M2 Sciences Cognitives Fondamentales et Appliquées

M1 Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales

Université Lumière Lyon 2

Atelier Data Science

Dossier

Approches statistiques et perspectives alternatives à l'analyse du questionnaire de personnalité NEO-PI-R

Julien BASTIAN

Paul FORET-BRUNO

Ugo ZENNARO

Février 2023

Avant-propos

Le but initial de l'Atelier Data Science était de travailler sur des données réelles ou hypothétiques en rapport avec le stage de fin de Master des étudiants de M2 Sciences Cognitives. Néanmoins, suite à l'impossibilité d'obtenir des données (car encore non récoltées) en rapport avec le stage de Paul Foret-Bruno, et à la difficulté pour trouver sur internet des données s'approchant du type de données initialement prévu, le choix a été fait de travailler sur un type de donnée différent mais tout aussi intéressant, présenté dans la section suivante.

Contexte théorique

Le présent dossier a pour but de proposer et appliquer des approches statistiques et des perspectives alternatives à l'analyse du questionnaire *Revised NEO Personality Inventory* (NEO-PI-R ; Costa & McCrae, 1992). Le NEO-PI-R est un questionnaire de personnalité basé sur un modèle descriptif très influent en psychologie de la personnalité, le modèle à cinq facteurs (aussi connu sous le nom du modèle des *Big Five*). Issu d'une longue tradition en psychologie de la personnalité tentant de décrire la structure universelle des traits de personnalité par un nombre minimum de facteur, le modèle à cinq facteurs a commencé à gagner de l'influence théorique suite aux travaux de Goldberg (1981) qui tenta de résumer la personnalité en cinq grands facteurs : le Neuroticisme (tendance à ressentir des émotions négatives comme l'anxiété ou la dépression), l'Extraversion (tendance à se sentir énergique et chercher la compagnie des autres), l'Ouverture à l'expérience (tendance à apprécier l'art et les idées nouvelles), l'Agréabilité (tendance à la coopération et l'entente avec les autres), et la Conscientieusité (tendance à l'auto-discipline et à l'organisation). Ce modèle fut ensuite repris, développé, et implémenté par Costa et McCrae (1985) qui ont créèrent le questionnaire NEO-PI à partir de celui-ci. Une version révisée du questionnaire fut proposée par ces mêmes auteurs quelques années plus tard et correspond à le version utilisée aujourd'hui, le NEO-PI-R (Costa & McCrae, 1992).

Le NEO-PI-R subdivise les cinq grandes dimensions présentées juste au-dessus en six facettes, de sorte à ce que chaque dimension rassemble les scores de six sous-dimensions. Par exemple, les facettes pour le Neuroticisme sont l'Anxiété, l'Hostilité, la Dépression, la Conscience de soi, l'Impulsivité et la Vulnérabilité. Chaque facette est évaluée à l'aide de huit items notés sur une échelle de Likert allant de 1 (pas du tout d'accord) à 5 (tout à fait d'accord), donnant un total de 240 items pour l'ensemble du questionnaire. Le NEO-PI-R est constitué

d'une version dédiée à l'auto-évaluation, et une autre dédiée à l'évaluation par une tierce personne. Les scores finaux correspondent donc à la combinaison de ces deux versions, et est censé donner une idée globale de la personnalité d'un individu.

L'analyse statistique d'un tel questionnaire est intéressante pour cet Atelier Data Science car sa forme permet d'utiliser des approches statistiques différentes et potentiellement innovantes, tout en évoluant dans le cadre théorique des sciences sociales et plus particulièrement de la psychologie de la personnalité. Nous avons ainsi pensé que travailler sur de telles données permettrait de nous familiariser à l'utilisation d'outils statistiques dédiés à l'analyse d'un dataset relativement volumineux, ce qui sera sans aucun doute bénéfique pour la suite de nos parcours respectifs.

