

# Analyse en Composantes Principales

---

Paul Chapron <sup>1</sup> & Yann Ménéroux <sup>1</sup>

2021-2022

<sup>1</sup>IGN-ENSG-UGE

**ENSG**  
Géomatique

ÉCOLE NATIONALE  
DES SCIENCES  
GÉOGRAPHIQUES

# Introduction

---

Techniques pour quantifier la liaison entre **deux** variables (quali ou quanti).

- corrélation
- régression linéaire
- $\chi^2$  (test d'indépendance)
- visualisation adéquate

La plupart des phénomènes intéressants (sociaux, spatiaux) sont **multi-factoriels**. Les données disponibles pour les décrire sont :

- partiellement **redondantes** : e.g. revenu et profession
- intrinsèquement **corrélées** : e.g. revenu et taille du logement
- parfois des proportions (somme à 1 ou 100%)

L'analyse **factorielle** cherche à réduire la **colinéarité** et le **nombre de dimensions** (=variables) qui décrivent une population ...

... en proposant de nouvelles variables **composites décorrélées**.



- Nom e.g. "Pikachu"
- Type 1  $\in \{Grass, Fire, Water, Bug, \dots\}$
- Type 2 idem
- HP : numérique
- Attack : numérique
- Defense : numérique
- Speed : numérique
- Special Attack : numérique
- Special Defense : numérique
- Generation : facteur  $\in \{1, 2, 3, 4, 5, 6\}$
- Legendary : booléen

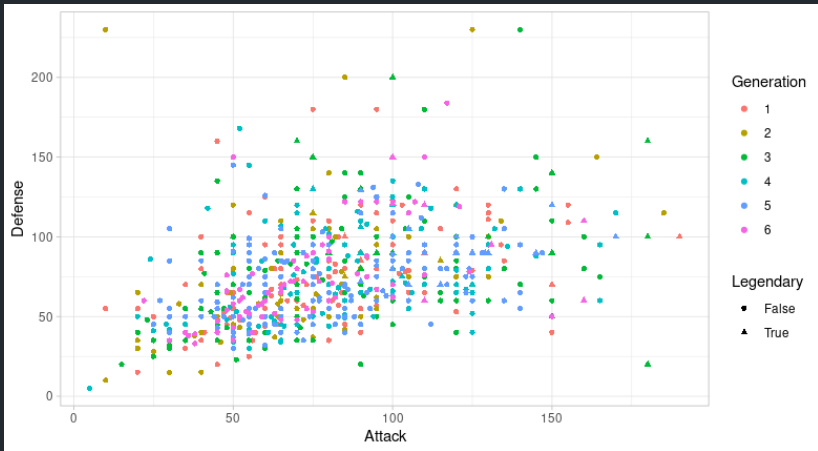
Existe-t-il des combinaisons qui **résumant bien** les caractéristiques des pokemons ? (moins de six!)

Comment les constituer ?

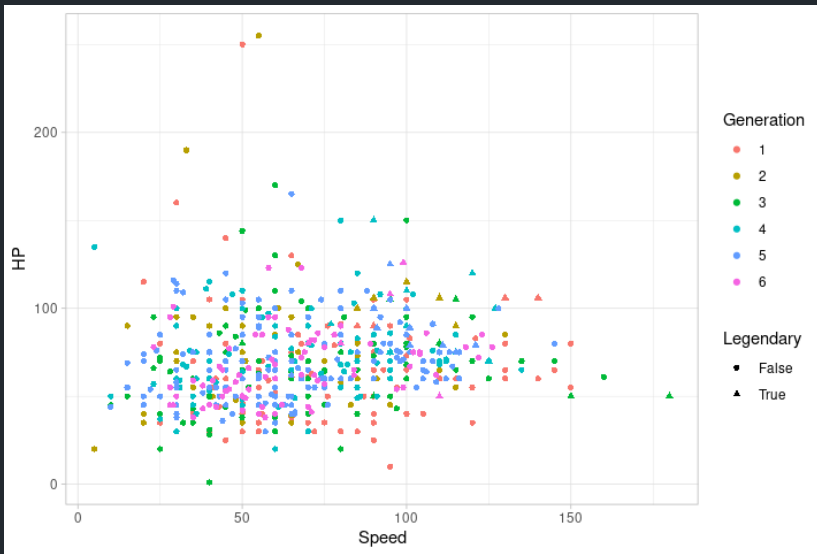
i.e. comment **combiner** les six variables numériques pour bien **expliquer leur variation** au sein de la population ?



