UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

Faculté SESG

Master statistique pour l'évaluation et la prévision

2022-2023

Cours d'Analyse de données

Analyse en composantes principales des données d'allocations doctorales.

Ismael Djoulde DIALLO

Introduction

Dans le contexte du financement de la recherche sur projets, l'agence nationale de recherche (ANR) se base sur un processus de sélection compétitif et rigoureux. En effet, pour un financement impartial des projets de recherche qui lui sont soumis, lors de l'examen de tout projet de recherche une notation allant de 1 à 5 est donnée à chacun des 5 critères d'évaluation du projet, notamment la clarté des objectifs du projet, son caractère novateur, la pertinence méthodologique, la compétence expertise et de la qualité EA (Equipe d'Accueil ?). L'avis final d'allocation ou non de financements à un projet de recherche dépendra de la note globale (ou moyenne globale) délibérée par le comité d'examen dudit projet.

Nous disposons de 62 projets de recherche examinés par les comités d'évaluation de l'ANR avec les notations de chacun des 5 critères susmentionnés.

Dans ce qui suit, nous nous proposons d'analyser la nature de la relation linéaire entre les 5 critères d'évaluation des projets de recherche. Au cas où il existerait une dépendance importante entre certains de ces critères i.e. que certains critères fournissent la même information que d'autres, nous résumerons les critères d'évaluation par de nouvelles composantes non corrélées de telles sortes qu'on perde le moins d'information possible sur les données. Les composantes retenues seront une combinaison linéaire des critères initiaux.

Pour ce faire, dans un premier temps nous effectuerons une analyse descriptive de notre jeu donnés afin de s'assurer que les notations des critères sont connues pour chaque projet sinon nous éliminerons les projets concernés. Ensuite, nous étudierons la corrélation linéaire des 5 critères du jeu de données retenu pour pouvoir procéder à la recherche des nouvelles composantes non redondantes.

1. Analyse descriptive du jeu de données

1. 1. Données manquantes

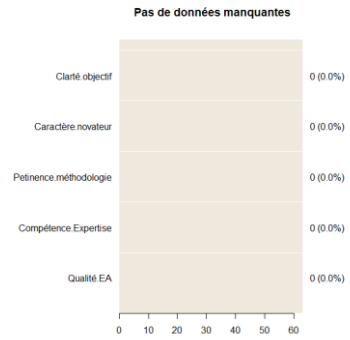


Figure 1 : Les données manquantes.

Lecture : Tous les critères d'évaluation pour chacun des 62 projets sont connus.

1. 2. Statistiques descriptives

Critères	Moyenne	Ecart-Type	Coefficient de variation	Minimum	Max
Clarté.objectif	4.016	0.859	0.213894422	2	5
Caractère.novateur	3.911	0.837	0.214011762	2	5
Petinence.méthodologie	3.677	0.971	0.264073973	1	5
Compétence.Expertise	4.379	0.803	0.1833752	2	5
Qualité.EA	4.444	0.573	0.128937894	3	5

Tableau 1 : Statistiques descriptives du jeu de données.

Lecture : les critères qualité EA (équipe d'accueil) et compétence expertise ont les plus grandes moyennes, avec respectivement 4,44 et 4,37 de moyenne alors que le critère pertinence méthodologique est celui qui a la plus faible moyenne soit 3,67.

Par ailleurs, en moyenne, la note du critère « pertinence méthodologie » d'un projet s'écarte 0.97 point de la moyenne dudit critère pour tous les projets. Le critère « pertinence méthodologie » présente une plus grande variabilité que tous les autres critères.

Pour une meilleure comparaison de variables différentes (présentant des moyennes différentes et ou exprimées dans différentes échelles), on préfère généralement le coefficient de variation à l'écart-type. Pour notre jeu de données, le coefficient de variation du critère « pertinence méthodologie » vaut plus du double de celui de la qualité EA, la variabilité du premier est plus élevée par rapport à sa moyenne. En revanche, le critère « clarté objectif » a une variabilité légèrement inférieure à celle du critère « caractère novateur » malgré qu'il ait un écart-type un légèrement plus élevé que celui du « caractère novateur », on peut en déduire une certaine homogénéité entre ces deux critères.

2. Corrélation entre les critères d'évaluation des projets de recherche

Ici, on veut s'assurer qu'il existe une corrélation minimale entre les critères d'évaluation.