Méthode

Dataset – Les données ont été obtenues à partir d'un set de données en libre accès sur *Harvard Dataverse* (https://dataverse.harvard.edu/) et proviennent d'une étude ayant utilisé le NEO-PI-R pour récolter les données de 857 individus (Goldberg, 2018)

Analyses statistiques – Nos analyses statistiques auront pour but de tenter de réduire le nombre de dimensions du questionnaires afin d'étudier s'il est possible d'analyser le questionnaire sous un angle innovant et potentiellement plus informatif.

Dans le data set, les scores des cinq grands traits ont été obtenus pour 857 individus ainsi que le score des six facettes correspondantes à chaque traits. On considérera ici que le score sur un trait résume l'information des facettes qui lui correspondent ; géométriquement, cela revient à avoir une position dans six dimensions puis de la résumer dans une seule qui sera le trait. Les facettes donnent donc un espace total à 30 dimensions et les traits un espace à cinq dimensions. L'objectif des analyses statistiques appliquées ici est d'isoler les 30 facettes puis de leur appliquer une méthode de réduction de dimensions pour générer un espace de moins de 30 dimensions et dans lequel les axes (i.e., les dimensions) ne seront plus un résumé monolithique d'un trait mais pourront être une combinaison de plusieurs facettes appartenant à des traits différents. Le résultat sera alors des "méta-traits" qui seront plus riches que les traits originaux. Par exemple, dans le modèle initial, un axe exprime initialement le score de Neuroticisme, alors que suite aux analyses un axe pourra exprimer à la fois le Neuroticisme à un extrême et l'Ouverture à un autre extrême (voire des combinaisons plus complexes). Utiliser un tel protocole permettra potentiellement de remettre en cause la construction des cinq grands

traits de personnalité en montrant qu'ils pourraient peut-être être construit de façon à refléter une information plus pertinente et riche sur les individus.

Notre analyse reposera sur trois approches statistiques, avec en premier lieu la méthode de l'analyse en composantes principales (ACP) qui est mathématiquement optimale pour la réduction de dimensions. L'ACP consiste en la création d'un petit nombre d'axes à partir de données en grande dimensions, qui seront chacun une combinaison des dimensions d'origine (aussi appelé réduction d'espace). La méthode est optimale au sens où elle conserve au mieux la variance dans les données. Cette variance peut être amalgamée avec « l'information » contenue dans les données : en effet, plus les données varient, plus il est possible de les différencier et donc plus elles contiennent de l'information. Du point de vue de la variance, l'espace initial en 100% et l'espace réduit en exprimera un certain pourcentage, au sein duquel chaque axe (i.e., chaque dimension) portera une partie de ce pourcentage. On parle ainsi de pourcentage de variance expliquée par un axe.

Des méthodes de clustering (création automatique de groupes) seront également utilisées pour connaitre les facettes qui se ressemblent le plus et sont le plus éloignées afin d'affiner notre analyse. Le clustering est une méthode qui permet de créer une forme de classification de données sans a priori. Ici, nous avons réalisé une forme de clustering appelée classification ascendante hiérarchique (CAH) et nous avons travaillé sur les facettes pour rendre apparente la proximité entre certaines (parfois étonnante) et la distance entre d'autres. La CAH rapproche itérativement les facettes entre elle. Ainsi, les facettes sont d'abord rapprochées de celles qui sont les plus similaires, puis le groupe obtenu est rapproché d'un autre groupe étant le plus similaire également. Cette classification est ascendante hiérarchique dans le sens où le graphique, appelé dendrogramme, se lit de bas en haut, et que plus le rapprochement se fait bas dans celui-ci, plus il correspond à une grande similarité entre deux éléments.

