Analyse multidimensionnelle et clustering portant sur des données concernant les Pokémons

Clément Malvy, Rayan Moissonnier, et Noé Lebreton
{clement.malvy,rayan.moissonnier,n.lebreton}@univ-lyon2.fr
Encadrant : Sabine Loudcher

1 Contexte et problématique d'étude

Pokémon est une franchise japonaise créée en 1996 par Satoshi Tajiri, présente tout particulièrement sous forme de jeu vidéo, édité par Nintendo. Le premier jeu Pokémon Rouge et Bleu, sorti en 1998 marque une entrée fracassante de cette licence qui reste, de nos jours, iconique du monde du jeu vidéo. Depuis, plus de 25 jeux issus de la série principale (il existe de nombreux dérivés) ont vu le jour, avec de plus en plus de nouvelles espèces.

En outre, il existe une grande communauté de joueurs stratégiques de ces jeux, où de nouvelles règles en plus de celles imposées par le jeu ont été créées dans le but d'équilibrer les combats et de les diversifier. En effet, lorsqu'il n'y avait que 151 Pokémon dans la première génération, les combats étaient très souvent similaires puisqu'une dizaine de Pokémon étaient stratégiquement viables.

Le monde Pokémon ne se résume donc pas seulement à des combats de monstres où les plus grands dominent. Chacune de ces centaines de créatures différentes possède des caractéristiques uniques. Le but de ce travail est de contribuer au "theorycrafting" du jeu Pokémon, ce terme se rapportant à l'analyse mathématique des règles ou mécaniques d'un jeu, souvent utilisé dans les jeux vidéo afin de découvrir des stratégies et tactiques optimales.

Lors ce projet, nous allons analyser les données des Pokémon afin d'étudier les interactions entre leurs caractéristiques. Puis, dans un deuxième temps, nous allons créer des groupes de créatures qui ont des caractéristiques proches afin de créer une tier liste (un regroupement des Pokémon en fonction de leur caractéristique).

Plus précisément, nous allons d'abord réaliser une analyse en composante principale pour mieux comprendre les données, c'est-à-dire étudier les interactions entre les caractéristiques mais aussi entre les Pokémon. Dans un deuxième temps, nous utiliserons une analyse factorielle des correspondances afin de comprendre l'interaction entre les variables **Attack** et **Weight_kg** (qui seront discréditées). Enfin, nous créerons des groupes de Pokémon grâce à des méthodes de clustering telles que la classification ascendante hiérarchique ou encore les k-means.

Nous utiliserons alors des outils d'analyse factorielle et de clustering, mais avant cela, nous devons décrire les variables de notre jeu de données.

2 Description des variables

Les Pokémon possèdent plusieurs types de caractéristiques, notre jeu de données comporte donc plusieurs types de variables. Tout d'abord il y a celles qui influent sur la puissance d'un Pokémon en combat, et celles qui n'ont aucune influence sur le combat, mais plutôt sur son apparence ou sa rareté.

Nous allons par la suite expliquer chacune des variables que nous conserverons pour notre analyse :

- **Name** : Variable qualitative / nom du Pokémon.
- **Hp** : Variable quantitative / points de vie du Pokémon.
- **Speed** : Variable quantitative / vitesse du Pokémon.
- **Attack** : Variable quantitative / attaque du Pokémon.
- **Sp_attack** : Variable quantitative / attaque spéciale du Pokémon.
- **Defense** : Variable quantitative / défense du Pokémon.
- **Sp_defense** : Variable quantitative / défense spéciale du Pokémon.
- **Against_X** : Variables quantitatives / ces variables donnent les dégâts subis par une attaque du type X. Par exemple, si against-bug = 2, le Pokémon subit le double de dégâts des attaques de type insecte. Une table des types figure sera en annexe pour mieux comprendre le fonctionnement.
- **Base_egg_steps** : Variable quantitative / nombre de pas nécessaire pour faire éclore les œufs de ce Pokémon.
- **Base_happiness** : Variable quantitative / bonheur du Pokémon lors de sa capture ou de sa naissance.
- **Experience_growth** : Variable quantitative / quantité d'expérience nécessaire pour atteindre le niveau maximal.
- **Generation** : Variable qualitative / génération à laquelle le Pokémon est apparu.
- **Is_legendary** : Variable booléenne / indique si le Pokémon est légendaire ou non.
- **Height_m** : Variable quantitative / taille du Pokémon en mètres.
- **Weight_kg** : Variable quantitative / poids du Pokémon en kilogrammes.
- **Pourcentage_male** : Variable quantitative / taux d'obtention d'un Pokémon mâle de cet espèce.
- **Type1 et Type2** : Variables qualitatives / il s'agit des types des Pokémon, certains n'en possèdent qu'un. (Il existe 18 types différents tel que Eau, Feu, Vol...).

3 Application des méthodes sur notre jeu de données

3.1 Prétraitement des données

Après une première exploration des données nous avons fait le constat qu'il manquait des valeurs sur 20 Pokémon concernant les variables **Weight_kg** et **Height_m**. Nous avons pris la décision de compléter ces informations manquantes directement dans le document csv. La source Poképédia qui référence toutes les spécificités des Pokémons nous a permis de remplir les informations manquantes.

