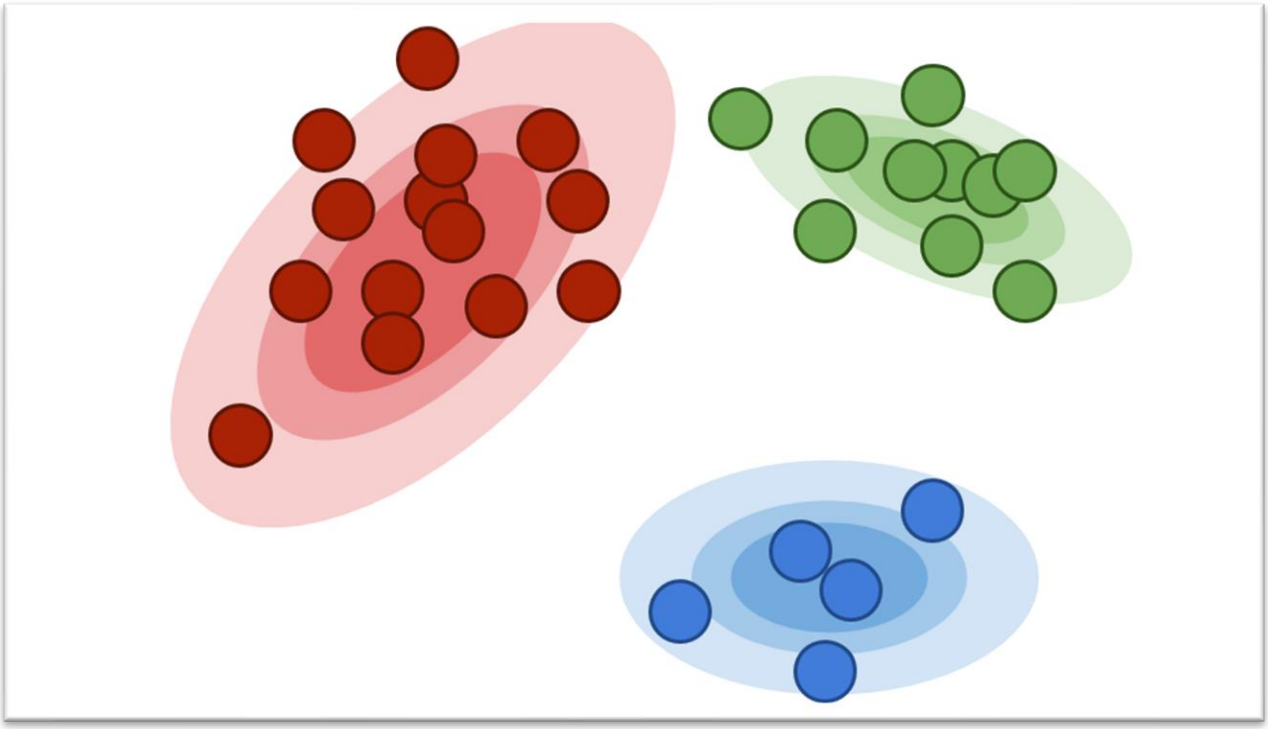
Analyse de données : TP2

L’analyse en composantes principales



AL NATOUR Mazen et HERVOUET Léo

AD 2023

Table des matières

0 – Introduction

L'Analyse en Composantes Principales (ACP) est une approche essentielle dans les méthodes d'analyse de données multidimensionnelles. Ce rapport vise à détailler notre compréhension de l'ACP en deux parties : la première sur un exemple pratique basé sur un tableau de notes d'étudiants, et la seconde est une application concrète de l'ACP à travers des données réelles issues de l'imagerie multispectrale, plus précisément le jeu de données Indian Pines.

Dans la première partie, nous nous sommes concentrés sur une analyse d'un tableau de notes attribuées à neuf étudiants dans différentes matières. Avec la bibliothèque scikit-learn, nous avons effectué l'ACP sur ces données, puis nous avons réalisé les représentations graphiques des résultats, les projections dans de nouveaux espaces vectoriels, et analysé les corrélations entre les variables initiales et les composantes principales.

La seconde partie de notre travail s'est axée sur le traitement de données réelles provenant de l'imagerie multispectrale du dataset Indian Pines. Avec l'ACP, notre objectif va être d'extraire les informations significatives tout en étant capables de visualiser les différentes classes de terrain.

À travers ce rapport, nous explorons les concepts liés à l'ACP, tout en mettant nos connaissances en application concrète sur des données réelles.

1 - ACP : Exemple simple

1.1 - Pré-étude des données

Nous commençons par charger les données du .csv, nous devons les transposer car nos tableaux sont par matières et nous voulons qu'ils soient par élèves.

