Université de Caen Normandie

IUT Grand Ouest Normandie

Bachelor Universitaire de Technologie

Science des Données

Campus de Lisieux

Science des Données 2 - SAE 4.02

Reporting d'une analyse multivarié

Thématique

Clustering des pratiques agricoles - cohorte Agrican 1960

Auteurs

MILON Louis
DIOP Mandir
UCENDO Sacha

Année universitaire 2024-2025

Table des matières

Introduction **3**

1. Calcul des matrices : ratio de pratique de l'activité professionnelle et ratio de pratique de la tâche agricole **4**

1.1. Calcul de la matrice des ratios de pratique de l'activité professionnelle 4

1.1.1. Calcul de la durée de l'activité professionnelle 4

1.1.2. Calcul de la durée de pratique de l'activité 4

1.2. Calcul de la matrice des ratios de pratique dans un secteur agricole 4

1.2.1 Calcul de la durée de pratique de la tâche agricole 5

2. Analyse en composante principale sur la matrice associée aux ratios de pratique de l'activité **6**

2.1. Éditons la matrice des corrélations 6

2.2. Réalisons une ACP sur les données 6

2.3. Éditons un tableau de synthèse des valeurs propres, pourcentage d'inertie et pourcentage d'inertie cumulée 7

2.4. Éditons le tableau des corrélations entre les variables et les axes factoriels 8

3. Classification automatique via la méthode des k-Means **9**

3.1. Identification du nombre optimal de clusters 9

3.2. Création des clusters 10

3.3 Caractéristiques des clusters 10

Commentaire du Tableau 6 : Caractéristiques démographiques par cluster 11

Conclusion **12**

Introduction

Notre étude porte sur la cohorte Agrican. Agrican dont “Agri” fait référence à l’agriculture et “Can” au cancer, s’intéresse à la santé des agriculteurs, mais aussi à celles de personnes qui travaillent dans les espaces verts, les coopératives agricoles, les forêts, le secteur maritime, et de nombreux secteurs connexes à l’agriculture. Notre population d’intérêt est constituée de 12310.

Au sein de cette population, nous allons nous interéssé à la sous-cohorte 1950 constituée d’agriculteurs ayant débuté leur carrière entre 1950 et 1970. Cet échantillon contient les informations suivantes :

- Informations agriculteurs : année de naissance, sexe, année de début de carrière dans l’agriculture
- Nom de chaque activité : codé en 0 ou 1 pour renseigner sur la pratique ou non de l’activité par l’agriculteur
- Nom de chaque tâche : codé en 0 ou 1 pour renseigner sur la pratique ou non de la tâche par l’agriculteur
- Année de début de début de chaque activité
- Année de fin de chaque activité
- Année de début de début de chaque tâche
- Année de fin de chaque tâche
- Statut tabagique, durée consommation tabac, nombre de paquet-année de tabac
- Ainsi que plusieurs autres variables

L’objectif de notre étude est de créer des profils agricoles en fonction des activités pratiquées durant leur carrière professionnelle et d’y apporter une description. Pour cela, nous allons calculer dans un premier temps quelques indicateurs en pourcentage nous permettant de reconstituer la carrière professionnelle de chaque agriculteur de la sous-cohorte. Ensuite, à partir des pourcentages en rapport aux activités, nous allons réaliser une analyse en composante principale (ACP) et réaliser un clustering de type k-means sur les données issues de l’ACP. Enfin, nous allons décrire les classes d’agriculteurs issues du clustering.

1. Calcul des matrices : ratio de pratique de l'activité professionnelle et ratio de pratique de la tâche agricole

1.1. Calcul de la matrice des ratios de pratique de l'activité professionnelle

Dans cette partie, il s'agit de calculer pour chacun des agriculteurs présents dans notre de jeu de données, le ratio de pratique de l'activité professionnelle. L'objectif est de connaître le pourcentage de la durée de chaque activité pratiqué par l'agriculteur par rapport à la durée de toute son activité professionnelle. Ainsi, le ratio de l'activité professionnelle se calcule comme suit :

$$\text{Ratio pratique activité} = \frac{\text{durée de pratique de l'activité}}{\text{durée de l'activité professionnelle}}$$

Afin de calculer ce ratio, nous allons d'abord calculer dans un premier la durée de pratique de chaque activité et la durée de l'activité professionnelle pour chaque agriculteurs de la cohorte 1960.

