BE Analyse de données – Epreuves du décathlon

Florent TACHENNE[1] Apolline BERNARD[2]

[1] Phelma, 3 Parvis Louis Néel, 38000 Grenoble, France

[2]Ense3, 21 Avenue des Martyrs, 38031 Grenoble, France

Résumé – Les scores totaux et par étapes des participants à un décathlon sont étudiés pour mieux comprendre comment se départagent les individus et les épreuves et pourquoi.

Abstract – The analysis of the scores by events of the individuals participating in a decathlon, makes it possible to better connect the figures with real natural explanations.

1 Analyse descriptive des données

Nous traçons une représentation synthétique de la distribution des valeurs pour chaque variable 1. On remarque que des valeurs aberrantes sont présentes parmis les variables saut à la perche et 1500m. En effet, en regardant le tableau de données de plus près les individus 58 et 59 réalise des score de 0 à la perche tandis que l'individu 53 réalise un score de 381 ce qui est très peu par rapport à l'avant dernier athlète 607 et la moyenne qui se situe dans les 600. Le but étant de réduire le nombre de données décrivant notre tableau, cela serait faux de résumer certains individus par un ou deux autres dont les score sont aberrants. Nous décidons donc de poursuivre l'analyse sans ces trois individus.

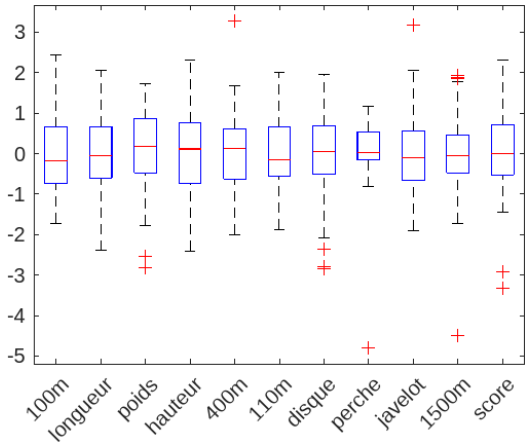
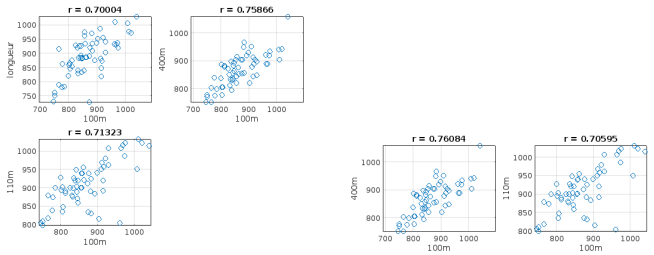


FIGURE 1 : Représentation synthétique de la distribution des valeurs pour chaque variable

Nous traçons alors les corrélations significatives pour chaque couple de donnée 2. Nous estimons que les valeurs de corrélations sont dites "significatives" lorsqu'elles dépassent 0.7.

On remarque un certain lien entre le score au 400m et au 100m, au 100m et au saut en longueur ainsi qu'au 100m et 110m. Cela ne semble pas irrationnel puisque le physique de "sprinteur" va aussi avec le saut en longueur. En revanche, lorsqu'on retire les individus 53, 58 et 59, on remarque que la corrélation entre les variables 100m et saut en longueur n'était



(a) Tout individus confondus (b) Sans les individus 53, 58 et 59

FIGURE 2 : Corrélations significatives

plus assez importante pour être significative. Nous verrons dans la partie suivante si le score au 100m permet de résumer le score dans d'autres disciplines.

2 Analyse en composantes principales

2.1 Nombres d'axes à retenir et qualité

Après calcul des composantes principales, nous traçons le spectre des valeurs propres 17

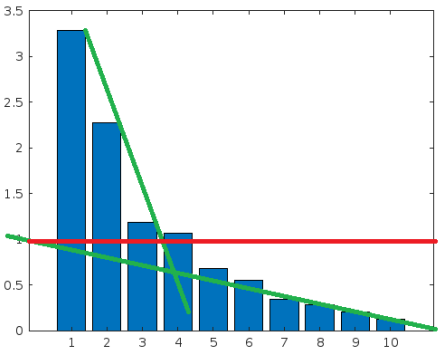


FIGURE 3 : Spectre des valeurs propres

On voit que le critère du coude nous fait garder que trois valeurs propres (vert) alors que le critère de Kayser (rouge) nous en fait garder quatre. Lorsqu'on calcule la qualité cu-

mulée, elle est de 0.7819 pour quatre axes. Comme elle est seulement de 0.675 pour 3 axes, nous gardons 4 axes. Sinon, nous manquerions de qualité pour réaliser notre analyse (les axes restants ne porteraient pas assez d'informations).