Voici nous données :

	Jean	Aline	Annie	Monique	Didier	Andreas	Pierre	Brigitte	Evelyne
Math	6.00	8.00	6.00	14.50	14.00	11.00	5.50	13.00	9.00
Sci	6.00	8.00	7.00	14.50	14.00	10.00	7.00	12.50	9.50
Fr	5.00	8.00	11.00	15.50	12.00	5.50	14.00	8.50	12.50
Latin	5.50	8.00	9.50	15.00	12.50	7.00	11.50	9.50	12.00
Dessin	8.00	9.00	11.00	8.00	10.00	13.00	10.00	12.00	18.00

Voici l'histogramme pour la matière 'Français' :

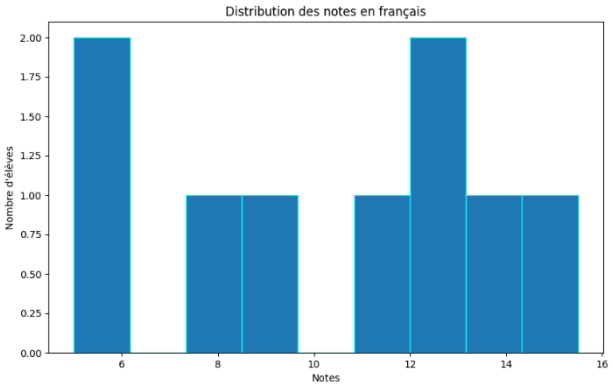


Figure 1 : Histogramme des notes de français

Voici l'histogramme pour la matière 'Latin :

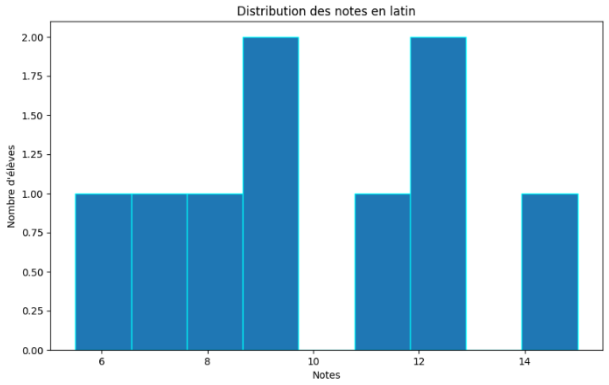


Figure 2 : Histogramme des notes de Latin

Nous pouvons voir que les notes de français se situent bien entre 5 et 15.5 et que les notes de Latin se situent également entre 5.5 et 15. Nous pouvons aussi constater que nos données sont plutôt homogènes.

Voici le nuage de points caractérisés par les deux variables ‘mathématique’ et ‘sciences’ :

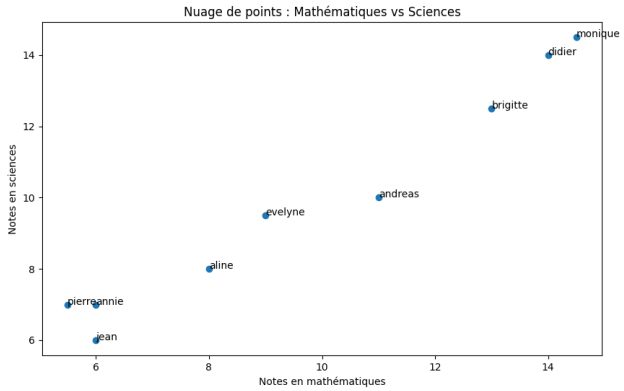


Figure 3 : Graphique, 'Math' vs 'Sci'

Avec ce graphique nous pouvons voir que plus un élève a une bonne note en Mathématiques alors plus il aura une bonne note en Sciences. Et l’inverse est aussi vrai des Sciences vers les Mathématiques. Ici les résultats sont homogènes.

Voici le nuage de points caractérisés par les deux variables ‘mathématique’ et ‘dessin :

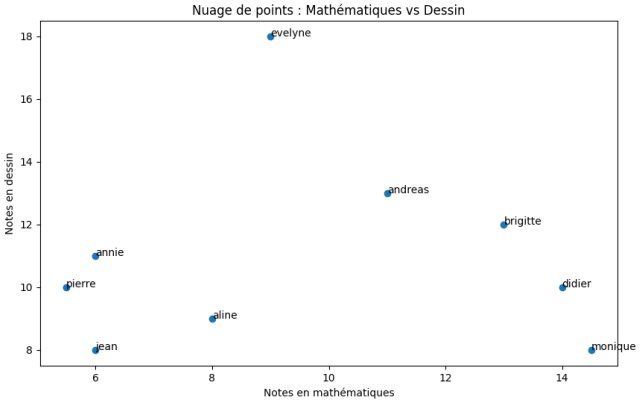


Figure 4 : Graphique, 'Math' vs 'Dessin'

Cette fois si nous pouvons voir que même si un élève a une bonne note en Mathématiques, cela ne veut pas dire qu’il aura une bonne note en dessin. Les résultats sont plutôt hétérogènes.