Enfin, on considérera aussi, plus partiellement, les auto-encodeurs qui sont une architecture de réseau de neurones adaptée pour la réduction de dimension. Il sera notamment possible de voir avec cette approche la différence de performance en fonction du nombre de dimensions choisi. Un auto-encodeur est constitué de deux parties. D'abord l'encodeur, qui est composé d'une couche d'entrée, avec dans notre cas 30 dimensions (30 facettes), puis d'une couche cachée de plus petite dimension, ici 15 neurones (donc 15 dimensions), et enfin d'une couche d'encore plus petite taille, appelée goulet d'étranglement, dont la taille est précisée plus loin. La deuxième partie est le décodeur. Elle utilise en entrée la couche de plus petite dimension de l'encodeur (i.e., le goulet d'étranglement), la fait suivre par une couche de plus grande dimension (ici 15 également) et se termine par une couche de sortie de dimension égale à

l'entrée initiale (soit 30). Cette architecture a donc une forme en « papillon » et repose sur un principe simple : dans l'encodeur est encodée l'information dans un nombre plus petit de dimensions, puis dans le décodeur ces dimensions sont décodées jusqu'à retrouver l'espace d'origine. Pour entraîner ce réseau de neurones, un vecteur lui est donné en entrée (ici un individu) et sa tâche est de l'encoder puis de le décoder, de sorte à minimiser la différence entre l'individu de départ et sa version encodée puis décodée. Dans la présente étude, 20 itérations par architecture avec à chaque fois 50 epochs et des batch de 64 individus ont été utilisées pour entrainer le réseau. À chaque itération, 70% des individus ont été sélectionnés pour l'entrainement et 30% pour le test. Les fonctions d'activations dont étaient équipées les couches tout au long du réseau étaient des leaky ReLU (Rectified Linear Unit), à l'exception de la sortie qui était une sigmoïde. La fonction de perte utilisée était l'entropie croisée et l'optimiseur Adam a été employé. Enfin, pour évaluer les performances lors du test, la norme de Frobenius, qui fait office de norme de référence lorsque l'on s'intéresse à des matrices, a été utilisée. De la même façon que l'ACP, les auto-encodeurs cherchent ainsi à réduire le nombre de dimensions, ici via la taille du goulet d'étranglement. Pour tester les performances de cette méthode, trois tailles de goulet d'étranglement ont été testées, à savoir 4, 5 et 6, et leur performances ont été comparées.

Toutes les analyses présentées ci-dessus ont été conduites sur R (v4.2.2). Le script entier sera joint au dossier et pourra être consultable.

Résultats

Analyse en composantes principales (ACP) – La représentation graphique de l'application de l'ACP aux données est visible en Figures 1a et 1b. L'utilisation de l'ACP a révèle que cinq dimensions (ou composantes) expriment 60% de la variance du set de données, et plus encore que quatre dimensions expriment 54% de la variance. Pour la suite du dossier et pour l'intérêt de l'exercice, on considérera ainsi que ces cinq dimensions estimées comme majeures par l'ACP correspondent à des méta-traits (cf. Méthode).

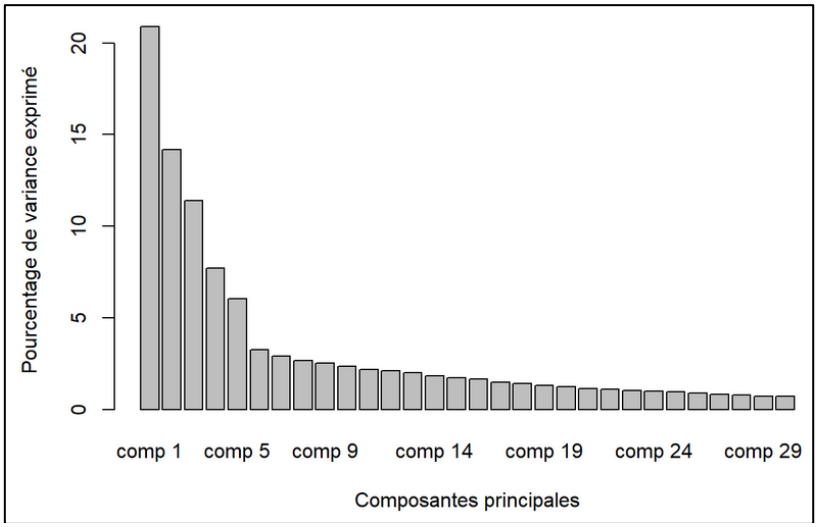

Figure 1a : Pourcentage de variance exprimé par chaque axe (ou composantes).