Suite à cette première étape, nous avons également remarqué la présence de 98 valeurs manquantes sur les 801 individus pour la variable **Percentage_male**. Cela est dû au fait que certain Pokémon soit asexué ou seulement femelle (codé en na sur les données), nous avons fait le choix de les recoder en 0. Nous pouvons simplement justifier ce choix, celui-ci est dû au fait qu'ils n'aient pas de pourcentage de mâle.

Afin de ne pas conserver des données inutiles à nos analyses nous avons supprimé les variables que nous n'utilisons pas :

- **Abilities** : gestion trop difficile car un Pokémon peut en avoir plusieurs et il en existe des centaines.
- **Japanese_name** : nom en japonais, il n'apporte rien à l'analyse.
- **Pokedex_number** : il s'agit d'une variable qui identifie un Pokémon mais son nom suffit.
- **Capture_rate** : il varie d'une génération à une autre pour les mêmes Pokémon, il est donc inutilisable.

Il y a deux variables **typex**, x ayant pour valeurs 1 et 2, le champ **type2** n'est pas toujours renseigné car les Pokémon n'ont pas systématiquement deux "caractéristiques élémentaires". Dans le but de nous servir du type en tant que variable supplémentaire lors d'une ACP, nous avons compilé ces 2 types en une seule variable du format **type1_type2** . Afin de minimiser, le nombre de modalités différentes, nous avons fait en sorte de ne pas accorder d'importance à l'ordre des types. Cela nous permet d'avoir 142 modalités différentes. Si nous avions accordé une importance à l'ordre il y aurait pu avoir jusqu'à 166 modalités différentes. Grâce à cette technique nous avons pu réduire le nombre de modalités de 24.

3.2 Analyses factorielles

Dans cette partie nous allons développer 2 différents types d'analyses factorielles pour pouvoir analyser au maximum les relations entre les individus et les variables de notre jeu de données.

3.2.1 Analyse en Composantes Principales

La première méthode choisie pour observer les relations présentes dans nos données est l'ACP. Ce choix provient d'un fait simple, étant donné que nos variables sont majoritairement de type quantitatif (environ 90% des variables) il était donc logique de choisir une méthode prenant en entrée des variables de ce type.

Avant toutes choses, nous nous sommes posé plusieurs types de questions comme:

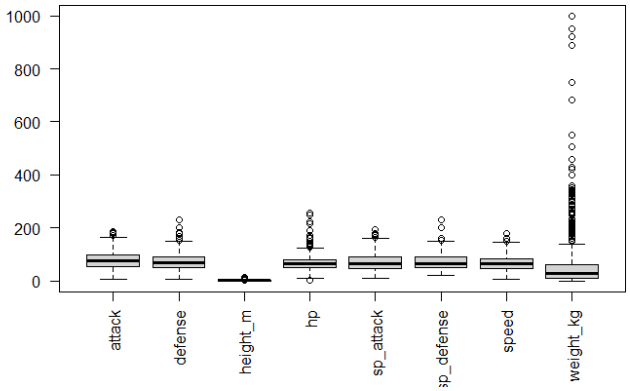
- Est-ce que certaines variables sont liées ?
- Quels sont les Pokémon qui se ressemblent ?
- Au contraire, quels sont les Pokémons qui sont différents ?
- Sur quelles informations pouvons-nous expliquer ces ressemblances/différences ?

Pour répondre à ces différentes questions, nous allons mettre en place une ACP sur certaines variables de notre jeu de données. Comme nous vous l'avons expliqué précédemment, le data frame est composé de 30 variables quantitatives ce qui veut dire, en d'autres termes, 30 potentiels variables pour réaliser cette méthode. Dans le cadre de cette analyse, nous avions pour but d'étudier les Pokémon notamment au niveau de leur puissance. Pour cela, nous nous sommes rendu compte que certaines variables disponibles n'étaient pas indispensables. Premièrement, la totalité des variables **Against_x** (avec x qui correspond au type du Pokémon) n'apporte pas d'informations pertinentes. De plus, les variables **base_egg_steps**, **base_happiness**, **experience_growth** et **percentage_male** sont également retirées pour cette analyse. Ces dernières n'ajoutent aucune plus-value dans notre objectif de comparer les caractéristiques de combat des Pokémon.

Pour réaliser cette Analyse en Composantes Principales nous avons donc fait le choix d'utiliser uniquement 11 variables (8 quantitatives et 3 qualitatives en tant que variables supplémentaires).