Regardons quelles variables sont les mieux décrites par les composantes principales gardées. Pour cela, nous représentons la qualité de chaque variable pour chaque composante principale 4. On remarque alors que toutes les variables ont une bonne qualité (>0,5) selon les composantes principales 1 à 4. La première composante principale est celle qui regroupe presque toute la qualité selon les variables 100m, longueur, 400m et 110m. Il y a donc un effet de taille sur cette composante principale.

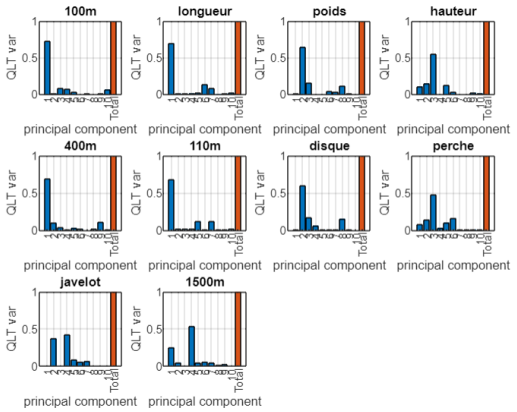


FIGURE 4 : Qualité des variables selon les axes principaux

2.2 Cercles de corrélation

Nous pouvons alors tracer les cercles de corrélations des individus dans les axes principaux retenus 5. Nous remarquons que les variables 100m, saut en longueur, 110m haies et 400m sont positivement corrélées selon la première composante principale (l'angle est faible et la corrélation s'approche du cercle unité). Aussi, les variables poids et disque sont positivement corrélées avec la deuxième composante principale. Enfin, les variables hauteur et perche sont positivement corrélées selon la troisième composante principale. En revanche, il n'y a pas de variables pour lesquelles on remarque une corrélation négative (angles proche de 180° et corrélation proche du cercle unité).

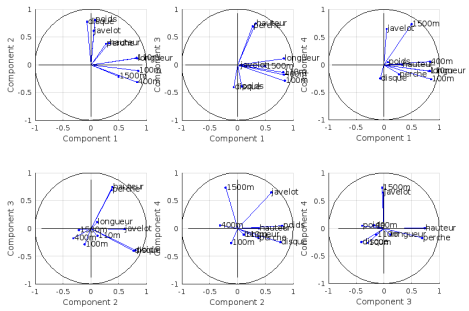


FIGURE 5 : Cercle de corrélations selon les plans retenus sans les individus 53, 58 et 59

2.3 Influence des individus

Intéressons nous maintenant à l'influence des individus dans l'ACP. Projetons les individus selon les composantes principales (Figure 6). Nous remarquons plusieurs individus qui ont été repérés en rouge après coup : 1, 2, 5, 49 et 52. Continuons l'analyse en regardant leur contribution suivant les axes (Figure 7). On peut constater que les individus 1, 2, 5, 49 et 52 ont une très forte contribution selon certains axes. Si leur contribution selon un axe est forte mais que la qualité de représentation selon ce même axe ne l'est pas assez, il faudra réaliser une étude supplémentaires sans cet individu.