# Attack vs. Defense



# Speed vs. HP



# L'inertie

---

L'inertie est l'équivalent **multi-dimensionnel** de la **variance** d'une variable.

C'est une notion centrale de l'ACP.

$$I = \frac{1}{n} \sum_{i=1}^n d^2(x_i, g)$$

Avec

- $n$  la taille de la population
- $x_i$  la valeur de la variable de l'individu  $i$
- $g$  le point moyen
- $d(x, y)$  une distance, souvent euclidienne :  $(x_i - g_i)^2$

L'inertie quantifie la **dispersion** du nuage de points

L'inertie est la "moyenne du carré des distances", ou encore la **somme des variances** des variables

Inertie faible  $\implies$  peu de variété dans les variables, individus semblables, faible quantité d'information

Soit une population  $P$  de  $n$  individus décrits par une variable  $X$   
l'inertie de la population est la **variance** de  $X$  :

$$I = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Le point moyen a pour "coordonnées"  $\bar{x}$

Soient  $X$  et  $Y$  deux variables qui décrivent des individus  $p_i$  de la population  $P$ , et  $g = (x_g, y_g)$  le point moyen de cette population, de coordonnées  $x_g = \bar{x}$  et  $y_g = \bar{y}$ .

L'inertie de  $P$  est :

$$I = \frac{1}{n} \sum_{i=1}^n (x_i - x_g)^2 + (y_i - y_g)^2$$

On reconnaît une somme de variances :  $I = \text{var}(X) + \text{var}(Y)$

Soient  $v$  variables , notées  $X^{(k)}$ ,  $k \in \{1, \dots, v\}$  qui décrivent les individus d'une population  $P$ , le point moyen de  $P$  est noté  $g$ .

L'inertie de  $P$  est :

$$I = \frac{1}{n} \sum_{k=1}^v \sum_{i=1}^n (x_i^{(k)} - x_g^{(k)})^2$$

on reconnaît

$$I = \sum_{k=1}^v \text{var}(X^{(k)})$$



# Espaces, vecteurs, axes, variables

---

L'ACP considère une population statistique décrite par plusieurs variables (continues).

Ces variables définissent un **espace vectoriel**, qu'on va appeler l'espace d'**origine**:

- un individu  $i$  est un **vecteur**
- la valeur de ses variables sont les **coordonnées** du vecteur dans cet espace.
- chaque variable est une **dimension** de cet espace. elle définit un **axe** de l'espace. (cf. axe des  $x$  dans un repère orthonormé)

Les variables étant potentiellement **corrélées**, les axes de l'espace de départ ne sont pas toujours (presque jamais) orthogonaux !

Les individus sont des vecteurs dans l'espace des variables  
mais également

Les variables sont des vecteurs dans l'espace des individus

L'ACP consiste à trouver de **nouveaux axes orthogonaux entre eux**, qui capturent le **plus d'inertie possible** de la population  $P$ .