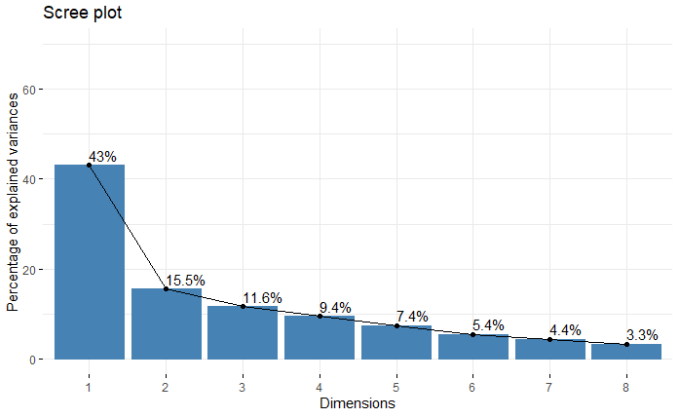
Après cette étape du choix des variables, il était nécessaire de choisir le type d'ACP (normée ou non). Pour cela nous avons étudié l'ordre de grandeur de nos variables, comme vous pouvez le voir sur le graphique ci-dessous nous avons certaines variables prenant des valeurs dites "extrêmes". Si nous ne prenons pas en compte la différence d'échelle entre les variables nos résultats seront biaisés. C'est pourquoi nous avons décidé de standardiser les variables (centrage et réduction) en appliquant une ACP normée.

Figure 1: Graphique contenant des boxplots pour chacune des variables quantitatives



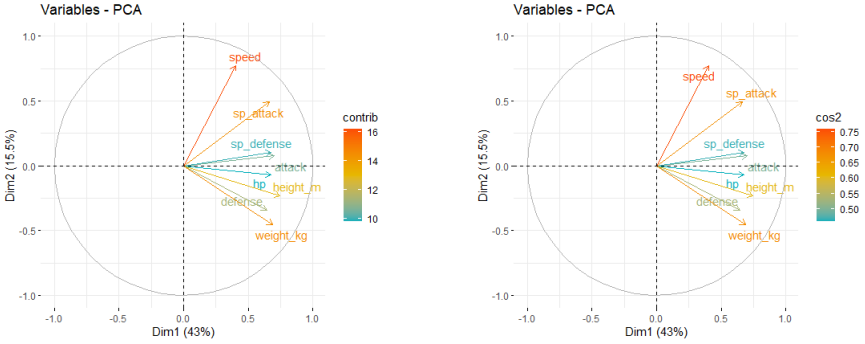
La première étape de l'interprétation des résultats de l'ACP était de choisir le nombre d'axes. Pour ce choix nous nous sommes basés sur le critère de Kaiser qui explique qu'il faut retenir uniquement les axes ayant une inertie supérieure à l'inertie moyenne. Étant donné que nous sommes dans le cadre d'une ACP normée, l'inertie moyenne vaut 1 et seulement les deux premiers axes ont une inertie supérieure à l'inertie moyenne. Comme vous pouvez le voir sur la figure 2, les deux premiers axes restituent plus de 58% de l'information ; c'est donc sur ces deux axes que nous allons nous focaliser pour l'interprétation.

Figure 2: Pourcentage d'inertie associée à chaque dimension de l'ACP



Suite au choix du nombre d'axes, nous avons décidé d'analyser les résultats concernant les variables grâce aux deux graphiques présentés ci-dessous. On peut tout d'abord observer que, sur l'axe 1, la totalité des variables est projetée du côté positif de l'axe. Nous pouvons constater que peu importe la mesure (contribution et cosinus carré) les trois variables les mieux représentées sont **height_m**, **attack** et **weight_kg**. Il est également important de noter que toutes les variables sauf **speed** permettent de décrire cet axe. En d'autres termes, cet axe permet de différencier les Pokémon les plus forts au niveau des statistiques du jeu.

Figure 3: Représentation des variables (à gauche selon la contribution et à droite selon le cosinus carré)



Concernant l'axe 2, nous pouvons constater que les variables sont projetées du côté positif mais également du côté négatif de celui-ci. Le côté positif de l'axe est majoritairement représenté par la variable **speed** qui contribue beaucoup à la création de cet axe et qui a un cosinus carré proche de 1. Pour parler du côté

négatif, il est principalement construit grâce à la variable **weight_kg**.

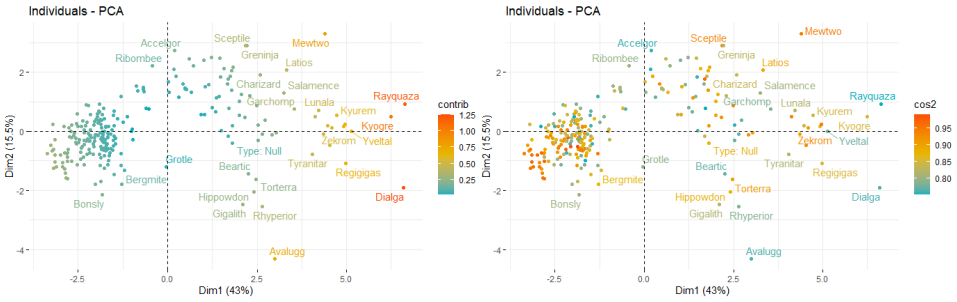
Pour conclure sur l'interprétation des variables sur les deux premiers axes, on constate que malgré le faible pourcentage d'inertie expliquée par l'axe 2 celui-ci oppose des variables de manière pertinente. Cet axe traduit donc le fait qu'un Pokémon lourd va être lent et inversement.

