GPS, capteurs, suivi en temps réel : analyser des données fonctionnelles dans le sport

Groupe Statistique et Sport de la SFdS

Marie Chion, **Sébastien Déjean**, Arthur Leroy, Christian Derquenne

Intervenant

- Ingénieur de recherche, statisticien, HDR
- Institut de Mathématiques

INSTITUT de MATHEMATIQUES de TOULOUSE

- Université de Toulouse

Université de Toulouse

- Coureur à pied
- Athletics Coaching Club Ramonville

ATHLETICS COACHING CLUB

Plan

- Analyse en Composantes Principales
- Classification non supervisée
- Validation et description des groupes (de classification)
- Maillage de Voronoï

Jeux de données

- `decathlon` : données disponibles dans le package FactoMineR, 41 lignes, 13 colonnes

- `records_athle` : records nationaux de 30 pays pour 9 épreuves d'athlétisme

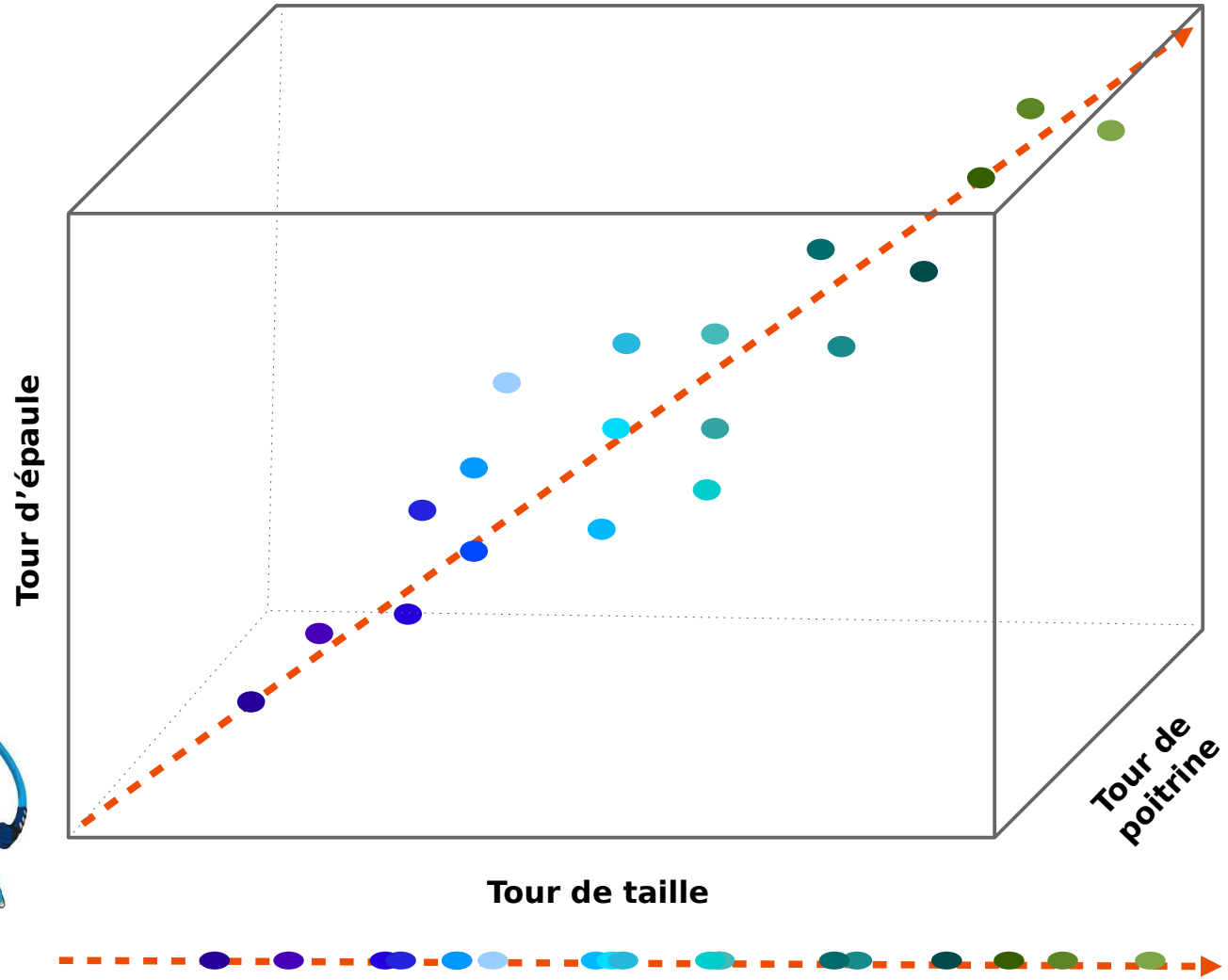
- `GPS_rugby` : position (x, y) de 15 joueurs de rugby pendant 9 secondes, fréquence 10 Hz, 1335 lignes, 4 colonnes

Analyse en Composantes Principales

Principe de l’ACP

Teasing: Utiliseriez-vous un carton de forme cubique pour emballer une canne à pêche ?

Principe de l’ACP



Do we need 3 dimensions to represent ‘standard’ individuals?

=

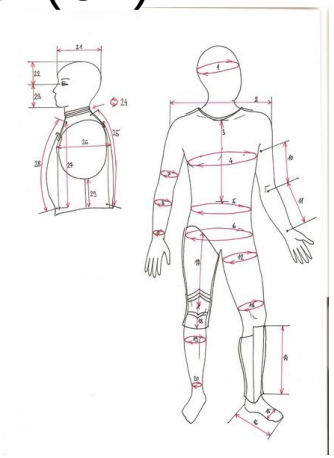
Do we need a cubic box to pack a fishing rod?

1ere composante principale «costautitude»

