Projet de traitement de donnée : Les personnalités

Charles Vin
Maeva Hermand

Décembre 2020

Table des matières

1 Introduction

Notre étude porte sur le thème des personnalités avec 19 719 données recueillies provenant d'une base de données de 2012. Il s'agit des résultats obtenus suite au test de personnalités "Big Five" [2] aussi connu sous le nom du "Model OCEAN" se trouvant sur un site internet. Il classe les personnalités sous cinq grandes catégories :

L'ouverture : appréciation de l'art, de l'émotion, de l'aventure, des idées peu communes ou des idées nouvelles, curiosité et imagination ;

La conscienciosité : autodiscipline, respect des obligations, organisation plutôt que spontanéité ; orienté vers des buts ;

L'extraversion : énergie, émotions positives, tendance à chercher la stimulation et la compagnie des autres ;

L'agréabilité : Une tendance à être compatissant et coopératif plutôt que soupçonneux et antagonique envers les autres ;

Le neuroticisme : contraire de stabilité émotionnelle : tendance à éprouver facilement des émotions désagréables comme la colère, l'inquiétude ou la dépression, vulnérabilité.

Ces catégories sont divisées en 10 questions chacune. Par exemple, *"I am the life of the party."* pour l'extraversion ou encore *"I get stressed out easily."* pour le neuroticisme. Les participants ont répondu grâce à une échelle de Likert (de 1 à 5) pour exprimer son degrés d'accord ou de désaccord.

1.1 Problématique

Nous allons travailler autour de ces deux questions :

Existe-il des grandes différences de personnalités entre les hommes et les femmes, entre les États-Unis et le monde ?

Les résultats du test sont-ils fiables d'un point de vue mathématique ?

1.2 Plan

Nous traiterons ce sujet sous 5 parties :

1. Tout d'abord, nous allons faire une présentation globale de la base de données et nous montrerons comment nous avons préparé les données.
2. Ensuite, nous allons chercher ce qu'on appelle "Le profil-type".
3. Nous allons ensuite tester la fiabilité du questionnaire et de la variable *Personality_ type* grâce à l'algorithme K-means
4. Dans cette partie, nous allons comparer les États-Unis et le monde, les hommes et les femmes.
5. Enfin, nous allons analyser nos résultats.

2 La base de donnée

2.1 Présentation de la base de donnée

La majorité des participants provient des États-unis (plus de 8000 personnes) ainsi qu'une partie provenant d'Asie. Les autres participants sont plutôt dispersés dans le monde. On remarque

aussi qu'une majorité de femme a répondu au test, environ 61% de femmes contre 39% d'hommes. Les participant ont en moyenne la 20aine, on peut justifier cela par le fait que ce test soit en ligne. On voit aussi que 63% des participants ont l'anglais en langue natale, et plus en détaille on peut dire qu'envions 66% des femmes et 58% des hommes parlent anglais de naissance. On peut également justifier ces observations en disant que ce test à été majoritairement effectué aux États-unis.

2.2 Préparation de la base de donnée

Dans une nouvelle base de donnée, nous avons fait la moyenne des dix questions pour chaque catégorie (E,N,A,C,O).Grâce à cela, nous avons pu définir une personnalité-type en prenant la catégorie où la moyenne à la catégories de question était la plus haute. Nous avons choisi cette manière de faire correspondre une personnalité à un profile mais nous savons pertinemment qu'un profile avec des réponses aux catégories parfaitement ou presque équilibrées aura un résultat de personnalité faussé.

3 Création du *Profile_ type*

Nous définissons le Profile-type comme le profil avec les caractéristiques les plus répandues chez les participants du test.
Pour l'obtenir, nous avons utilisé plusieurs méthodes de calculs en fonction de la variable. Pour ce qui est de la race, de la langue natale, du genre, de la préférence manuelle (droitier ou gaucher), du pays et de comment le participant a trouvé le test, nous avons choisi d'effectuer un **maximum**.
Pour ce qui est de l'âge et, comme dis précédemment, pour chaque catégorie de question, nous avons effectué une **moyenne**.
En ce qui concerne le type de personnalité, nous avons donc effectué une **moyenne** de chaque catégorie de question (E,N,A,C,O) et nous avons ensuite pris le **maximum**.

