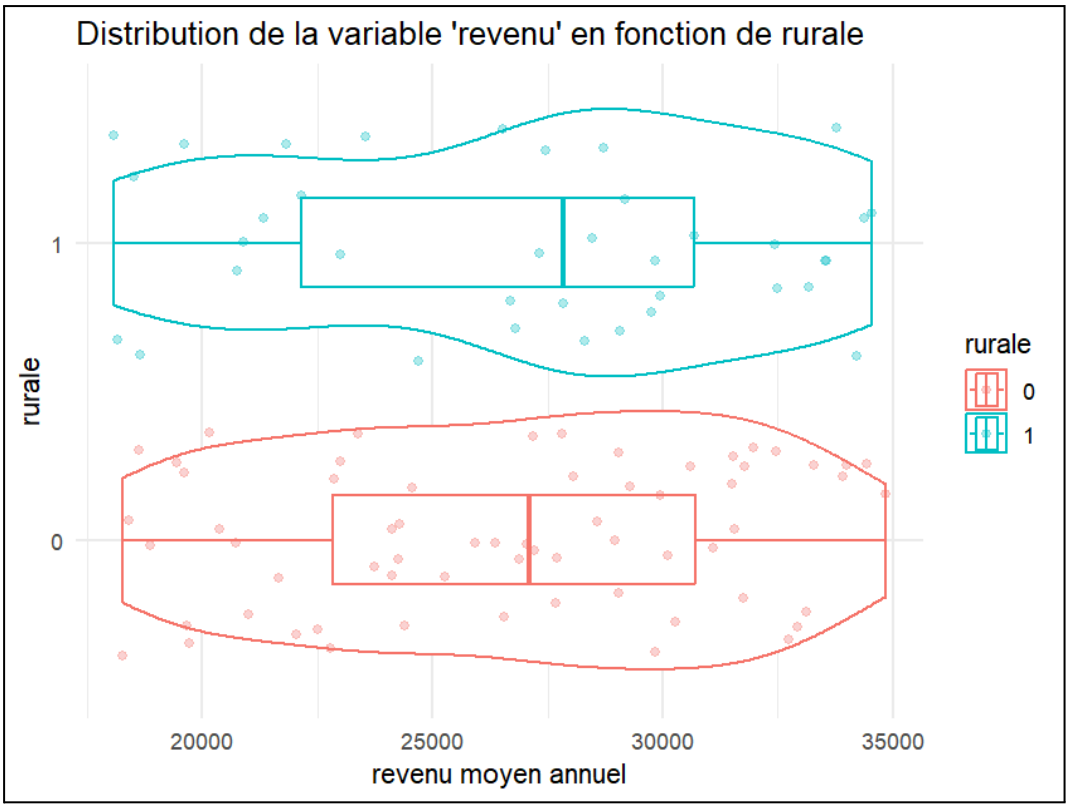
Les p-values indiquent généralement que les données ne suivent pas une distribution normale (p-value inférieure à 0.05 ou 0.10), en particulier pour les modalités de certaines variables. L'hypothèse de normalité est rejetée au seuil de 5 % et même à 10 % pour les données de plusieurs modalités. Ainsi, la distribution normale de la variable quantitative n'est pas vérifiée pour chaque modalité.

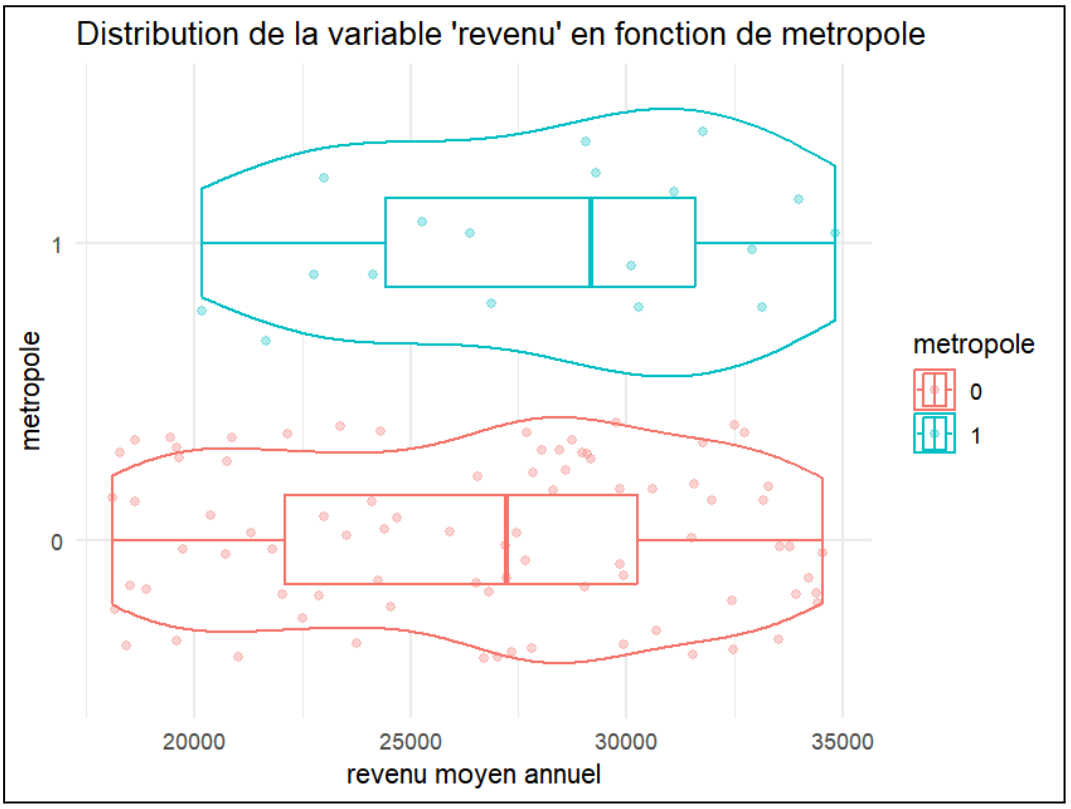
Par conséquent, les tests de Student et la méthode ANOVA, qui reposent sur l'hypothèse de normalité, n'ont pas pu être appliqués à toutes les associations entre une variable quantitative et une variable qualitative. Il était donc préférable d'utiliser le test non paramétrique de Wilcoxon-Mann-Whitney pour comparer les distributions des modalités des variables qualitatives en fonction de chaque variable quantitative.

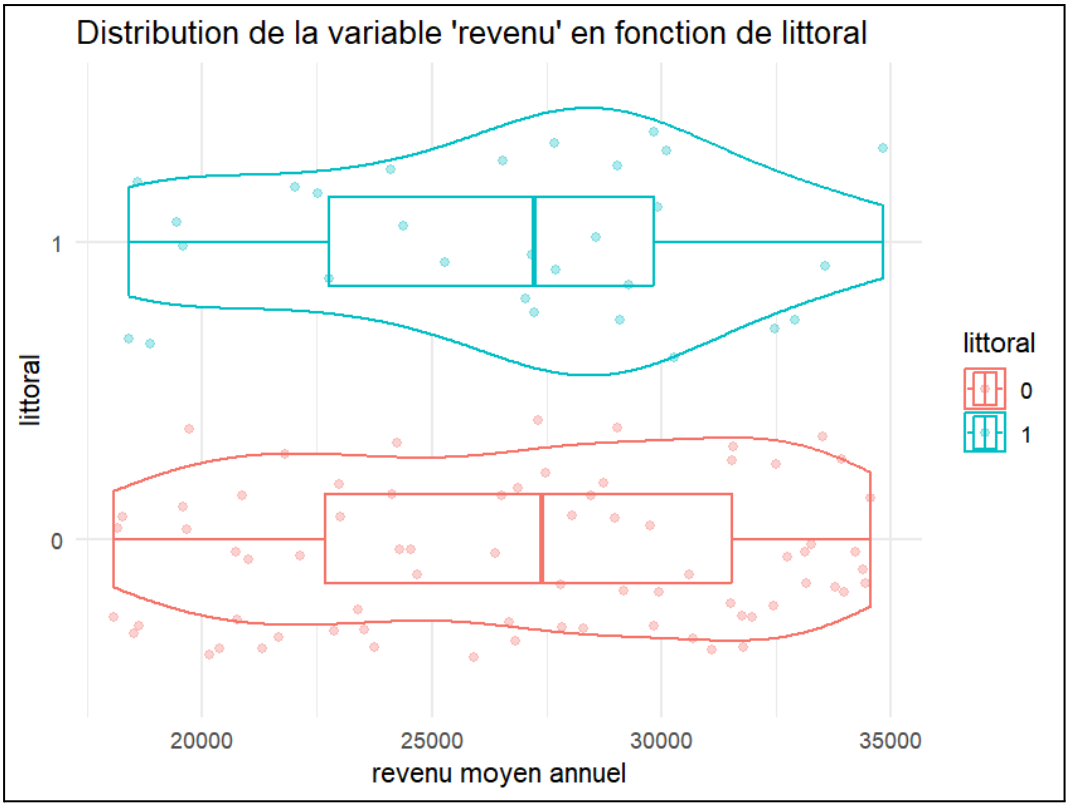
Concernant la variable littoral, aucun test ne révèle de différence significative entre les modalités 0 et 1, quelle que soit la variable quantitative, sauf pour la variable mortalité ($p=0.03935$). Pour la variable rurale, les tests ne montrent pas de différence significative entre les modalités 0 et 1 pour toutes les variables quantitatives, à l'exception de mortalité ($p=0.006175$). Quant à la variable metropole, plusieurs tests indiquent une différence significative, notamment pour population ($p=0.01101$), social ($p=0.01281$), et surface ($p=0.001559$).

Bien que certains groupes présentent des différences statistiquement significatives au seuil de 5 %, la majorité des modalités associées aux variables quantitatives ne montrent pas de différences significatives.

Figure 7 : Distribution du revenu en fonction des modalités de chaque variable qualitative.







D'après les graphiques, habiter dans une zone rurale est associé à des revenus légèrement inférieurs, tout comme résider près d'un littoral. En revanche, vivre dans une métropole semble significativement augmenter le niveau de revenu, probablement en raison des opportunités économiques et des emplois mieux rémunérés que ces zones offrent.

IV - Analyse en Composantes Principales (ACP)

Procédons maintenant à une analyse en composantes principales (ACP).

Il s'agit d'une méthode de réduction de dimensionnalité qui transforme des variables initiales corrélées en nouvelles variables non corrélées appelées composantes principales. Ces composantes, des combinaisons linéaires des variables d'origine, capturent successivement le maximum de variance des données. La première étape consiste à centrer et réduire les variables, puis à calculer la matrice de covariance ou de corrélation. Les composantes principales sont obtenues par décomposition en valeurs propres, où chaque composante explique une proportion de la variance totale. L'objectif est de retenir les premières composantes qui capturent l'essentiel de l'information, facilitant ainsi la visualisation, l'exploration et l'analyse des données complexes.

L'analyse en composantes principales est particulièrement pertinente pour analyser les disparités économiques et sociales entre les départements français. Face à un grand nombre de variables susceptibles d'influencer le revenu annuel moyen par territoire, comme l'accès à l'éducation, l'emploi, ou les caractéristiques démographiques, l'ACP permet de réduire la complexité des données tout en conservant l'essentiel de l'information. En identifiant des axes synthétiques qui expliquent les principales sources de variabilité, elle facilite la mise en évidence des facteurs structurants des inégalités territoriales. De plus, cette approche offre une représentation visuelle des relations entre les départements et les facteurs économiques et sociaux, permettant ainsi de mieux comprendre les dynamiques sous-jacentes.

Il est important de noter que cette simplification des données peut entraîner une certaine perte d'information.