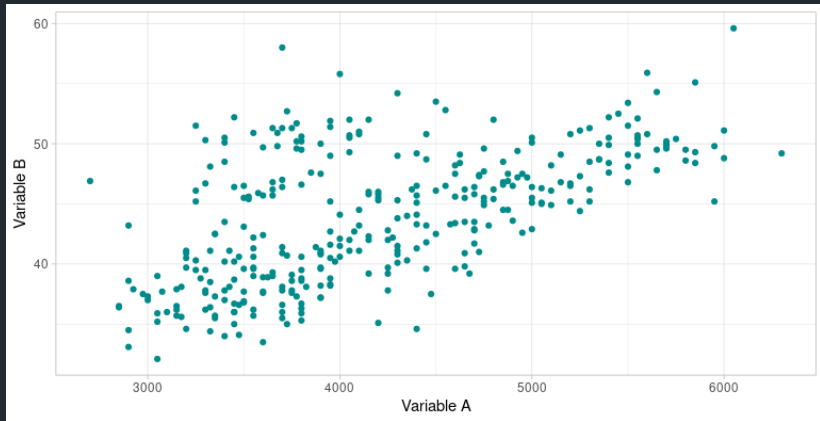
Ces axes définiront un nouvel espace : l'**espace d'arrivée**

On trouve ces axes en combinant (linéairement), les variables de la population  $P$ , par exemple :

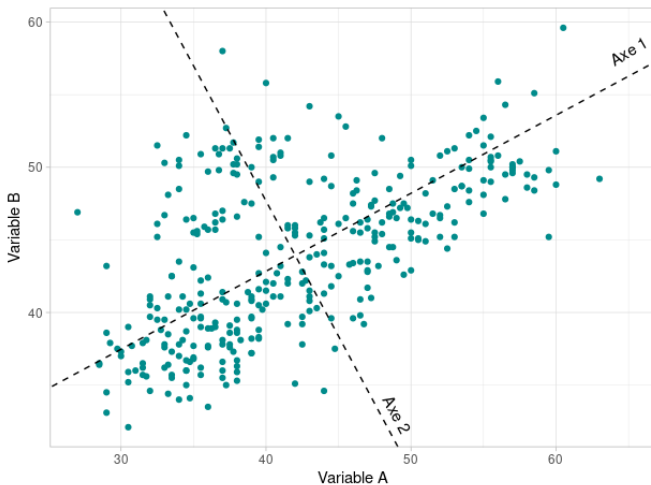
$$\text{axe}_1 = \alpha X + \beta Y + \gamma Z$$

La composition de ces combinaisons (les valeurs de  $\alpha, \beta, \gamma$ ) pour chaque axe est donnée en résolvant un système d'équations algébriques