Notre Profile-type est donc une femme, d'origine caucasienne d'Europe de 26 ans et 4 mois, qui parle anglais de naissance, qui est droitière, qui habite aux États-unis et qui vient d'une autre page sur le même site. On lui trouve 3.1 comme moyenne aux réponses du type E et N, 3.2 aux réponses du type A et C, et 3.3 aux réponses du type O. Elle aurait donc un profil ouvert. On ne trouve pas les mêmes résultat en comptant le nombre de profil avec la variable *Personality_ type* car la moyenne prend en compte toutes les personnalités alors que *Personality_ type* peut ignorer une moyenne élevé au profit d'une autre même légèrement plus élevé.

4 K-Means

4.1 Qu'est-ce que K-Means

K-Means ou K-moyenne en français, est un algorithme de clustering, c'est à dire de partitionnement de donnée. Il est utilisé en l'apprentissage non-supervisé ou semi-supervisé. Cette branche de l'apprentissage machine vise à extraire des classes ou groupes d'individus représentés par des points de n-dimension et ayant des caractéristiques communes. Contrairement à l'apprentissage supervisé, il ne nécessite pas d'éléments déjà classés.

Il est utilisé dans la **classification par cluster**, ce que nous allons faire ici. Ou bien en quantification vectorielle, comme par exemple réduire la palette de couleur d'une vidéos. Des

applications ont été faite dans l'astronomie, dans la vision par ordinateur ou bien en marketing.

4.2 Déroulement de l'algorithme

L'algorithme classique, parfois appelé algorithme de Lloyd-Max, est de complexité NP (exponentielle). Il existe d'autres algorithmes beaucoup plus rapides (complexité Polynomial). Celui-ci se décompose en 3 grandes étapes :

1. **L'initialisation** : Il y a plusieurs méthodes d'assignation faisant varier la qualité des résultats. La plus commune place les clusters de manière aléatoire.
2. **L'assignation** : on assigne chaque point à son cluster le plus proche.
3. **Le calcul des points moyens** : On fait le barycentre de tous les points d'un cluster pour ensuite bouger ce même cluster vers ce barycentre.

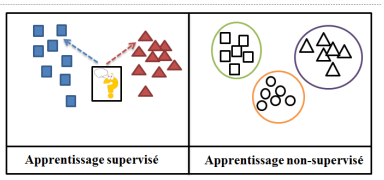


FIGURE 1 – Source : Wikipedia[3]

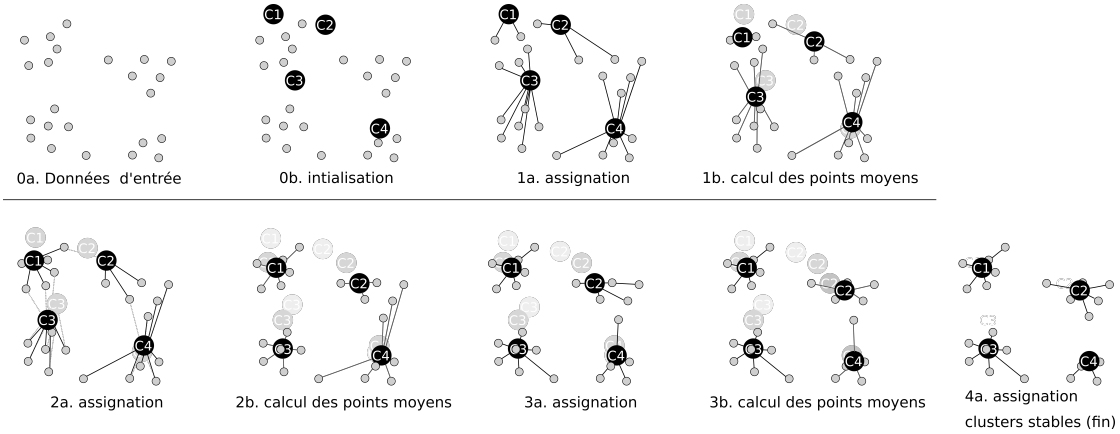


FIGURE 2 – Déroulement de l'algorithme K-Means (source : Wikipedia [3])

4.3 Démarche

Nous avons choisi d'utiliser K-Means dans le but de **vérifier la fiabilité** de la variable *personality_type*, mais également la fiabilité du questionnaire. Le questionnaire n'étant pas labellé originalement, il n'était pas possible d'utiliser un algorithme d'apprentissage supervisé (ou semi-supervisé) nécessitant des éléments déjà classés. Il fallait donc regrouper les participants en **5 classes** grâce à l'algorithme et d'observer si on retrouvait les même résultats qu'avec *personality_type*. L'algorithme k-means nous a semblé être le plus simple et le plus direct parmi les algorithmes d'apprentissages non-supervisés.