A - Valeurs propres et nombre d'axes

Tableau 4 : Tableau des valeurs propres

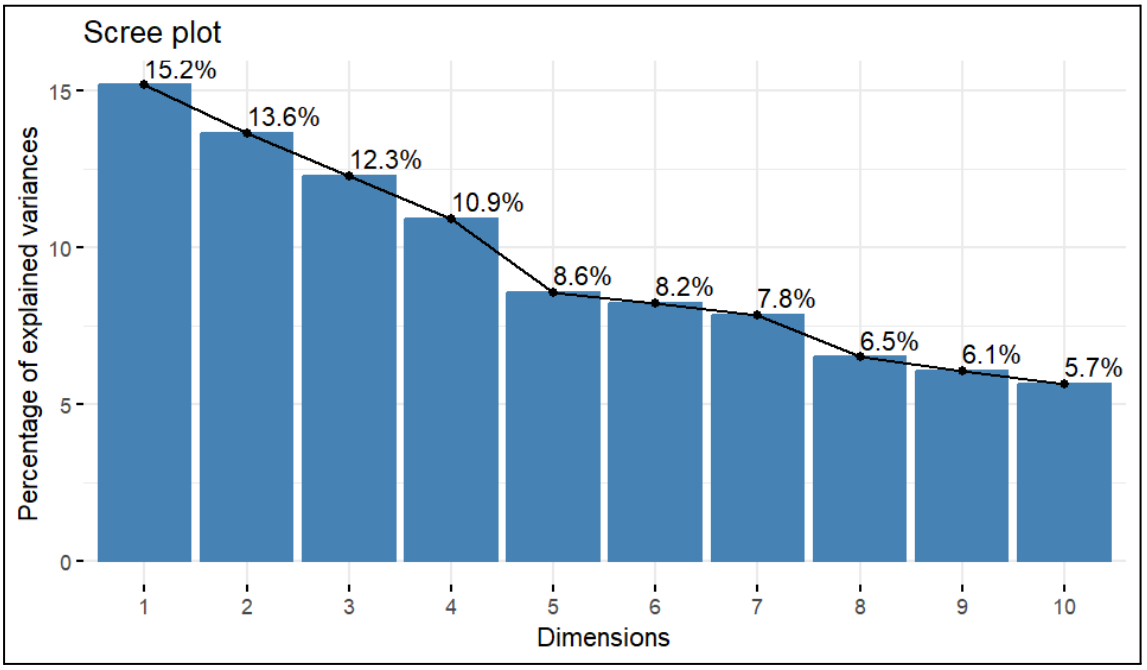
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	1.6684868	15.168062	15.16806
comp 2	1.5009254	13.644776	28.81284
comp 3	1.3505569	12.277790	41.09063
comp 4	1.1990163	10.900148	51.99078
comp 5	0.9426506	8.569551	60.56033
comp 6	0.9051982	8.229074	68.78940
comp 7	0.8617362	7.833965	76.62337
comp 8	0.7148513	6.498649	83.12202
comp 9	0.6660621	6.055110	89.17713
comp 10	0.6229999	5.663636	94.84076
comp 11	0.5675163	5.159239	100.00000

Quatre composantes possèdent une valeur propre supérieure à 1. Elles sont donc retenues d'après le critère de Kaiser. La première composante principale a une valeur propre de 1,67 et explique 15,17 % de la variance totale des données. La deuxième composante principale, avec une valeur propre de 1,50, explique 13,64 % de la variance totale. Ensuite,la troisième composante, ayant une valeur propre de 1,35, explique 12,28 % de la variance. Enfin, la quatrième composante a une valeur propre de 1,19 et explique 10,90 % de la variance totale des données. Combinées, les quatre premières composantes expliquent 51,99 % de la variance totale expliquée, ce qui indique qu'elles capturent une part significative de l'information contenue dans les données. Les composantes suivantes ont des valeurs propres inférieures à 1, ce qui signifie qu'elles expliquent une part de variance plus faible. Ainsi, les quatre premières composantes sont significatives et doivent être prises en compte dans l'analyse.

Cependant, en complément de ces quatre premières composantes, il peut être pertinent de considérer également les trois composantes suivantes (composantes 5, 6 et 7). Bien qu'elles aient des valeurs propres inférieures à 1, elles contribuent ensemble à expliquer 24,63 % supplémentaires de la variance, portant le pourcentage cumulé de variance expliquée à 76,62 %. Ce seuil est souvent considéré comme satisfaisant dans les études sur les disparités économiques et sociales, car il permet de capturer une part importante de l'information tout en limitant la perte liée à la réduction de dimensionnalité. Retenir ces trois

composantes supplémentaires pourrait donc enrichir l'analyse en tenant compte de facteurs potentiellement significatifs pour le sujet étudié.

Figure 8: Graphique des valeurs propres



La courbe montre une forte décroissance pour les quatre premières composantes principales, ce qui indique qu'elles expliquent une proportion importante de la variance totale (51,99 %). Après la quatrième composante, la courbe s'aplatit, reflétant une contribution marginale plus faible des composantes suivantes. Toutefois, les composantes 5, 6 et 7 apportent encore une part non négligeable de variance expliquée, ajoutant 24,63 % à la variance cumulée et portant le total à 76,62 %, un seuil souvent jugé suffisant pour des analyses robustes en sciences sociales. La visualisation confirme que les quatre premières composantes doivent être retenues selon la règle de Kaiser (valeurs propres > 1). En complément, inclure les composantes 5, 6 et 7 peut être justifié pour capturer une plus grande part d'information et affiner l'interprétation des disparités économiques et sociales.

Ainsi, nous réduisons la dimension de 11, correspondant aux 11 variables quantitatives de notre modèle, à une dimension de 7, voire 4, en sélectionnant les composantes principales qui expliquent la majeure partie de la variance des données. Cette réduction entraîne une certaine perte d'information, mais permet de simplifier le modèle tout en conservant les aspects les plus significatifs des données. Nous poursuivons donc notre analyse en examinant les contributions des variables, qui reflètent l'importance relative de

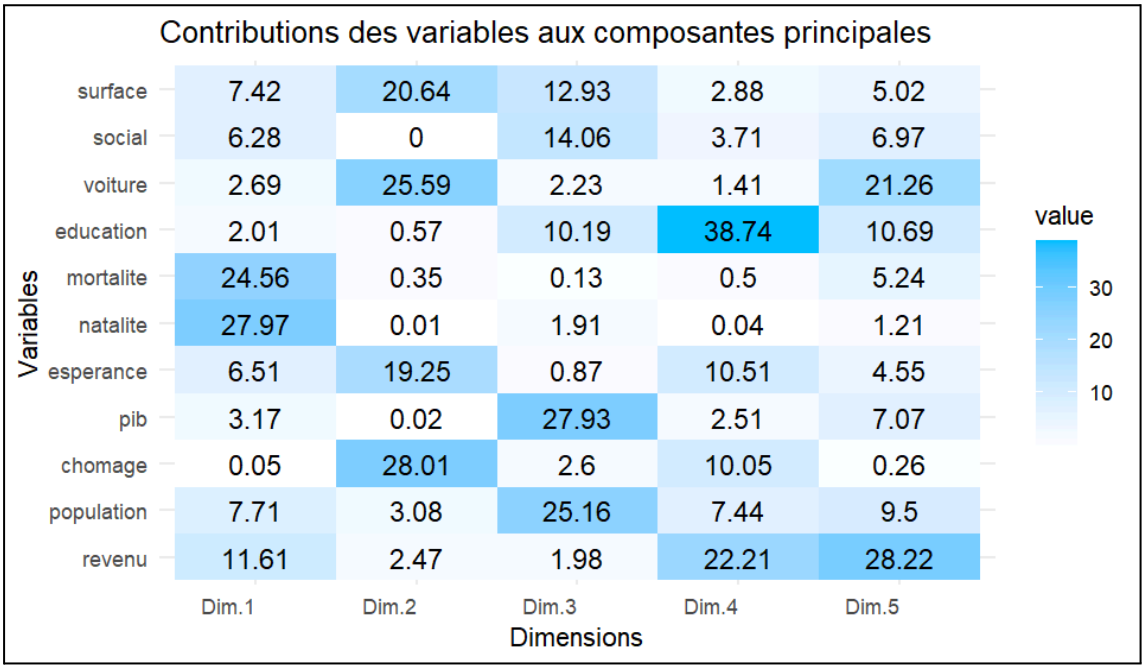
chaque variable dans la définition des axes principaux. Nous étudierons également les coordonnées des variables sur les 4 axes principaux, qui indiquent comment chaque variable se projette dans l'espace des composantes. Enfin, l'analyse des cosinus carrés des variables permettra de mesurer la qualité de la représentation de chaque variable dans l'espace réduit, en évaluant la proportion de variance expliquée par chaque composante.

B - Contributions, corrélations, cosinus carrés

1. Contributions

Premièrement, les contributions des variables aux différentes composantes principales montrent dans quelle mesure chaque variable influence chaque dimension. Plus la contribution est élevée, plus la variable joue un rôle important dans la composante principale correspondante. Ces contributions permettent de comprendre l'importance relative de chaque variable dans la capture de la variance expliquée par chaque composante.

Figure 9 : Contributions des variables



La dimension 1 est principalement influencée par la natalité (27,97%) et la mortalité (24,56%), suggérant que cet axe capture les dynamiques démographiques des départements. La dimension 2, quant à elle, est dominée par le chômage (28,01%) et la possession de voiture (25,59%), reflétant les inégalités socio-économiques. La dimension 3 est fortement

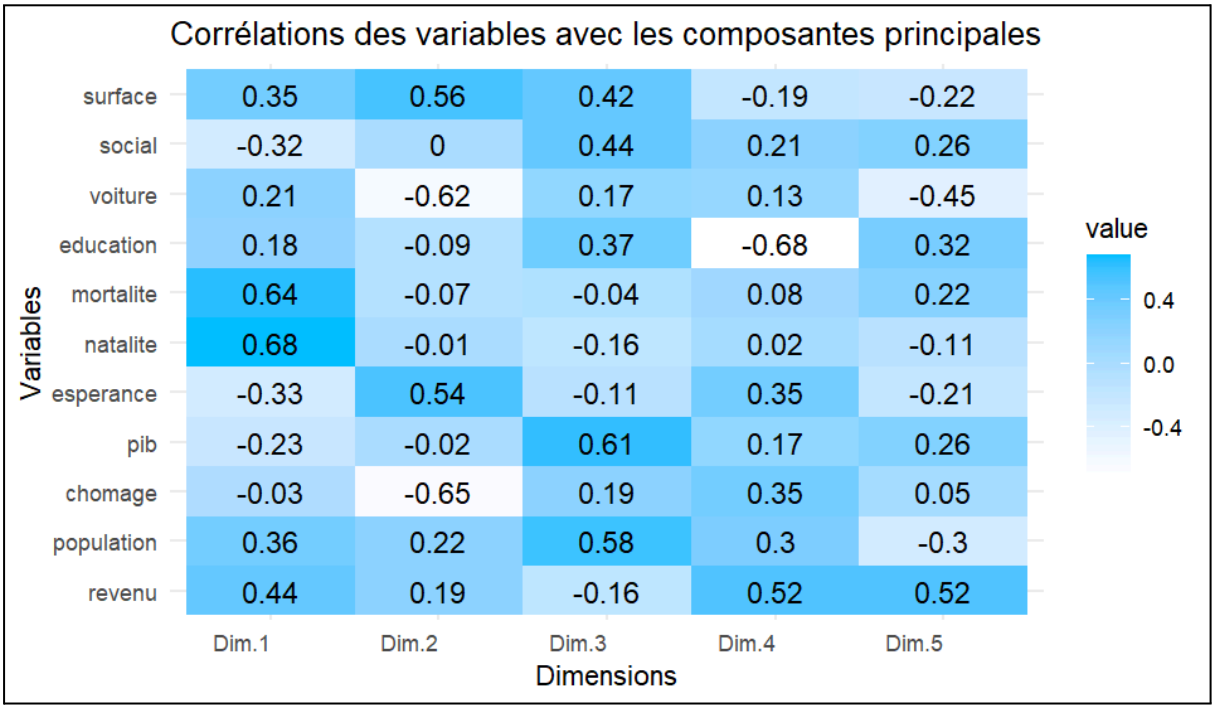
associée au PIB (27,93%) et au nombre d'habitants (25,16%), exprimant la concentration économique et démographique des départements. Enfin, la dimension 4 est principalement déterminée par le taux d'éducation (38,74%) et le revenu moyen annuel (22,21%), mettant en évidence l'importance du capital humain. Ces interprétations aident à identifier les principaux axes de variation dans les données, facilitant ainsi la compréhension des relations entre les variables étudiées.