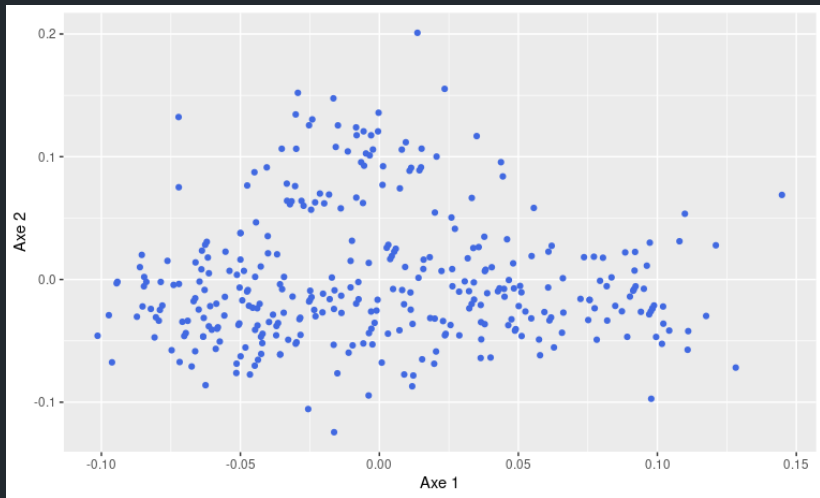
## Espace de départ



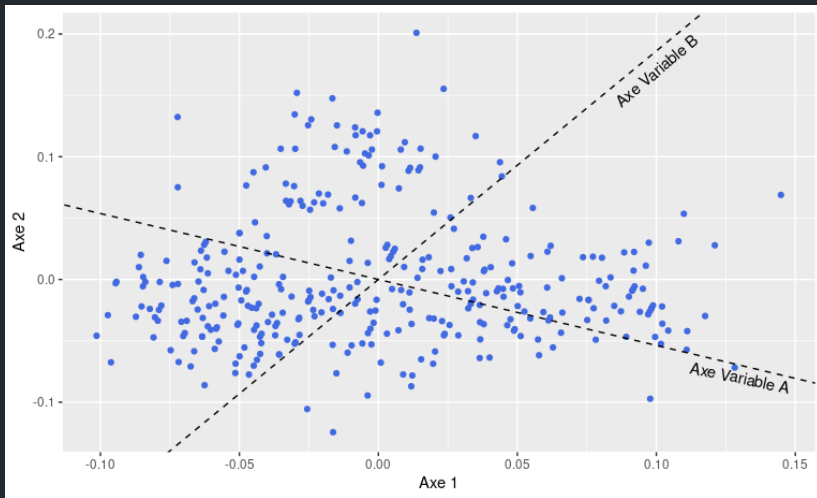
Espace de départ + Les axes de l'espace d'arrivée



## Espace d'arrivée



Espace d'arrivée + les axes de l'espace de départ





espace de départ  $\rightarrow$  ACP  $\rightarrow$  espace d'arrivée

Les **axes** sont les **vecteurs propres** de la matrice de corrélation de  $P$ ,  $Cor_P$ . On peut les calculer (ouf !)

l'ACP est le calcul d'une transformation linéaire qui **re-projette** des vecteurs-individus dans un nouvel **espace** – l'espace d'arrivée – constitué par les **nouveaux axes**.

On appelle ces axes **composantes**, elles sont **linéairement indépendantes** et forment une **base** de l'espace d'arrivée.

L'ACP capture l'inertie de  $P$  en créant des composantes (les vecteurs propres de  $Cor_P$ ).

Il y a autant de composantes possibles que de dimensions de l'espace de départ.

L'intérêt de l'ACP est de pouvoir se limiter à **quelques** composantes pour à la fois

- capturer suffisamment l'inertie ( $\approx$  l'information) de  $P$
- réduire la dimensionnalité ( $\approx$  complexité) de  $P$

L'inertie capturée par une composante  $k$  est sa valeur propre ,  $\lambda_k$

On ordonne les composantes par valeur propre décroissantes:

La 1<sup>ère</sup> composante correspond au vecteur propre de plus grande valeur propre, elle capture la plus grande proportion d'inertie

La 2<sup>nde</sup> composante correspond au vecteur propre de la seconde plus grande valeur propre , elle capture la seconde plus grande proportion d'inertie

et ainsi de suite.

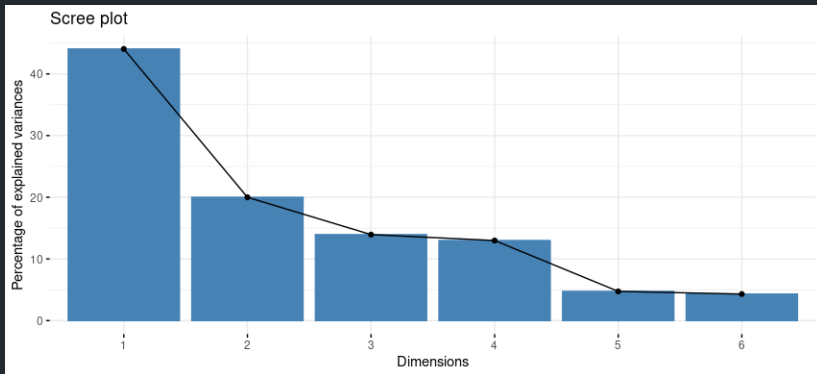
Si les variables sont centrées et réduites, leur somme vaut  $Dim(P)$

# Interpréter les résultats d'une ACP

---

- **Dimensionnalité** : L'essentiel de l'inertie est-elle exprimée en peu de dimensions dans l'espace d'arrivée ?
- **Colinéarité des variables** : Comment les variables de l'espace de départ sont-elles corrélées entre elles et aux axes de l'espace d'arrivée ?
- **Contribution** : À quel point Individus et Variables contribuent aux axes de l'espace d'arrivée ?
- **Représentation** : Les Individus et Variables sont-ils-elles bien représentés par les axes de l'espace d'arrivée ?