1.2 - Calcul de l'ACP

L'analyse en Composantes Principales (ACP) est un outil mathématique qui aide à simplifier des informations denses. Cette méthode permet de récupérer les informations les plus importantes pour pouvoir les comprendre plus facilement sans être submergé par trop d'informations.

Par exemple, sur un graphe en 2D avec une profondeur de 100, nous pouvons appliquer l'ACP pour nous retrouver avec moins d'informations. Par exemple faire l'ACP pourrait nous donner seulement 10 composantes pour nous représenter le même graphe ce qui nous permettrait de prendre moins de place au niveau stockage et aussi de prendre moins de temps de calcul si nous devons en effectuer.

1.3 – Représentation

Voici la variance cumulée selon le nombre d'axes :

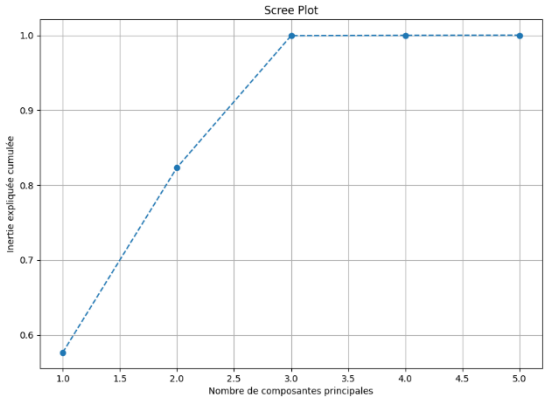


Figure 5 : Courbe d'inertie en fonction du nombre d'axes

Dans notre cas nous pouvons voir que c'est à partir de 3 axes que la courbe d'inertie commence à se stabiliser, donc en utilisant la méthode du coude nous devons choisir 3 comme nombre d'axes à conserver pour garder le maximum d'informations.

Voici la projection des individus sur le plan **E1 U E2** :

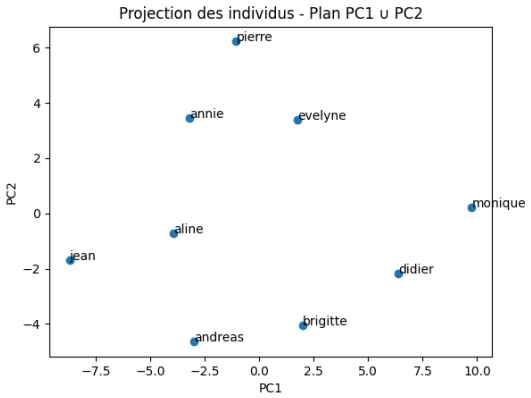


Figure 6 : Projection sur le plan E1 U E2

Voici le cercle des corrélations entre E1 et E2 :

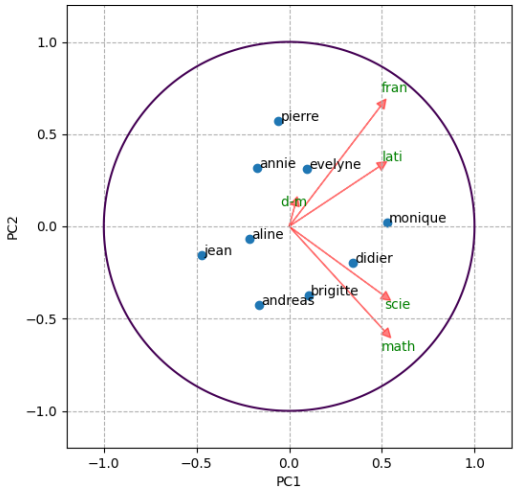


Figure 7 : Cercle des corrélations entre E1 et E2

Sur notre projection nous pouvons voir que l'axe E1 représente la moyenne de l'élève sans la matière 'dessin', nous pouvons voir cela car les directions des flèches vont toutes dans le même sens sur l'axe E1. L'axe E2 quant à lui, représente la corrélation entre la matière 'français' et 'latin'. Et cette corrélation est négative pour les matières 'Math' et 'Science' sur l'axe E2.

Donc sur la projection nous pouvons effectivement voir que quand un élève à une bonne moyenne (sans la matière 'dessins') alors plus la position de l'élève sera élevée sur l'axe PC1 (E1). Nous pouvons aussi voir les forces et faiblesses des élèves grâce à l'axe PC2 (E2). Si un élève est haut dans le graphique alors il est bon en 'français' et en 'latin', s'il est bas alors il est bon en 'math' et en 'science' et s'il se trouve au milieu alors il est aussi fort dans les sciences qu'en littérature.