Un petit exemple

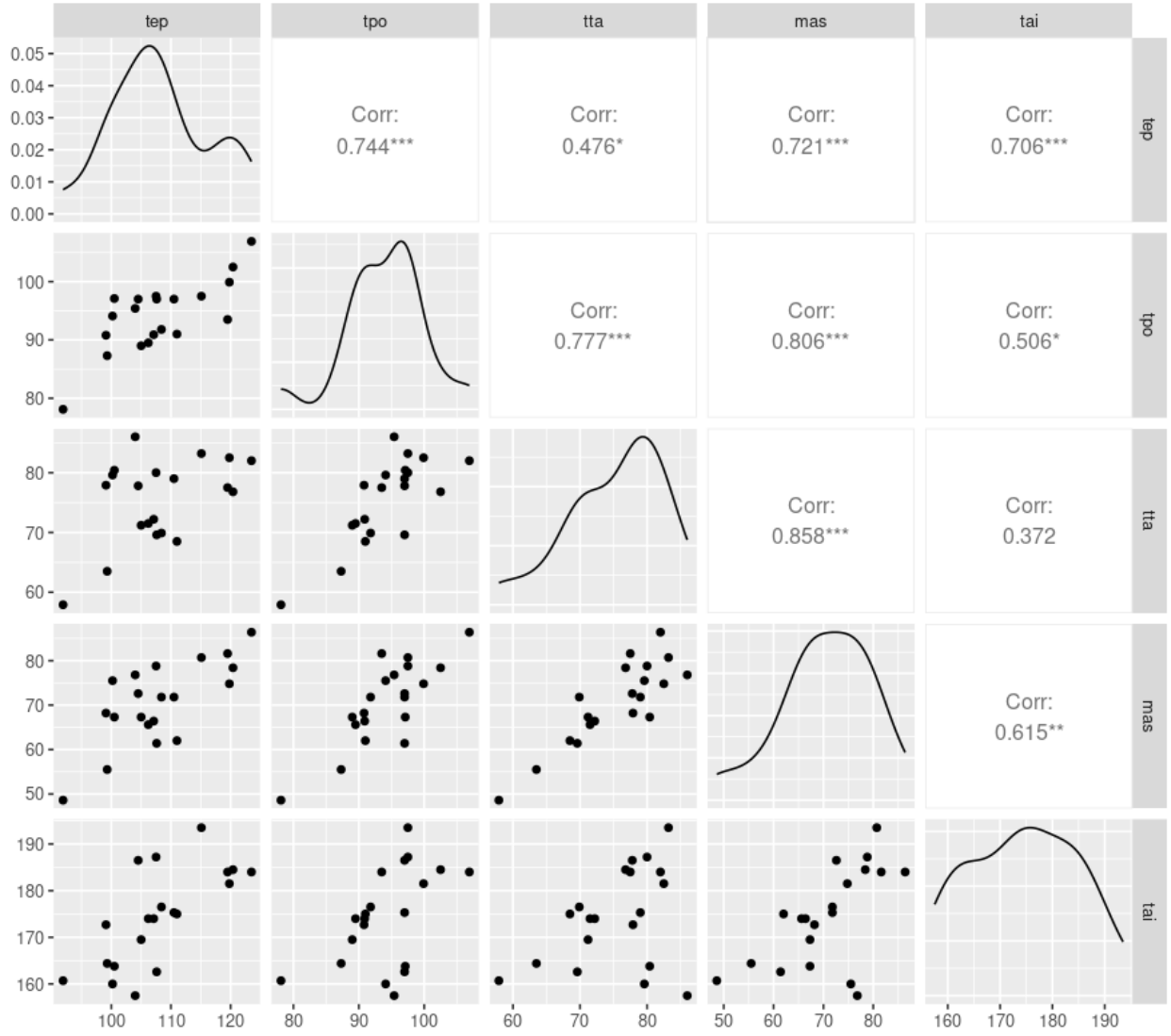
- 20 individus
- 5 variables

```
tep : tour d'épaule (cm)
tpo : tour de poitrine (cm)
tta : tour de taille(cm)
mas   : masse (kg)
tai   : taille (cm)
```



Id	tep	tpo	tta	mas	tai
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

Un petit exemple



Le coeur de l'ACP

Coefficients de combinaison linéaire (*loadings*)

	PC1	PC2	PC3	PC4	PC5
tep	0.45	0.40	-0.57	-0.43	0.36
tpo	0.47	-0.21	-0.46	0.67	-0.27
tta	0.43	-0.57	0.33	-0.01	0.61
mas	0.49	-0.18	0.24	-0.50	-0.65
tai	0.38	0.66	0.55	0.33	0.05

```
PC1 =  0.45*tep + 0.47*tpo + 0.43*tta + 0.49*mas + 0.38*tai
PC2 =  0.40*tep - 0.21*tpo - 0.57*tta - 0.18*mas + 0.66*tai
...
```

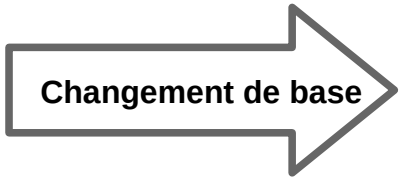
? D'où sortent ces coefficients ? décomposition en éléments propres de la matrice de covariance, ou décomposition en valeurs singulières de la matrice de données, ou par algorithme itératif (en cas de données manquantes)

Résultats numériques et graphiques

Données centrées-réduites

Id	tep	tpo	tta	mas	tai
I1	-0.22	-0.77	-0.54	-0.54	-0.03
I2	0.30	0.46	0.51	0.13	0.09
I3	0.85	0.54	1.10	1.09	1.83
I4	-0.43	0.46	0.34	0.22	1.16
I5	-0.07	0.54	0.65	0.89	1.23
I6	1.42	0.93	1.00	0.46	0.68
I7	1.86	2.08	0.93	1.71	0.92
I8	1.49	1.36	0.20	0.85	0.97
I9	0.36	-0.52	-0.96	-0.93	0.06
I10	1.38	-0.11	0.30	1.19	0.92
I11	-0.37	-0.85	-0.58	-0.35	-0.46
I12	-0.95	-0.01	0.60	0.53	-1.37
I13	-1.08	-0.55	0.36	-0.26	-0.16
I14	-0.05	0.46	-0.81	-0.99	-1.12
I15	-0.49	0.20	1.49	0.67	-1.61
I16	0.04	-0.39	-0.76	0.13	0.20
I17	-1.06	-1.12	-1.66	-1.63	-0.95
I18	-1.95	-2.63	-2.45	-2.37	-1.31
I19	-0.12	-0.54	-0.44	-0.45	-0.03
I20	-0.91	0.48	0.71	-0.35	-1.01
Mean	**0**	**0**	**0**	**0**	**0**
Var.	**1**	**1**	**1**	**1**	**1**

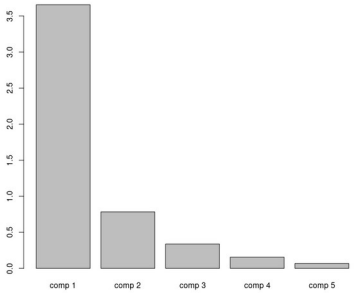
	PC1	PC2	PC3	PC4	PC5
tep	0.45	0.40	-0.57	-0.43	0.36
tpo	0.47	-0.21	-0.46	0.67	-0.27
tta	0.43	-0.57	0.33	-0.01	0.61
mas	0.49	-0.18	0.24	-0.50	-0.65
tai	0.38	0.66	0.55	0.33	0.05



	PC1	PC2	PC3	PC4	PC5
I1	-1.00	0.46	0.16	-0.16	0.14
I2	0.69	-0.24	-0.13	0.14	0.22
I3	2.41	0.63	0.92	0.04	0.24
I4	0.75	0.27	0.86	0.78	-0.15
I5	1.45	0.15	0.92	0.35	-0.28
I6	2.05	0.18	-0.43	0.00	0.63
I7	3.50	0.09	-0.81	0.02	-0.38
I8	2.24	0.71	-0.68	0.16	-0.20
I9	-0.96	1.03	-0.48	-0.01	0.29
I10	1.68	0.82	0.16	-1.00	-0.01
I11	-1.20	0.12	0.07	-0.39	-0.06
I12	-0.45	-1.77	0.11	-0.33	-0.40
I13	-0.80	-0.60	0.86	0.18	0.14
I14	-1.10	-0.23	-1.33	0.48	-0.06
I15	0.24	-2.34	-0.05	-0.55	0.17
I16	-0.36	0.66	0.05	-0.28	-0.43
I17	-2.97	0.43	-0.35	0.23	-0.10
I18	-4.97	0.74	0.23	-0.14	-0.04
I19	-0.75	0.38	0.04	-0.09	0.12
I20	-0.45	-1.51	-0.11	0.57	0.15
Mean	**0**	**0**	**0**	**0**	**0**
Var.	**3.66**	**0.78**	**0.34**	**0.16**	**0.07**