Les contributions des variables aux différentes dimensions reflètent leur rôle dans la formation des axes du cercle de corrélation. Plus la valeur est importante, plus la variable est éloignée de l'origine des axes, et donc plus elle contribue à la composante principale.

2. Corrélations

Deuxièmement, les corrélations des variables avec les composantes principales indiquent la direction et la force de la relation entre chaque variable et les composantes principales. Une corrélation élevée positive ou négative (proche de 1 ou -1) signifie une relation forte, la variable contribue de manière significative à l'explication de la composante, tandis qu'une corrélation faible (proche de 0) indique une relation plus faible et suggère une influence moindre. Ces corrélations aident à interpréter la façon dont les variables sont reliées aux dimensions principales.

Figure 10 : Corrélations des variables



La natalité (0,68) et la mortalité (0,64) sont fortement corrélées avec la dimension 1, ce qui suggère de nouveau que cette composante reflète des dynamiques démographiques. Cette situation peut être due aux conditions socio-économiques défavorisées de la population du département, qui présente des taux de natalité et de mortalité élevés en raison de la pauvreté. Le chômage (-0,65) et la possession de voiture (-0,62) sont négativement corrélés avec la dimension 2, indiquant que les départements présentant des taux de chômage élevés tendent également à avoir une proportion plus faible de ménages possédant une voiture. Cette relation peut suggérer que, dans les départements avec un taux de chômage élevé, les populations ont moins de ressources financières et peuvent donc manquer de moyens pour investir dans une voiture. Par conséquent, cette dimension semble bien capturer les inégalités socio-économiques au sein des départements. Le PIB (0,61) et la population (0,58) sont les variables les plus corrélées avec la dimension 3, soulignant l'importance de la taille économique et démographique. Cette corrélation positive indique que cette composante capte les variations des départements à la fois plus peuplées et économiquement plus développées. En effet, des zones à forte densité de population, comme les grandes villes, ont généralement des niveaux de PIB plus élevés en raison de la concentration d'activités économiques, d'industries et de services. Ainsi, la dimension 3 reflète la concentration économique et démographique des départements. Enfin, le taux d'éducation (-0,68) et le revenu moyen annuel (0,52) montrent une forte corrélation avec la dimension 4, mettant en évidence l'impact du capital humain et de la richesse sur cette composante. Le taux d'éducation est négativement corrélé, ce qui pourrait indiquer que des niveaux d'éducation plus élevés sont associés à des caractéristiques opposées de cette dimension. En revanche, le revenu étant positivement corrélé, cela suggère que des niveaux de revenu plus élevés sont associés à des facteurs liés à cette composante. En effet, dans certaines régions, les départements avec un taux d'éducation élevé peuvent être associés à des départements plus ruraux ou moins développés économiquement, où les revenus sont moins élevés, même si le taux d'éducation est plus élevé. Ces départements peuvent avoir moins d'opportunités économiques, malgré un niveau d'éducation plus élevé.

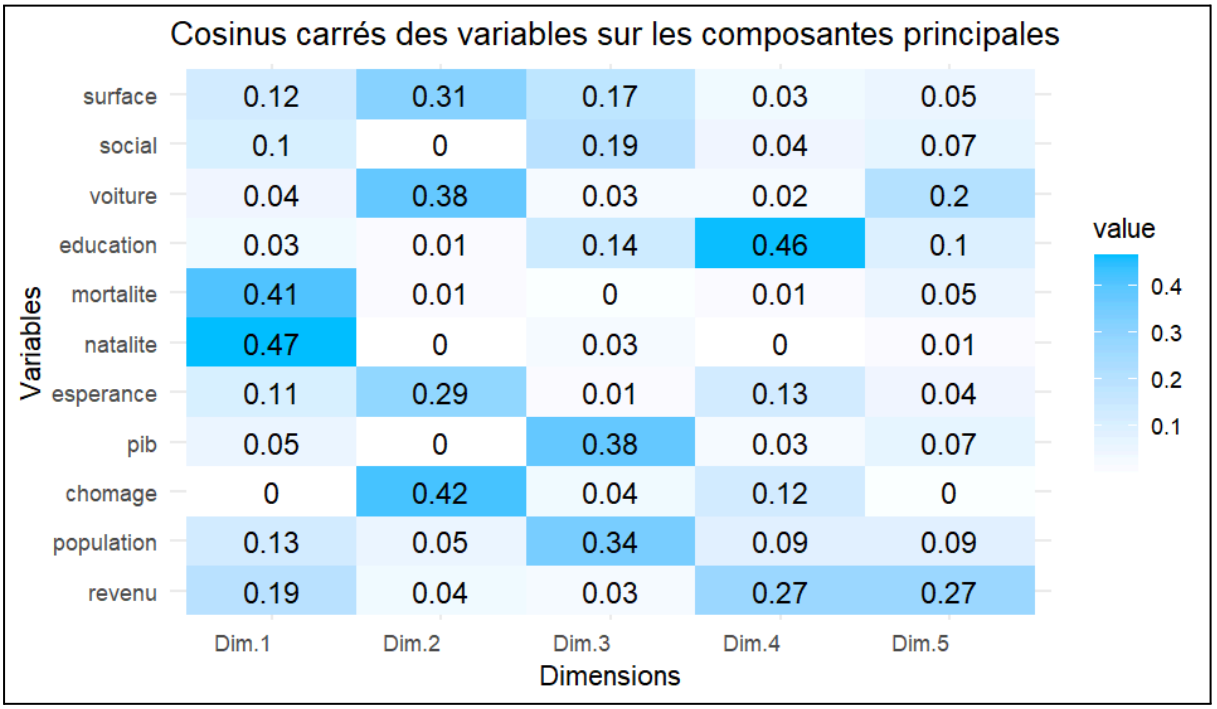
Les corrélations des variables avec les composantes principales permettent ainsi de déterminer dans quelle mesure chaque variable influence les axes principaux de variation dans les données, représentés par l’abscisse et l’ordonnée du cercle de corrélation de l’ACP. En effet, les coordonnées des variables dans le cercle de corrélation correspondent aux

corrélations des variables avec les deux premières composantes principales du cercle. Une corrélation élevée indique donc une forte contribution d'une variable à une composante.

3. Cosinus carrés

Les cosinus carrés des variables sur les composantes principales mesurent la proportion de la variance expliquée par une variable pour chaque composante. Une valeur élevée (proche de 1) indique que la variable contribue fortement à la formation de la composante, tandis qu'une valeur faible (proche de 0) signifie une influence moins importante.

Figure 11 : Cosinus carrés des variables



La dimension 1 est principalement influencée par la natalité (0.47) et la mortalité (0.41), ce qui suggère qu'elle capte principalement les dynamiques démographiques des départements. La dimension 2 est dominée par le taux de chômage (0.42) et la possession de voiture (0.38), mettant en évidence les inégalités socio-économiques. La dimension 3 est surtout expliquée par le PIB (0.38) et le nombre d'habitants (0.34), indiquant que cette composante est liée à la taille économique et démographique des départements. Enfin, la dimension 4 est davantage marquée par le taux d'éducation (0.46), reflétant une dimension

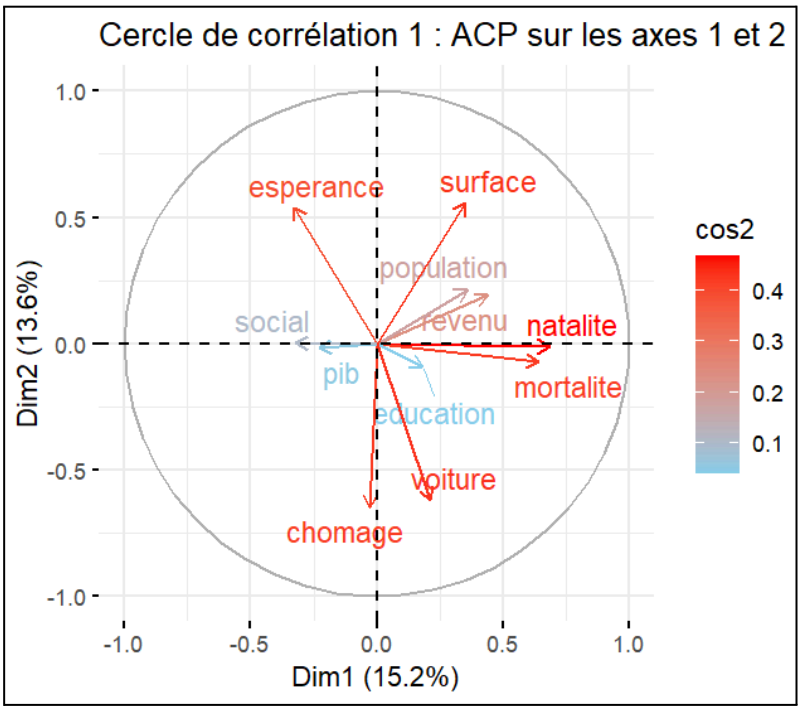
liée au capital humain, avec des contributions plus faibles des autres variables, comme le revenu (0.27).

Les cosinus carrés inférieurs à 0,5 indiquent que la contribution de chaque variable à la formation des composantes principales est modérée. Cela signifie qu'aucune variable ne domine complètement la variance expliquée par les dimensions. Bien que certaines variables exercent une influence notable sur certaines composantes (comme la natalité et la mortalité pour la dimension 1, ou le taux de chômage et la possession de voiture pour la dimension 2), les contributions sont réparties, chaque dimension représentant une combinaison linéaire de plusieurs variables. En conséquence, les cosinus carrés inférieurs à 0,5 montrent qu'aucune variable n'a une influence prédominante sur les dimensions, chaque composante principale étant une combinaison d'informations provenant de plusieurs variables.

C - Cercle de corrélation

Projetons désormais nos variables sur les différents cercles de corrélation des composantes principales afin de confirmer les résultats de notre analyse précédente.