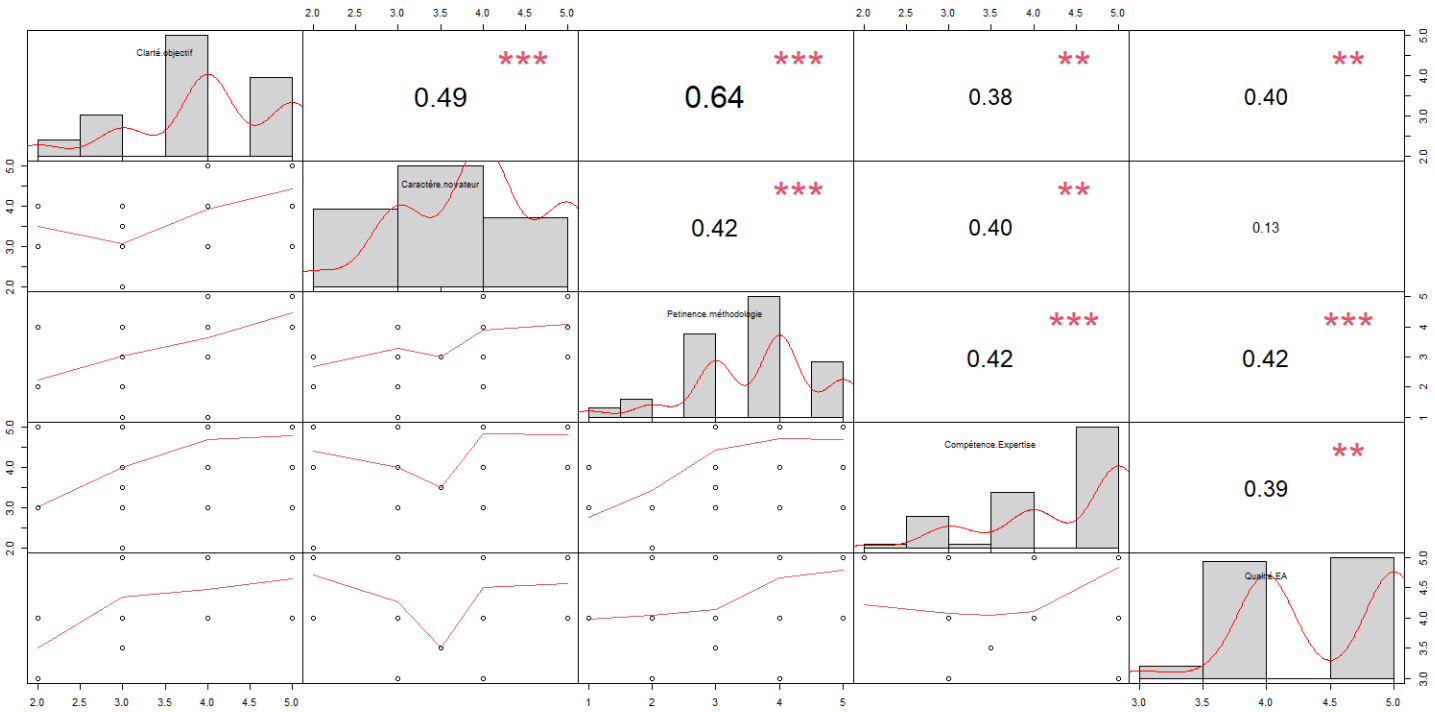


Figure 2 : Nature des corrélations entre les critères d'évaluation d'un projet de recherche.

Sur le graphe ci-dessus, on peut observer les corrélations existantes entre les différents critères. Pour chaque paire de critères, on peut lire l'intensité de la relation dans la partie supérieure de la diagonale du graphique et observer dans sa partie inférieure le nuage de points de la paire. La corrélation entre la paire de critères « Clarté de l'objectif » et « Pertinence méthodologique » vaut 0.64. Il s'agit d'une corrélation positive, ce qui signifie que ces deux critères évoluent dans le même sens, les deux critères augmentent ou diminuent simultanément. Toutefois, il est nécessaire de préciser, puisqu'il y a souvent confusion entre causalité et corrélation, que ce n'est pas parce qu'un projet de recherche a eu une notation élevée pour la clarté de son objectif que le comité d'examen lui a attribué une note élevée pour sa pertinence méthodologique. Cette corrélation pourrait être due à une simple coïncidence ou être expliquée par l'influence qu'aurait les autres critères sur la paire « Clarté de l'objectif » et « Pertinence méthodologique ».