Pour lancer une analyse k-means, nous avons choisis d'utiliser la **bibliothèque python open source Scikit-learn**. Dans un premier temps, nous avons représenté chaque participant sous la forme d'un point de dimension 50 en utilisant ses 50 réponses. Puis dans un second temps, nous avons utilisé les 5 moyennes par type de question pour représenter un participant.

4.4 Résultats

En comptant le nombre de point correspondant à chaque type de personnalité puis en prenant le maximum de chaque cluster (les autres chiffres peuvent être considérés comme des erreurs, des points en marge de plusieurs clusters), on obtient ces résultats :

— En utilisant les **50 questions**, on tombe en général sur 2 clusters représentant le type N et 3 représentant le type O. Mais on peut observer un cluster assez hétérogène entre le type O et le type A.

— On retrouve plus de diversité en utilisant les **5 moyennes**, avec 1 cluster de représentant le type A, 2 représentant le type O et 2 représentant le type N. Et un cluster assez hétérogène pouvant représenter le type A, C et O.

Les tables 1 et 2 montrent des exemples de résultats, en effet les clusters changent de nombre à chaque exécution de l'algorithme. Les chiffres retrouvés sont à peu de chose près les **mêmes à chaque fois**. Dans ces deux tables, les maximums sont en rouge et les valeurs remarquables en orange.

TABLE 1 – Exemple de Résultats avec le vecteur de dimension 50

Clusters	0	1	2	3	4
A	711	265	997	313	1002
C	521	303	593	411	727
E	433	156	725	482	894
N	1965	3007	34	234	460
O	1020	446	1601	1241	1178

TABLE 2 – Exemple de Résultat avec le vecteur de dimension 5

Clusters	0	1	2	3	4
A	204	791	201	1017	1075
C	102	564	282	1184	423
E	105	734	159	704	988
N	3457	0	1831	114	298
O	305	1791	312	2114	964

Pour pousser un peu plus l'analyse des résultats, nous avons observé ces chiffres sous **forme de pourcentage**. Après avoir transformer les résultats initiaux (en rouge dans les tables 2 et 1) sous forme d'un pourcentage, nous avons classé et renommé les clusters dans l'ordre croissant. Il a suffit de les faire coïncider ensuite avec les taux de répartitions des personnalités dans toute l'étude en gardant un ordre croissant (variable *personalities* dans la table 3).

Les deux k-means restent assez proches dans leurs taux, ils trouvent pratiquement les **mêmes pourcentages** pour les points de type O et N. Mais en comparaison avec les taux initiaux, on constate une large sous représentation du profil O dans les k-means au profit du type N. Les 3 autres profils (C,E,A) semblent correspondre.

4.5 Discutions

En apparence, les tables 2 et 1 nous montre des résultats totalement différents que ceux que l'on trouve avec la variable *Personality_ type*. Certains cluster sont sur-représentés comme **O et N** alors que d'autre ne le sont pas du tout. Mais cet effet doit venir du fait qu'il y ait beaucoup

TABLE 3 – Les résultats sous forme de taux

	C	E	A	O	N
kmean5	10.416061	17.379498	18.038987	20.861216	33.304238
kmean50	12.875014	13.838149	18.067087	22.141038	33.078711
personalites	12.957047	13.641665	16.674274	27.820883	28.906131

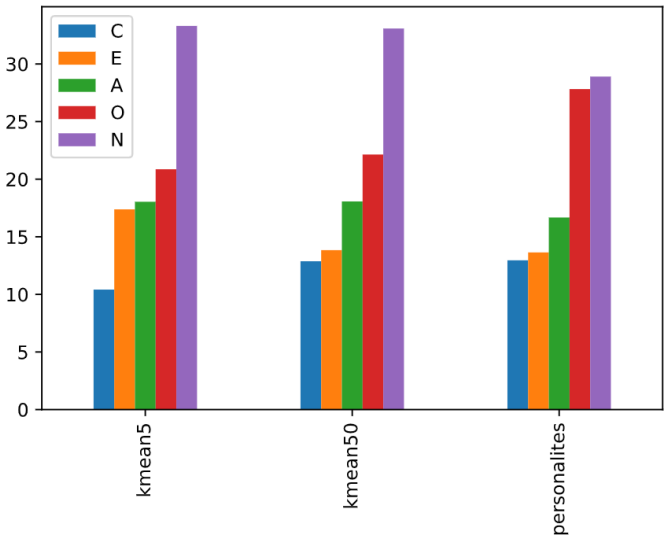


FIGURE 3 – Répartitions initiale des personnalités dans l'étude, avec les résultats des deux k-means

plus de profil O et N dans l'étude et donc naturellement plus de points O et N. L'observation sous la forme d'une "répartition croissante des taux" permet d'éviter ce problème.