Exemples :

- **Monique** : nous pouvons voir sur le graphe qu'elle est proche de 0 pour l'axe PC2 (E2), ce qui est normal car elle est aussi forte en sciences qu'en littérature (math : 14.5, sci : 14.5, fran : 15 et latin : 15). Elle est aussi la plus haute sur l'axe PC1 (E1) car c'est elle qui a la meilleure moyenne.
- **Pierre** : nous pouvons voir sur le graphe qu'il est haut sur l'axe PC2 (E2) ce qui est normal car comme nous l'avons vu PC2 représente la différence entre les sciences et la littérature, comme ici Pierre est bon en littérature et beaucoup moins bon en sciences lors il se retrouve en haut de l'axe. Aussi comme sa moyenne de sciences et de littérature est inférieure à 10 alors il se retrouve en dessous de 0 sur l'axe PC1 (E1).

Voici la projection des individus sur le plan **E1 U E3** :

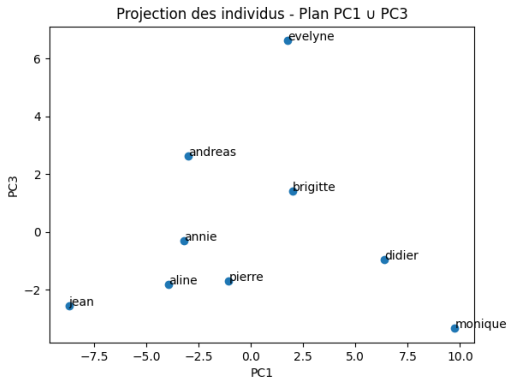


Figure 8 : Projection sur le plan E1 U E3

Voici le cercle des corrélations entre E1 et E3 :

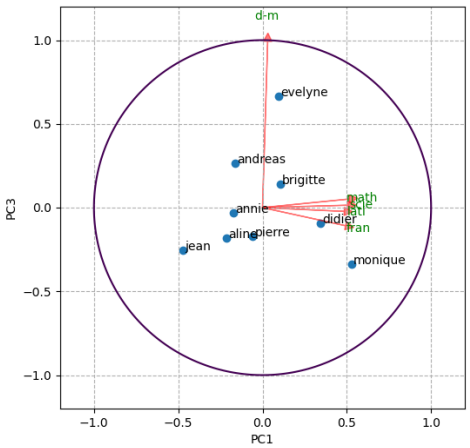


Figure 9 : Cercle des corrélations entre E1 et E2

Dans ce cas-ci, nous avons toujours l'axe E1 qui représente la moyenne (sans la matière 'dessin') comme pour E1 U E2. Ici c'est E3 qui change et représente seulement la corrélation avec la matière 'dessin'. Nous pouvons voir que sur le graphique, sur l'axe PC3 (E3), plus l'élève est haut alors plus il a une bonne moyenne en 'dessin'.

Exemples :

- **Monique** : Comme sur le graphe précédant nous pouvons voir qu'elle est toujours la plus haute sur l'axe PC1 (E1) et pour l'axe PC3 (E3) nous pouvons voir qu'elle est tout en bas, ce qui est normal puisse qu'elle a la pire moyenne en 'dessin'.
- **Evelyne** : Comme elle a la meilleure moyenne en 'dessin' forcément elle se retrouve tout en haut de l'axe PC3 (E3), aussi comme sa moyenne de sciences et littérature est supérieure à 10 alors elle se retrouve au-dessus sur l'axe PC1 (E1).

Voici une Heatmap pour nous permettre de mieux visualiser les composantes principales :

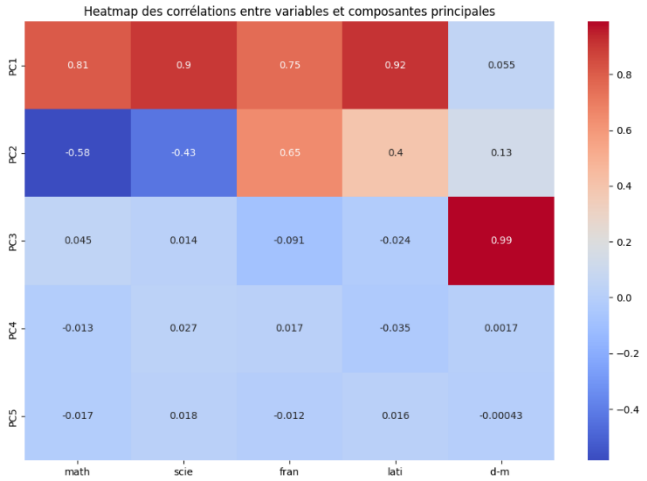


Figure 10 : Heatmap des composantes principales en fonction des matières

PC1 (E1) : les matières 'math', 'scie', 'fran' et 'lati' on un très grand impact et sont tous dans la même direction. Ici 'd-m' est négligeable car proche de 0.