Nous retenons qu'il existe une redondance d'informations entre les critères d'évaluation des projets, quoi qu'elle soit faible pour certains critères. De ce fait, dans ce qui suit nous tenterons de résumer les critères par de nouvelles composantes non redondantes.

3. Analyse en composantes principales (ACP)

Nous rappelons que l'ACP cherche à réduire des données multidimensionnelles, expliquées donc par plusieurs variables quantitatives, à quelques dimensions dans lesquelles la variation des données est maximale i.e. en conservant le plus d'information possible des variables initiales. Dans notre cas, les 62 projets sont présentés en 5 dimensions (critères) et nous recherchons des composantes principales de nombre idéalement inférieur à 5 qui condensent la plus grande partie de l'information contenue dans les 5 critères initiaux et qui se prêtent à l'interprétation.

Il est plus souvent effectué une transformation pour standardiser les données lorsque les variables actives de l'ACP ne sont pas exprimées dans les mêmes unités afin d'éviter tout biais qui pourrait en résulter. Dans notre cas, même si les critères sont exprimés dans la même unité, nous allons tout de même les normaliser pour les rendre comparables (plus haut nous avons constaté que les moyennes et les écart-types des critères n'étaient pas les mêmes).

3. 1. Qualités et contribution des variables aux composantes principales

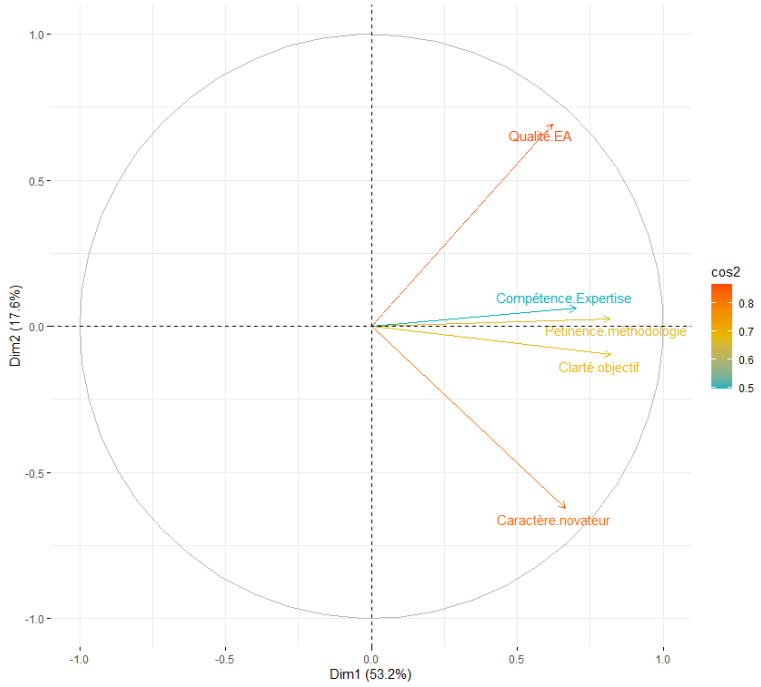


Figure 3 : Corrélation et qualité de représentation des variables dans le plan formé par les 2 1ères composantes.

Lecture : Sur le graphique, il est représenté les 5 critères dans le plan formé par les composantes qui portent le plus d'information, notamment Dim1 et Dim2. Tous les critères étant positivement corrélés entre eux et la Dim1 (on parle d'effet taille), ils sont donc regroupés dans les quadrants 1 et 4 du cercle. La distance de chaque critère de l'origine du cercle représente son influence dans la construction des dites composantes, ainsi plus grande est cette distance (couleur rouge ici) mieux est la qualité de la représentation du critère par les composantes, c'est notamment le cas du « caractère novateur » et de la « qualité EA » qui sont proches du pourtour du cercle (distance maximale) contrairement au critère « compétence expertise » qui lui est moins éloigné de l'origine et donc relativement moins bien représenté par les deux composantes.

3. 2. Correction de l'effet taille

Pour atténuer l'effet taille évoqué au [3. 1] nous représentons les variables dans l'espace formé par Dim2, l'axe qui permet de distinguer les critères de taille semblables, et Dim3.