Et en effet, la Figure 3 permet de retrouver des résultats qui coïncident beaucoup plus avec la variable *Personality_ type*, notamment avec le k-means de dimension 50. Celui-ci retrouve quasiment au dixième près les **mêmes proportions** des profils C et E. Le profil A obtient également, à quelques unités près, les mêmes résultats qu'avec la catégorisation manuelle des questions. Mais dans cette analyse, le k-means 50 a beaucoup plus de mal à discerner les profils O et N qui pourtant sont à l'opposé d'un point psychologique. Nous y voyons qu'une explications : un certain nombre de personne se trouve **à la bordure entre ouverture et neuroticisme**, et ce n'est pas traduit par *Personality_ type*. Une analyse statistique plus poussée sur ces deux profils pourrait éclairer la question mais cela reste peu probable les profils étant opposés.
La méthode k-means permettrait donc de retrouver certain résultat obtenue avec *Personality_ type*, validant donc la pertinence de cette variable ainsi que des catégories de question faite par les psychologue. Mais certains paramètres restent à étudier comme la sur-représentation des profils N.

5 Diverses comparaisons

On a vu que pour les États-unis comme pour le monde, la proportion de type de personnalité A, C et E sont sensiblement presque identiques avec 1% d'écart entre les personnalités A et E et 3% entre la personnalité C. Ce n'est que plus pour la personnalité O avec 4% d'écart, plus précisément 30% aux États-unis contre 26% dans le monde. Quand à la personnalité N, nous avons 25% aux États-unis contre 32% dans le monde.

Pour les **hommes**, on a remarqué qu'aux États-unis comme pour le monde, la proportion de type de personnalité A, C et E sont sensiblement presque identiques, avec entre 1% et 2% d'écart. Quand à la personnalité N, nous avons un écart un peu plus important avec 7%. Et nous remar-

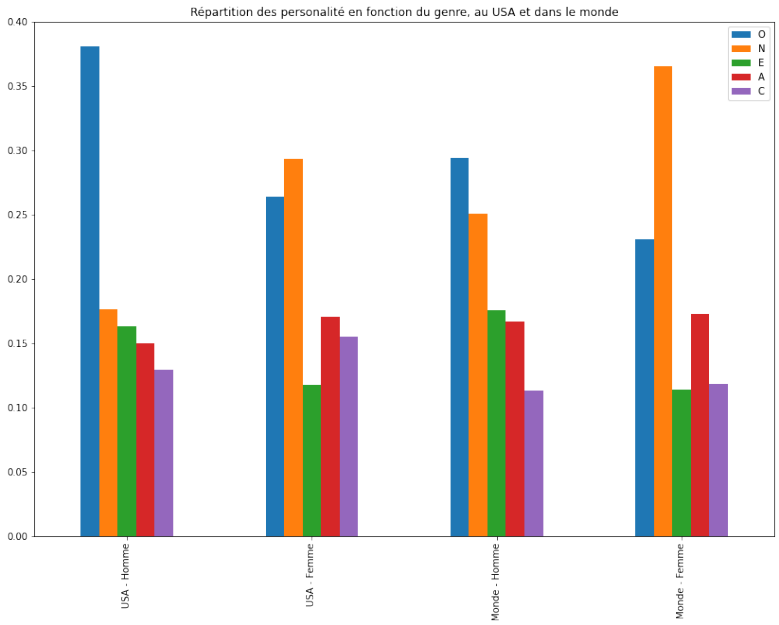


FIGURE 4 – Répartition des personnalités entre les hommes et les femmes aux États-unis et dans le monde.

quons pour la personnalité O une plus grande différence avec 9%. Pour les **femmes**, dans la figure 4, on a vu qu'aux États-unis comme pour le monde, la proportion de type de personnalité A, E et O sont très proches. Nous avons le même taux de personnalité A aux États-unis et dans le monde, 1% d'écart de personnalité E et pour la personnalité 0, nous avons 3% d'écart. Avec la personnalité C, nous avons 4%. Quand à la personnalité N, nous avons un écart légèrement plus important de 8%.