1.1.1. Calcul de la durée de l'activité professionnelle

La durée de l'activité professionnelle s'obtient comme suit :

$$\text{Durée de l'activité professionnelle} = \text{année de fin de pratique} - \text{année de début de pratique}$$

Chaque agriculteur pouvant pratiqué plusieurs activités, l'année de fin de pratique correspond à l'année de fin la plus récente de l'ensemble des activités réalisés et l'année de début de pratique correspond à l'année de début la plus ancienne de l'ensemble les les activité réalisés.

1.1.2. Calcul de la durée de pratique de l'activité

La durée de pratique d'une activité correspond à la durée pendant laquelle chaque agriculteur a réalisé l'activité durant sa carrière. Elle s'obtient comme suit :

$$\text{Durée de pratique de l'activité} = \text{année de fin de pratique de l'activité} - \text{année de début de pratique de l'activité}$$

Chaque activité contenant plusieurs tâches, l'année de fin de pratique de l'activité correspond à l'année de fin la plus récente de l'ensemble des tâches réalisées au sein de l'activité et l'année de début de pratique de l'activité correspond à l'année de début la plus ancienne de l'ensemble des tâches réalisées au sein de l'activité. Nous allons ainsi réalisé cette opération pour chaque activité par agriculteur.

Disposant désormais de la durée de pratique de chaque activité ainsi que de la durée de l'activité professionnelle pour chaque agriculteur de la cohorte 1960, nous allons pouvoir calculer le ratio de pratique de chaque activité pour chaque activité.

1.2. Calcul de la matrice des ratios de pratique dans un secteur agricole

Le ratio de pratique de la tâche agricole (Ratio pratique tâche agricole_2) représente la proportion du temps consacré à une tâche agricole spécifique par rapport au temps total dédié à l'activité agricole associée à cette tâche.

$$\text{Ratio pratique tache agricole} = \frac{\text{durée de pratique de la tâche agricole}}{\text{durée de l'activité professionnelle}}$$

Disposant déjà de la durée de l'activité professionnelle pour chaque agriculteur, nous allons calculer la durée de pratique de chaque tâche agricole pour chaque agriculteur.

1.2.1 Calcul de la durée de pratique de la tâche agricole

Pour calculer la durée de pratique des tâches agricoles, nous allons effectuer la soustraction entre les dates de fin et les dates de début. Soit l'opération suivante :

Durée de pratique de la tâche agricole = année de fin de pratique de la tâche−année de début de pratique de la tâche

Etant donnée que nous disposons maintenant de la durée de pratique de chaque tâche agricole, nous pouvons à présent déterminer le ratio ci-dessus.

2. Analyse en composante principale sur la matrice associée aux ratios de pratique de l'activité

2.1. Éditons la matrice des corrélations

Table 1: Extrait de la matrice des corrélations (Partie 1)

	Bov	Mou	Coc	Che	Vol	Prai	Vigne	Mais	Ble
Bov	1.000	0.018	0.206	0.171	0.102	0.620	-0.288	0.302	0.381
Mou	0.018	1.000	0.135	0.109	0.147	0.112	-0.040	0.010	0.086
Coc	0.206	0.135	1.000	0.341	0.397	0.176	-0.051	0.056	0.217
Che	0.171	0.109	0.341	1.000	0.254	0.168	0.042	-0.075	0.137
Vol	0.102	0.147	0.397	0.254	1.000	0.121	-0.017	0.044	0.134
Prai	0.620	0.112	0.176	0.168	0.121	1.000	-0.159	0.379	0.494
Vigne	-0.288	-0.040	-0.051	0.042	-0.017	-0.159	1.000	-0.095	-0.123
Mais	0.302	0.010	0.056	-0.075	0.044	0.379	-0.095	1.000	0.544
Ble	0.381	0.086	0.217	0.137	0.134	0.494	-0.123	0.544	1.000

Table 2: Extrait de la matrice des corrélations (Partie 2)