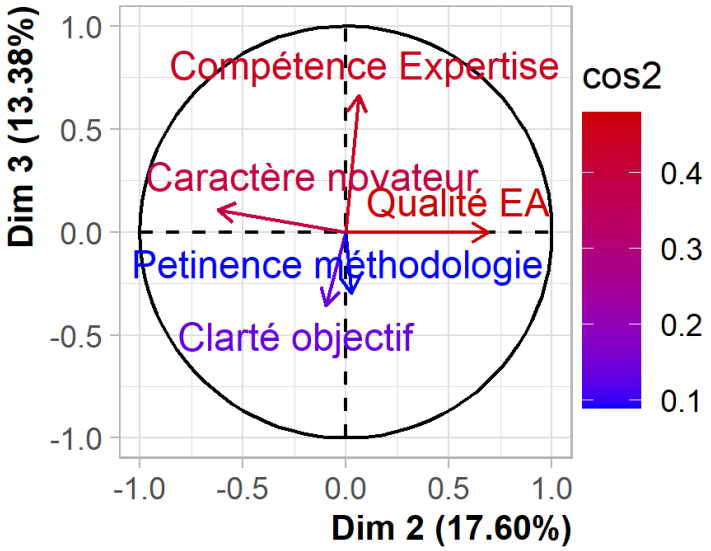


Figure 4 : Qualité de la représentation des variables dans le plan formé par les 2ème et 3 -ème composantes.

Les critères 'Compétence Expertise', 'Caractère novateur ' et 'Qualité EA' sont les mieux représentés dans ce plan alors que le critère 'Pertinence méthodologique' est le moins représenté.

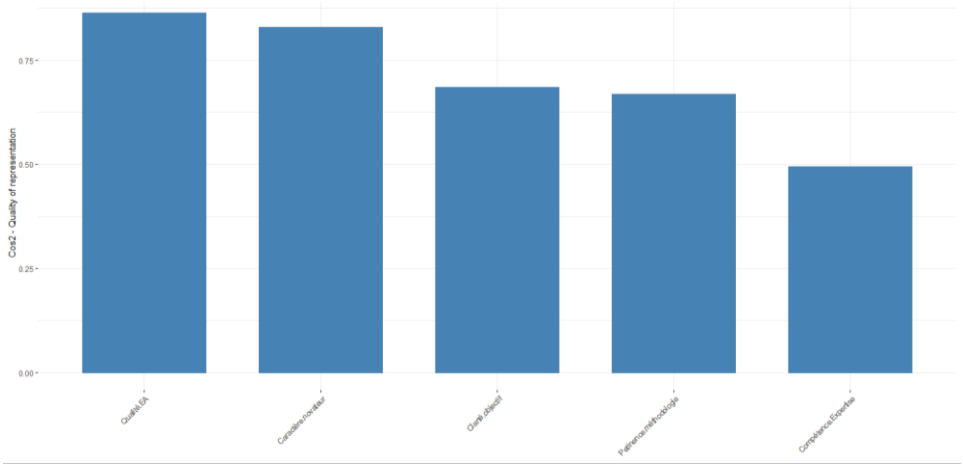


Figure 4 : Qualité de la représentation des variables par les 2 premières composantes.

Lecture : La qualité de la représentation des variables dans une ACP est mesurée par le cosinus carré (cos2) allant de 0 à 1. Hormis le critère « compétence Expertise » qui a un cos2 inférieur à 0.5, tous les autres critères semblent être bien représentés par les deux premières composantes. C'est aussi la meilleure représentation à deux dimensions qu'on peut avoir.

3. 3. Valeurs propres et distribution de l'inertie

Les valeurs propres mesurent la quantité de variance expliquée par chacune des composantes principales, les premières composantes ont les valeurs les plus élevées. L'examen des valeurs propres permet de choisir les composantes principales à retenir (Kaiser 1961). La méthode graphique de sélection de composantes

principales proposée par Cattell arrête l'extraction des composantes au nombre de composantes pour lequel se manifeste le changement de pente.

L'inertie mesure l'information contenue dans les données et la façon dont cette information est repartie entre les différentes variables qu'elles contiennent. Comme nous l'évoquions en l'entame de la partie [4] nous recherchons des dimensions dans lesquelles la variation (dispersion) des données est la plus élevée possible. L'information contenue dans les données sera donc captée par des composantes dont il faudra choisir le nombre.