Le **scree plot** montre la proportion d'inertie capturée par les différentes composantes. La valeur propre associée aux vecteurs propres (axes) est proportionnelle à l'inertie capturée.



**Idéalement**, les premières (2 ou 3 ) composantes capturent une partie significative (e.g.  $\gtrsim 50\%$  de l'inertie de  $P$ )

Cela signifie que les composantes **résumement bien** l'information contenue dans les variables de  $P$ , en **peu de dimensions**.



Pour profiter du "résumé" de l'ACP, il faut se limiter à un certain nombre de composantes pour définir l'espace d'arrivée dans lequel on reconstituera les individus et variables.

Heuristiques du choix du nombre:

- On garde les  $q$  axes que l'on sait **interpréter** : 2 ou 3 !
- "coude" dans le scree-plot.
- ne conserver que les  $\lambda > 1$  ou  $\lambda > 2$
- Karlis-Saporta-Spinaki : conserver les  $\lambda$  t.q.  $\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$
- Gavish & Donoho (2014) :  $\lambda = \frac{4\sigma\sqrt{n}}{\sqrt{3}}$  avec  $\sigma$  le bruit estimé dans les données.

Avec  $\lambda$ , les valeurs propres associées aux axes,  $p$  le nombre de variable de  $P$ , et  $n$  la taille de  $P$

En pratique , si on sélectionne  $q$  composantes, il faudra projeter les individus et les variables dans  $C_q^2$  plans pour les visualiser.

Si  $q = 3$ , il faut 3 graphiques  $\{(q_1, q_2), (q_2, q_3), (q_1, q_3)\}$ .

Si  $q = 4$ , il en faut 6 !

On sait passer de l'espace de départ à l'espace d'arrivée : On peut **projeter** les variables et les individus dans l'espace d'arrivée

De cette projection on tire beaucoup d'information utiles:

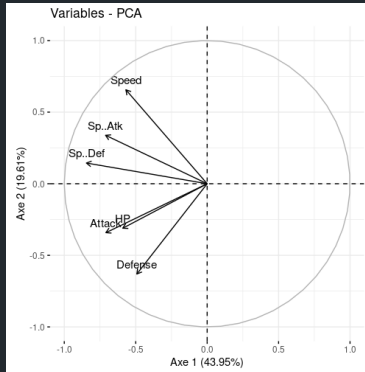
- corrélations de variables
- contribution / représentation des variables
- contribution / représentation des individus
- regroupements d'individus

# Projection des variables

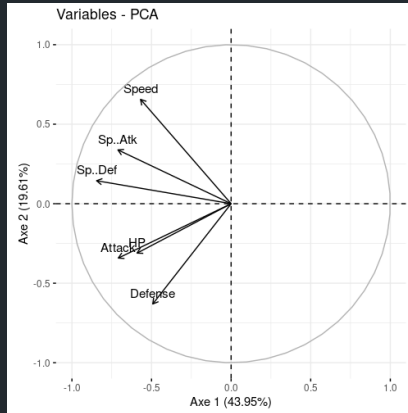
---

Rappel : les variables sont des vecteurs dans l'espace des individus.