	Pois	Bet	Tou	Col	Tabac	Arb	PdT	LegChamp	Serres
Pois	1.000	0.273	0.123	0.314	-0.015	-0.001	0.120	0.042	-0.018
Bet	0.273	1.000	-0.058	0.129	0.046	0.093	0.400	0.073	-0.060
Tou	0.123	-0.058	1.000	0.415	0.033	-0.026	-0.043	-0.020	-0.035
Col	0.314	0.129	0.415	1.000	0.027	0.016	0.065	-0.018	-0.033
Tabac	-0.015	0.046	0.033	0.027	1.000	0.061	0.090	-0.004	0.014
Arb	-0.001	0.093	-0.026	0.016	0.061	1.000	0.195	0.002	-0.004
PdT	0.120	0.400	-0.043	0.065	0.090	0.195	1.000	0.184	0.003
LegChamp	0.042	0.073	-0.020	-0.018	-0.004	0.002	0.184	1.000	0.330
Serres	-0.018	-0.060	-0.035	-0.033	0.014	-0.004	0.003	0.330	1.000

Extrait de la matrice des corrélations

La matrice de corrélation révèle majoritairement des liens faibles entre les activités agricoles. Toutefois, certaines corrélations modérées se dégagent, comme entre **Bov** et **Prai** (0.620), **Coc** et **Vol** (0.397), ou encore **Maïs** et **Blé** (0.544). Ces associations suggèrent des pratiques fréquemment liées ou complémentaires. À l'inverse, des cultures comme la **Vigne** ou les **Serres** restent peu corrélées aux autres.

2.2. Réalisons une ACP sur les données

Étant donné que nous disposions de 13 variables ou activités et de 12 310 unités statistiques ou agriculteurs, nous savons en avance que le nuage des individus se fera dans un espace à 13 dimensions tandis que le nuage des variables se fera dans un espace à 12 310 dimensions avec la sélection des 5 activités d'élevage et les 8 cultures les plus fréquentes.

2.3. Éditons un tableau de synthèse des valeurs propres, pourcentage d'inertie et pourcentage d'inertie cumulée

Table 3: Valeurs propres, pourcentage d'inertie et pourcentage d'inertie cumulée

	composante	inertie	inertie_perc	inertie_cumulee
comp 1	comp 1	3.06	23.53	23.53
comp 2	comp 2	1.88	14.43	37.96
comp 3	comp 3	1.15	8.84	46.80
comp 4	comp 4	1.04	7.99	54.79
comp 5	comp 5	0.97	7.44	62.23
comp 6	comp 6	0.96	7.42	69.65
comp 7	comp 7	0.81	6.23	75.88
comp 8	comp 8	0.72	5.55	81.43
comp 9	comp 9	0.63	4.86	86.29
comp 10	comp 10	0.57	4.37	90.65
comp 11	comp 11	0.52	4.01	94.67
comp 12	comp 12	0.37	2.83	97.50
comp 13	comp 13	0.33	2.50	100.00

Grâce à cette sortie, nous pourrons déterminer le nombre d'axes factoriels que nous allons retenir. En faisant usage du critère des 80% d'inertie cumulée, nous allons retenir 7 axes factoriels. Si nous sélectionnons les axes selon le principe des valeurs propres, c'est-à-dire les axes factoriels pour lesquels la variance est supérieure à 1, nous allons retenir 4 axes factoriels. Cependant, en utilisant ce critère, nous allons récupérer seulement environ 57% de l'information initiale, ce qui est peu. Nous allons donc privilégier le critère des 80% d'inertie cumulée pour sélectionner le nombre optimal d'axes factoriels.

L'inertie projété sur l'espace de dimension 7 est de 75.88%.

2.4. Éditons le tableau des corrélations entre les variables et les axes factoriels