Figure 12 : Analyse en composantes principales sur les axes 1 et 2

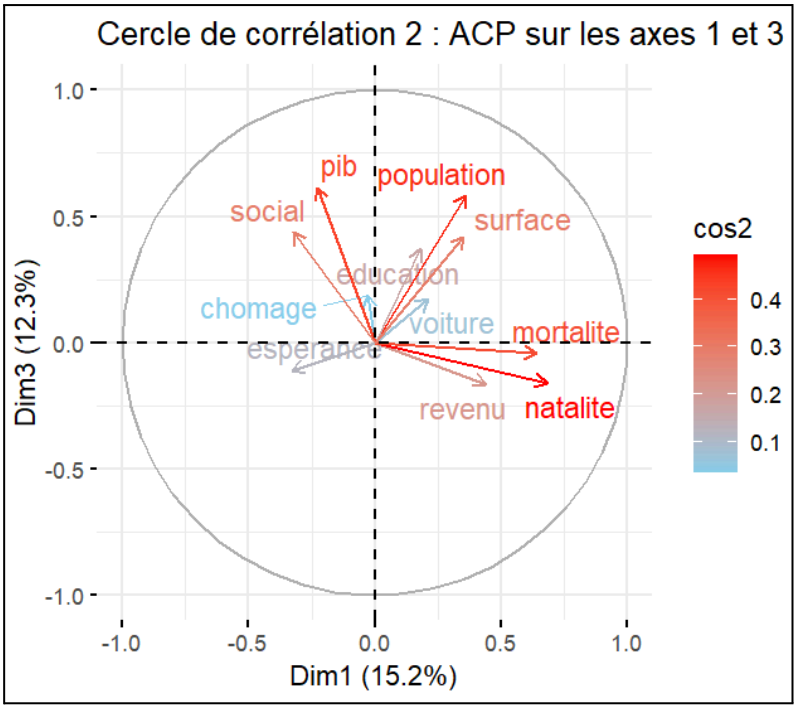


Ce premier cercle des corrélations montre la projection des variables (représentées par des vecteurs partant de l'origine) sur les deux premiers axes de l'ACP. Les première et deuxième dimensions correspondent respectivement à l'axe des abscisses et à celui des ordonnées. Elles présentent une inertie de 15,2 % et 13,6 %, ce qui signifie qu'elles expliquent respectivement 15,2 % et 13,6 % de la variance totale des données.

Les variables les plus éloignées de l'origine et proches du cercle de corrélation, telles que natalite, mortalite, chomage, voiture, esperance et surface, sont les mieux représentées sur le graphique et les plus fortement corrélées avec les deux composantes. Parmi ces variables, natalite et mortalite, alignées le long de l'axe de la première dimension (avec un faible angle par rapport à cet axe), sont positivement corrélées à la première composante. Ensuite, chomage et voiture, principalement alignées le long de l'axe de la deuxième dimension, sont négativement corrélées à la deuxième composante. Enfin, la variable surface est une combinaison positive des deux premières composantes, ce qui indique qu'elle est influencée positivement par ces deux dimensions. En revanche, la variable espérance est une combinaison négative de la première composante et positive de la seconde, ce qui suggère qu'elle est inversément liée à la première composante et positivement associée à la deuxième.

Par ailleurs, les variables qui forment un angle faible entre elles, telles que natalite et mortalite, sont très fortement corrélées positivement. Cela suggère qu'elles mesurent des aspects similaires des données. Si l'angle entre deux variables est d'environ 90°, comme celui entre natalite et chomage, il n'existe pas d'association linéaire entre celles-ci. En outre, si les points sont opposés et que l'angle est d'environ 180°, comme entre esperance et voiture, cela signifie que ces deux variables sont très fortement corrélées négativement. Cette dispersion des vecteurs et leur proximité respective autour de chaque axe des deux premières composantes principales montrent que ces variables sont bien représentées par ces dimensions, ce qui confirme que celles-ci expliquent une part significative de la variance totale.

Figure 13: Analyse en composantes principales sur les axes 1 et 3



Le deuxième cercle de corrélation concerne les axes 1 et 3, avec une inertie totale de 27,5% pour ces deux axes.

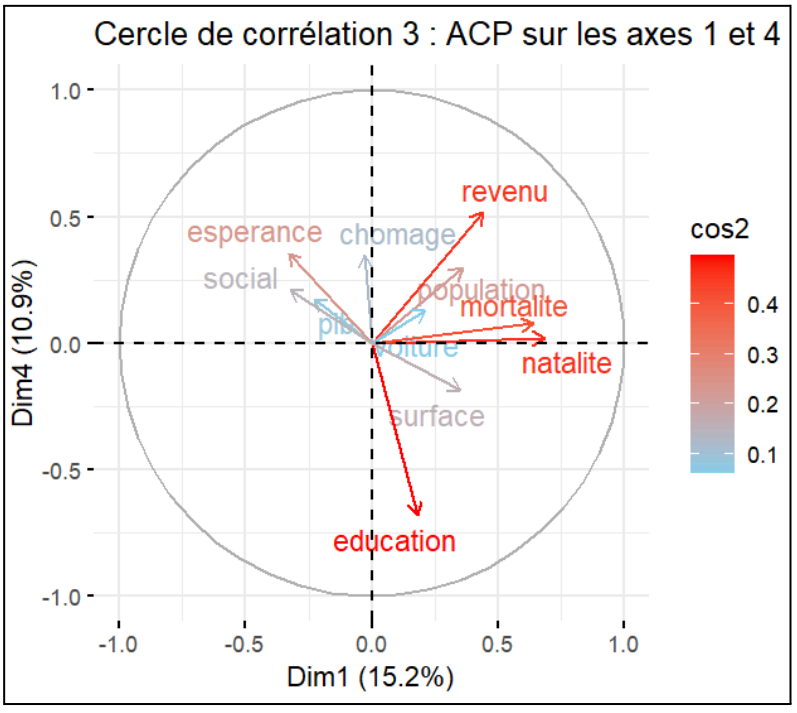
Examinons la variable PIB comme exemple. D'après la matrice des contributions, la contribution du PIB à la dimension 1 est relativement faible (3.17%), indiquant que cette variable n'a qu'un impact modéré sur cette composante principale. En revanche, sa contribution à la dimension 3 est beaucoup plus élevée (27.93%), soulignant le rôle significatif du PIB dans la formation de cette dimension. C'est pourquoi, dans ce deuxième cercle de corrélation, la projection du PIB sur l'axe de la dimension 3 est relativement éloignée de l'origine, indiquant une influence importante sur cette composante, mais celle sur l'axe de la dimension 1 est plus proche de l'origine, soulignant sa faible contribution à cette dimension.

Ensuite, en se basant sur la matrice des corrélations, la corrélation du PIB avec la dimension 1 est faible et négative (-0.23), ce qui signifie qu'une augmentation du PIB est faiblement associée à une diminution de cette composante. Par contre, la corrélation avec la dimension 3 est plus élevée et positive (0.61), indiquant une association positive plus marquée entre le PIB et cette composante. Nous observons ainsi, dans le cercle de corrélation, que la position du PIB forme un angle relativement grand (se rapprochant de 90°)

avec l'axe de la dimension 1, illustrant la faible corrélation négative, et un angle plus petit avec l'axe de la dimension 3, suggérant une forte corrélation positive.

Enfin, selon la matrice des cosinus carrés, le cosinus carré du PIB sur la dimension 1 est très faible (0.05), indiquant que sa contribution à la variance expliquée par cette dimension est négligeable. En revanche, le cosinus carré sur la dimension 3 est plus élevé (0.38), montrant que le PIB joue un rôle plus important dans la variance expliquée par cette composante. Dans le cercle de corrélation, nous constatons donc que la projection du PIB sur l'axe de la dimension 3 est plus éloignée de l'origine, confirmant son influence modérée, et que sa projection plus proche de l'origine sur l'axe de la dimension 1, reflète son faible impact sur cette dimension.

Figure 14: Analyse en composantes principales sur les axes 1 et 4



Le troisième cercle de corrélation se rapporte aux axes 1 et 3 qui expliquent ensemble 26,1% de la variance totale.

Pour mieux comprendre, considérons la variable éducation en exemple. D'après la matrice des contributions, le taux d'éducation contribue modérément à la dimension 1 (2,01 %) et de manière beaucoup plus significative à la dimension 4 (38,74 %). Dans le cercle de

corrélation, nous observons que cette contribution se traduit ainsi par une projection de la variable sur l'axe 4 relativement éloignée de l'origine, indiquant une forte implication de l'éducation dans cette dimension, tandis que sa projection sur l'axe 1 reste plus proche de l'origine, signalant une influence plus faible.

Selon la matrice des corrélations, le taux d'éducation présente une faible corrélation positive avec la dimension 1 (0,18) et une forte corrélation négative avec la dimension 4 (-0,68). Dans le cercle de corrélation, nous constatons donc que la variable "éducation" est relativement proche de l'axe de la dimension 4, mais dans la direction opposée à celle des variables fortement corrélées positivement avec cette dimension, ce qui reflète sa corrélation négative. Sa position éloignée de l'axe des abscisses de la dimension 1, formant un angle proche de 90°, indique une corrélation plus faible avec cette première dimension.

Enfin, la matrice des cosinus carrés révèle que le cosinus carré pour la dimension 1 est très faible (0,03), indiquant que l'éducation contribue peu à expliquer la variance de cette dimension. C'est pourquoi sa projection sur l'axe 1 dans le cercle de corrélation est très proche de l'origine. En revanche, pour la dimension 4, le cosinus carré est plus élevé (0,46), montrant que l'éducation joue un rôle plus significatif. Dans le cercle de corrélation, sa projection sur l'axe 4 est donc plus éloignée de l'origine, reflétant une contribution plus importante à cette dimension (annexe 1, 2 et 3).

E - Définition des variables latentes

Tableau 5 : Contributions des variables (en %) selon les axes du cercle

	Corrélations positives		Corrélations négatives	
	Variables	Contributions importantes	Variables	Contributions importantes
Axe 1	natalite mortalite	27,97 24,56		
Axe 2	surface esperance	20,64 19,25	chomage voiture	28,01 25,59
Axe 3	pib population	27,93 25,16		
Axe 4	revenu	22,21	education	38,74

Dans le contexte de l'analyse en composantes principales (ACP), les variables latentes ou composantes principales sont des variables synthétiques construites à partir des variables initiales de la base de données. Plus précisément, elles sont des combinaisons linéaires pondérées des variables initiales, calculées pour maximiser la variance expliquée par les données dans un espace de dimension réduite. Chaque composante principale résume l'information de plusieurs variables en une seule dimension, la première capturant la plus grande part de la variance totale, suivie par les autres, toutes orthogonales entre elles. Les coefficients associés aux variables indiquent leur contribution à chaque composante, ce qui facilite l'interprétation des dimensions latentes sous-jacentes, telles que des aspects économiques ou démographiques, selon les variables dominantes. Ces variables latentes permettent ainsi de résumer l'essentiel des informations contenues dans les données en regroupant les variables initiales en des indicateurs plus globaux et synthétiques.

Le tableau ci-dessus trie les variables en fonction des différentes dimensions ou axes, ce qui permet de définir les variables latentes associées. Chaque axe est associé aux variables initiales qui contribuent de manière significative à la variance expliquée par chaque

composante (avec des cosinus carrés supérieurs à 0,20 et des contributions supérieures à 15%) et selon la direction de leur corrélation (positive ou négative) avec cette composante. La sélection des variables les plus importantes et la direction de leur corrélation ont ensuite été validées par leur position sur les cercles de corrélation correspondants.

Nous pouvons ainsi définir les variables latentes associées aux différents axes. Tout d'abord, les variables natalite et mortalite présentent des corrélations positives et des contributions importantes à l'axe 1 (27,97 % pour natalite et 24,56 % pour mortalite). Cette première composante capture principalement les variations des taux de naissance et de décès, suggérant que la première variable latente peut ainsi être définie comme le dynamisme démographique.

Ensuite, les variables surface, esperance, chomage et voiture ont une contribution importante à l'axe 2 (respectivement de 20,64 %, 19,25 %, 28,01 % et 25,59 %). Parmi celles-ci, surface et esperance sont positivement corrélées avec cette deuxième dimension, alors que chomage et voiture sont négativement corrélées. Ainsi, l'augmentation de la surface verte par habitant et de l'espérance de vie a un impact positif, tandis que la hausse du taux de chômage et de la part des ménages possédant une voiture a un impact négatif. Cette variable latente semble donc refléter la qualité de vie dans le département. Une faible part de ménages possédant une voiture peut indiquer un environnement urbain où les réseaux de transports publics sont bien développés, réduisant la dépendance à la voiture. Cela conduit à moins de congestion, une meilleure qualité de l'air et une diminution de la pollution sonore. De plus, dans les zones urbaines, la densité des infrastructures et des services permet une accessibilité aisée à pied ou en transports publics, ce qui favorise une meilleure qualité de vie.