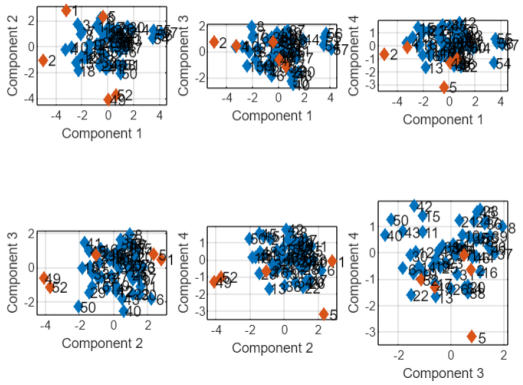


FIGURE 6 : Projection des individus dans les plans factoriels retenus

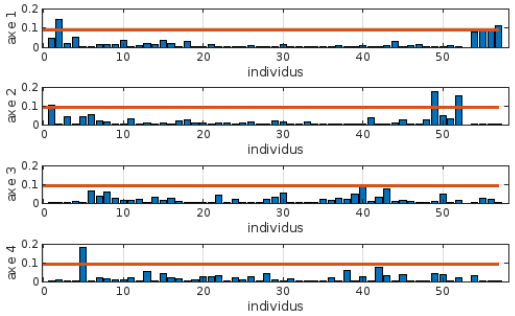


FIGURE 7 : Contribution des individus pour chaque composante principale avec un seuil égal à $5w_i$

Nous pouvons voir la qualité de représentation des individus par les axes principaux (Figure 8) avec en rouge celle des individus repérés suite à la projection dans les plans factoriels. Nous avons tracé un seuil haut à 0.8 et le seuil bas à 0.5. Nous remarquons que les individus 1 et 5 repérés précédemment n'ont pas une bonne qualité selon les axes retenus. Cela signifie que ces individus ne sont pas proches dans l'espace de départ. Seuls les individus 2, 49 et 52 ont une bonne qualité de représentation selon le premier ou le deuxième axe.

On peut aussi représenter la qualité sur les cercles de corrélation pour les axes qui nous intéressent (1er et 2nd) a la Figure 9. On observe que les individus 49 et 52 sont orientés dans le même sens que les variables disque et poids. Si l'on regarde dans notre tableau de données, on remarque en effet que ces individus ont fait les pires scores en lancer de disque et de poids. Ils sont donc bien proches dans l'espace de départ.

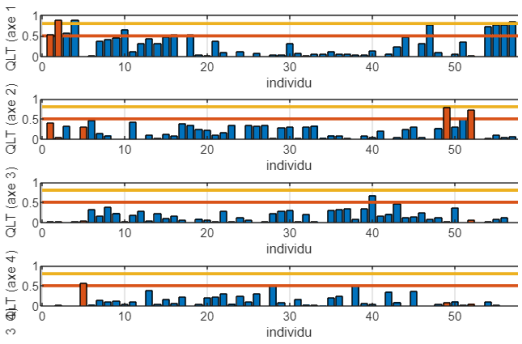


FIGURE 8 : Qualité de représentations des individus pour chaque axe principal retenu

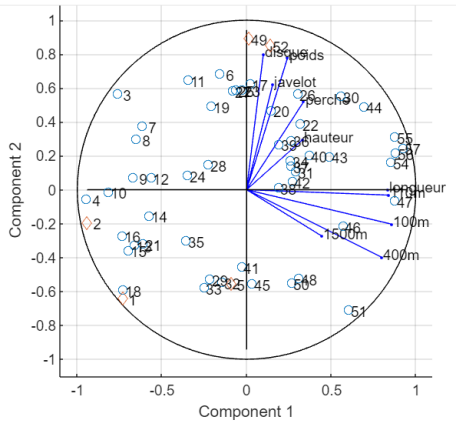


FIGURE 9 : Cercle de corrélations dans le plan 1, 2 et qualité d'observation des individus (ronds bleus)

Tentons de refaire l'ACP sans les individus de mauvaise qualité (1 et 5). Traçons les cercles de corrélations sans ces individus (Figure 13 et regardons si cela change quelque chose.

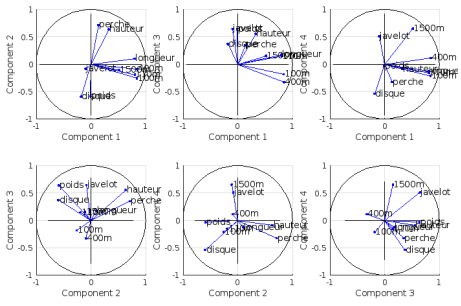


FIGURE 10 : Cercle de corrélations selon les axes retenus sans les individus 1, 5, 53, 58 et 59

Tout d'abord, nous pouvons voir que la variable 400m et les variables de sprint ainsi que la perche avec la hauteur sont moins corrélées. Peut être que ces individus étaient juste excellents en tout et cela ne permettait pas de représenter la dynamique globale des individus. Enfin, on observe sur le plan 1,2 que les variables perche et hauteur sont corrélées négativement avec les variables poids et disque. Il semblerait donc qu'être un bon lanceur ne permet pas d'être un bon sauteur et vice-versa.