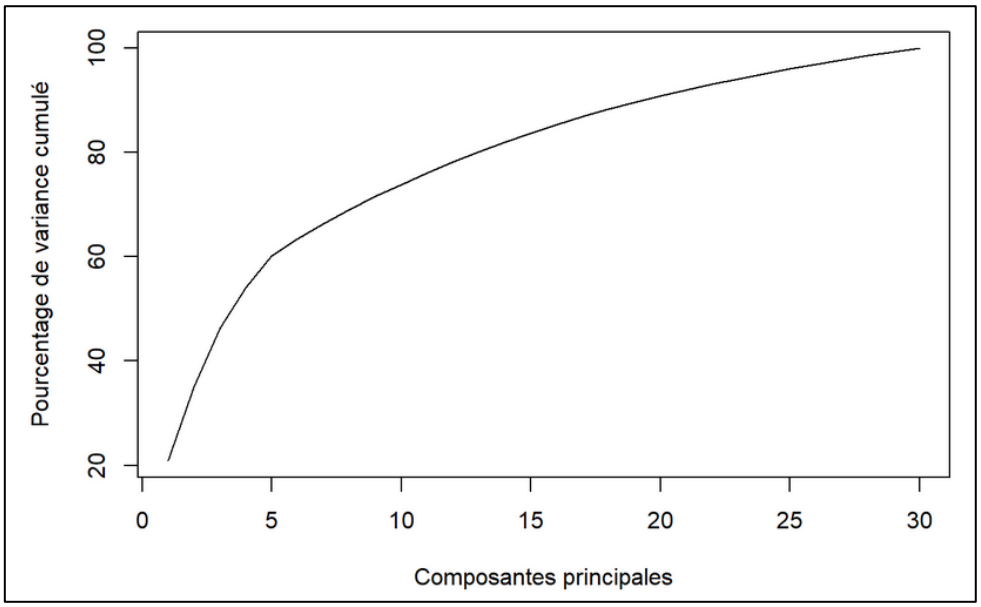

Figure 1b : Pourcentage de variance cumulé exprimé par les axes (ou composantes).

Afin de déterminer de quelles combinaisons de facettes d'origines sont constitués ces méta-traits, la corrélation de chaque facette avec chacun des cinq méta-traits a été effectuée et est représentée sous forme de matrice de corrélations en Figure 1c.

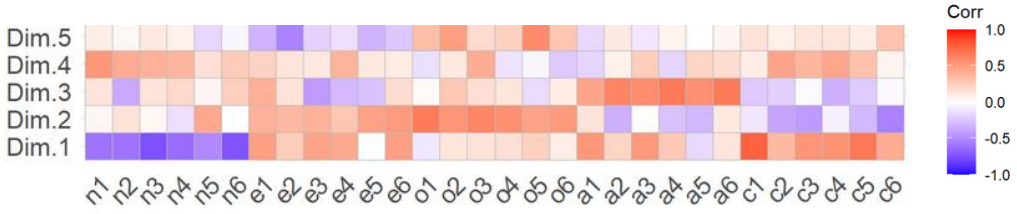

Figure 1c : Matrice de corrélation des facettes avec les cinq méta-traits exprimant le plus de variance.

La première chose à noter est que les méta-traits (i.e., dimensions) 1, 2 et 3 sont le plus corrélés aux facette n1 à n6 (négativement), o1 à o6 et a1 à a6 (positivement), respectivement. Cela suggère que ces méta-traits issus de l'ACP correspondent assez fidèlement aux traits d'origine Neuroticisme, Ouverture et Agréabilité. Cependant, il est intéressant de noter que le méta-trait 1 est également fortement associé positivement aux facettes c1 à c6 appartenant au trait d'origine Conscientieusité. Il ressort donc de l'ACP qu'un méta-trait permettrait de rassembler l'information des deux traits d'origine Neuroticisme et Conscientieusité, et qu'un haut score sur ce méta-trait traduirait une faible propension aux Neuroticisme et une haute propension à la Conscientieusité.

