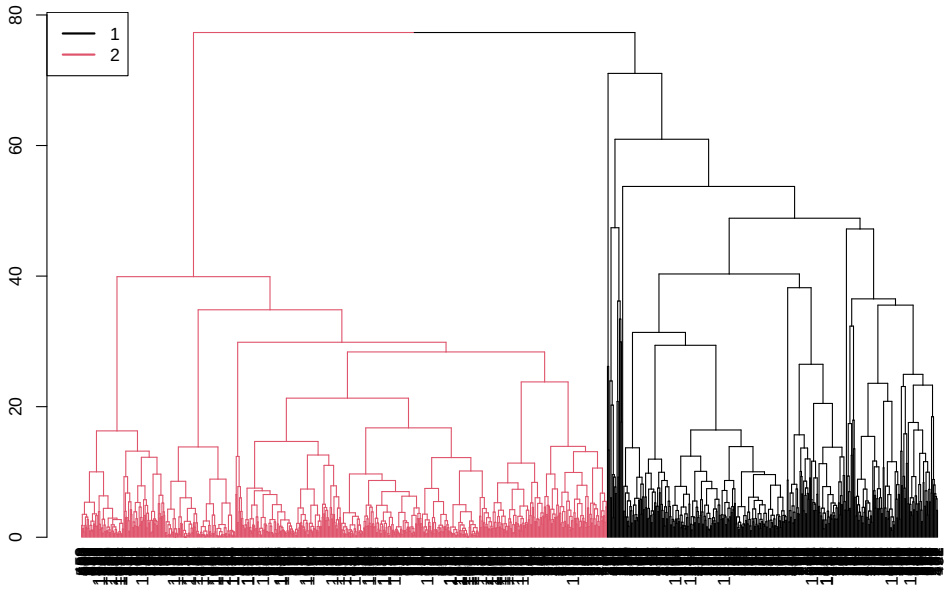
Examen Appretissage non supervisé

Ugo ZENNARO

Données et Méthodes

Pour comprendre la base de données nous allons dans un premier temps étudier toutes les variables quantitatives continues, qui représentent en majorité le comportement bancaire des clients, en visualisant celle-ci et en essayant d'y trouver des groupes. Cela nous permettra de caractériser différents profils de clients dont on étudiera les caractéristiques qualitatives ensuite.