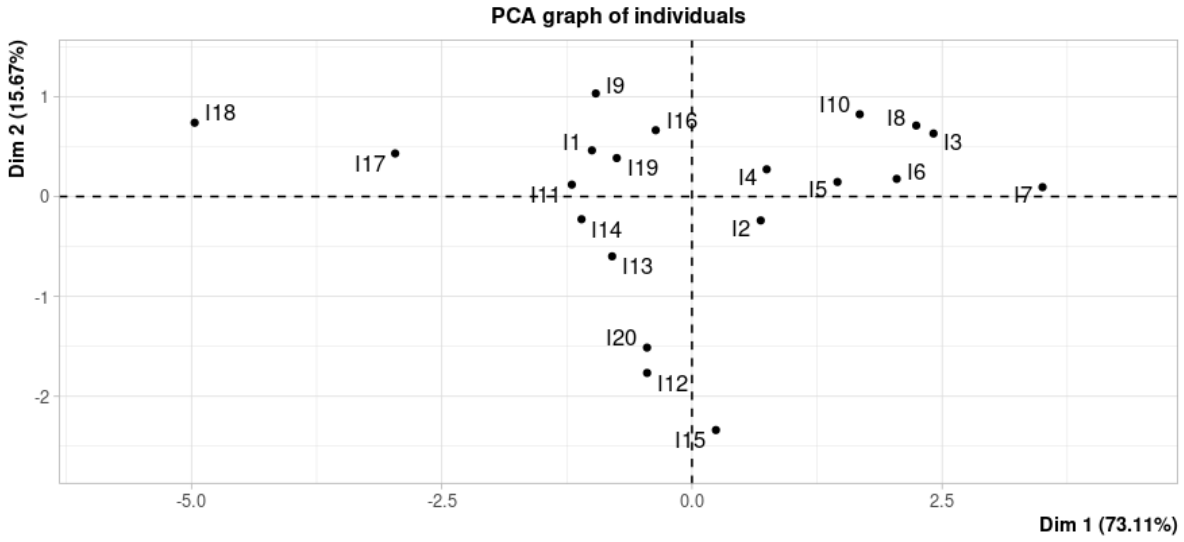
Résultats numériques et graphiques

	PC1	PC2	PC3	PC4	PC5
Variance	3.656	0.783	0.338	0.156	0.068
% of var.	73.110	15.668	6.756	3.112	1.353
Cumulative % of var.	73.110	88.779	95.535	98.647	100.000



Représentation des individus

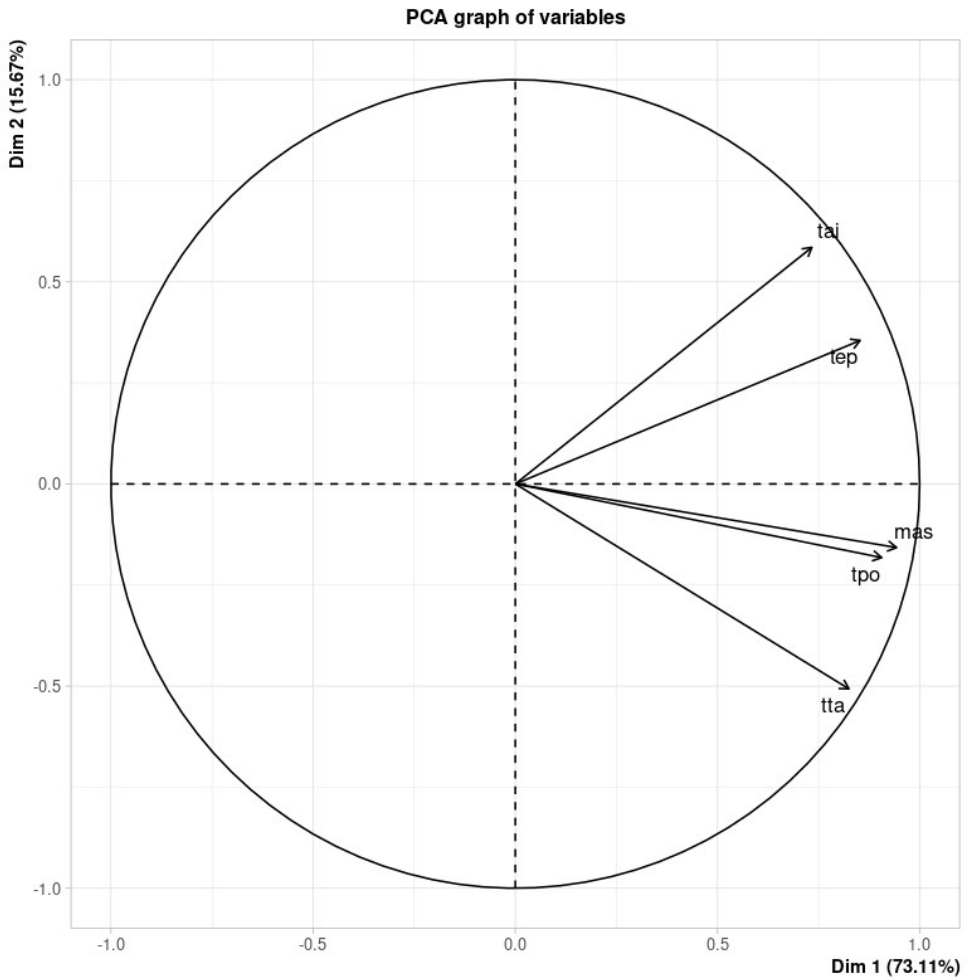
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
I1	-1.00	0.46	0.16	-0.16	0.14
I2	0.69	-0.24	-0.13	0.14	0.22
I3	2.41	0.63	0.92	0.04	0.24
I4	0.75	0.27	0.86	0.78	-0.15
I5	1.45	0.15	0.92	0.35	-0.28
I6	2.05	0.18	-0.43	0.00	0.63
I7	3.50	0.09	-0.81	0.02	-0.38
I8	2.24	0.71	-0.68	0.16	-0.20
I9	-0.96	1.03	-0.48	-0.01	0.29
I10	1.68	0.82	0.16	-1.00	-0.01
I11	-1.20	0.12	0.07	-0.39	-0.06
I12	-0.45	-1.77	0.11	-0.33	-0.40
I13	-0.80	-0.60	0.86	0.18	0.14
I14	-1.10	-0.23	-1.33	0.48	-0.06
I15	0.24	-2.34	-0.05	-0.55	0.17
I16	-0.36	0.66	0.05	-0.28	-0.43
I17	-2.97	0.43	-0.35	0.23	-0.10
I18	-4.97	0.74	0.23	-0.14	-0.04
I19	-0.75	0.38	0.04	-0.09	0.12
I20	-0.45	-1.51	-0.11	0.57	0.15