En outre, les variables pib et population présentent des corrélations positives et des contributions importantes à l'axe 3 (27,93 % pour pib et 25,16 % pour population). Cette troisième dimension reflète principalement la richesse économique des départements à travers la variable pib et la taille de la population via population. Cette variable latente peut donc indiquer la concentration économique et démographique des départements.

Enfin, les variables revenu et education ont une contribution importante à l'axe 4 (respectivement de 22,21 % et 38,74 %). Le revenu moyen annuel impact positivement cette quatrième dimension, tandis que le taux d'éducation l'influence négativement. La quatrième variable latente peut être définie comme une mesure de la tension entre le revenu moyen

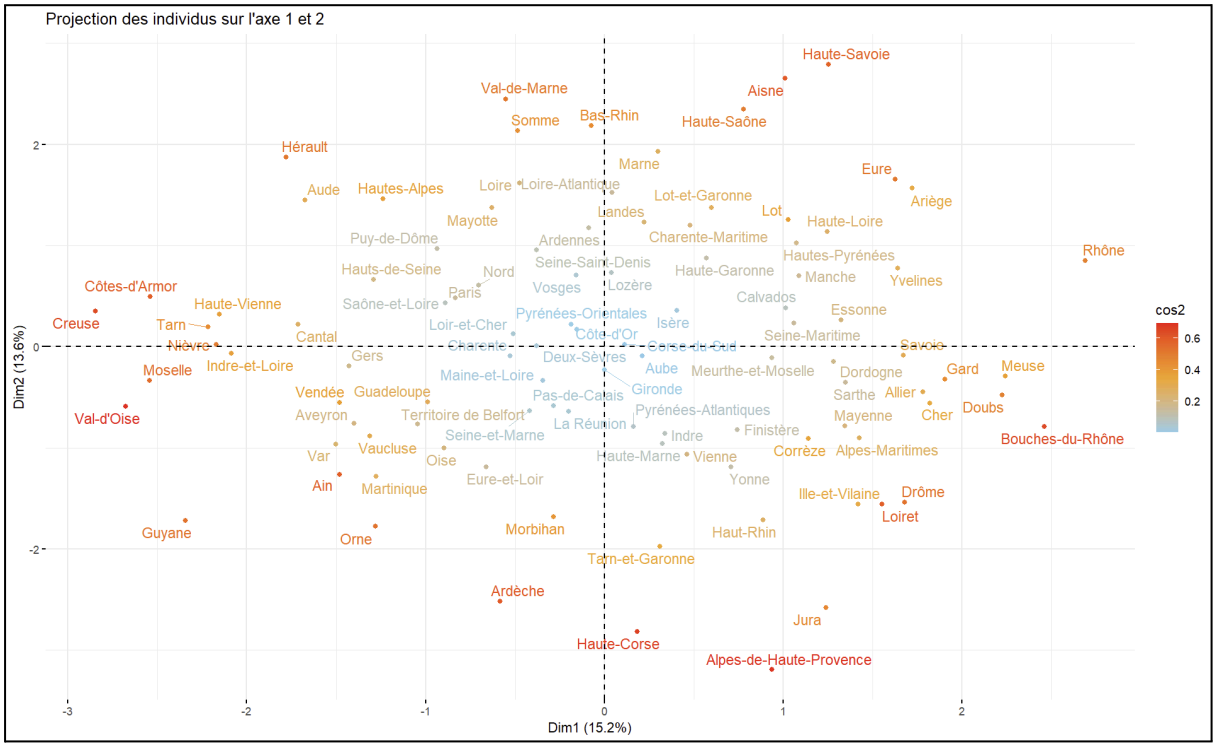
annuel et le taux d'éducation dans les départements. Elle met en évidence la relation inverse entre les niveaux de revenus et d'éducation, où des départements avec des revenus moyens plus élevés tendent à avoir des taux d'éducation plus bas, et inversement. Cette variable latente peut donc refléter un déséquilibre socio-économique, indiquant potentiellement des départements où la richesse financière ne correspond pas nécessairement à un niveau d'éducation élevé parmi la population.

La prochaine étape consiste à projeter les individus dans le plan factoriel afin d'évaluer la pertinence des variables latentes définies précédemment.

F - Projection des individus sur le plan factoriel et interprétation

La projection des individus sur le plan factoriel, formé par les axes 1 et 2, permet d'analyser la répartition des départements en fonction des deux premières composantes principales.

Figure 15: Projection des individus sur le plan factoriel des axes 1 et 2



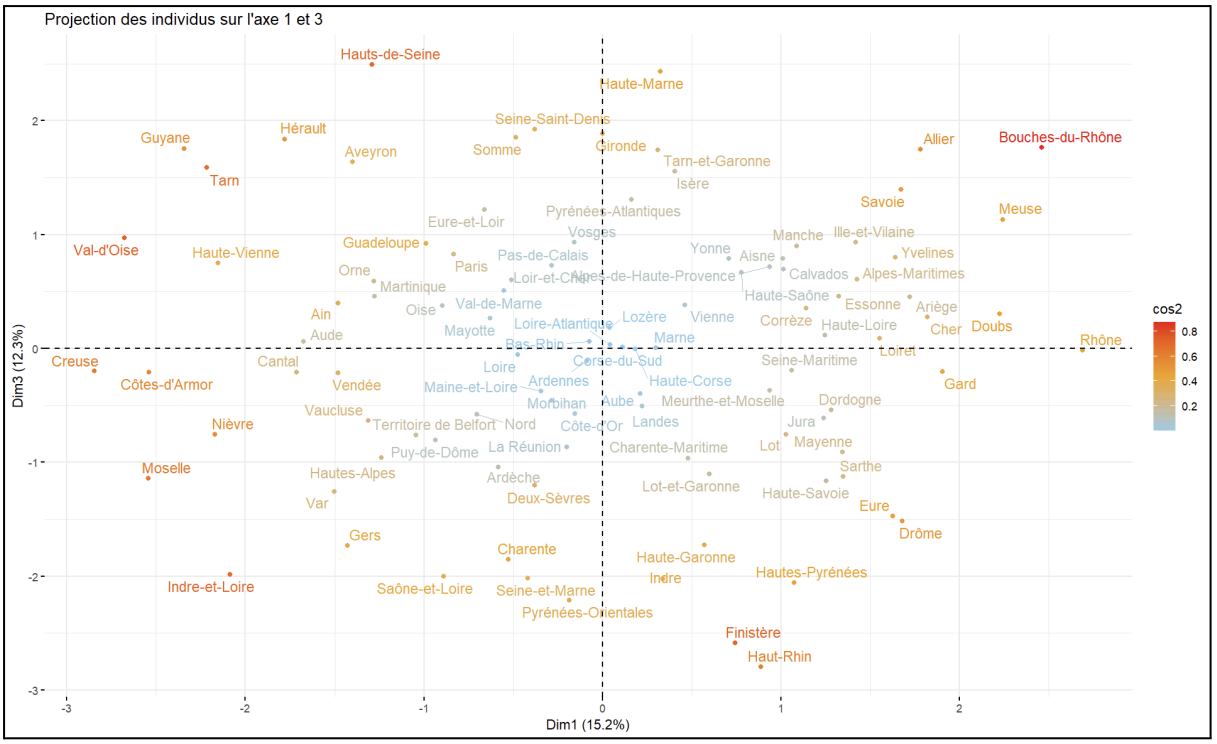
Les deux dimensions ici expliquent respectivement 15,2 % et 13,6 % de la variance totale des données. Les départements proches de l'origine indiquent une influence limitée des variables associées à ces axes. À l'inverse, les départements éloignés de l'origine, situés dans les extrémités des quadrants, montrent une forte corrélation avec les variables dominantes des axes. Par exemple, les départements dans le quadrant supérieur droit présentent des caractéristiques liées aux variables fortement corrélées positivement aux dimensions 1 et 2.

Les couleurs des points indiquent la qualité de la représentation (cos^2), avec des teintes rouges/orangées traduisant une bonne qualité. Cette analyse visuelle facilite l'identification des groupes homogènes de départements et des dimensions qui expliquent leurs disparités.

Dans un premier temps, nous observons que les départements les plus à gauche ont des niveaux faibles de natalité et de mortalité. Par exemple, le département du Rhône est classé 18ème en termes de natalité et 3ème en termes de mortalité. Par conséquent, la dimension 1, définie par la première variable latente, reflète effectivement le dynamisme démographique dans ces départements, caractérisé par des taux relativement faibles de natalité et de mortalité.

Dans un second temps, nous constatons que les départements situés en haut du graphique présentent une qualité de vie relativement élevée. Par exemple, la Haute-Savoie se classe 10e en termes de surface verte par habitant, 7e en espérance de vie, 86e pour le taux de chômage (élevé) et 98e pour la part des ménages possédant une voiture. La deuxième variable latente reflète effectivement cette qualité de vie, avec une corrélation positive pour la surface verte par habitant et l'espérance de vie, et une corrélation négative pour le taux de chômage et la part de ménages possédant une voiture.

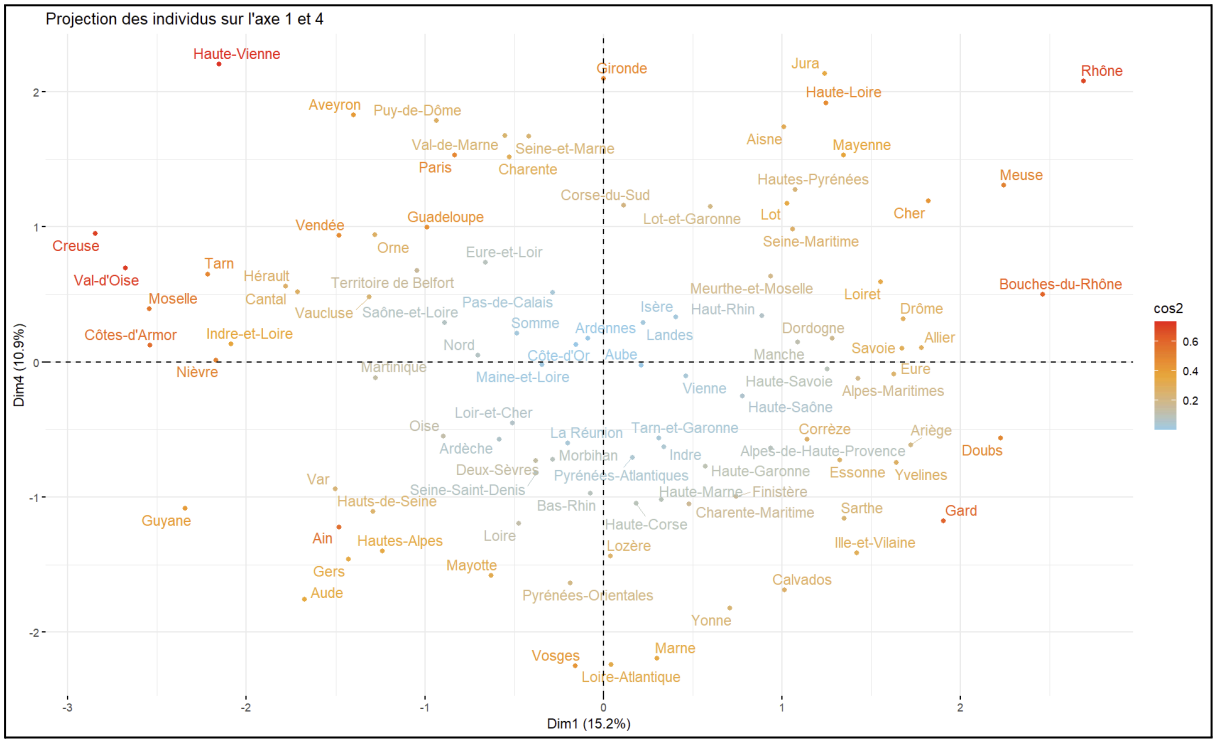
Figure 16: Projection des individus sur le plan factoriel des axes 1 et 3



Sur cette projection, nous observons que les départements situés dans la partie haute du graphique sont fortement associés à un PIB élevé et à une grande population. Par exemple, le département des Hauts-de-Seine occupe la 8ème position en termes de PIB et est le 24ème département le plus peuplé. Les départements situés en bas du graphique, tels que le Haut-Rhin, montrent une position opposée, indiquant des caractéristiques moins fortement corrélées avec les indicateurs principaux de cet axe. Le Haut-Rhin se classe ainsi au 100ème rang en termes de PIB et au 90ème rang parmi les départements les moins peuplés. La troisième variable latente, représentée par l'axe 3, reflète effectivement la concentration économique et démographique des départements.