2.4 Données supplémentaires

Regardons maintenant les résultats de l'ACP avec les individus 53, 58 et 59, retirés au début de l'analyse car ils présentaient des scores aberrants. La projection dans les plans factoriels (Figure 11 nous permet de voir comment se situent les individus par rapport au reste des athlètes. On remarque bien les individus 58 et 59 qui avaient fait un score de 0 à la perche. Cela ne s'avère pas vrai pour le n°53 qui avait un score très faible par rapport aux autres.

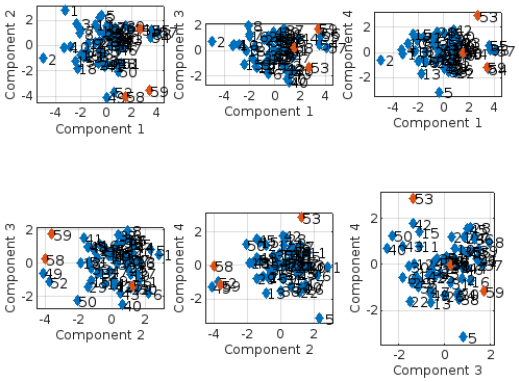


FIGURE 11 : Projections des individus non retenus dans l'ACP (rouge) selons les axes principaux retenus

Regardons la qualité de représentation de ces individus (Figure 12). Tous ces athlètes ont une mauvaise qualité de représentation sur les axes retenus et notamment l'axe 3 et 4. Traçons les cercles de corrélation avec ces individus et regardons ce que cela change (Figure 13. On ne remarque pas de grande différence en retraçant les cercles de corrélations avec les individus supplémentaires. Seules quelques valeurs de corrélations ont légèrement changé. Cependant, cela n'est pas assez pour changer les conclusions de l'ACP.

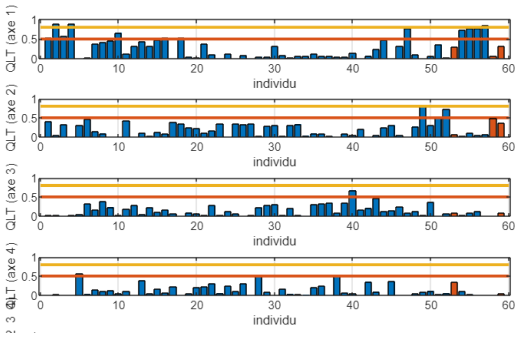


FIGURE 12 : Qualité de représentations des individus non retenus dans l'ACP (rouge) pour chaque axe principal retenu

Observons désormais ce qui en aurait été de l'ACP si l'on avait la variable score. Nous traçons les cercles de corrélation avec cette variable supplémentaire (Figure 14. Nous pouvons voir que la variable score n'atteint pas des valeurs assez proches du cercle unité pour pouvoir conclure sur sa corrélation avec une autre variable. Outre cela, on remarques que l'ajout de cette variable change la corrélation entre la hauteur et la perche (diminution). En somme, nous avons bien fait de la retirer de l'ACP.

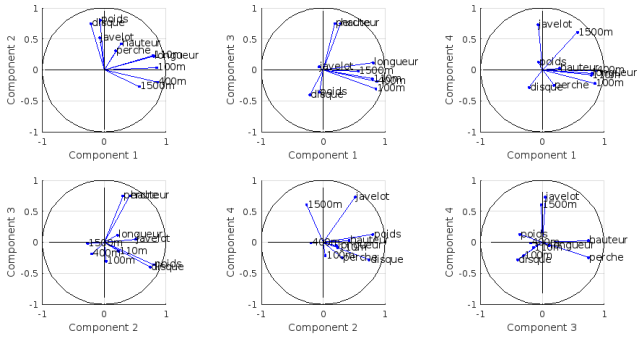


FIGURE 13 : Cercle de corrélations selon les plans retenus sans les individus 1, 2, 5, 49 et 52