Ce pattern est en revanche difficilement observable pour les méta-traits 4 et 5, où les corrélations sont plus diverses et réparties sur un plus grand nombre de facettes ne correspondant pas aux traits d'origines. On peut par exemple noter que le méta-trait 5 semble être associé le plus aux facettes e2, o5 et o2, tandis que le méta-trait 4 semble être le plus associé aux facettes n1 à n4 et c2 et c4. Cela suggère que ces méta-traits reflètent une information diverse, plus riche et dépassant les bordures des traits d'origines.

Nous pouvons enfin noter qu'aucune facette de l'Extraversion n'est fortement corrélée avec un méta-trait.

Clustering – La CAH réalisée sur les facettes et sur les traits d'origines est représentée graphiquement sous forme de dendogrammes en Figure 2a (facettes) et Figure 2b (traits).

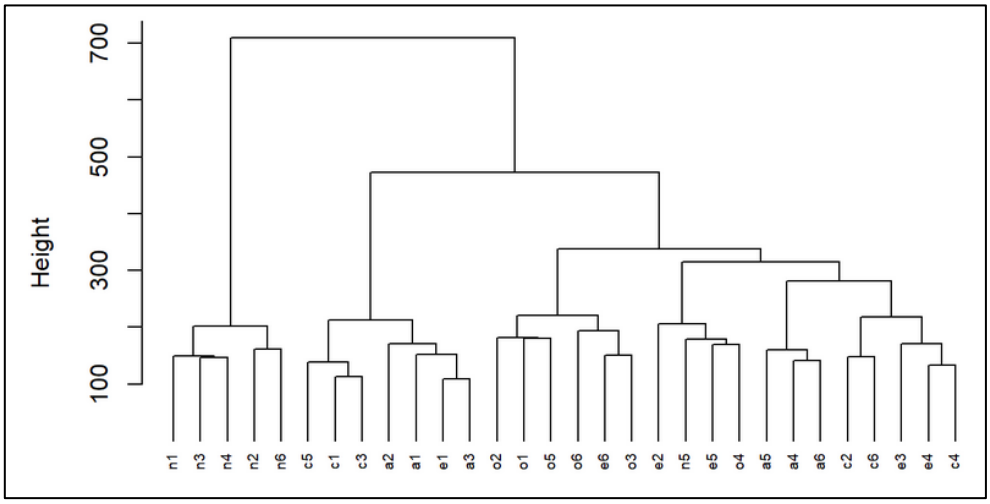

Figure 2a : Classification ascendante hiérarchique des facettes.

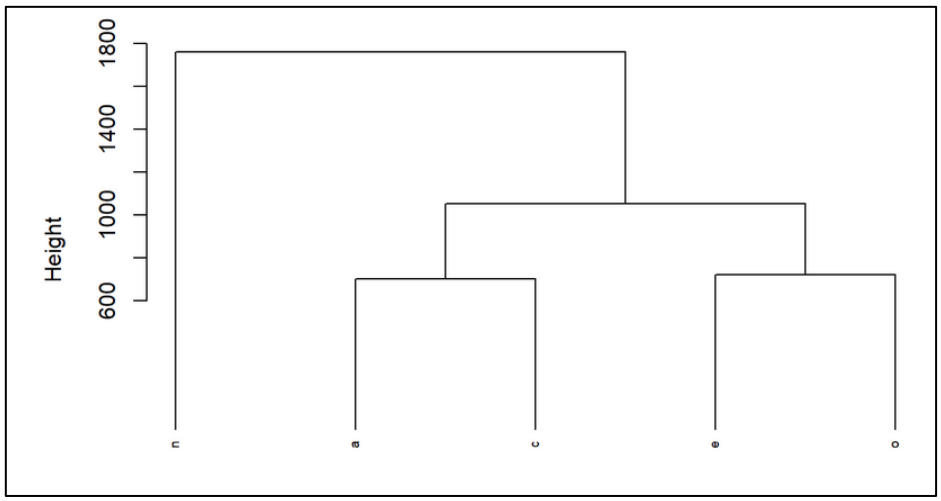

Figure 2b : Classification ascendante hiérarchique des traits.