Après avoir analysé les résultats concernant les variables nous nous sommes orientés vers l'interprétation des individus sur les deux premiers axes factoriels. Sur la figure 4 nous pouvons observer la projection des individus, nous allons tout d'abord étudier l'axe 1. Le côté positif de cet axe a été construit (contribution) par plusieurs Pokémon comme Rayquaza, Dialga ou encore Kyogre et les individus les mieux représentés (cos2) sont Arceus et Xerneas. Ce sont des Pokémon légendaires pour la plupart qui ont pour particularité de posséder des caractéristiques de combats très élevées, relativement aux autres.

Pour parler du côté négatif, les Pokémon les mieux représentés sur cette partie de l'axe sont Ditto, Barboach et Pidgey. Ces trois Pokémon sont très faibles, ils possèdent les pires caractéristiques globales du jeu selon l'axe 1, ce qu'on nous pouvons confirmer d'après notre expérience de jeu.

Concernant l'axe 2 qui restitue 15,5% de l'information, les individus représentant le mieux le côté positif sont Raichu, Electabuzz et Galvantula (cos2) mais aussi Alakazam et Phermosa (contribution). Ces différents Pokémon sont réputés pour leur incroyable statistique en vitesse. Concernant le côté négatif de cet axe, ce sont Groltle et Rhyhorn qui sont les Pokémon les mieux représentés et Steelix, Cosmoem et Celesteela qui contribuent le plus à la construction de cet axe. Ceux-ci sont connus pour être des Pokémon très lourds, presque une tonne pour Cosmoem par exemple.

Figure 4: Représentation des individus (à gauche selon la contribution et à droite selon le cosinus carré)



Pour conclure sur cette ACP, nous avons pu voir que les deux premiers axes restituent 58% des données initiales. Le premier permet de différencier les Pokémon ayant des statistiques élevées de ceux qui ont des statistiques plus faibles. Le second, quant à lui, permet de repérer les Pokémon lourds des Pokémon rapides. Cela nous a aidé à mieux comprendre les interactions entre les variables et les individus.

3.2.2 Analyse Factorielle des Correspondances

Le deuxième objectif de cette partie était d'étudier les correspondances entre les modalités en colonnes et en lignes. Pour répondre à cet objectif, nous avons fait le choix de mettre en place une Analyse Factorielle des Correspondances. Dans ce but, nous avons besoin de 2 variables qualitatives, nous avons donc effectué plusieurs tests sur différentes variables à seulement 2 modalités, le fait que le Pokémon soit légendaire ou non, mais cela n'a pas fonctionné dû au degré de liberté qui serait de 1. Pour pallier ce problème nous avons donc décidé de créer des catégories d'individus en créant manuellement des classes avec une variable quantitative.

La première variable a été créée à partir de la variable quantitative **weight_kg**, nous l'avons donc discréditée en 3 intervalles recodés avec les étiquettes suivantes :

- léger ((0,55])
- normal ((55,100])
- lourd ((100,1000])

La deuxième variable provient de la variable **attack** qui a été découpée en 5 intervalles recodés avec les labels :

- très faible ((5,55])
- faible ((55,70])
- moyen ((70,90])
- fort ((90,120])
- très fort ((120,185])

Avec ces deux nouvelles variables qualitatives, nous allons pouvoir savoir si un Pokémon lourd est également un Pokémon avec une attaque forte ou inversement si un Pokémon lourd est un Pokémon avec une attaque faible.

Lors de la mise en place d'une AFC, il est important de vérifier certains points comme le fait d'avoir aucun croisement entre les modalités des deux variables avec un effectif inférieur à 5. Pour cela nous avons créé le tableau de contingence entre ces deux variables, celui-ci ne comporte aucun croisement avec un effectif inférieur à 5.

Pour qu'une AFC soit pertinente il est nécessaire que les deux variables étudiées soient dépendantes car l'objectif est de décrire les liens entre celles-ci. Si les variables sont indépendantes, il n'y aura aucun effet multivarié donc aucune pertinence dans la mise en place de cette méthode. Pour vérifier cela nous avons effectué un test du χ^2 pour obtenir une réponse concernant la dépendance des variables. La p-valeur obtenue est inférieure au seuil alpha fixé à 5%, en d'autres termes nous rejetons H0 l'hypothèse d'indépendance des variables.

Avant de faire l'AFC il est intéressant d'étudier les profils lignes et colonnes pour avoir un avis a priori sur les répartitions des individus entre les modalités des deux variables.

Figure 5: Tableaux des profils lignes et des profils colonnes

	léger	normal	lourd	Total
très faible	89.4	7.5	3.1	100.0
faible	87.2	8.1	4.7	100.0
moyen	75.0	12.5	12.5	100.0
fort	45.2	28.9	25.9	100.0
très fort	35.8	13.6	50.6	100.0
Ensemble	71.1	13.8	15.1	100.0