CAH

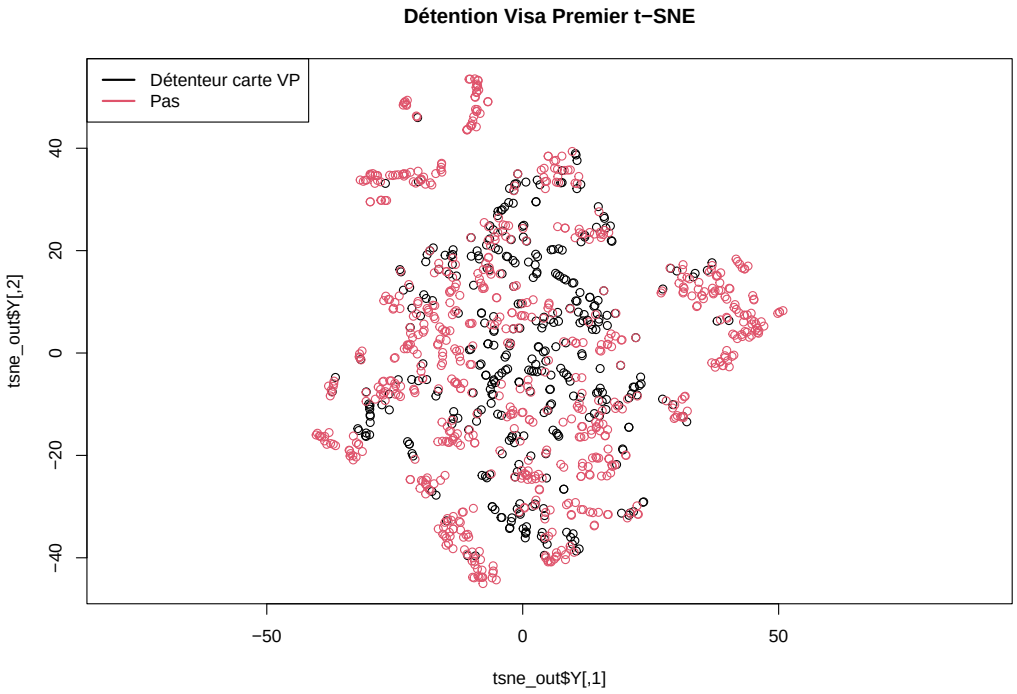


Pour commencer un dendrogramme permet d'avoir un point de vu global sur le jeu de données et permet d'envisager 2 groupes dont le groupe 1 est particulièrement hétérogène (celui de droite).

t-SNE

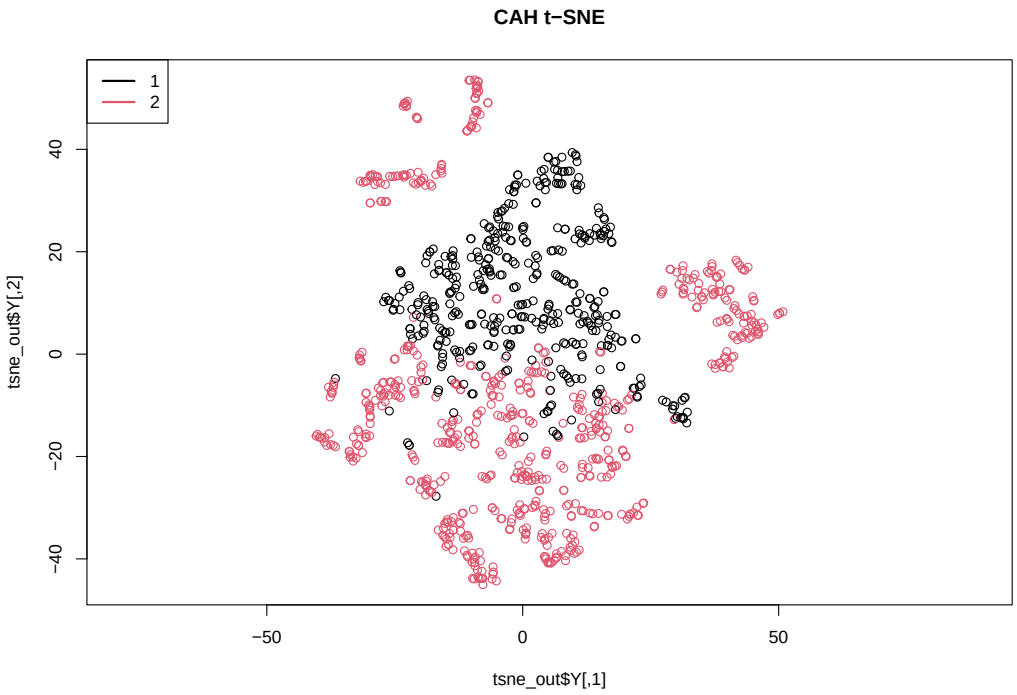
Maintenant nous allons étudier les représentations du jeu de données proposés la méthode t-SNE. Nous y visualiserons plusieurs distinctions (des différentes méthodes de clusterings et de la détention de la carte VisaPremier) qui nous permettront de comprendre la configuraion du jeu de données.

t-SNE VisaPremier



Ici on a du mal à avoir des intuitions sur notre jeu de données, mais on peut voir des groupes d'individus vers l'extérieur du nuage qui ne sont pas en possession de la carte VisaPremier.