On peut projeter les variables dans l'espace d'arrivée :



Si les variables sont centrées et réduites lors de l'ACP, on peut les représenter dans un cercle de corrélation et évaluer visuellement leur corrélation

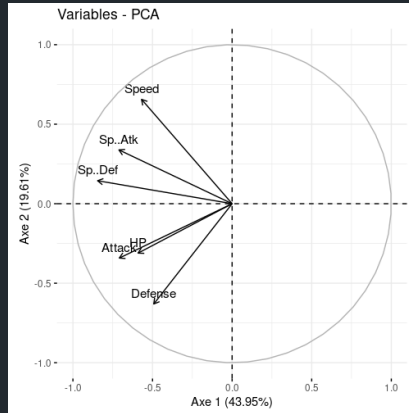


Variable  $\leftrightarrow$  Flèche

Coordonnées de la variable  $\leftrightarrow$  **corrélation linéaire** avec les composantes

Proximité au cercle  $\leftrightarrow$  qualité de **représentation** de la variable

Angle des variables  $\leftrightarrow$  corrélation des variables entre elles



- la corrélation de Defense avec l'Axe 1 est de -0.5
- Attack et HP sont très corrélées
- Speed et Defense sont (linéairement) indépendantes

Ici : regroupement de variables ? Oui !

La **contribution** d'une variable  $v$  à l'inertie de l'axe  $k$  est la coordonnée carrée de  $v$  sur l'axe  $k$  divisée par son inertie.

$$ctrb_{vk} = \frac{c_{vk}^2}{\lambda_k}$$

Plus la contribution d'une variable est élevée , plus elle est importante pour expliquer la variabilité de  $P$



La **qualité de représentation** d'une variable  $v$  par l'axe  $k$  est la coordonnée carrée de  $v$  sur l'axe  $k$  :

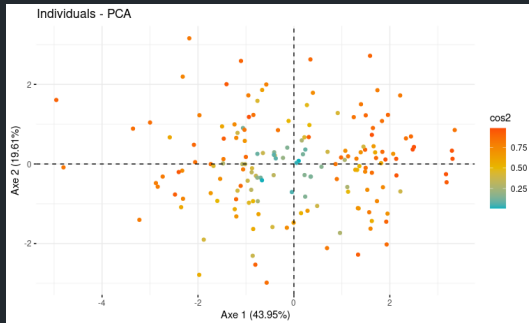
$$ctrb_{vk} = c_{vk}^2$$

# Projection des individus

---

Rappel : les individus sont des vecteurs dans l'espace des variables de  $P$ .

On peut projeter les individus dans l'espace d'arrivée :



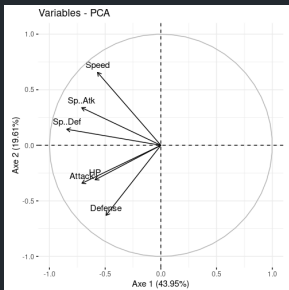
Si les variables sont centrées et réduites lors de l'ACP, on peut représenter les individus dans le cercle de corrélation.

## Avantages

- Réduit la dimensionnalité
- Regroupe les variables et les individus

## Limites

- Composantes difficiles à interpréter en elles-mêmes
- Que faire si  $p$  est grand et si les premières composantes capturent peu d'inertie ?



- L'Axe 1 "prend tout" : c'est la puissance générale des pokémon, une sorte de **score global**
- L'Axe 2 sépare les variables en **deux groupes** : celle du combat "standard" (Attack, Defense, HP) et celles du combat "spécial/rapide" (Sp. .Atk, SP. .Def, Speed)
- On pourrait être tenté de diviser les pokemons en "Costauds classiques" vs. "Ninjas spéciaux" .

Merci à Anh Le , <https://anhqle.github.io/gotta-plot-them-all/>

<http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique-73-acp-analyse-en-composantes-principales-avec-r-l-essentiel>

- This is important

- This is important
- Now this



- This is important
- Now this
- And now this

- This is really important
- Now this
- And now this

**Mono message sur une diapo**

$$A = \sum_{i=1}^n \left( 1 + \frac{1}{x_i} \right)^{\alpha}$$