Cette analyse à l'aise du clustering nous permet de noter plusieurs choses. Tout d'abord, on peut remarquer la formation de cinq groupes commençant à se distinguer de manière significative à partir de la hauteur 150 sur le dendogramme des facettes (Figure 2a). Les groupes seront ici lus de gauche à droite sur le graphique. À l'exception du premier groupe (à gauche), qui rassemble la majorité des facettes du Neuroticisme, et du troisième, qui rassemble la majorité des facettes de l'Ouverture, chaque groupe rassemble des facettes appartenant à différents traits d'origines. Notamment, le deuxième groupe rassemble trois facettes de la Conscientieusité et trois facettes de l'Agréabilité, suggérant que certaines caractéristiques de ces traits peuvent être rapprochés. Les quatrième et cinquième groupe sont constitués d'une diversité de facettes appartenant pour la plupart aux traits d'origine Agréabilité, Extraversion et Conscientieusité.

On peut également remarquer que la CAH a regroupé à une hauteur plus importante les trois derniers groupes, suggérant une relative similarité entre ces derniers. Enfin, il apparaît avec la CAH que le premier groupe, qui rassemble cinq des six facettes du Neuroticisme, est le groupe le plus distinct des autres, car il n'est regroupé avec eux qu'à la hauteur maximum.

Ce dernier résultat est en accord avec la CAH réalisée sur les grands traits (Figure 2b). En effet, sur cette CAH, le trait Neuroticisme reste distinct des autres traits jusqu'à la hauteur maximale dans la hiérarchie. L'Agréabilité et la Conscientieusité sont rassemblées quant à elles dès la première hauteur, ce qui est très similaire au deuxième groupe formé par la CAH sur les facettes, qui rassemblait trois facettes de l'Agréabilité et trois de la Conscientieusité. Enfin, l'Ouverture et l'Extraversion sont ici également regroupées dès le début. Ce dernier résultat n'est pas incompatible avec la CAH sur les facettes, qui avait regroupé à une hauteur intermédiaire (environ 300) les deux derniers groupes, qui contenaient cinq des six facettes de

l'Extraversion, avec le troisième groupe qui contenait également cinq des six facettes de l'Ouverture. Globalement, la CAH sur les traits supporte donc la CAH sur les facettes (qui est plus informative).

Auto-encodeurs – Les performances de chaque architecture de réseau de neurones auto-encodeurs utilisé sont représentées graphiquement dans les Figures 3a et 3b.

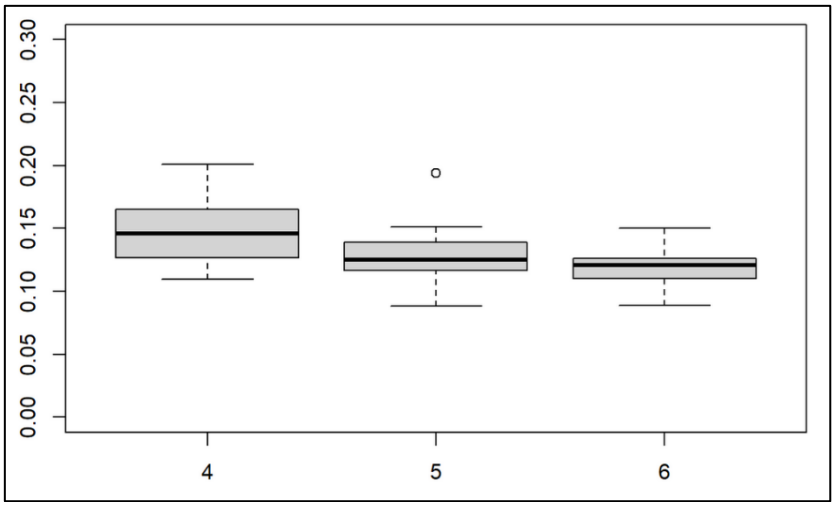

Figure 3a : Boîtes à moustaches représentant la performance de chaque architecture (en norme de Frobenius) en fonction de le taille de chaque goulet d'étranglement.