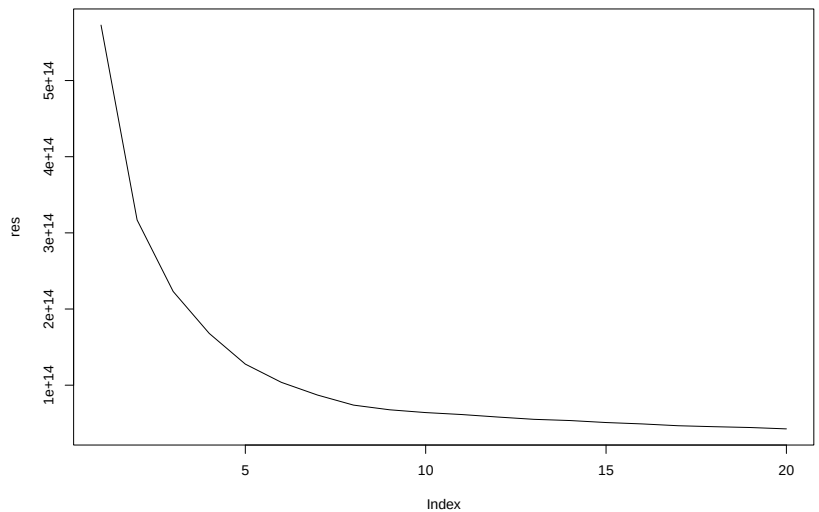
t-SNE CAH



On voit que la CAH retrouve un découpage qui distingue beaucoup plus le centre et la périphérie que la séparation proposé par la détention de la carte mais rejoint l'idée d'une structure centrale et une autre périphérique.

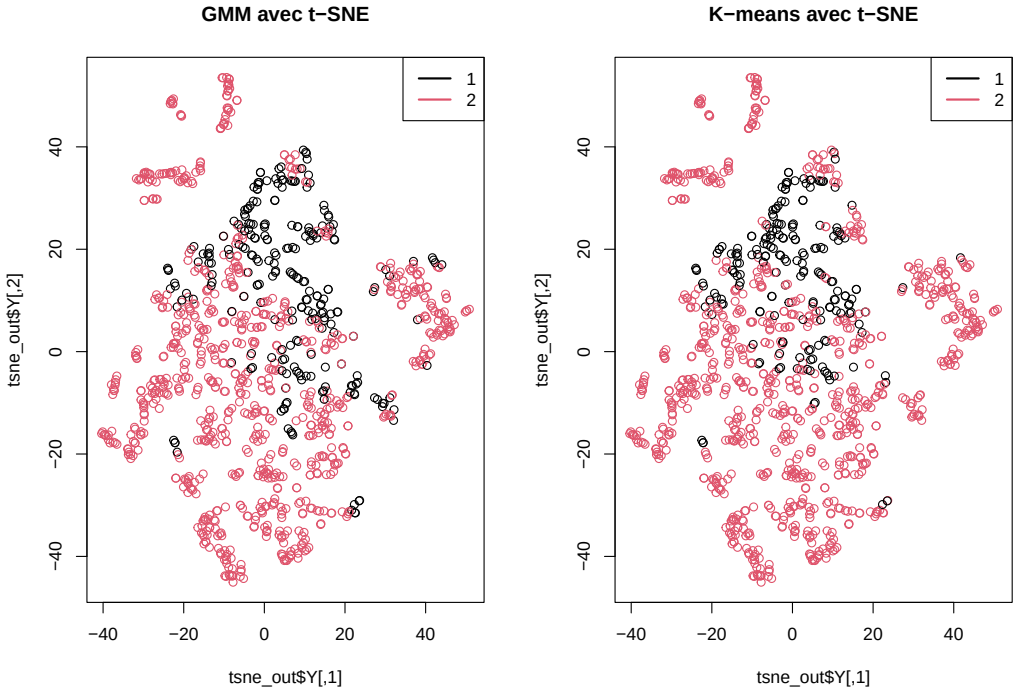
Nombre de groupes

Sachant que, plus on fait de groupes, plus la variance de la population de chacun de ces groupes sera faible, on cherche le plus petit nombre de groupe à partir duquel la variance ne décroit pas brusquement.



Là on voit qu'il n'y a pas de nombre de groupes qui se distingue avec la méthode de k-means car il n'y a pas de cassure à partir de laquelle la variance dans chaque groupe diminue moins. On peut tout de même tester avec 2 groupes sur les méthodes k-means et GMM dans la continuité de ce qu'on a vu jusqu'à maintenant.

t-sne K-means / CAH



Là encore on retrouve notre structure qui différencie le centre et la périphérie du nuage de points. Avant de comparer sur plus d'axes nos méthodes avec l'ACP on peut comparer les résultats de groupes qu'elles proposent pour déterminer si certaines sont redondantes.