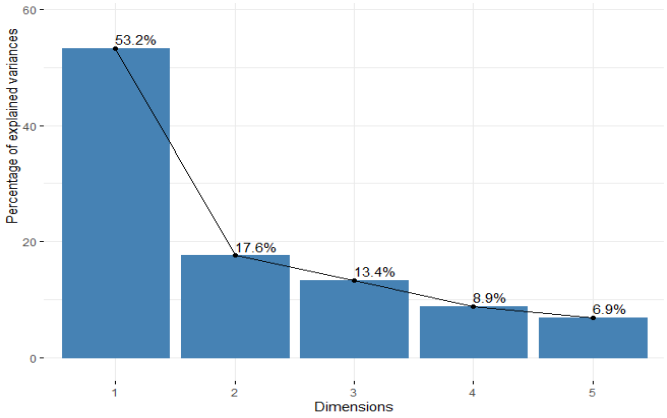
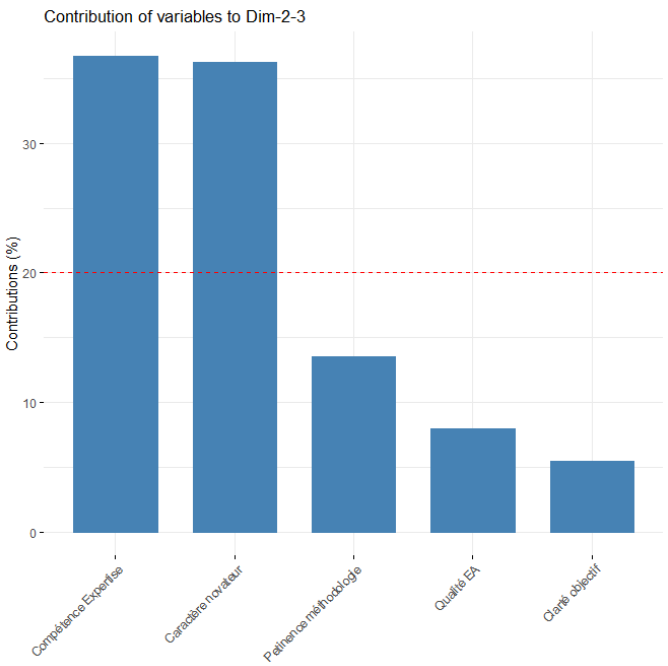


Figure 5 : Graphique des valeurs propres des composantes principales.

L'ACP que nous avons réalisée nous montre que les 2 premières composantes de l' analyse expriment **70.8%** (dont 53.2 % provenant de la Dim.1 et 17.8% de la Dim.2) de l'inertie totale du jeu de données ; cela signifie que 70.80% de la variabilité totale du nuage des individus (ou des critères) peut être représentée dans l'espace de dimensions DIM.1-Dim.2. Cette valeur est supérieure à la valeur référence de **54.66%**, la variabilité expliquée par ce plan est donc significative.

Dans la partie suivante, nous nous restreindrons aux deux premières composantes.

3. 4. Contribution des variables aux composantes DM.2 et DM.3



	Dim.2	Dim.3
Clarté objectif	1.081	19.009
Caractère novateur	43.969	1.757
Petinence méthodologie	0.076	13.319
Compétence Expertise	0.429	65.915
Qualité EA	54.445	0.000

La ligne en pointillé rouge sur le graphe correspond à la contribution moyenne attendue de chaque variable à la formation de l'espace de dimensions DM.2 et DM.3, les contributions des critères « compétence Expertise » et « caractère novateur » sont les plus importantes dans la constitution de cet espace. La contribution du critère « qualité EA » à DM.2 est de 54.45% alors que son apport à l'axe DM.3 est nul.

3. 5. Classification des projets

Lorsqu'on procède à une ACP sur individus, l'effet taille n'est pas observable. Nous réalisons donc une classification sur le plan formé par les 2 premières composantes.