Par ailleurs, les départements situés aux extrémités des quadrants, comme les Bouches-du-Rhône en haut à droite, montrent une forte corrélation avec les variables latentes des dimensions 1 et 3, représentant le dynamisme démographique et la concentration économique et démographique. Les Bouches-du-Rhône se distinguent par leur classement élevé concernant le taux de natalité (25ème), le taux de mortalité (13ème), le PIB départemental (33ème) et la population totale (14ème).

Figure 17: Projection des individus sur le plan factoriel des axes 1 et 4



Cette troisième projection permet de confirmer ou non la définition de la quatrième variable latente représentée par l'axe des ordonnées. Les départements situés en haut contribuent fortement et positivement au 4ème axe, tandis que ceux situés en bas y contribuent fortement et négativement. D'une part, la Gironde, en haut, est le 28ème département où le revenu moyen annuel est le plus élevé, tandis que la Loire-Atlantique est le 62ème ; d'autre part, la Gironde est le 62ème département où le taux d'éducation est le plus élevé, alors que la Loire-Atlantique est le 3ème. Le classement des départements explique leur position sur cette projection, confirmant que la quatrième variable latente reflète un certain déséquilibre socio-économique, caractérisé par une tension entre le revenu moyen annuel et le taux d'éducation dans les départements.

D'autre part, les départements situés aux extrémités des quadrants présentent des caractéristiques fortement corrélées avec les variables latentes. Par exemple, le Rhône, situé en haut à droite du graphique, est classé 18ème pour le taux de natalité, 3ème pour le taux de mortalité, 20ème pour le revenu moyen annuel et 83ème pour le taux d'éducation. La position et le classement du département expliquent ainsi les particularités des variables latentes.

Ces analyses visuelles facilitent l'identification des groupes homogènes de départements et des dimensions qui expliquent leurs disparités.

V - Régression linéaire multiple

A - Sélection des variables explicatives significatives

Nous allons maintenant effectuer une régression linéaire multiple de notre base de données. Notre modèle de base se compose de 15 variables explicatives.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \beta_8 X_{8i} + \beta_9 X_{9i} + \beta_{10} X_{10i} + \beta_{11} X_{11i} + \beta_{12} X_{12i} + \beta_{13} X_{13i} + \varepsilon_i$$

Avec : X1i = population, X2i = chomage, X3i = pib, X4i = esperance, X5i = natalite, X6i = mortalite, X7i = education, X8i = voiture, X9i = sociaux, X10i = surface, X11i = littoral, X12i = rurale, X13i = metropole

Notre fonction se base sur le critère d'information d'Akaike (AIC), proposé par Hirotugu Akaike en 1973, pour choisir notre modèle. Par sa formule AIC = −2*log(L) + 2*k, avec L, la vraisemblance du modèle, et k, le nombre total de paramètres estimés, ce critère permet de comparer les modèles en évaluant lequel s'ajuste le mieux aux données tout en évitant le surajustement. Les trois méthodes de sélection minimisent ainsi l'AIC pour nous proposer le meilleur modèle.

Après avoir effectué la fonction step, nous obtenons AIC = 1715.91 pour la méthode ascendante, AIC = 1715.91 concernant la sélection descendante et AIC = 1715.91 pour la méthode à double sens. Nous retenons donc le modèle ayant un critère d'Akaike de 1715.91, car, par définition, plus ce critère est faible, meilleur est le modèle et pour ce cas présent les 3 trois sont les mêmes.

Tableau 6: Tableau récapitulatif de notre modèle après stepwise

	Variables
Y	revenu
X_1	mortalite
X_2	natalite

X_3	metropole1
X_4	education

B - Choix du meilleur modèle

Après sélection de notre modèle grâce à la méthode Stepwise, nous pouvons effectuer une régression linéaire de notre modèle.

Tableau 7: Régression linéaire multiple de notre modèle (MCO)

	Estimate	Std. Error	t value	Pr(> |t|)
Constante	20366.8	5311.8	3.834	0.000225 ***
mortalite	757	359.7	2.105	0.031 *
natalite	397.3	229.8	1.729	0.087 .
metropole1	2152.4	1253.5	1.717	0.089 .
education	-69.6	50.1	-1.389	0.167
Multiple R-squared: 0.1141		Adjusted R-squared: 0.07716		p-value: 0.01934

Après vérification des hypothèses, ce tableau nous permettra d'interpréter le test de Fisher. Celui-ci nous permettra de valider ou non l'hypothèse H_0 selon laquelle aucune des variables indépendantes n'a d'effet significatif sur la variable dépendante. Dans ce cas présent, la p-value inférieure à 0,05 nous permettrait de rejeter H_0 ; ce qui signifierait qu'il y a au moins une variable qui explique le revenu.

Ensuite, nous pourrons effectuer une vérification de la significativité individuelle de nos variables explicatives avec un test de Student. En effet, il apparaît que la variable mortalite serait significative au seuil de 5 % et les variables natalite et metropole1 le

serait au seuil de 10 %. Nous effectuerons l'interprétation des coefficients après validation des tests statistiques vérifiant les hypothèses de la méthode des MCO.

C - Test statistiques

Tableau 9: Test de Kolmogorov-Smirnov

Test de Kolmogorov-Smirnov	
p-value	0.7724

Une p-value supérieure à 0,05 indique que l'hypothèse nulle ne peut pas être rejetée, ce qui signifie que les résidus du modèle suivent une distribution normale. Cela garantit que l'une des hypothèses clés des moindres carrés ordinaires (MCO) est respectée, assurant ainsi la validité des estimations du modèle (annexe 4).

Tableau 9: Test de Ramsey

Test de Ramsey	
p-value	0.6885

Une p-value supérieure à 0,05 indique que l'hypothèse nulle ne peut pas être rejetée, ce qui signifie que le modèle est bien spécifié et que la forme fonctionnelle linéaire choisie est appropriée pour représenter la relation entre les variables explicatives et la variable dépendante (annexe 5).

Tableau 10 : Test de Breusch-Pagan

Test de Breusch-Pagan	
p-value	0.1117

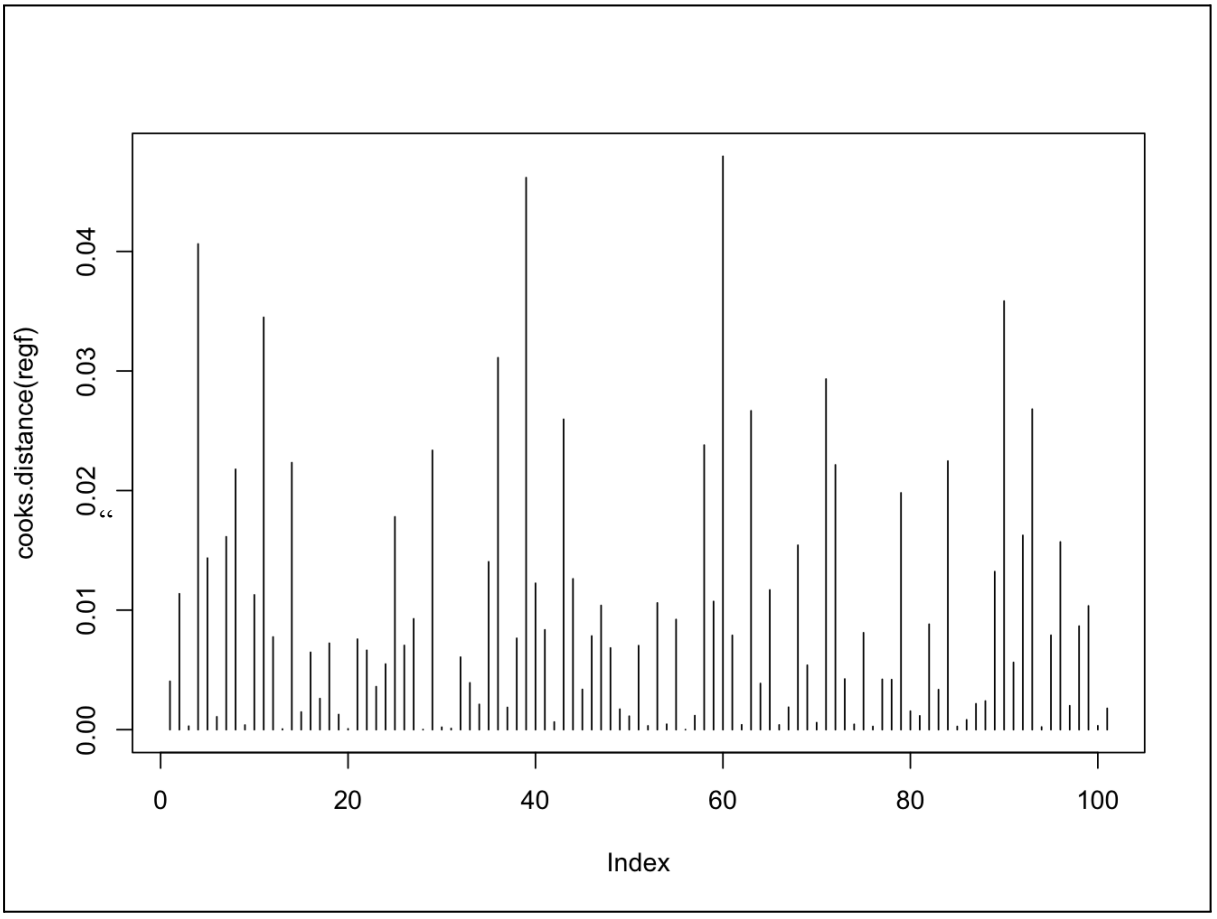
Une p-value supérieure à 0,05 indique que l'hypothèse nulle ne peut pas être rejetée au seuil de 5 %, ce qui suggère que les résidus du modèle sont homoscédastiques.

Tableau 11: Test de VIF

Variables	Vif
mortalite	1.05
natalite	1.04
metropole	1.02
education	1.02

Les VIF montrent des valeurs proches de 1 pour chaque variable. Il n'existe donc pas de problème de multicolinéarité entre les variables explicatives du modèle. Par conséquent, aucune des variables n'a besoin d'être supprimée.

Figure 18: Distance de cook



D’après le graphique ci-dessus, étant donné que la distance de Cook est nettement inférieure à 1, aucune observation n'exerce une grande influence sur les estimations des paramètres du modèle ou sur les prédictions. Il n'est donc pas nécessaire de supprimer des observations.

En conclusion, toutes les hypothèses des Moindres Carrés Ordinaires (MCO) semblent être respectées, à savoir : la normalité des résidus, la forme fonctionnelle linéaire appropriée, l'absence de multicolinéarité, l’homoscédasticité des erreurs et l'absence d'autocorrélation des résidus. Nous pouvons donc passer à l’interprétation des résultats.

D - Interprétations des résultats significatifs

Taux de mortalité (+757) ⇒ Une augmentation de 1 pour 1000 du taux de mortalité est associée à une augmentation moyenne de 757 euros du revenu annuel, toutes choses égales par ailleurs.