Table 4: Corrélation entre les variables et les composantes principales

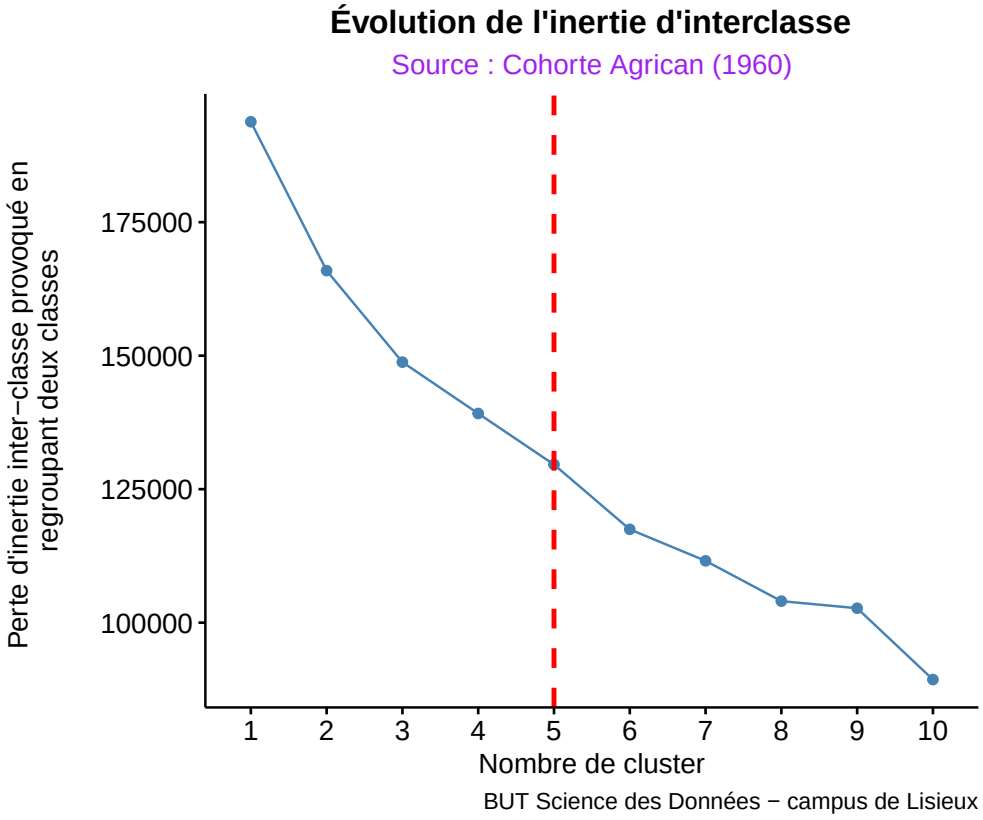
Activite	Dim.1	Dim.2	Dim.3	Dim.4
Bovins	0.64	-0.43	-0.14	-0.21
Moutons et Chèvres	0.21	0.14	-0.30	-0.37
Cochons	0.57	0.40	-0.19	-0.08
Chevaux	0.45	0.43	-0.22	-0.03
Volailles	0.44	0.42	-0.28	-0.07
Prairies	0.69	-0.41	-0.09	-0.02
Blé ou orge	0.72	-0.30	0.20	0.26
Maïs	0.45	-0.53	0.16	0.31
Vigne	-0.23	0.26	-0.22	0.76
Pommes de terre	0.51	0.49	0.30	0.09
Betteraves sucr.	0.58	0.26	0.31	0.13
Arboriculture	0.17	0.36	-0.07	0.18
Cultures légumières	-0.06	0.31	0.75	-0.24

L'analyse du tableau montre que plusieurs activités agricoles présentent une forte corrélation avec la première composante principale, comme les **Prairies** (0.69), le **Blé ou orge** (0.72) ou encore les **Bovins** (0.64). Ces variables influencent donc fortement la construction de l'axe 1. La composante 2 est surtout corrélée négativement aux **Maïs** (-0.53) et **Bovins** (-0{,}43), et positivement aux **Cochons** (0{.}40) et **Volailles** (0.42), suggérant une opposition entre types d'élevage. Les axes 3 et 4 sont moins fortement corrélés aux variables, à l'exception de **Cultures légumières** avec l'axe 3 (0.75) et **Vigne** avec l'axe 4 (0.76). Cela indique une spécificité marquée de ces deux activités sur ces dimensions. Ainsi, les premiers axes résument l'essentiel de la variance, en opposant grandes cultures, élevages, et cultures spécialisées.

3. Classification automatique via la méthode des k-Means

3.1. Identification du nombre optimal de clusters

Afin de déterminer le nombre optimal de classe dans lesquelles nous allons répartir les agriculteurs, nous allons commencer par éditer la courbe d'évolution de l'inertie d'interclasse. Celle-ci permet d'examiner comment l'inertie interclasse évolue lorsque nous fusionnons des clusters, passant de 10 clusters à un seul.



De manière générale, nous constatons que plus il y a de classes, plus l'inertie interclasse est faible ce qui signifie qu'il y a moins de différence au sein des clusters, et inversement.