Résultats numériques et graphiques

	PC1	PC2	PC3	PC4	PC5
tep	0.85	0.35	-0.33	-0.17	0.09
tpo	0.91	-0.18	-0.27	0.26	-0.07
tta	0.83	-0.51	0.19	0.00	0.16
mas	0.94	-0.16	0.14	-0.20	-0.17
tai	0.73	0.59	0.32	0.13	0.01

Représentation des variables




```
   cor(tep, PC1) =  0.85
   cor(tep, PC2) =  0.35
   …
```

```
   cor(tpo, PC1) =  0.91
   cor(tpo, PC2) = -0.18
   ...
```

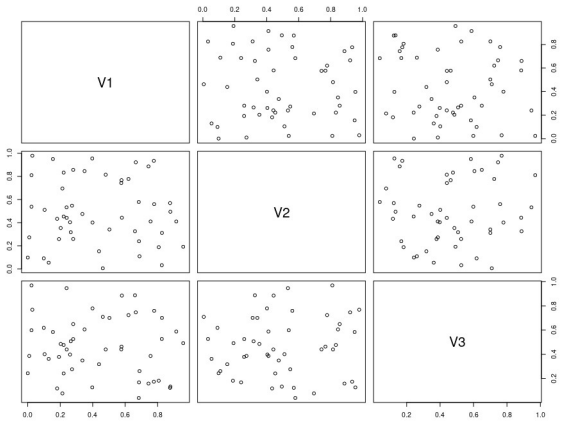
Id	tep	tpo	tta	mas	tai
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
I1	-1.00	0.46	0.16	-0.16	0.14
I2	0.69	-0.24	-0.13	0.14	0.22
I3	2.41	0.63	0.92	0.04	0.24
I4	0.75	0.27	0.86	0.78	-0.15
I5	1.45	0.15	0.92	0.35	-0.28
I6	2.05	0.18	-0.43	0.00	0.63
I7	3.50	0.09	-0.81	0.02	-0.38
I8	2.24	0.71	-0.68	0.16	-0.20
I9	-0.96	1.03	-0.48	-0.01	0.29
I10	1.68	0.82	0.16	-1.00	-0.01
I11	-1.20	0.12	0.07	-0.39	-0.06
I12	-0.45	-1.77	0.11	-0.33	-0.40
I13	-0.80	-0.60	0.86	0.18	0.14
I14	-1.10	-0.23	-1.33	0.48	-0.06
I15	0.24	-2.34	-0.05	-0.55	0.17
I16	-0.36	0.66	0.05	-0.28	-0.43
I17	-2.97	0.43	-0.35	0.23	-0.10
I18	-4.97	0.74	0.23	-0.14	-0.04
I19	-0.75	0.38	0.04	-0.09	0.12
I20	-0.45	-1.51	-0.11	0.57	0.15

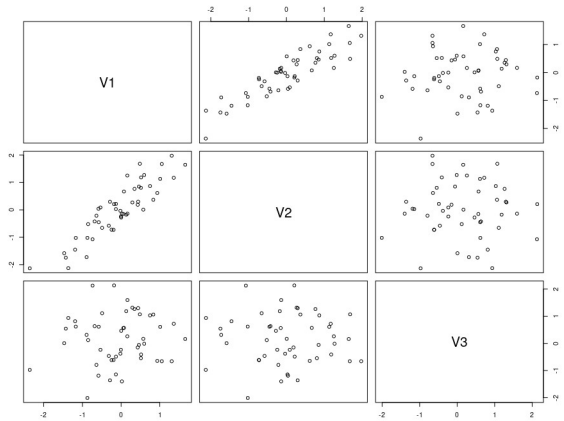
Exemples simulés

3 jeux de données : 50 observations, 3 variables (V1 – V2 - V3)

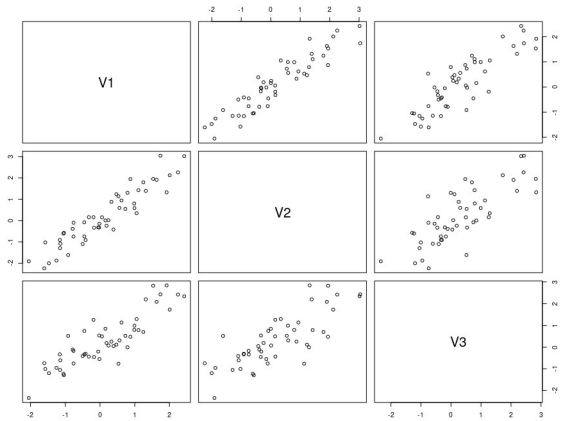
Cas 1)
{V1} - {V2} - {V3}



Cas 2)
{V1 - V2} - {V3}



Cas 3)
{V1 - V2 - V3}



Matrices de corrélation de Pearson

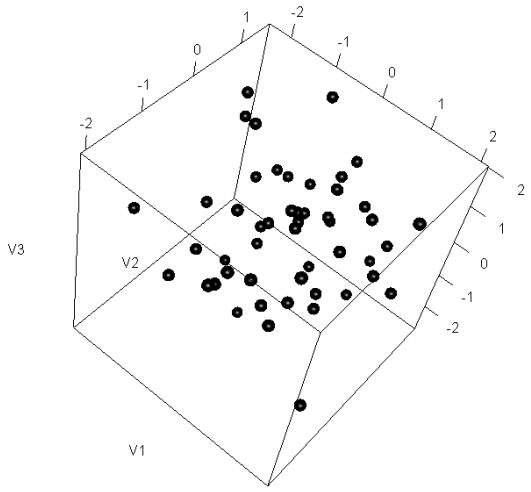
1)	**V1**	**V2**	**V3**
V1	1.00	-0.05	-0.12
V2	-0.05	1.00	0.06
V3	-0.12	0.06	1.00

2)	**V1**	**V2**	**V3**
V1	1.00	0.90	0.08
V2	0.90	1.00	-0.01
V3	0.08	-0.01	1.00

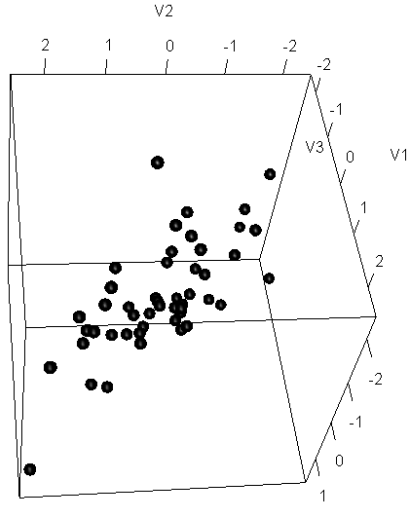
3)	**V1**	**V2**	**V3**
V1	1.00	0.93	0.87
V2	0.93	1.00	0.79
V3	0.87	0.79	1.00