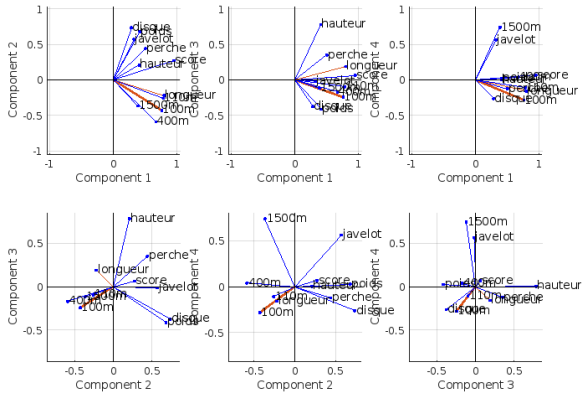


FIGURE 14 : Cercle de corrélations selon les plans retenus avec la variable score

3 Classification

Il est question ici, de classification hiérarchique des données utilisée pour le calcul de l'ACP. Dans un premier temps on cherche à classifier les décathloniens selon différents profils, les variables étant ici les différentes épreuves.

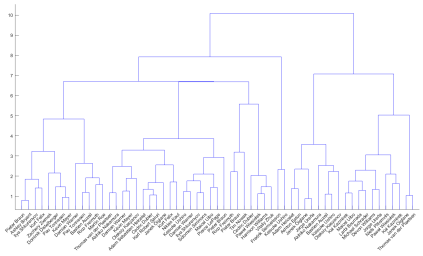


FIGURE 15 : Dendrogramme de classification des décathloniens par méthode de Ward

On observe différentes manières de couper le dendrogramme, il n'y à pas de coupure franche à effectuer. Pour orienter le choix du taux on peut s'appuyer sur l'évolution des indices obtenus.

On observe en figure 16 un coude qui permet de déterminer la coupure la plus adapter à 0.4. Un groupe de classe est réservé aux deux individus ayant raté le saut à la perche, ils ne sont

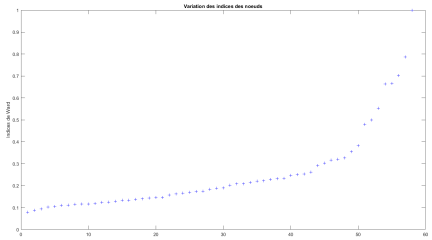


FIGURE 16 : évolution des indices de Ward

regroupés avec les autres qu'à l'avant dernier niveau, au delà du seuil choisi. En outre les décathloniens peuvent être rassemblés selon quelques profils correspondant aux groupes de classe : sprinter, lanceur, sauteur, coureur de fond. Par exemple un individu regroupé avec lui-même au premier niveau de classification à gardé le même profil au cours des années.

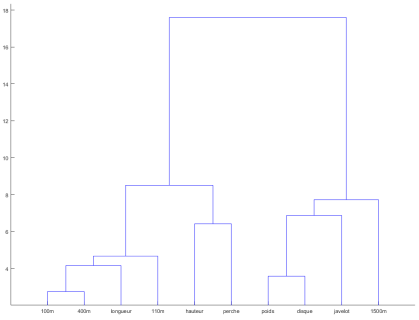


FIGURE 17 : Dendrogramme de classification des épreuves par méthode de Ward

Ici, on peut prévoir une coupure franche à 0.9. Les groupes obtenus correspondent à deux types de dispositions physiques, le premier groupe de classe regroupe les épreuves de sprint, tandis que l'autre groupe de classe regroupe plutôt les lancés.

Les différents outils mis en oeuvre dans cette analyse permettent de faire parler de jeu de donnée en donnant des informations utiles sur des individus sortant du lots, où bien sur des caractéristique physiques requises pour le décathlon.

4 Bibliographie

La bibliographie est insérée avec BibTEX. Dans l'article, utilisez la commande `\bibliography`. Le style de bibliographie n'est pas à préciser car il est automatiquement défini dans la classe : il s'agit de `plain-fr` (en français) ou `plain` (en anglais).

Références

[1] M. GOOSSENS, F MITTELBACH et A. SAMARIN : *The LaTeX Companion*. Addison-Wesley, 1994.

[2] L. LAMPORT : *LaTeX User's Guide and Reference Manual*. Addison-Wesley, 1994.