PC2 (E2) : 'fran' et 'lati' on un très grand impact et 'math', 'scie' on un très grand impact mais négatif. Les deux s'annulent ce qui fait que quand une personne est aussi forte en sciences quand littérature alors sont PC2 est proche de 0. Ici aussi 'd-m' n'a pas un très grand impact.

PC3 (E3) : 'd-m' a un très grand impact positif et les autres matières sont proche de 0 donc elles ont un impact presque nulle. Cette composante est donc indexée sur la moyenne en 'd-m'.

2 - Données réelles

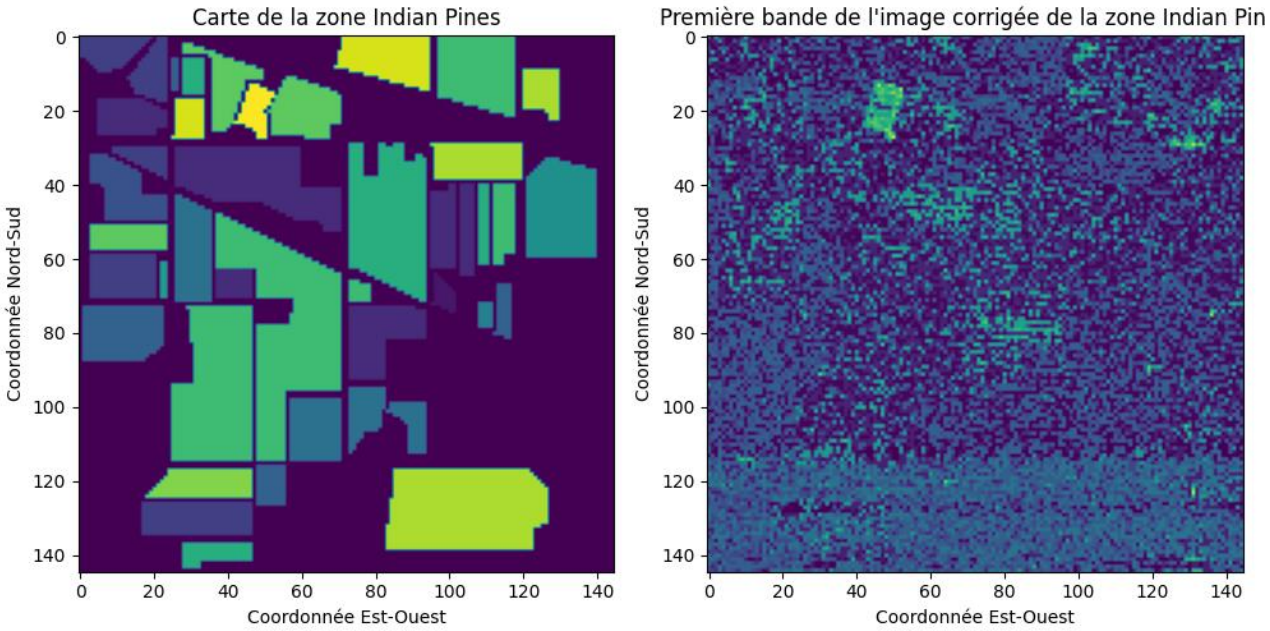


Figure 11: Affichage des zones Indian Pines

La première image « Carte de la zone Indian Pines » représente une carte colorée de la zone Indian Pines, avec différentes couleurs indiquant divers types de terrains/ caractéristiques géographiques. Les axes X et Y sont étiquetés "Coordonnée Est-Ouest" et "Coordonnée Nord-Sud", fournissant ainsi les coordonnées spatiales.

La deuxième image « Première bande de l'image corrigée de la zone Indian Pines » représente les données de la première bande spécifique de la 1ère image. Les couleurs dans cette image indiquent l'intensité du signal ou un autre paramètre mesuré dans cette bande spécifique.

Les deux images sont alignées en termes d'échelle et de coordonnées, facilitant ainsi une comparaison visuelle directe. C'est utile pour analyser les caractéristiques de la zone Indian Pines sous différentes perspectives.

Voici le nombre d’axes en fonction de l’inertie :