Taux de natalité (+397.3) ⇒ Une augmentation de 1 pour 1000 du taux de natalité est associée à une augmentation de 397.3 unités du revenu annuel, toutes choses égales par ailleurs.

Vivre dans une métropole (+ 2152.4) ⇒ Une augmentation de 1 unité de la variable métropole est associée à une augmentation de 2152.4 unités du revenu annuel, toutes choses égales par ailleurs.

E - Variables latentes et sélection des variables explicatives significatives

Tableau 12: Tableau récapitulatif de notre modèle

	Variables
Y	revenu
X_1	variable latente 1
X_2	variable latente 2
X_3	variable latente 3
X_4	littoral1

Après sélection de notre modèle grâce à la méthode Stepwise, nous pouvons effectuer une régression linéaire de notre modèle.

Tableau 13: Régression linéaire multiple de notre modèle (MCO)

	Estimate	Std. Error	t value	Pr(> |t|)
Constante	26336.4	409.7	64.275	< 2e-16 ***
X_1	1738.6	266.9	6.514	3.47e-09 ***
X_2	801.9	280.2	2.862	0.00518 **
X_3	-762.7	297.1	-2.567	0.01180 *
X_4	1541.6	781.0	1.974	0.05129 .
Multiple R-squared: 0.5427		Adjusted R-squared: 0.5186		p-value: 7.627e-15

Après vérification des hypothèses, nous pouvons interpréter nos résultats. Tout d'abord, nous constatons que la p-value du test de Fisher est bien inférieure à 0,05, ce qui indique que notre modèle global est statistiquement significatif, suggérant qu'au moins un de nos coefficients est significativement différent de zéro.

Ensuite, les p-values du test de Student pour l'ensemble des variables explicatives sont inférieures aux seuils conventionnels. Les quatre variables explicatives sont donc statistiquement significatives aux seuils de 0,1 %, 1 %, 5 %, et 10 %, respectivement.

Dans un premier temps, la variable latente (X1), le dynamisme démographique, mesuré par les taux élevés de natalité et de mortalité, a un effet positif et significatif sur le revenu moyen annuel dans les départements français. Chaque augmentation d'une unité de ce dynamisme est associée à une hausse moyenne de 1738,6 euros du revenu annuel.

Dans un second temps, la variable latente (X2) montre que la qualité de vie a un effet positif significatif sur le revenu moyen annuel. Chaque unité d'amélioration de la qualité de vie augmente le revenu moyen annuel de 801,9 euros en moyenne. Cette relation, statistiquement significative, confirme que les départements avec une meilleure qualité de vie (plus de surface verte, une espérance de vie plus élevée, un taux de chômage plus faible, et une moindre dépendance à la voiture) tendent à avoir des revenus moyens plus élevés.

Puis, la variable latente (X3), représentant la concentration économique et démographique des départements, montre un effet négatif significatif sur le revenu moyen annuel. Chaque unité d'augmentation de cette concentration réduit le revenu moyen annuel de 762,7 euros en moyenne. Cette relation, statistiquement significative, peut être expliquée par des inégalités de revenu plus marquées pour une partie de la population. Les zones urbaines, souvent plus peuplées et économiquement dynamiques, peuvent être caractérisées par une concentration de personnes ayant des revenus moyens plus bas, en particulier pour les travailleurs dans des secteurs moins rémunérés ou dans des emplois précaires. Cette dynamique peut réduire le revenu moyen annuel global, même si la région est économiquement active.

Enfin, notre dernière variable explicative indique que les départements situés sur le littoral ont un revenu moyen annuel supérieur de 1541,6 € par rapport à ceux qui ne le sont pas.

Comme nous pouvons l'observer, il existe des différences entre la régression linéaire multiple avec nos variables initiales et celle avec nos variables latentes issues de l'analyse en composantes principales. En effet, pour la régression linéaire multiple classique, nous avions en significatif les variables de la mortalité, de la natalité et du fait de vivre en métropole. Ces variables isolées permettaient de capturer des aspects spécifiques des variations du revenu moyen annuel, mais sans prendre en compte les relations complexes et les corrélations entre les variables explicatives.

En revanche, l'analyse en composantes principales permet de réduire la dimensionnalité en regroupant ces variables dans des axes latents (ou composantes principales), ce qui simplifie l'interprétation tout en intégrant davantage de complexité. Cette approche révèle des facteurs latents, comme le dynamisme démographique, la qualité de vie et la concentration économique et démographique, qui offrent une vue d'ensemble plus globale. Ces composantes principales permettent de mieux comprendre les relations sous-jacentes entre les variables explicatives et le revenu moyen annuel.

Ainsi, bien que la régression linéaire multiple soit utile pour identifier l'effet direct de variables spécifiques, elle peut manquer de cohérence lorsque les variables explicatives sont fortement corrélées ou redondantes. En comparaison, les variables latentes capturent ces interdépendances, ce qui en fait un outil puissant pour analyser des phénomènes complexes dans des contextes où les variables sont fortement interreliées.

Cependant, la régression utilisant les variables latentes de l'ACP présente également des limites, notamment en ce qui concerne l'interprétation des composantes principales, qui sont des combinaisons linéaires des variables d'origine et peuvent parfois être difficiles à relier directement à des phénomènes observables. En conclusion, ces deux approches sont complémentaires : la régression linéaire multiple avec les variables initiales permet une analyse plus ciblée, tandis que celle avec les variables latentes offre une vision globale, particulièrement utile dans des contextes où les interactions entre variables jouent un rôle clé.

VI - Conclusion

La visée de notre projet était de comprendre les disparités économiques et sociales entre les départements français et d'identifier les facteurs influençant le revenu moyen annuel par territoire. Pour cela, nous avons mobilisé plusieurs approches analytiques, notamment une Analyse en Composantes Principales et une régression linéaire multiple, qui ont permis de réduire la complexité des données et de dégager des enseignements clés.

L'ACP a révélé quatre dimensions principales expliquant 52 % de la variance totale. La première dimension met en évidence le dynamisme démographique, principalement influencé par des variables comme la natalité et la mortalité. La seconde dimension reflète la qualité de vie, liée à l'espérance de vie, à la surface verte par habitant, au taux de chômage ainsi qu'à la part de ménages véhiculés. La troisième dimension capture la concentration économique et démographique des départements, fortement corrélée au PIB et à la population. Enfin, la quatrième dimension souligne l'impact du capital humain et des tensions socio-économiques, avec des relations entre le revenu moyen annuel et le taux d'éducation. Ces dimensions permettent de visualiser les disparités entre départements, où le Rhône se distingue par son dynamisme démographique, la Haute-Savoie par sa qualité de vie, les Hauts-de-Seine par sa concentration économique et démographique, et la Gironde par ses tensions socio-économiques, marquées par un revenu moyen annuel élevé en contraste avec un taux d'éducation relativement plus faible.

La régression linéaire multiple met en évidence des différences entre le modèle utilisant les variables initiales et celui exploitant les variables latentes issues de l'ACP. Ces deux approches diffèrent dans leur capacité à expliquer la variable dépendante. D'une part, la régression classique identifie des variables spécifiques, telles que le taux de mortalité, le taux de natalité et la présence d'une métropole, mais son pouvoir explicatif reste limité, avec un R^2 ajusté de 11,4 %. Ce faible pouvoir explicatif peut être attribué à l'absence de prise en compte des corrélations entre les variables explicatives. Bien qu'elle permette d'identifier des effets directs, cette approche devient moins cohérente lorsque les variables sont corrélées entre elles. D'autre part, la régression utilisant les variables latentes issues de l'ACP, qui synthétisent les phénomènes complexes et capturent les dépendances entre les variables initiales, offre un pouvoir explicatif plus élevé, avec un R^2 ajusté de 0,5427. Cette approche permet d'obtenir une vue d'ensemble plus complète des relations entre les facteurs et le

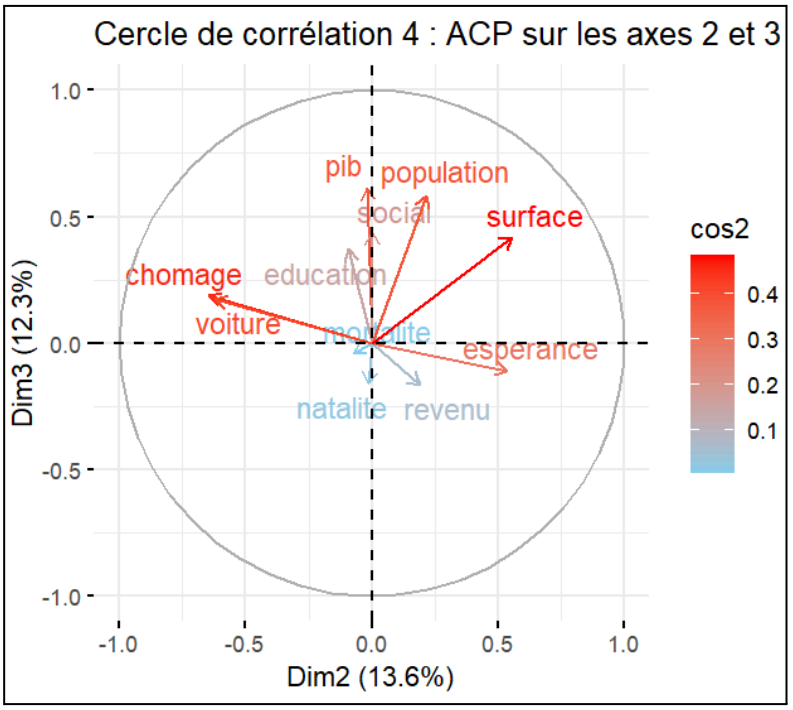
revenu moyen annuel, en tenant compte des interactions entre les variables. Les deux approches sont donc complémentaires : la régression classique permet une analyse ciblée des effets directs, tandis que l'ACP fournit une perspective globale en intégrant les relations entre les variables explicatives.

Après analyse de nos résultats, nous pouvons donner certaines recommandations pour améliorer le revenu annuel par département.. Il serait pertinent d'investir davantage dans l'amélioration de la qualité de vie en facilitant l'accès aux soins de santé, notamment dans les départements les plus défavorisés. Le développement économique des zones rurales et littorales pourrait être soutenu par la diversification des activités économiques et par la création de nouvelles opportunités professionnelles (délocalisation d'entreprise). Enfin, renforcer les initiatives éducatives dans les départements où le niveau de revenu est élevé mais où le taux d'éducation est faible pourrait contribuer à réduire les déséquilibres socio-économiques.