À partir du graphique représentant l'évolution de la perte d'inertie inter-classe en fonction du nombre de clusters, on observe une forte décroissance initiale de l'inertie entre 1 et 4 clusters, puis une diminution plus progressive à partir de 5 clusters. Ce type de courbe est caractéristique de la méthode du coude, qui consiste à identifier le point à partir duquel l'ajout de nouveaux clusters n'apporte qu'un gain marginal en termes d'explication de la variance entre les groupes.

Dans ce graphique, le coude semble se situer autour de 4 ou 5 clusters, ce qui indique qu'au-delà de ce nombre, les gains d'information deviennent moins significatifs.

Par conséquent, on peut partir sur l'hypothése de choisir 5 clusters.

Effectifs des clusters

Source : **Cohorte Agrican 1960**

Classe	Effectif
1	2677
2	808
3	840
4	1110
5	795
6	4808
7	671
8	601

3.2. Création des clusters

Ayant préalablement déterminé le nombre optimal de cluster, nous pouvons maintenant réaliser le clustering et affecter chaque agriculteur à une classe de telle sorte que nous ne retrouvons dans chaque classe que des agriculteurs ayant une activité similaire. Ainsi, grâce à la fonction kmeans, nous avions pu obtenir la segmentation suivante :

Dans le suite de notre analyse, nous allons nous intéresser aux clusters dans lesquels nous retrouvons le plus d'agriculteurs. Nous allons donc éditer un tableau qui renseigne sur le ratio moyen de pratique de l'activité et la valeur de la statistique de test (v-test) pour les clusters 2, 4, 5 et 6 et aussi, sur le ratio moyen de pratique de l'activité obtenu sur la sous-cohorte 1960.

3.3 Caractéristiques des clusters

Table 5: Caractéristiques des clusters

	Moy c5	v-test5	Moy c6	v-test6	Moy c4	v-test4	Moy c2	v-test2	Moyenne
Arboriculture	15.40	8.63	7.47	-4.73	5.41	-12.04	4.65	-10.77	10.25
Bettraves	3.62	-16.61	1.02	-20.77	3.83	-24.24	2.62	-20.78	14.87
Blé ou orge	16.77	-32.79	9.41	-41.24	87.67	63.15	5.81	-51.06	47.66
Bovins	7.16	-59.75	9.52	-58.06	82.80	35.32	94.08	41.62	61.28
Chevaux	6.76	-7.45	3.26	-13.65	4.83	-16.00	8.02	-6.08	11.12
Cochons	6.39	-10.58	1.47	-18.84	7.26	-13.63	7.03	-10.97	12.92
Autres légumières	20.68	30.33	2.11	-8.99	2.23	-12.80	1.57	-11.49	6.31
Maïs	10.78	-25.55	5.43	-32.43	67.45	62.51	13.37	-25.81	32.20
Moutons/chèvres	6.06	0.63	1.76	-9.37	5.71	-0.25	3.84	-5.14	5.78
Pommes de terre	12.38	-4.84	3.90	-16.91	5.41	-21.62	3.83	-19.25	15.86
Prairies	13.31	-47.55	15.13	-46.32	85.45	44.06	61.68	4.83	57.75
Vignes	3.51	-27.95	96.99	76.84	20.87	-12.95	7.48	-27.06	28.73
Volailles	13.82	-0.33	3.53	-15.95	6.79	-16.11	6.98	-12.13	14.04

Le tableau 5 présente les caractéristiques des clusters d'agriculteurs identifiés. On observe une forte spécialisation des clusters : cluster2 dominé par les bovins (94.08), cluster3 par le blé/orge (87.67) et le maïs (67.45), cluster4 par les prairies (85.45), cluster5 par les vignes (96.99), et cluster6 par plusieurs cultures (arboriculture, betteraves, etc.). Les valeurs v-test associées confirment la significativité statistique de ces

spécialisations. Les moyennes générales révèlent une prédominance des bovins (61.28) et des prairies (57.75) dans l'ensemble de l'échantillon étudié.