Exemples simulés

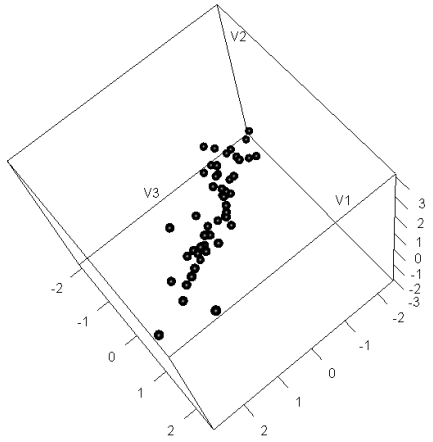
Cas 1)



Cas 2)



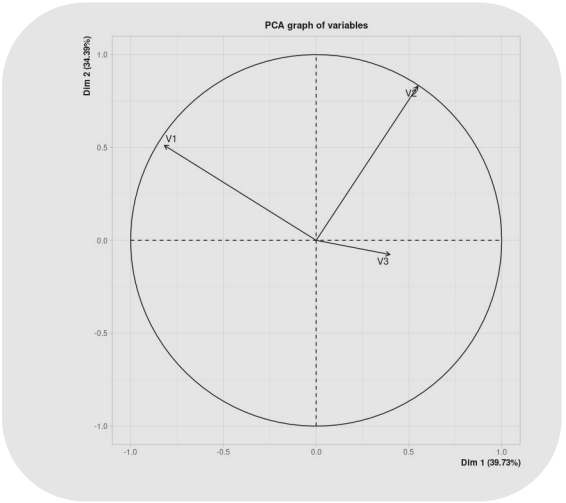
Cas 3)