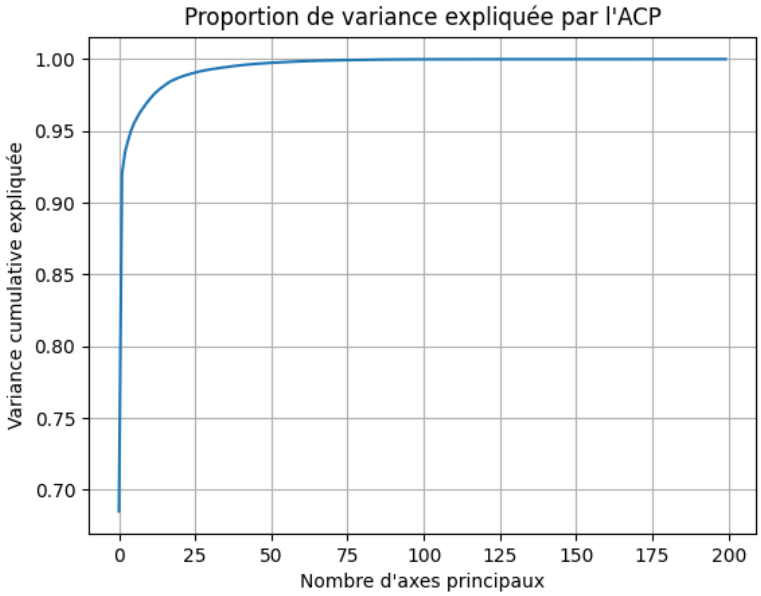


Figure 12 : Proportion de variance expliquée par l'ACP

En calculant le ratio de variance expliquée pour chaque composante principale ainsi que la variance cumulative, on peut tracer une courbe croissante représentant la proportion cumulative de la variance expliquée par l'ACP en fonction du nombre d'axes principaux. Cette courbe montre une évaluation du nombre nécessaire d'axes principaux pour conserver une grande partie de la variance des données. Dans notre cas la courbe atteint 95% de la variance expliquée avec 5 axes, cela signifie que ce nombre d'axes principaux est suffisant pour conserver cette proportion élevée d'information. Et pour avoir 99% des informations il nous faudrait prendre 25 axes.

Voici l’image d’origine refaite :