Nous avons ensuite observé **la répartition des personnalité en fonction de l'âge** dans la figure 6. On peut remarquer de **nettes variations inversées** sur les deux profils opposés O et N en fonction de l'âge. En effet, il semblerait que plus l'âge augmente plus il y aurait de profil O, et de moins en moins de profil N. Les variations des autres profils reste négligeable. On retrouve cette effet au **Etat-Unis et dans le monde**.

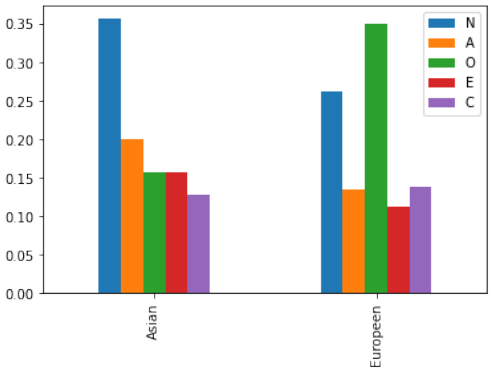


FIGURE 5 – Comparaison Asie/Europe en fonction des personnalités.

En comparant **les participants d'origine Européenne à ceux d'origine Asiatique**, on remarque qu'il y a 1% en écart de personnalité C. On a 5% écart pour la personnalité E et 6% pour la personnalité A. Un écart un peu plus important est présent pour la personnalité N avec 10%. Mais l'écart le plus important et remarquable est celui de la personnalité O avec 19%. On retrouve cette différence sur la Figure 7a et 5.

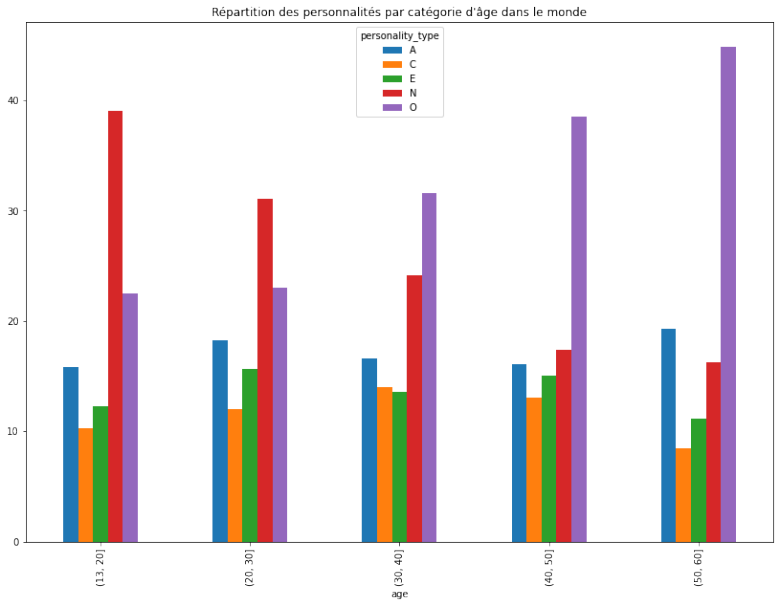
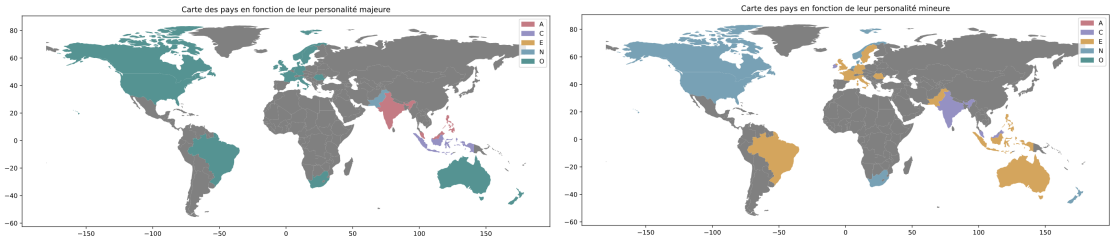


FIGURE 6 – Répartition des personnalités en fonction de l'âge dans le monde.