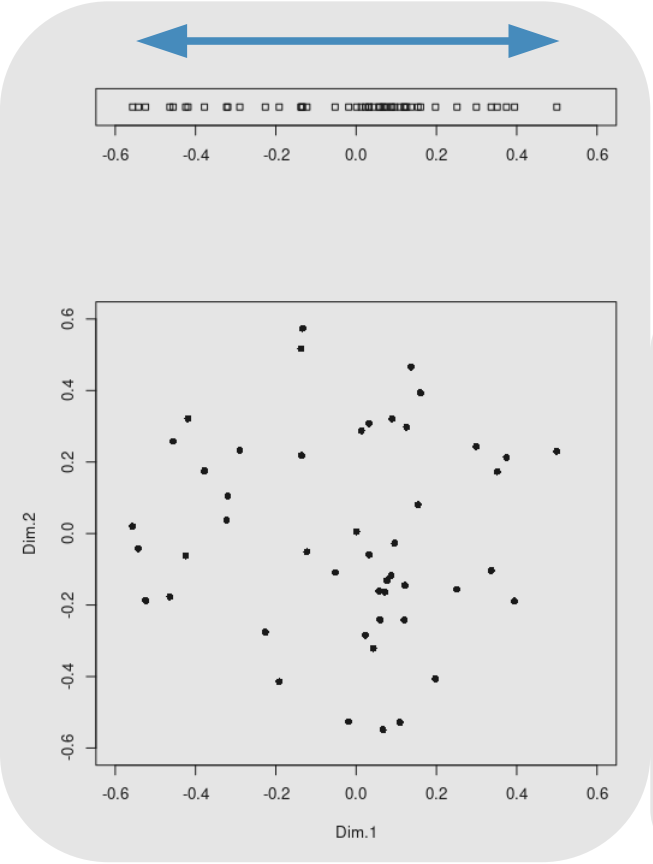


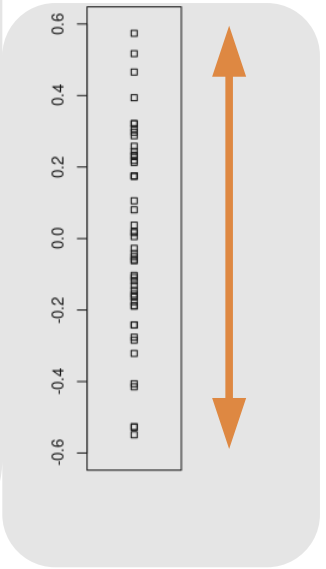
Exemples simulés

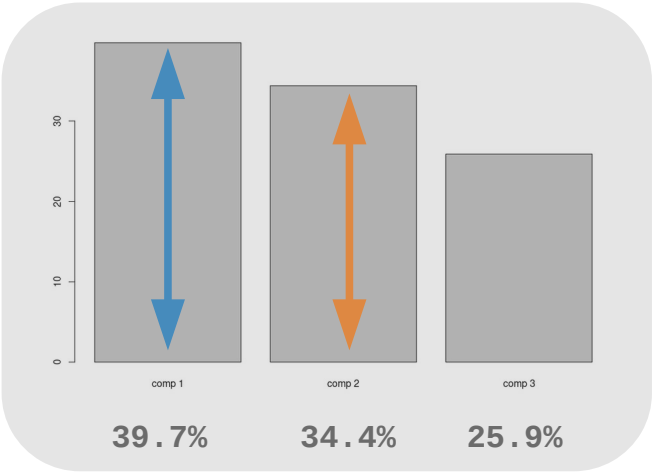
Loadings

	Dim.1	Dim.2	Dim.3
V1	-0.23	0.14	0.07
V2	0.15	0.23	-0.03
V3	0.10	-0.02	0.22





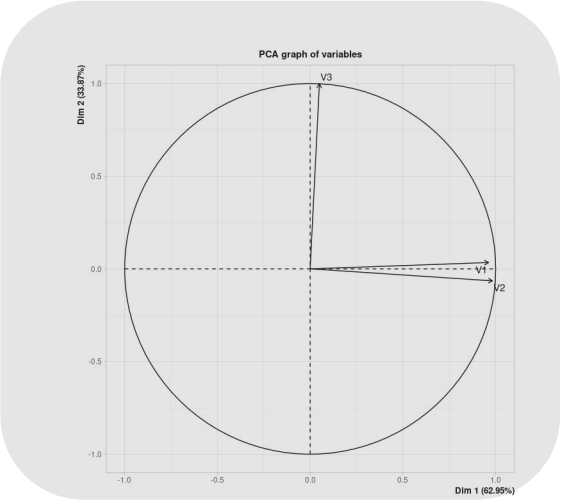


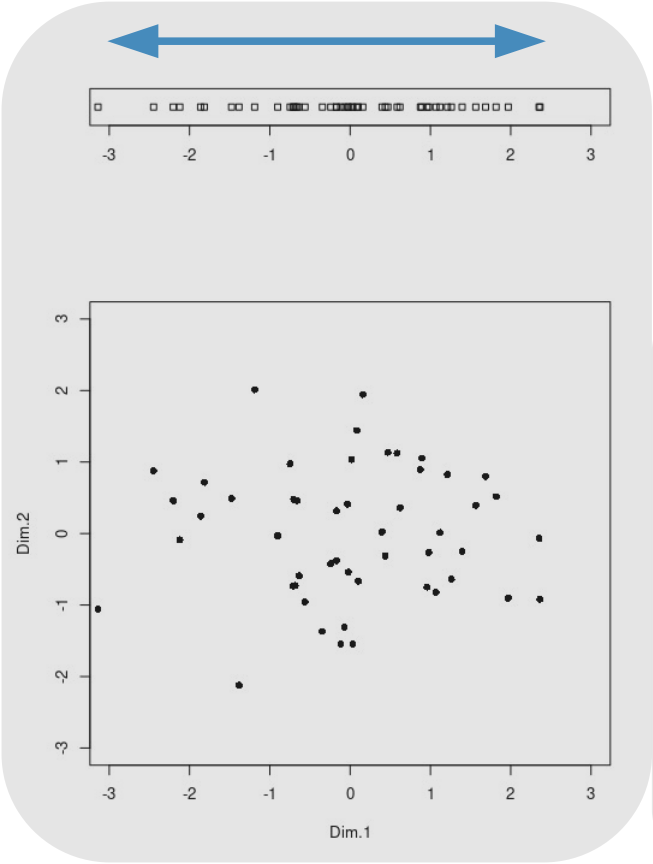


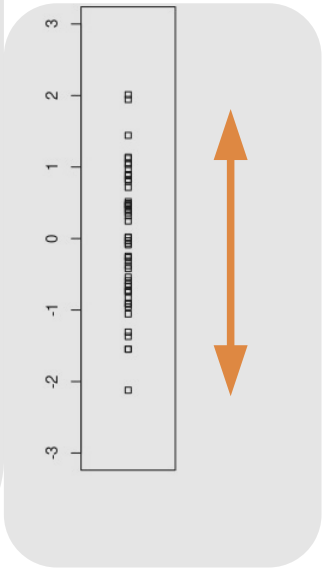
Exemples simulés