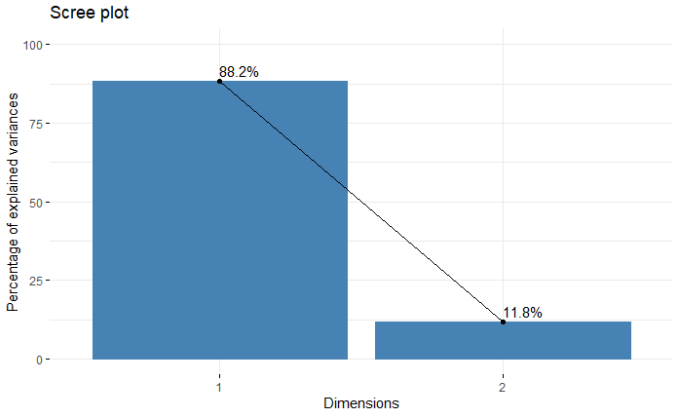
	léger	normal	lourd	Ensemble
très faible	35.6	15.5	5.8	28.4
faible	22.8	10.9	5.8	18.6
moyen	23.3	20.0	18.3	22.1
fort	13.2	43.6	35.8	20.8
très fort	5.1	10.0	34.2	10.2
Total	100.0	100.0	100.0	100.0

Concernant les profils lignes, on peut tout d'abord constater que la classe "léger" de la variable **weight_kg** modifiée est la classe la plus représentée. Nous pouvons également voir qu'il y a une certaine cohérence entre la combinaison des modalités des deux variables, par exemple 89,4% des Pokémon "très faible" sont également "léger". Au contraire, plus de la moitié (50,6%) des Pokémon "très fort" sont des Pokémon "lourd".

Pour maintenant parler des profils colonnes, étant donné que ces profils correspondent aux fréquences relatives des catégories de la variable **weigh_kg** modifiée dans chaque groupe de la variable **attack** transformée les différences sont moins élevées car le pourcentage est calculé sur 5 modalités (contre 3 précédemment). Globalement, on constate que les Pokémon "très faible" sont les plus présents (28,4%) dans le jeu. De plus, d'une manière générale, les Pokémon "léger" sont rattachés à une classe inférieure ou égale à "moyen" pour la variable **attack** modifiée. Nous pouvons observer exactement l'inverse de cette relation avec la classe "lourd" de la variable **weight_kg** transformée.

L'étape suivante concerne le choix du nombre d'axes, pour pouvoir faire ce choix nous avons analysé l'inertie pour chaque axe. Le graphique présenté ci-dessous permet donc de retenir uniquement le premier axe, c'est donc celui-ci que nous allons interpréter dans la suite de cette partie.

Figure 6: Pourcentage d'inertie associée à chaque dimension de l'AFC



Pour parler de l'interprétation sémantique de l'axe 1 avec les profils lignes (figure 7), le profil "très fort" ainsi que le profil "fort" caractérisent le côté positif de l'axe 1 tandis que celui des "très faible" mais aussi des "faible" caractérisent le côté négatif du premier axe. En d'autres termes, le profil des "très fort" s'oppose au profil des "très faible". On peut donc conclure que la distribution du poids des Pokémon très forts est différente que celle des Pokémon très faibles.

Figure 7: Tableau représentant les résultats des profils lignes selon la contribution

	coord	contrib	cos2
très fort	0.92432082	40.1188484	0.8607733
fort	0.52010981	26.1893386	0.7981587
très faible	-0.40429523	21.5100323	0.9921663
faible	-0.36318554	11.3171426	0.9918741
moyen	-0.09205595	0.8646381	0.9988575

Nous pouvons maintenant s'intéresser des profils colonnes (figure 8) sur l'axe 1, le profil des "lourd" définit le côté positif de l'axe 1 par opposition au profil des "léger" qui caractérise le côté négatif de l'axe 1. Ces différentes interprétations veulent donc dire que les Pokémon lourds n'ont pas la même puissance d'attaque que les Pokémon légers.

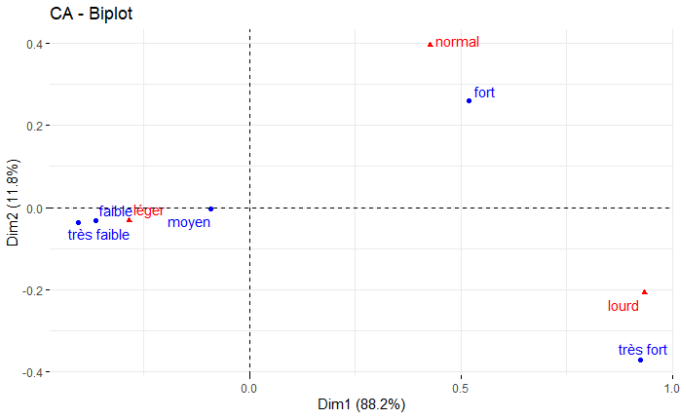
Figure 8: Tableau représentant les résultats des profils colonnes selon la contribution

	coord	contrib	cos2
lourd	0.9346204	61.78006	0.9526632
léger	-0.2841374	26.53736	0.9875614
normal	0.4280200	11.68258	0.5395111

Étant donné que l'axe 1 restitue plus de 88% de l'information, nous avons fait le choix de ne pas analyser le deuxième axe.