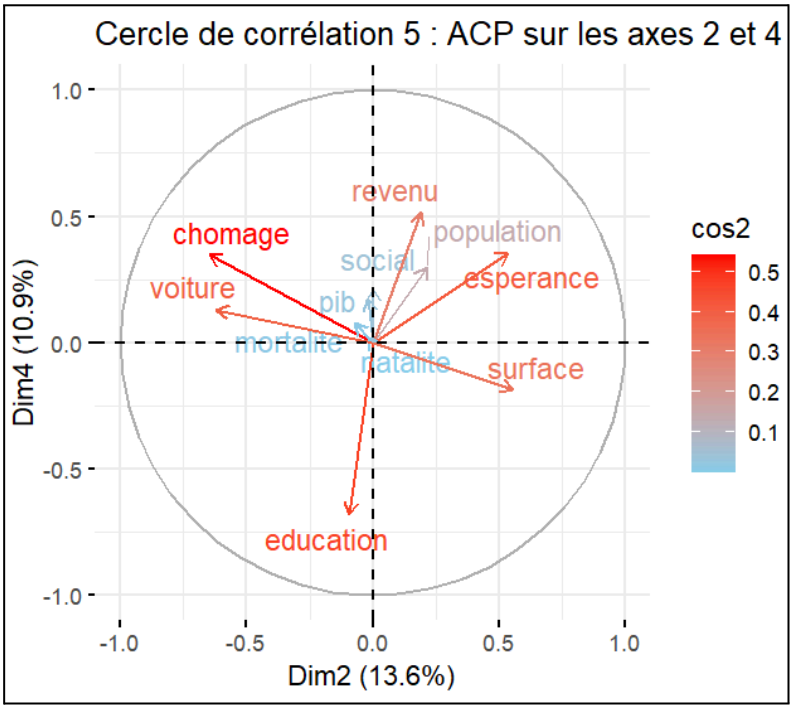
En conclusion, cette étude met en lumière les facteurs clés des disparités territoriales en France. Étendre cette analyse à d'autres échelles géographiques, comme les pays de l'Europe, permettrait d'approfondir cette compréhension et d'enrichir les perspectives sur les dynamiques économiques et sociales à l'échelle européenne

VII - Annexes

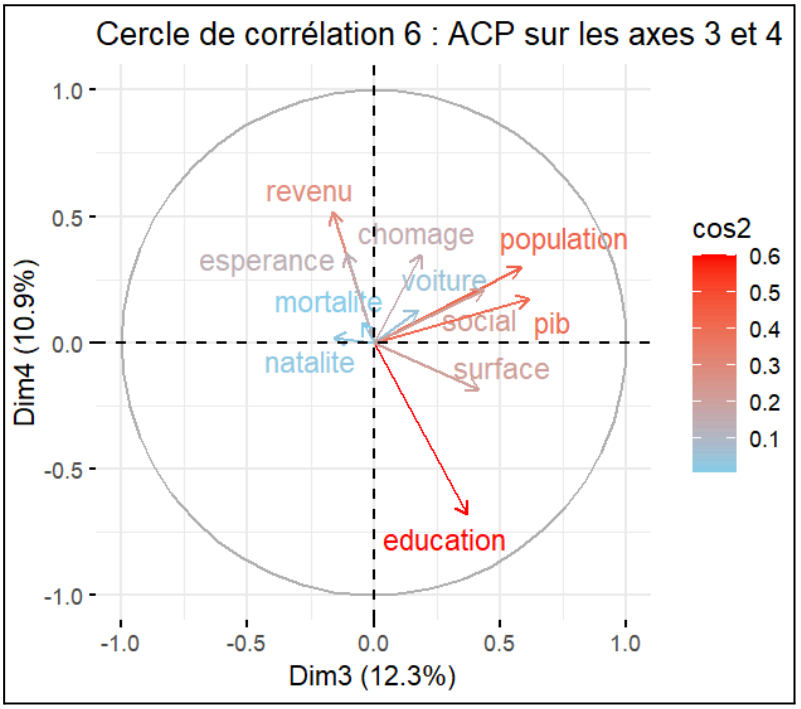
Annexe 1 : Analyse en composantes principales sur les axes 2 et 3



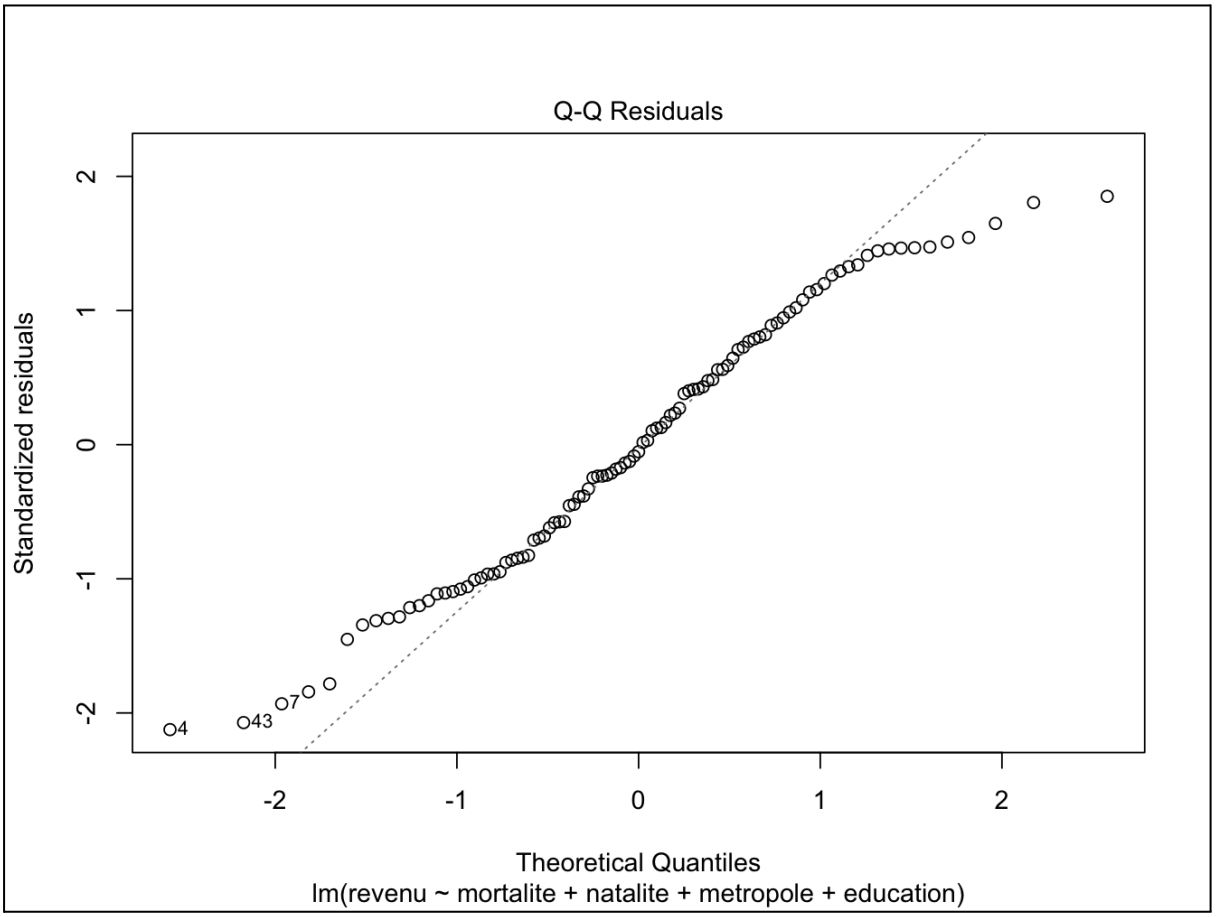
Annexe 2 : Analyse en composantes principales sur les axes 2 et 4



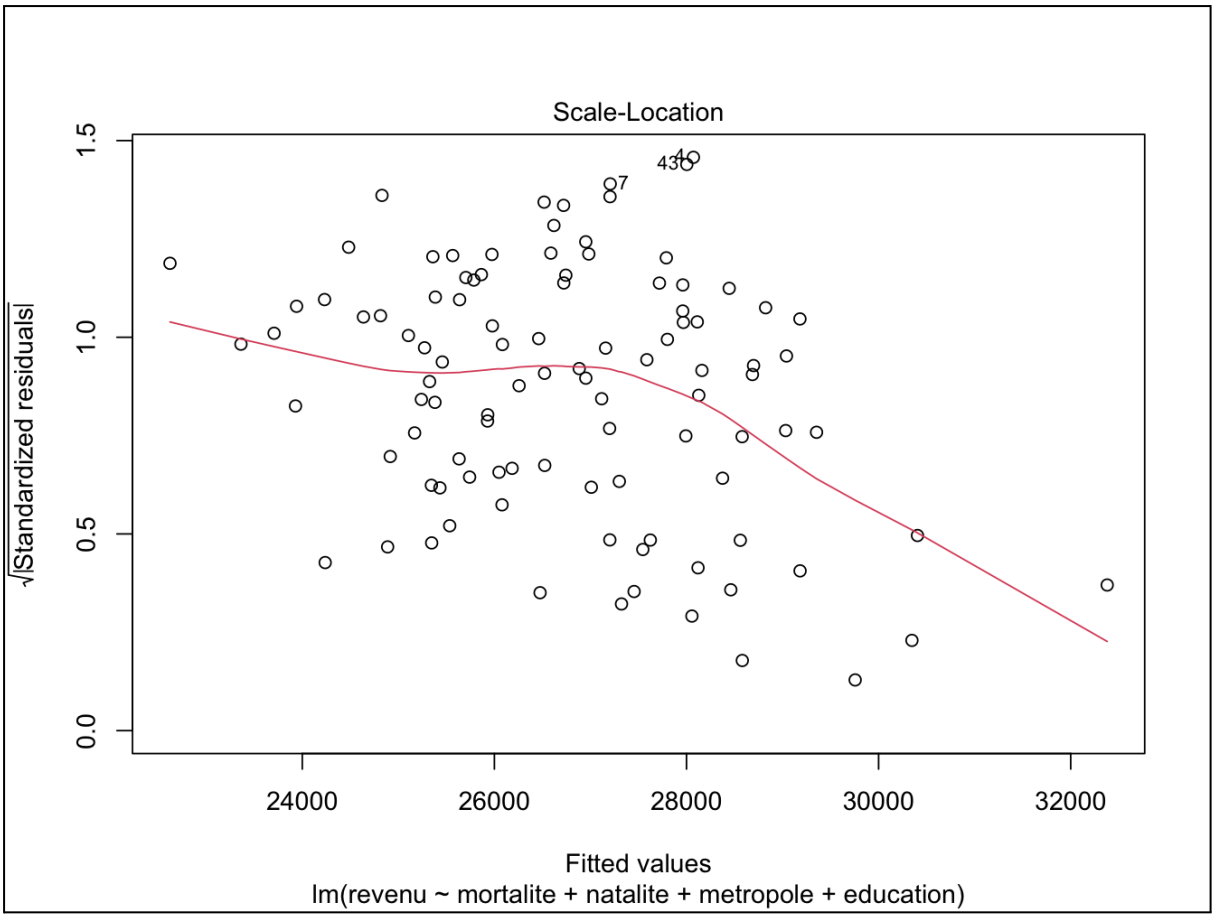
Annexe 3 : Analyse en composantes principales sur les axes 3 et 4



Annexe 4: Q-Q plot des résidus



Annexe 5: Graphique de la distribution des résidus en fonction des valeurs prédites



VIII - Bibliographie

INSEE. (2023). Recensements de la population. INSEE - Recensements de la population.

INSEE. (2023). Taux de chômage par département. INSEE - Indicateurs du marché du travail.

INSEE. (2023). Revenus et niveaux de vie par département. INSEE - Statistiques sur les revenus.

INSEE et Eurostat. (2022). Comptes économiques régionaux et locaux. Eurostat.

INSEE. (2023). Espérance de vie à la naissance et par département. INSEE - Espérance de vie.

INSEE. (2018). Taux de natalité par département. INSEE - Indicateurs démographiques.

INSEE. (2023). Indicateurs de mortalité par département. INSEE - Statistiques sur la mortalité.

Ministère de l'Éducation Nationale. (2023). Données sur l'éducation par département. Portail des données de l'Éducation Nationale.

INSEE. (2023). Équipements des ménages : Proportion possédant une voiture. INSEE - Statistiques sur les équipements.

DREES. (2022). Données sur les aides sociales et prestations. DREES.

IGN. (2023). Superficie des départements. Institut Géographique National.

DREAL. (2023). Indicateurs environnementaux pour les départements côtiers. DREAL Portail.

Banque Mondiale. (2022). Indicateurs de développement économique et social. Banque Mondiale.

OCDE. (2019). Indicateurs territoriaux. Organisation de Coopération et de Développement Économiques.

Eurostat. (2022). Statistiques sur les régions d'Europe. Eurostat.

Observatoire des Inégalités. (2024). Rapport sur les inégalités territoriales en France.

PNUD. (2024). Rapport sur le développement humain. Programme des Nations Unies pour le Développement.

IX - Table des matières