Indices de Rand

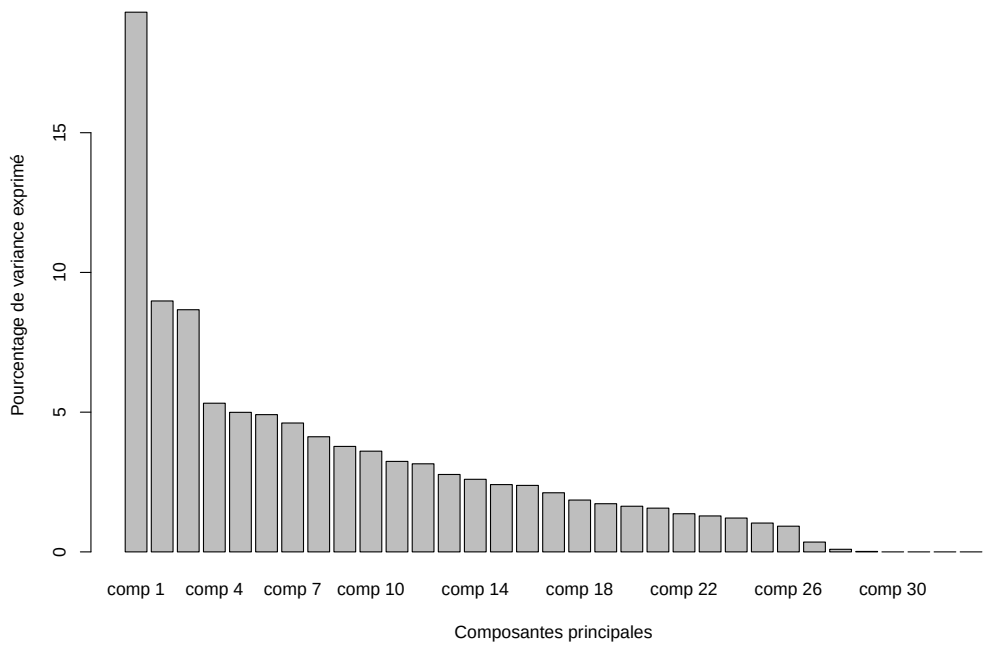
	CAH	KM	GMM	VP
CAH	1.00000000	0.2769231	0.3277941	0.09470765
KM	0.27692311	1.0000000	0.5718001	0.17625768
GMM	0.32779408	0.5718001	1.0000000	0.21958434
VP	0.09470765	0.1762577	0.2195843	1.00000000

Les valeurs affichées sont des indices entre 0 et 1 déterminants à quel point les groupes crées sur chaque méthode sont similaires. On voit que les classes proposées par les K-means et les mélanges de gaussiennes sont les plus similaires et que la distinction faite par la détention de la carte VistaPremier se rapproche le plus du regroupement proposé par les mélanges de gaussiennes.