Suite à l'interprétation divisée en fonction des profils, nous allons faire une analyse simultanée des profils lignes et profils colonnes. On peut observer plusieurs phénomènes sur la figure 9, comme le fait que les modalités des deux variables sont représentées de manière logique. Pour revenir aux questions que nous nous sommes posées précédemment, un Pokémon ayant une attaque très forte est également un Pokémon lourd et pas inversement. De plus, on constate qu'un Pokémon léger a une attaque soit faible soit très faible.

Figure 9: Graphique représentant simultanément les profils lignes (bleu) et les profils colonnes (rouge)



Pour conclure sur cette AFC, celle-ci nous a permis de nous conforter dans nos idées de départ qui était que les modalités de ces deux variables étaient corrélées entre elles de manière cohérente, d'après la logique du jeu.

3.3 Classification non supervisée

Dans cette partie nous allons développer deux différentes méthodes de classification afin de créer des groupes de Pokémon qui se ressemblent d'un point de vue de leurs caractéristiques de combat, c'est-à-dire, leur attaque, défense, points de vie, attaque spéciale, défense spéciale, vitesse, taille et poids. Nous le rappelons, le but de ces groupes est de créer des groupes équilibrés de Pokémon pour les tournois. Il ne faudrait pas que votre Pikachu se fasse terrasser par un Leviator !

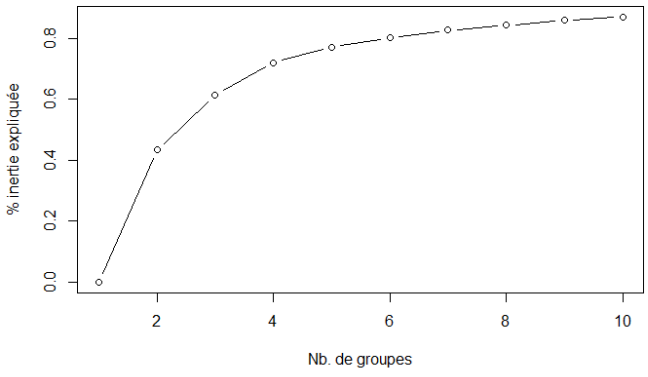
Les variables sur lesquelles nous allons effectuer nos méthodes de clustering ne sont pas à la même échelle, par exemple certaines sont en kg, en points ou en d'autres mesures. Nous avons le choix de centrer et réduire les données, la standardisation de celle-ci ramène la moyenne de tous les individus de chaque variable à 0. Il est important de préciser qu'il est possible d'obtenir des moyennes négatives par classe.

3.3.1 Classification Ascendante Hiérarchique

Nous avons d'abord choisi cette première méthode très classique parmi celles permettant le clustering. En effet, elle possède des avantages tels que la non-nécessité d'un nombre de classe préétablie mais aussi la rapidité par rapport à la CDH (Classification Descendante Hiérarchique).

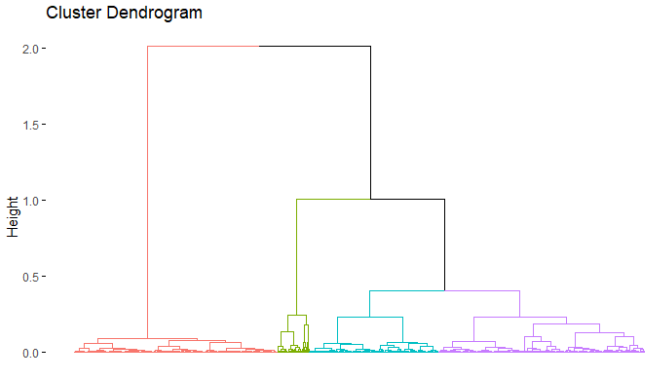
Grâce au graphique de restitution de l'inertie en fonction du nombre de classes, ci-dessous, nous avons choisi de conserver la partition qui sépare en quatre groupes les individus.

Figure 10: Graphique d'inertie restituée en fonction du nombre de groupes



Nous pouvons constater qu'à partir de quatre classes, nous ne gagnons plus assez d'informations, en d'autres termes, le pourcentage d'inertie expliquée n'augmente plus assez. Nous allons donc conserver la partition qui sépare les Pokémon en 4 groupes.

Figure 11: Dendrogramme des Pokémon



Ci-dessus se trouve le dendrogramme. En raison du nombre d'individus, il est difficile à exploiter mais nous pouvons quand même constater qu'il y a un groupe (vert) de faible effectif comparé aux trois autres. Dans un deuxième temps, nous allons donc rentrer dans les détails de ces groupes afin de savoir comment ils ont été construits.

Dans un premier temps, nous allons donc expliquer les groupes avec les différentes variables du jeu de données. Ainsi les individus du groupe 1 possèdent des caractéristiques faibles globalement, toutes négatives (vu qu'elles ont été centrées et réduites). Quant au deuxième groupe, les Pokémon ont des caractéristiques moyennes (proches de 0) sauf la vitesse qui est très élevée. Dans le troisième groupe, toutes les caractéristiques sont moyennes, proches de 0. Et enfin, dans le dernier groupe, les individus ont des caractéristiques globalement élevées.

Les classes peuvent être illustrées par des individus particuliers. A ce titre, deux types d'individus particuliers sont proposés :

- Les parangons, les individus les plus proches du centre de la classe.
- Les individus caractéristiques c'est-à-dire les individus les plus éloignés des centres des autres classes.