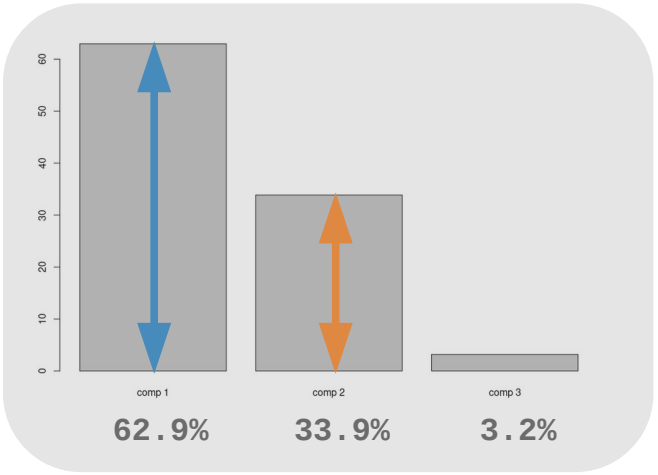
Loadings

	Dim.1	Dim.2	Dim.3
V1	0.77	0.03	0.22
V2	0.97	-0.06	-0.17
V3	0.05	0.91	-0.02





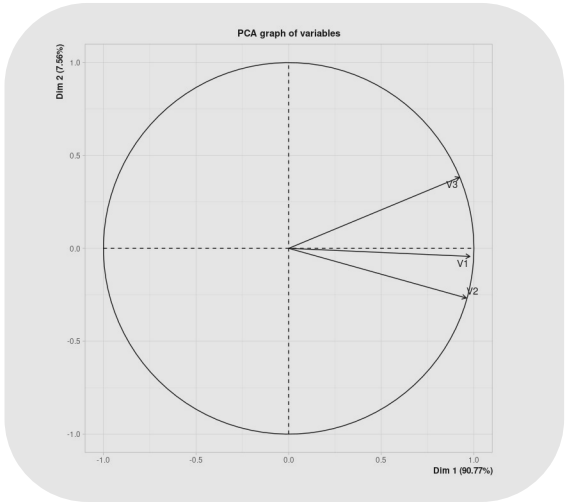


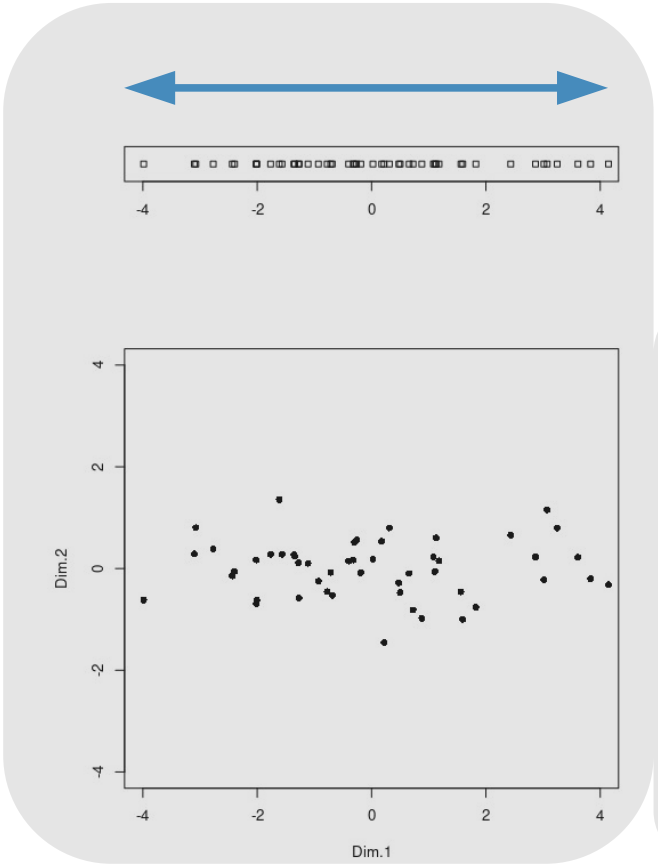


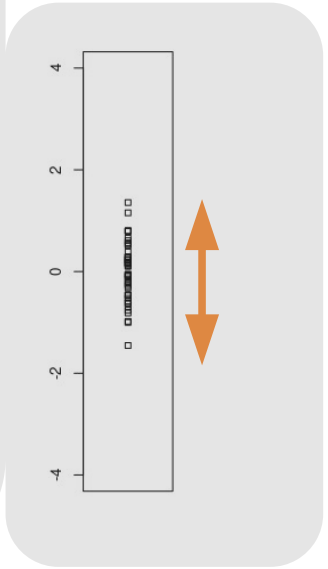
Exemples simulés

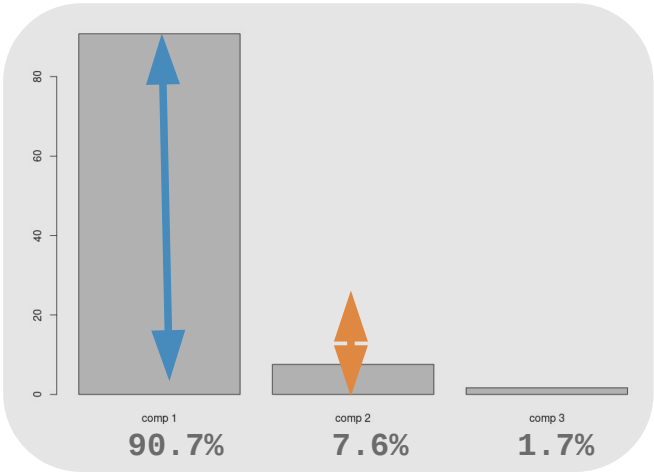
Loadings

	Dim.1	Dim.2	Dim.3
V1	1.07	-0.05	0.22
V2	1.23	-0.34	-0.13
V3	1.07	0.44	-0.07