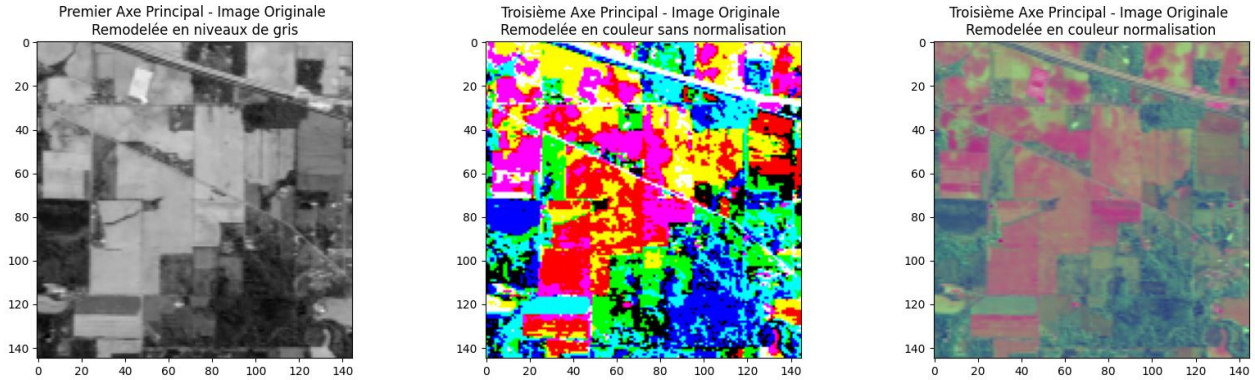


Figure 13 : Images d'origine reconstituées

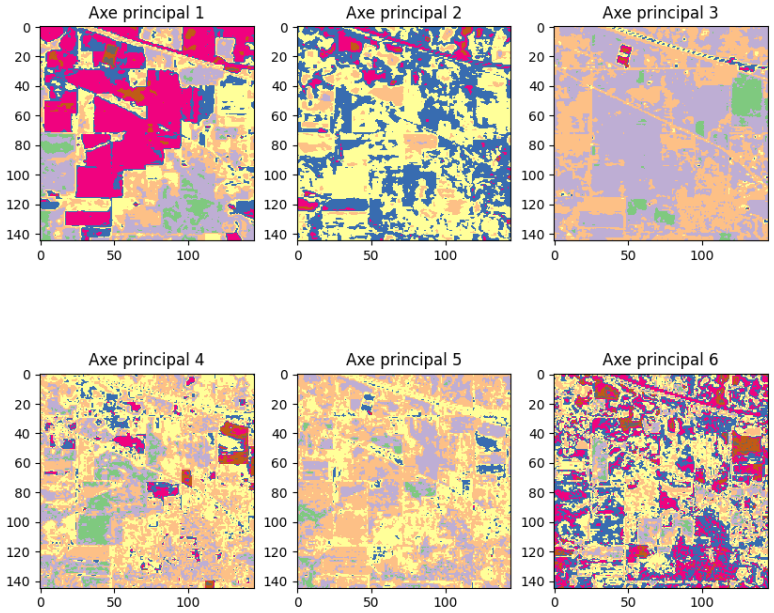


Figure 15 : Six premiers axes principaux

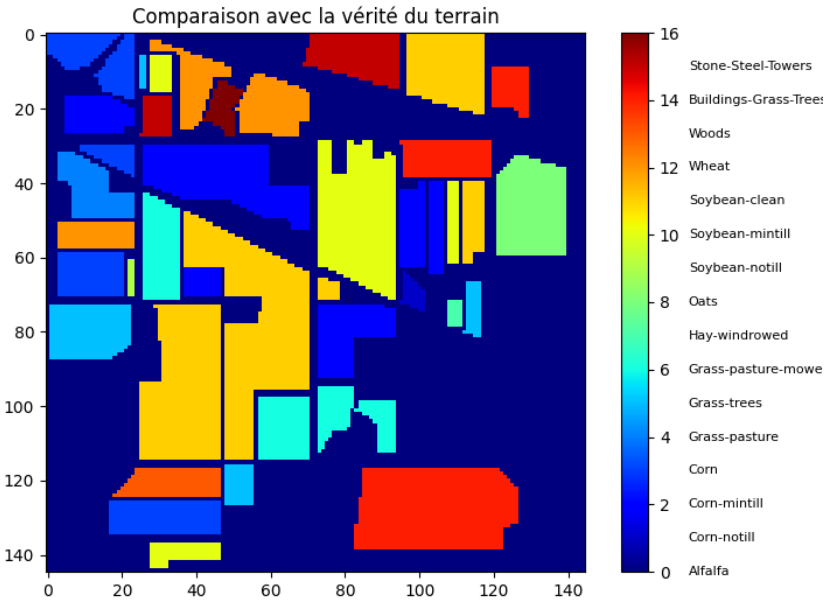


Figure 14 : Terrain de vérité

En appliquant l'Analyse en Composantes Principales (ACP) aux données et en traçant la variance cumulative expliquée par les axes principaux, nous avons observé qu'il était nécessaire d'utiliser environ 5 axes pour expliquer plus de 95% de la variance, ce qui permet de conserver 95% de l'information. Ensuite, nous avons sélectionné le premier axe principal et visualisé son aspect sous forme d'image en niveaux de gris. De plus, nous avons représenté la projection des données sur les 5 premiers axes principaux en couleur. Enfin, nous avons comparé ces résultats avec la vérité du terrain, à savoir une carte des classes de couverture du sol, et affiché la signature spectrale du premier axe principal.

Il est à noter que le premier axe principal semble refléter un contraste entre les zones cultivées et les zones naturelles. La projection des données sur les 5 premiers axes principaux montre une séparation significative entre les différentes classes de couverture du sol, suggérant ainsi que l'ACP peut servir de phase préliminaire pour la classification des images hyper-spectrales. En comparant la vérité du terrain avec les images en niveaux de gris ou en couleur, nous pouvons ainsi identifier les zones où l'ACP a bien ou mal capturé la variabilité des données

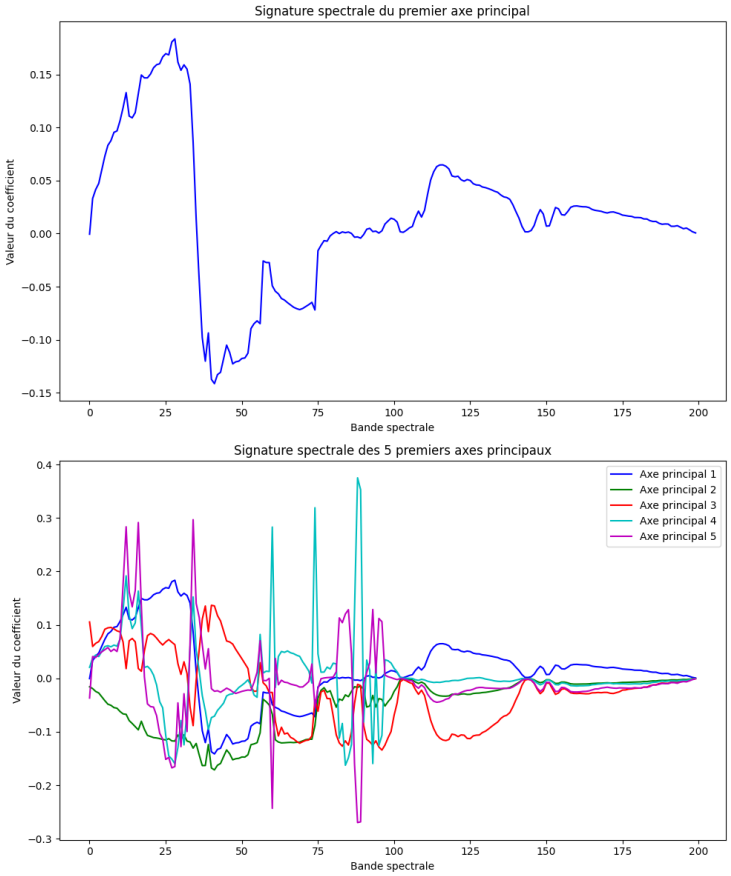


Figure 16 : Signature spectrale des axes principaux

On observe une variation du coefficient du premier axe principal en fonction de la bande spectrale. Ce coefficient représente le poids de chaque bande spectrale dans la combinaison linéaire qui constitue le premier axe principal, capturant ainsi la plus grande variance dans les données et résumant efficacement les différences entre les pixels de l'image.