(a) Personnalité majeure

(b) Personnalité mineure

FIGURE 7 – Carte des personnalités

6 Analyse des résultats

On remarque que chez tous les participants, l'agréabilité, l'extraversion et la conscienciosité sont, à peu de chose près, a même proportions, et que l'ouverture et la négativité sont plus présents. Chez les participants aux États-unis, comme chez les participants dans le monde, le neuroticisme et l'ouverture sont presque présents en même quantité. On observe que l'ouverture domine au Etats-Unis contrairement au reste du monde ou le neuroticisme l'emporte. Le monde serait légèrement plus négatif par rapport aux États-unis, et donc que les États-unis sont légèrement plus ouvertes que le monde. La figure 7b illustre cela et montre également que cette ouverture existe dans les pays Scandinave confirmant par ailleurs leur réputation.

On peut dire que les **hommes dans le monde** sont, à peu de chose près, aussi ouverts que négatifs (25% de N et 29% de O). Cependant, en plus faible quantité, ils sont aussi extravertis qu'agréables. Quand aux **hommes aux États-unis**, avec un écart entre 1 et 5%, on peut dire qu'il y a un équilibre entre tous les profils sauf pour l'ouverture qui profite d'une grande

dominance. On observe donc en général peu de différence entre les hommes des États-unis et ceux du monde.

On peut dire que les **femmes dans le monde** sont aussi consciencieuses qu'extraverties (1% d'écart). Cependant, en plus grande quantité elles sont, à peu de chose près, aussi ouvertes qu'agréables (23% de O et 17% de A). On remarque que la personnalité dominante est N, assez loin devant les autres personnalités pour dire que les femmes dans le monde sont en générale plus négatives. On peut dire que les **femmes aux États-unis** sont, à peu de chose près, aussi ouvertes que négatives (26% de O et 29% de N). Cependant, en plus faible quantité, elles sont aussi consciencieuses qu'agréables (1% d'écart), elles sont aussi légèrement moins extraverties. On observe donc en général peu de différence entre les femmes des États-unis et celles du monde, les différences principales sont un équilibre entre le négatif et l'ouverture aux États-unis et une dominance flagrante de négativité dans le monde.

Avec un taux identique pour l'agréabilité 3% d'écart pour l'ouverture, pour la négativité 4% et pour l'extraversion et la consciencionsité 5%, on peut remarquer que les **femmes aux États-unis** ont une répartition de personnalité étrangement proche de celle des **hommes dans le monde**.

Quand à la **répartition des personnalités en fonction de l'âge**, le déclin clair du neuroticisme au profit de ouverture dans le monde et aux USA pourrait s'expliquer par **la sagesse** acquise à travers l'âge.

La différence de répartition des personnalités entre les continents asiatique et européen se retrouve également dans d'autre étude [1], ce qui valide nos résultats.

7 Conclusion

Même si les différences ne sont pas significatives, on peut observer que les États-Unis ont une plus grande ouverture en comparaison avec le monde entier, contrant l'image de l'américain nerveux véhiculé par les médias, pour mettre en avant celle de l'américain accueillant et ouvert. On remarque cependant que l'agréabilité est en taux constant pour toutes les catégories. Les plus grandes variables de changements dans les personnalités sont l'ouverture et la négativité, notamment entre les hommes et les femmes. Un grand facteur de modification de la personnalité est l'âge, négligeant le neuroticisme au profit d'une ouverture.

Les **États-unis** sont en assez grandes dominances, on a pu le remarquer surtout dans le profil-type ou dans la répartition des pays dans le test. Cela peut créer un déséquilibre en sous-représentant les autres pays et fausser certains de nos résultats, la participation dans les autres pays étant trop pauvre pour en tirer des conclusions statistiquement correctes. L'écart global entre les pourcentages reste néanmoins faible, ce qui prouve que le monde entier est assez équilibré en terme de personnalités. Les peu de différences observées pouvant être justifiées par la culture, la religion, l'éducation ou encore plein d'autres facteurs.

Références

[1] Leong FT Cheung FM, van de Vijver FJ. Toward a new approach to the study of personality in culture. *The American Psychologist*, 66(7) :593–603, 2011.

[2] Wikipedia.org. Big five personality traits. https ://en.wikipedia.org/wiki/Big_Five_personality_traits.

[3] Wikipedia.org. K-means. https ://en.wikipedia.org/wiki/K-means_clustering.