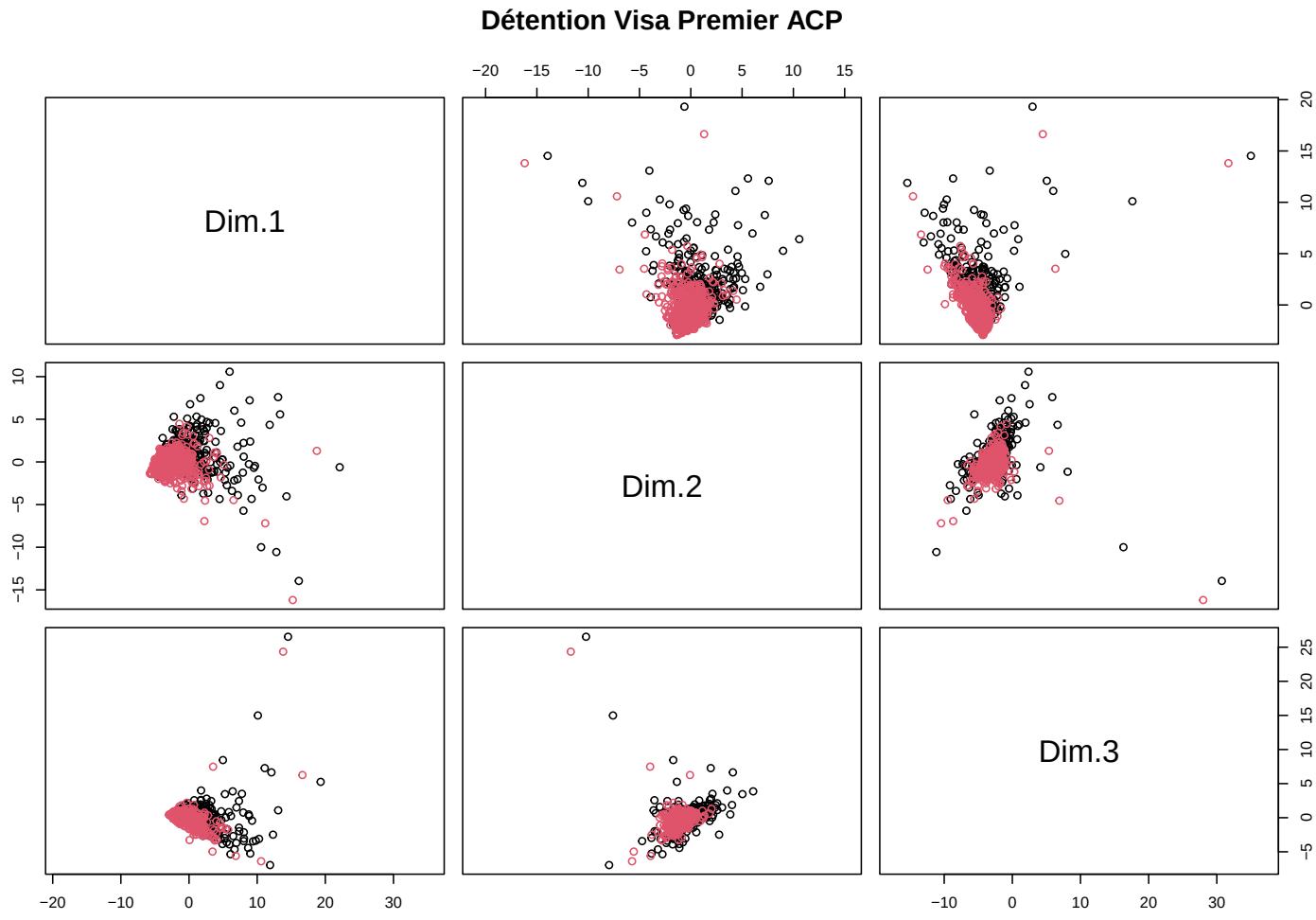
ACP

Maintenant nous allons étudier l'ACP pour essayer de trouver des représentations, et les combinaisons de variables associés, qui proposent des distinctions pertinentes.

ACP - Nombre d'axes



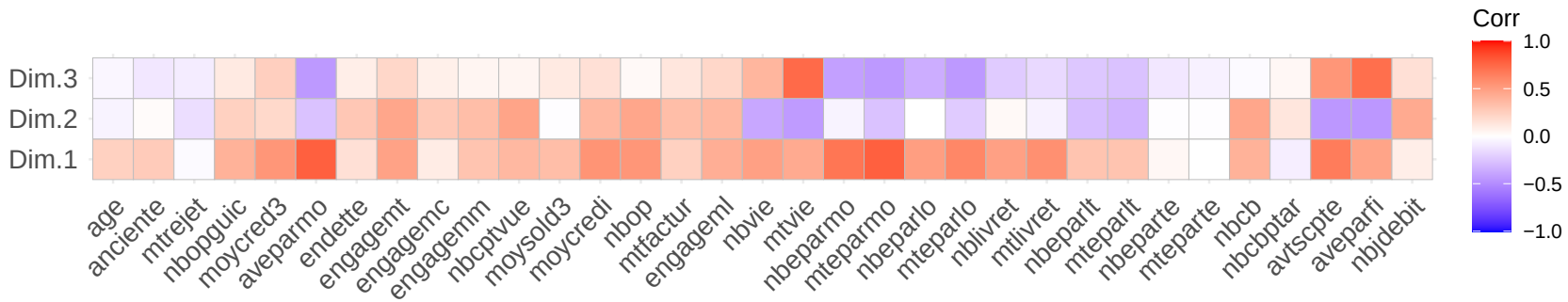
On voit une variance qui se réparti sur beaucoup d'axe. On regardera seulement les 3 premiers après lesquels l'évolution de la variance exprimé par chaque axe semble trop linéairement décroissante.