On peut noter que le coefficient atteint une valeur maximale vers la 25ème bande spectrale, suggérant ainsi l'importance significative de cette bande dans la distinction des pixels de l'image. Cette bande correspond à la région du proche infrarouge, où la végétation reflète le rayonnement solaire. On peut donc supposer que l'image contient de nombreuses zones végétalisées « forêts, etc.. », et que le premier axe principal est particulièrement efficace pour les différencier des autres types de surfaces.

En revanche, le coefficient atteint une valeur minimale vers la 50ème bande spectrale, indiquant que cette bande a peu d'importance dans la distinction des pixels de l'image. Cette bande correspond à la région de l'infrarouge moyen, où l'eau absorbe le rayonnement solaire. On peut donc supposer que l'image contient peu de zones aquatiques, et que le premier axe principal ne permet pas de les détecter efficacement.

En fin, on a la signature spectrale des cinq axes principaux. Ce sont les mêmes axes permettant de conserver 95% de l'information. Ainsi, avec ce graphe, on peut confirmer qu'à la fin, les cinq convergent vers zéro, et donc on conservera l'information correspondante.

3 – Bonus

Nous utilisons maintenant K-Means pour transformer tous les résultats de l'ACP en une forme plus visuelle. Ici, nous utilisons K-Means avec un nombre de clusters égal à 5 pour une meilleure visualisation des parties. Après l'utilisation de K-Means sur notre résultat de l'ACP, nous obtenons les labels de la segmentation que nous devons reformer, ce qui va nous donner notre image finale.

Voici l'image des données de la vérité du terrain et le résultat que nous avons obtenu grâce à K-Means :

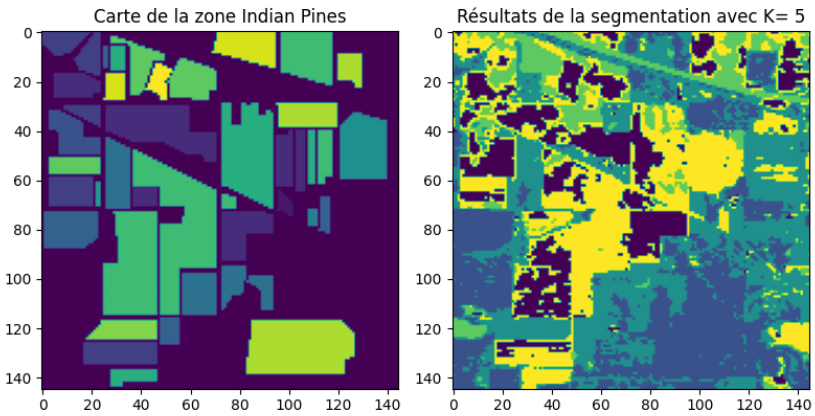


Figure 17 : Vérité du terrain et résultat de la segmentation de l'ACP

Nous pouvons voir que K-Means représente bien les grandes zones majeures et que nous nous rapprochons beaucoup du terrain de vérité, mais avec de petites imperfections. Nous avons choisi K=5 car en dessous de 5, l'image ne serait pas assez détaillée pour les différentes zones, et au-dessus de 5, nous aurions eu un découpage en trop de petites zones, ce qui ne nous aurait pas permis de voir correctement les grandes zones comme sur l'image de vérité.

4 - Conclusion

En résumé, nous avons étudié l'Analyse en Composantes Principales (ACP) dans deux contextes différents. La première partie a démontré l'efficacité de l'ACP dans l'analyse des performances scolaires, en mettant en évidence les forces et faiblesses des étudiants. La seconde partie nous a permis d'appliquer l'ACP à des données hyperspectrales, réduisant la complexité des informations tout en préservant les plus cruciales. Les images reconstituées ont illustré la capacité de l'ACP à révéler des structures significatives, démontrant ainsi son potentiel dans la classification d'images.

En conclusion, ce TP a souligné la puissance de l'ACP en tant qu'outil d'analyse de données polyvalent dans des domaines variés. Son application réussie dans l'évaluation des notes d'élèves et l'analyse d'images hyperspectrales montre sa pertinence dans le domaine de l'informatique, offrant une approche plus simple pour interpréter des données complexes.