Table 6: Caractéristiques démographiques par cluster

Variable	Type	Cluster_2	Cluster_4	Cluster_5	Cluster_6
TabagismeF	Moyenne	0.53	0.46	0.58	0.55
TabagismeF_sd	'Ecart-type	0.50	0.50	0.49	0.50
TabagismeF_na	% NA	2.60	1.70	2.20	1.90
NbPaquetAnneeF	Moyenne	15.91	15.05	18.29	18.10
NbPaquetAnneeF_sd	'Ecart-type	15.00	13.95	18.53	18.09
NbPaquetAnneeF_na	% NA	57.70	62.60	55.50	57.70
carriereAgriDeb	Moyenne	1958.97	1961.02	1960.45	1959.76
carriereAgriDeb_sd	'Ecart-type	6.03	5.93	6.04	6.15
carriereAgriDeb_na	% NA	0.00	0.00	0.00	0.00
duree_carriere	Moyenne	29.01	36.69	29.64	33.96
duree_carriere_sd	'Ecart-type	16.08	11.48	15.77	14.30
duree_carriere_na	% NA	0.00	0.00	0.00	0.00

Commentaire du Tableau 6 : Caractéristiques démographiques par cluster

Le tableau 6 intitulé **"Caractéristiques démographiques par cluster"** présente les moyennes, écarts-types et taux de valeurs manquantes (% NA) pour plusieurs variables liées au tabagisme, au nombre de paquets-années, au début de carrière agricole et à la durée de carrière, répartis selon les clusters identifiés.

On observe que **le cluster 4** regroupe les agriculteurs ayant en moyenne **la plus longue durée de carrière (36,69 ans)**, suivi du **cluster 6 (33,96 ans)**. À l'inverse, **le cluster 2** présente la durée moyenne de carrière la plus courte (**29,01 ans**).

Concernant le **tabagisme**, c'est **le cluster 5** qui affiche la **proportion moyenne la plus élevée de fumeurs (0,58 soit 58 %)**, suivi du **cluster 6 (55 %)**, du **cluster 2 (53 %)**, tandis que **le cluster 4 affiche le taux le plus bas (46 %)**.

En ce qui concerne la **consommation moyenne annuelle (NbPaquetAnneeF)**, elle est la plus élevée dans les **clusters 5 (18,29)** et **6 (18,10)**, légèrement plus faible dans le **cluster 2 (15,91)**, et la plus basse dans le **cluster 4 (15,05)**.

Enfin, les années moyennes de **début de carrière agricole** sont relativement proches entre les clusters, autour de **1960**, avec une légère avance pour le **cluster 2 (1958,97)** et un léger retard pour le **cluster 4 (1961,02)**.

Ces données suggèrent une **association entre l'ancienneté dans la profession, la durée de carrière et le comportement tabagique**, les groupes ayant une carrière plus longue ou plus ancienne présentant généralement un **taux de tabagisme plus élevé**.

Conclusion

En utilisant les informations fournies sur les agriculteurs de la cohorte 1960, nous avons réalisé deux analyses distinctes. Premièrement, nous avons effectué une Analyse en Composantes Principales (ACP) sur les données dans le but de créer des variables synthétiques qui résumeraient l'information des 18 variables initiales. Deuxièmement, nous avons utilisé la méthode des k-Means pour effectuer un clustering de notre cohorte en 5 classes distinctes.

L'analyse menée sur la cohorte Agrican 1960 nous a permis d'identifier des profils d'agriculteurs en fonction de la répartition de leurs pratiques professionnelles à travers différentes activités agricoles. Grâce à une Analyse en Composantes Principales (ACP), nous avons pu réduire la dimensionnalité du jeu de données tout en conservant environ 76 % de l'information initiale. Cette étape a facilité la mise en œuvre du clustering via la méthode des k-means.

L'observation de la courbe de l'inertie interclasse nous indique un **point d'inflexion notable autour de 4 ou 5 clusters**, à partir duquel les gains en homogénéité intra-classe deviennent faibles. Bien que l'analyse initiale ait été réalisée avec 8 clusters, une **segmentation en 5 groupes** aurait pu constituer un **compromis plus pertinent** entre complexité du modèle et lisibilité des résultats, tout en maintenant une bonne séparation des profils agricoles.

L'étude des clusters formés révèle une spécialisation marquée de certains groupes (ex. : vignes dans le cluster 5, bovins dans le cluster 2), ainsi que des différences significatives dans les trajectoires professionnelles et les comportements (notamment vis-à-vis du tabac). Ces résultats mettent en lumière la **diversité des parcours agricoles** au sein de la sous-cohorte étudiée, tout en illustrant la **richesse descriptive offerte par les méthodes multivariées**.

Ces analyses peuvent constituer une base solide pour des études épidémiologiques futures, en explorant par exemple la relation entre type d'activité agricole et santé, ou en intégrant de nouvelles variables contextuelles pour affiner les profils.