Ici la seule distinction qui semble pertinente à observer est que les détenteurs de la carte VP semblent être plus avancés sur l'axe 1.

Interprétation des axes

Avec les corrélations suivantes on comprend déjà que l'axe 1 est croissant positivement de presque toutes les variables. On peut donc déduire que les individus détenteurs de la carte ont, de manière générale, une activité bancaire plus importante.



Analyse

Proportions d'individus par classe:

	CAH	KM	GMM	VP
1	0.386194	0.1791045	0.2238806	0.3339552
2	0.613806	0.8208955	0.7761194	0.6660448

On voit ici que la classe 1 (la classe centrale selon le nuage de point proposé par t-SNE) pour chaque classification à une proportion d'individus moindre et que les détenteurs de la carte VisaPremier constituent un tiers de la clientèle. On regarde maintenant ce qui constitue les profils crées.

Individus moyens

Table 1: K-means

age	moycred3	endette	moysold3	mtfactur	nbcb
48	91	9	24080	55992	2
41	16	5	7611	16178	1

Table 2: Mixture de Gaussiennes

age	moycred3	endette	moysold3	mtfactur	nbcb
47	85	11	21976	64463	2
41	13	4	7241	11341	1

Table 3: Classification ascendante hiérarchique

age	moycred3	endette	moysold3	mtfactur	nbcb
46	56	11	13124	39622	1
40	13	2	8948	13045	1

On semble avoir les groupes centraux qui correspondent à une clientèle plus riche, légèrement plus âgée. Celle ci a en moyenne plus de carte de crédits et semble avoir un taux d'endettement plus grand.

Analyse sociale

On regarde via les groupes crées par la CAH les distinctions sociales qui transparaissent de la distinction centre/périphérie proposée par t-SNE.

	Pagri	Part	Pcad	Pemp	Pinc	Pouv	Pret	Psan
1	0	21	196	110	2	34	10	41
2	1	10	249	178	0	51	11	158

	F.	Fcel	Fdiv	Fmar	Fsep	Fuli	Fveu
1	14	107	36	243	5	5	4
2	18	253	50	303	9	20	5

	.	A	B	C	D	E
1	16	181	138	31	46	2
2	117	25	165	187	122	42

État clientelle VisaPremier

Table 4: Visa Premier

age	moycred3	endette	moysold3	mtfactur	nbcb
45	62	7	17352	53036	2
42	13	5	7155	8404	1

	Pagri	Part	Pcad	Pemp	Pinc	Pouv	Pret	Psan
1	0	29	229	63	0	7	9	21
2	1	2	216	225	2	78	12	178

	F.	Fcel	Fdiv	Fmar	Fsep	Fuli	Fveu
1	13	89	30	211	4	8	3
2	19	271	56	335	10	17	6

	.	A	B	C	D	E
1	23	104	107	51	45	28
2	110	102	196	167	123	16

Si on voulait que la clientèle de la carte VisaPremier tende vers le groupe 1 (le plus aisé) proposé par la classification ascendante hiérarchique, on pourrait dire à première vu qu'il faudrait essayer d'attirer un peu plus d'employé et un peu moins de cadres.

Conclusion

Toutes les méthodes de clusterings semblent suggérer une distinction plus ou moins similaire dans le jeu de données. D'ailleurs il semblerait même que ces distinctions proposent un profil de la clientèle cible de la carte VisaPremier qui soit plus pertinent que la vraie clientèle de la carte. Il semblerait intéressant de pousser ces études et chercher des effets de causalité qui justifient l'adhésion à la carte ou non pour comprendre ce qui peut entraîner une rétissance à la carte chez le publique cible, et comment et pourquoi des individus en dehors de cette cible se retrouvent à souscrire à ce service. Aussi il y aurait possiblement des clusters plus évidents en étudiant de manière distinctes les différentes CSP et situations familiales.