Le parangon principal du groupe 1 est Snorunt qui en effet est un Pokémon très faible en combat, il est également léger. L'individu caractéristique est Magicarpe.

Concernant le deuxième groupe, son principal parangon est Infernape, qui est un Pokémon réputé pour être très rapide et son individu caractéristique est Pheromosa qui possède une des plus grosses vitesses du jeu (mais des autres statistiques moyennes).

Pour le troisième groupe, Poliwrath est le Pokémon le plus proche du centre de classe, et Shuckle le Pokémon représentant le mieux ce groupe, il s'agit tous deux de Pokémon moyens en termes de statistiques.

Quant au quatrième groupe, il regroupe les Pokémon les plus forts, ayant des caractéristiques globales élevées mais aussi ceux qui ont un poids très fort. Ce poids fort compense les autres statistiques de combat du Pokémon. Le parangon principal de ce groupe est Regigigas qui est un des Pokémon qui a les plus grosses statistiques de combat du jeu. Son individu caractéristique est Guzzlord, qui est un Pokémon légendaire redoutable.

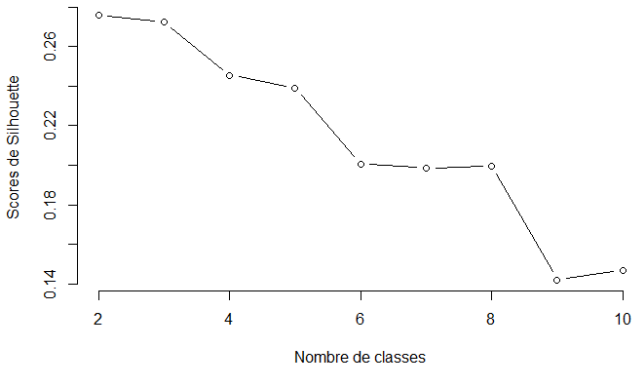
Les résultats de cette CAH sont très pertinents. Les Pokémon sont regroupés en 4 groupes en fonction de leur caractéristique, et cette partition est très satisfaisante. En effet nos connaissances de jeu nous confirment que ces groupes sont équilibrés et pourraient faire office d'un bassin (choix parmi un groupe) de Pokémon pour un tournoi.

3.3.2 Méthode des K-means

Nous avons fait le choix de choisir une seconde méthode afin de faire un comparatif avec la CAH. La méthode des k-means est une stratégie très répandue dans le domaine du data mining. Contrairement à la CAH, les temps de calcul des k-means sont réduits avec une faible complexité. Le seul inconvénient de cette méthode est de définir un nombre de classes a priori.

Pour déterminer le nombre optimal de classe, nous avons dû calculer les silhouettes pour chaque partition en faisant varier le nombre de classes.

Figure 12: Silhouette du nombre de classes avec la méthode des k-means



Dans l'optique de comparer cette méthode avec celle de la CAH nous avons fait le choix de garder 4 classes. De plus, d'après le critère de silhouette, le choix du nombre de groupes semble cohérent. Nous pouvons justifier ce choix car ce critère doit être maximisé, la silhouette de la partition faite avec 4 classes obtient une valeur correcte vis-à-vis de celles obtenues avec 2 et 3 groupes.

Suite au choix du nombre de classes, nous allons interpréter les résultats de chaque classe. Les effectifs des groupes sont de 326, 212, 227 et 36. Nous constatons donc que seul un groupe contient peu d'individus. Avec la figure 13, nous allons pouvoir décrire chaque classe grâce aux moyennes des variables standardisées. Pour être plus précis, les lignes du tableau ci-dessous représentent les centres de chaque classe.

Figure 13: Tableau des moyennes des variables standardisées des 4 classes

	attack	defense	height_m	hp	sp_attack	sp_defense	speed	weight_kg
1	-0.7035255	-0.66972576	-0.51579378	-0.6781761	-0.6796697	-0.7377226	-0.5356673	-0.4140984
2	0.4649935	-0.03830298	0.05211935	0.1369752	0.6576977	0.2572317	1.1800862	-0.1547880
3	0.3670257	0.78959666	0.20911181	0.6012762	0.2111248	0.6604376	-0.3946868	0.2125758
4	1.3182176	1.31145513	3.04530803	1.5432495	0.9504189	1.0012535	0.3900879	3.3210116

Le cluster 4 est celui avec le plus petit effectif (36 individus), celui-ci maximise les moyennes de toutes les variables sauf celle de la variable **speed**. En d'autres termes, cette classe contient les Pokémon les plus forts du jeu et c'est donc pour cela que son effectif est faible.

Les classes 1 et 3 (ayant pour effectif 326 et 227) regroupent respectivement les Pokémon très faibles et moyens. Pour parler plus précisément de la classe 2, toutes ses moyennes sont négatives ce qui traduit le fait que les Pokémon de cette classe sont très faibles. En ce qui concerne la classe 3, c'est celle qui rassemble les Pokémon normaux, ni trop forts ni trop faibles.