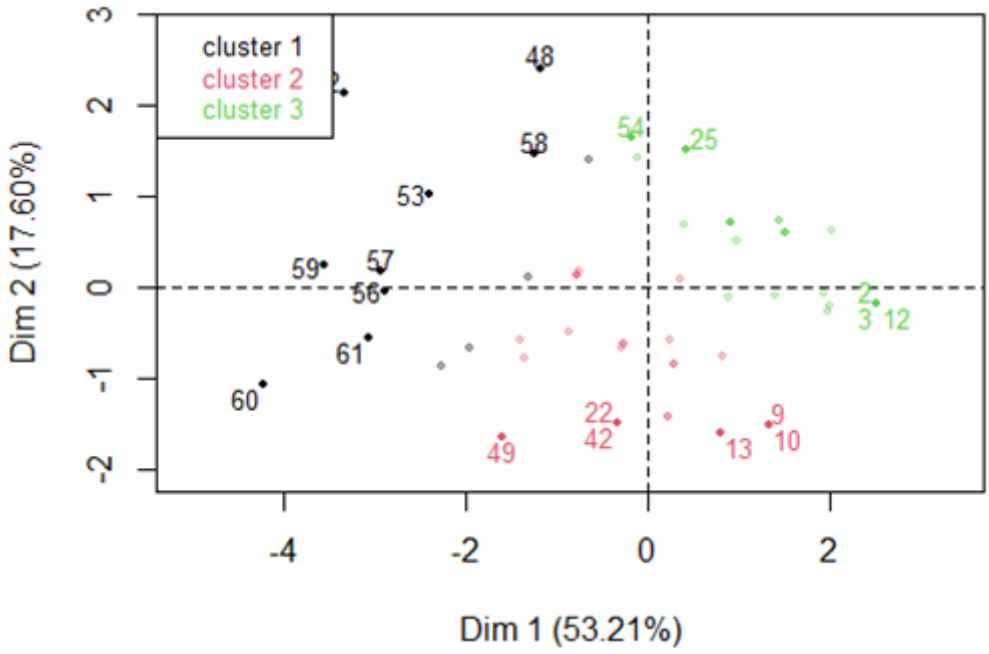


Figure 4 - Classification ascendante hiérarchique des projets dans l'espace formé par Dim.1 et Dim.2.

Lecture : le graphique montre les similitudes entre projets en termes de notations dans le nouvel espace, plus ils sont proches plus la structure de leur notation est identique. Ainsi, les projets similaires forment un cluster (classe) homogène. On a pu constituer trois classes disjointes coloriées différemment sur le graphique.

A l'image de la représentation des variables expliquées en [4], les projets les mieux représentés dans l'espace formé par Dim.1 et Dim.2 sont ceux les plus éloignés du centre, c'est notamment le cas des projets 13 ;10 ;12 ;48 ;59 ;57 ;60 ;61 ; 49.

La **classe 1** est composée d'individus tels que *48*, *53*, *56*, *57*, *58*, *59*, *60*, *61* et *62*. Ce groupe est caractérisé par de faibles valeurs pour les critères *Petinence.méthodologie*, *Compétence.Expertise*, *Clarté.objectif*, *Caractère.novateur* et *Qualité.EA* (de la plus extrême à la moins extrême).

La **classe 2** est composée d'individus tels que *9*, *10*, *13*, *22*, *42* et *49*. Ce groupe est caractérisé par de faibles valeurs pour le critère *Qualité.EA*.

La **classe 3** est composée d'individus tels que *2*, *3*, *12*, *25* et *54*. Ce groupe est caractérisé par de fortes valeurs pour les critères *Qualité.EA*, *Compétence.Expertise*, *Petinence.méthodologie*, *Clarté.objectif* et *Caractère.novateur* (de la plus extrême à la moins extrême).

Conclusion

L'analyse en composantes principales que nous avons effectuée ici nous a permis de visualiser les corrélations entre les 5 critères de sélection de projets de recherche par l'ANR pour l'allocation de financements. Ensuite, en raison de l'existence avérée (quoique globalement faible) d'associations linéaires entre les critères nous les avons tous retenus afin de procéder à la recherche de composantes principales non corrélées qui sont des combinaisons linéaires des critères initiaux. En raison des avantages d'interprétabilité qu'offrent les espaces en deux dimensions nous avons fait le choix de retenir les 2 composantes principales qui, à elles seules, résument 70,8% de la variabilité des données.

Il est important de préciser qu'il n'existe pas de méthodes objectives de sélection du nombre de composantes principales à retenir. Tout dépend du contexte de l'étude, l'interprétabilité recherchée par l'analyste, de la taille des données…

Les composantes principales sont utilisées à des fins de modélisations. De ce fait, au vu de la petite taille du jeu de données et de l'importance de tous les critères dans notre cas, il est intéressant de se demander si la réduction des 5 critères initiaux à deux nouvelles variables est pertinente et vaut réellement une perte d'information d'environ 30%.

Références

Agence nationale de la recherche : Processus de sélection.

https://anr.fr/fr/lanr/nous-connaitre/processus-de-selection/ [Consulté le 14 octobre 2022].

Sthda : Méthodes des Composantes Principales dans R, Guide Pratique.

http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/73-acp-analyse-en-composantes-principales-avec-r-l-essentiel/ [Consulté le 14 octobre 2022].