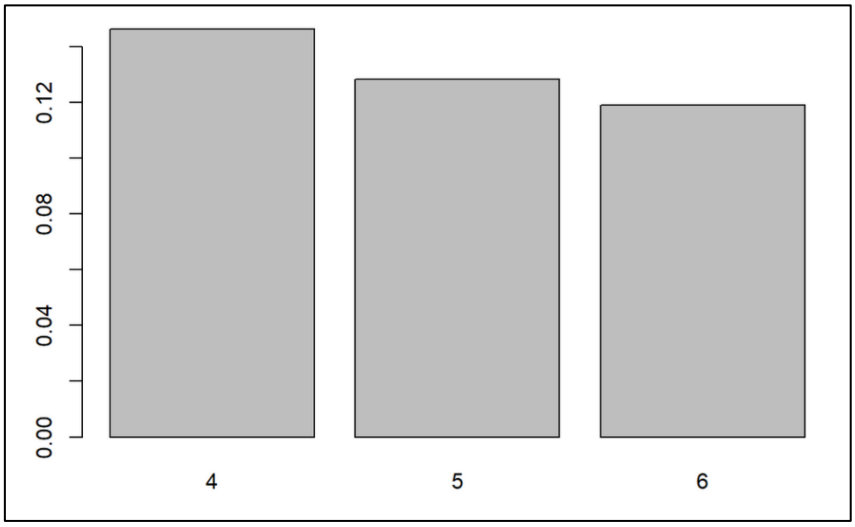

Figure 3b : Histogramme représentant la performance moyenne de chaque architecture (en norme de Frobenius) en fonction de le taille de chaque goulet d'étranglement.

On peut noter suite à l'application des auto-encodeurs qu'aucune architecture n'apparaît comme significativement meilleure que les autres. Ce résultat est en lui-même une information intéressante car il suggère que les performances sont similaires lorsque quatre ou six dimensions

sont utilisées. Ceci est en accord avec le premier résultat de l'ACP montrant que six dimensions expliquent 65% de la variance des données tandis que quatre dimensions en expliquent 54%.

Discussion

Dans le présent dossier, nous avons cherché à appliquer des approches statistiques alternatives pour l'analyse du questionnaire de personnalité NEO-PI-R. L'analyse en composantes principales (ACP), la classification ascendante hiérarchique (CAH) et les réseaux de neurones auto-encodeurs ont été utilisés sur un set de données en libre accès provenant de 857 individus ayant rempli le questionnaire.

L'utilisation de l'ACP a permis de mettre en évidence que 60% de la variance du jeu de données étaient expliqués par cinq dimensions tandis que 54% étaient expliqués par quatre dimensions, suggérant que l'utilisation de quatre dimensions seulement permet de conserver une grande partie des informations contenues dans les données. L'ACP a ensuite révélé que les méta-trait 2 et 3 étaient corrélés le plus aux facettes Ouverture et Agréabilité, respectivement, suggérant que ces facettes délimitent un aspect de personnalité bien distinct mathématiquement parlant, ce qui est en accord avec la structure initiale du questionnaire NEO-PI-R Il est en revanche ressorti de l'ACP qu'un méta-trait semblait inclure des facettes de deux traits d'origine majeurs que sont le Neuroticisme et la Conscientieusité. En effet, ce méta-trait était fortement corrélé négativement aux facettes du Neuroticisme et fortement corrélé positivement aux facettes de la Conscientieusité. En d'autres termes, il ressort de cette analyse une association forte et proportionnellement inverse entre la propension à être Consciencieux et la propension au Neuroticisme. Ainsi, il apparaît que plus l'on est auto-discipliné et organisé, moins l'on présente de tendance à ressentir des émotions négatives comme l'anxiété et la dépression. Ce résultat est intéressant car il suggère l'existence d'un méta-trait plus riche en termes de contenu informatif que ceux du questionnaire. Enfin, l'ACP suggère que les facettes de l'Extraversion sont plus informatives lorsqu'elles sont réparties dans plusieurs méta-traits que regroupé sous un seul trait. En effet, aucun des méta-traits ne présentait de forte corrélation avec les facettes de l'Extraversion, mais plutôt une corrélation répartie de manière « diffuse. »