Enfin, la classe 2 (212 individus) représente les Pokémon ayant une vitesse rapide, cela dû au fait qu'il maximise la variable **speed**.

La deuxième étape au niveau de l'interprétation des classes est celle concernant les individus. Comme pour la CAH, il est intéressant d'étudier pour chaque classe les parangons ainsi que les individus caractéristiques. Etant donné que la fonction kmeans de Rstudio ne donne pas ce type d'information nous avons décidé de les calculer "manuellement". Pour cette partie nous avons fait le choix de calculer uniquement les parangons, c'est-à-dire les individus les plus proches du centre de la classe. Pour cela nous avons calculé, pour chaque individu, la distance au carré entre le Pokémon et le centre de sa classe.

Voici la liste des parangons obtenus avec notre méthode et cela pour chaque classe

- Classe 1 : Mime Jr.
- Classe 2 : Drapion
- Classe 3 : Exploud
- Classe 4 : Reshiram

Grâce à notre maîtrise du jeu mais aussi les résultats obtenus concernant les variables, nous pouvons affirmer que ces Pokémon sont bien représentatifs des différentes classes.

L'application de la méthode des k-means nous a permis de constituer 4 groupes de Pokémon, qui vis-à-vis de notre connaissance du jeu nous semblent cohérents. Pour vous donner un exemple, Reshiram est l'un des plus forts Pokémon du jeu, il est légendaire et il se retrouve dans la classe regroupant les Pokémon les plus forts du jeu.

3.3.3 Comparaison des 2 méthodes

Ces deux méthodes de classification nous ont donc permis d'obtenir deux partitions. Le but de cette partie est de comparer les résultats. Pour ce faire, nous allons déjà comparer l'inertie restituée en fonction du nombre de classes ainsi que la silhouette pour les deux méthodes. Puis dans un second temps, regarder si les individus ont tendance à être classés dans la même classe pour la CAH et le K-means.

Lors de nos deux regroupements, nous avons choisi de regrouper en quatre groupes les Pokémon. Ces choix ne sont pas le fruit du hasard mais bien effectués grâce aux graphiques de restitution d'inertie ou des silhouettes. En outre, cela nous permet de comparer plus facilement les résultats, nous voulons maintenant savoir si les deux méthodes classent les mêmes individus dans les mêmes groupes.

Nous avons donc obtenu un dataframe comportant le nom du Pokémon ainsi que les classes auxquelles il a été attribué pour chacune des méthodes. Le tableau de contingence nous indique les résultats suivants.

Figure 14: Répartition des individus en fonction des classes obtenues avec la CAH et les K-means

	K-means			
CAH	1	2	3	4
1	326	0	0	0
2	0	212	0	0
3	0	0	227	0
4	0	0	0	36

Nous constatons donc que les deux partitions obtenues sont les mêmes. En effet, chacune des classes est similaire. Nous sommes donc satisfaits de ces méthodes de classification, cela est rassurant lorsque nous trouvons les mêmes résultats avec deux méthodes différentes.

4 Conclusion

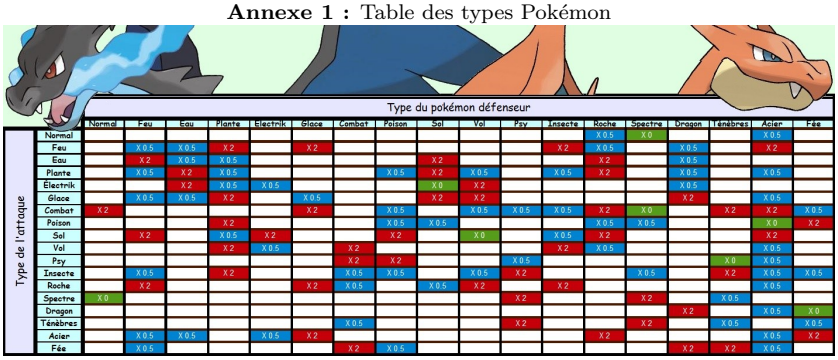
Les méthodes factorielles nous ont aidé à comprendre l'interaction entre individus et variables. Ainsi, l'ACP, où nous avons retenu deux axes, le premier pour identifier les Pokémon avec des caractéristiques élevées et le second qui opposait la vitesse et le poids, nous a permis d'analyser globalement ces interactions. Quant à l'AFC, elle nous a permis de confirmer notre hypothèse selon laquelle les Pokémon les plus lourds seraient aussi les plus forts. Ces deux variables sont belles et bien liées positivement.

Concernant les méthodes de clustering, nous avons obtenu quatre groupes qui permettaient de séparer les Pokémon faible, moyen, fort et ceux ayant une vitesse très élevée. Les deux méthodes donnaient les mêmes résultats ce qui nous conforte dans leur pertinence.

Les résultats de nos analyses nous ont permis de mieux comprendre les données, d'en faire sortir des informations telles que les liens entre les individus mais aussi des variables.

5 Annexe

Annexe 1 : Table des types Pokémon



Ci-dessus se trouve la table des types Pokémon avec leurs résistances aux attaques des 18 types. Par exemple une attaque Dragon fera le double de dégâts aux Pokémon de ce même type.