Les résultats de la CAH sur les facettes (et dans une moindre mesure sur les traits) vont globalement dans la direction des résultats de l'ACP. En effet, la CAH a fait ressortir plusieurs groupes rassemblés en fonction de leur similarité. De manière intéressante, le premier groupe (rassemblant cinq facettes du Neuroticisme) et le troisième groupe (rassemblant cinq facettes

de l'Ouverture) sont assez similaires aux méta-traits 1 (Neuroticisme et Conscientieusité) et 2 (Ouverture), respectivement. La seule différence entre le premier groupe et le méta-trait 1 est l'absence des facettes de la Conscientieusité dans le premier groupe. En revanche, la CAH n'a pas permis de retrouver un groupe similaire au méta-trait 3 (Agréabilité), ou seulement partiellement ; en effet, dans ce groupe, trois facettes de l'Agréabilité ont été regroupées avec trois facettes de la Conscientieusité. Enfin, la CAH confirme de plus un des résultats de l'ACP suggérant que les facettes de l'Extraversion sont plus informatives lorsqu'elles sont réparties dans plusieurs méta-traits ou groupes que lorsqu'elles sont rassemblées sous un seul et même trait. En effet, les facettes de l'Extraversion étaient ici réparties dans quatre groupes différents. Ainsi, l'Extraversion semble, suite à ces deux analyses, plutôt arbitraire et ne reflétant pas ou très peu une information pertinente.

Enfin, l'utilisation d'auto-encodeurs a permis de montrer que la performance des architecture des réseaux de neurones étaient très similaires lorsque le nombre de dimensions était de quatre, de cinq ou de six. En d'autres termes, il semble que l'utilisation de quatre dimensions soit équivalent à l'utilisation de cinq ou six dimensions, Ces résultats sont très intéressants au vu de ceux obtenus par l'ACP et la CAH, car ces deux approches ont révélé que les facettes de l'Extraversion ne pouvaient pas être rassemblées sous un seul et même groupe mais répartissaient plutôt leur pouvoir explicatif dans plusieurs groupes. Il se pourrait ainsi que le trait Extraversion soit un trait ne présentant pas vraiment de délimitation mathématique et que l'inclusion de ces facettes dans plusieurs autres méta-traits soit plus pertinent. Ceci pourrait expliquer pourquoi la performance des auto-encodeurs avec quatre dimensions ne diffère pas significativement de ceux avec cinq ou six dimensions.

Conclusion

En conclusion, l'utilisation d'approches statistiques alternatives a permis de mettre en perspective la construction du questionnaire de personnalité NEO-PI-R et plus généralement du modèle des Big Five, en suggérant une nouvelle façon de construire les grands traits de personnalité par l'inclusion des facettes de l'Extraversion au sein de plusieurs méta-traits, ainsi que par une association des facettes de la Conscientieusité avec le Neuroticisme ou l'Agréabilité. De futures analyses statistiques utilisant des approches plus puissantes sur un échantillon plus important pourraientt permettre d'apporter encore plus de perspectives sur l'analyse de ce questionnaire.

Références

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory manual* (FL: Psychological Assessment Ressources). Odessa.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO-Five-Factor Inventory (NEO-FFI)* (FL: Psychological Assessment Ressources). Odessa.

Goldberg, L. (2018). *(3) NEO-PI-R* [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/HE6LJR

Goldberg, L. R. (1981). Language and individual differences : The search for universals in personality lexicons. *Review of Personality and Social Psychology*, *2*(1), 141-165.