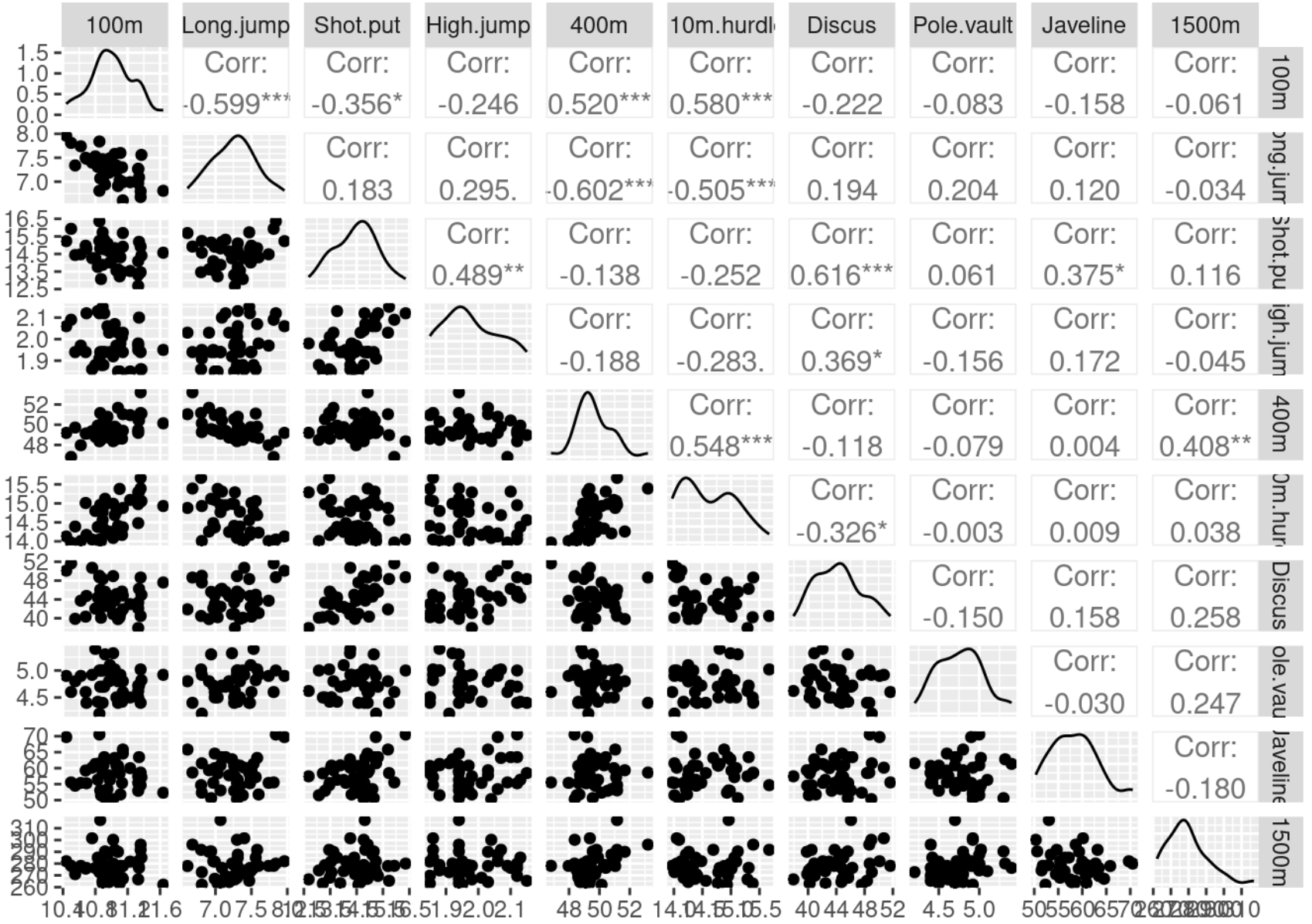






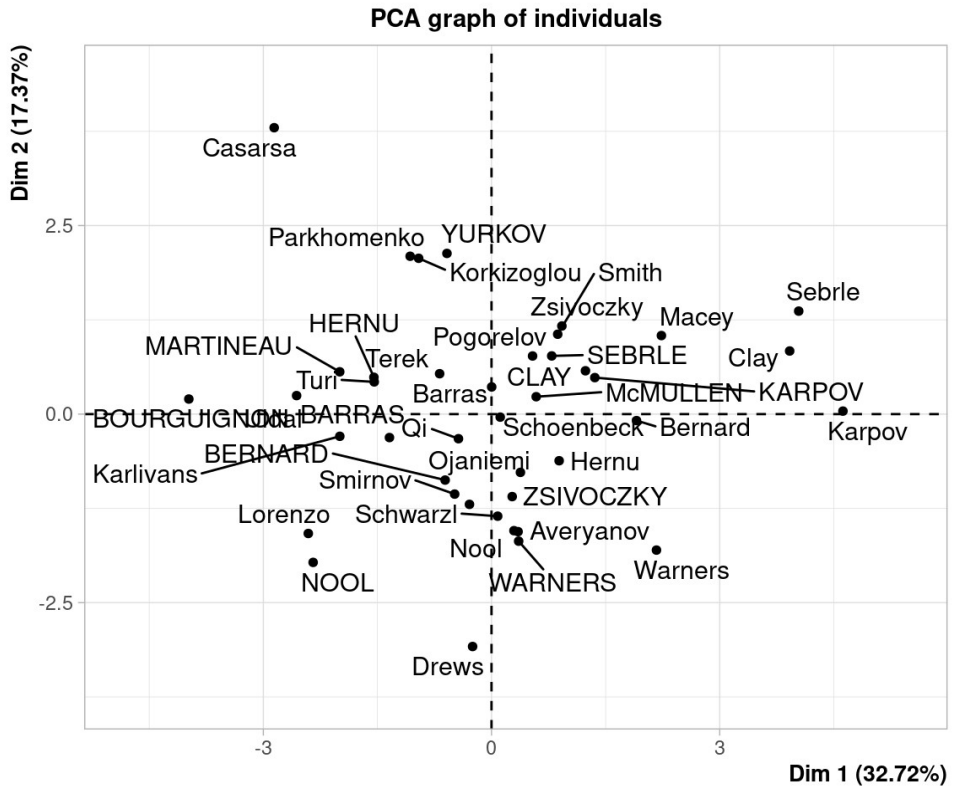
Exemple : les données decathlon

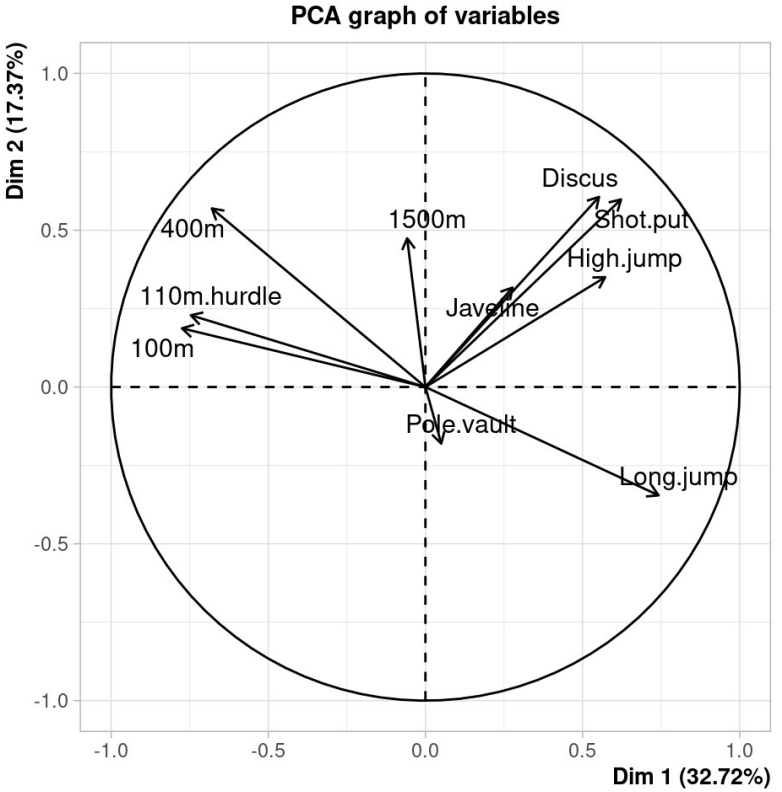
```
R> library(FactoMineR)
R> data(decathlon)

R> library(GGally)
R> ggpairs(decathlon)
```



Exemple : les données decathlon

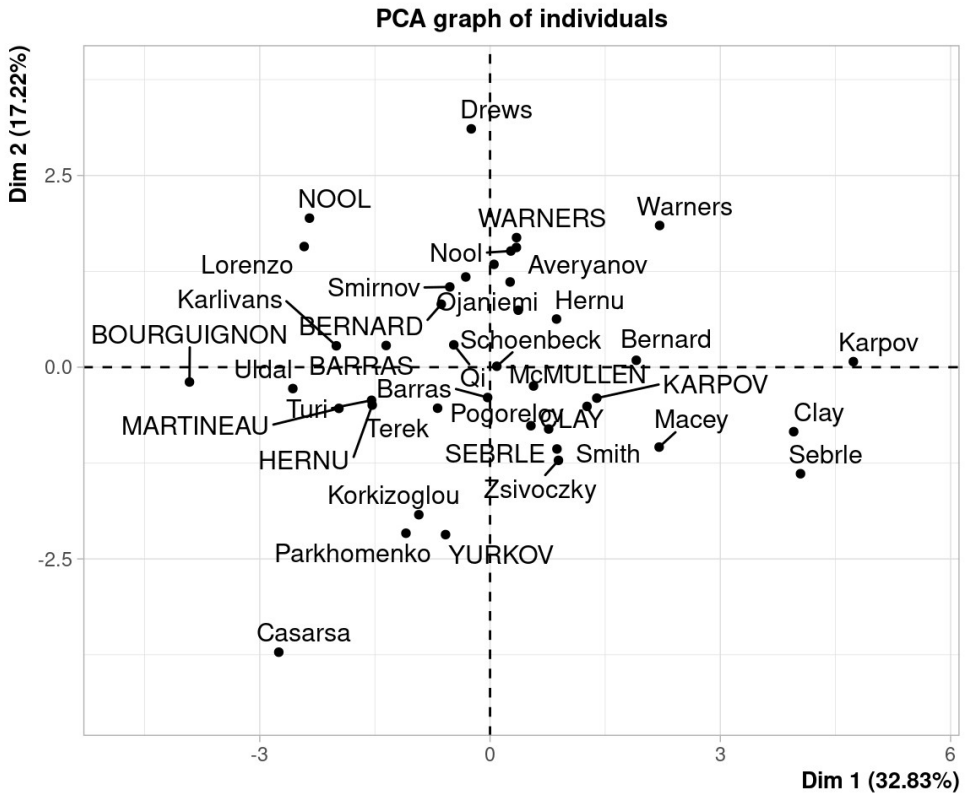
```
R> PCA(decathlon)
```

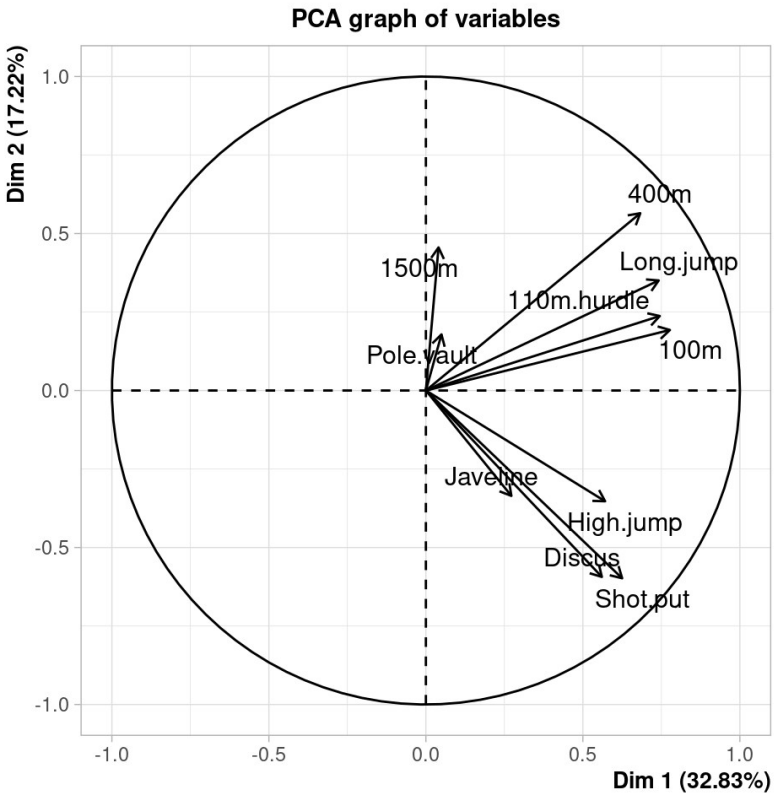




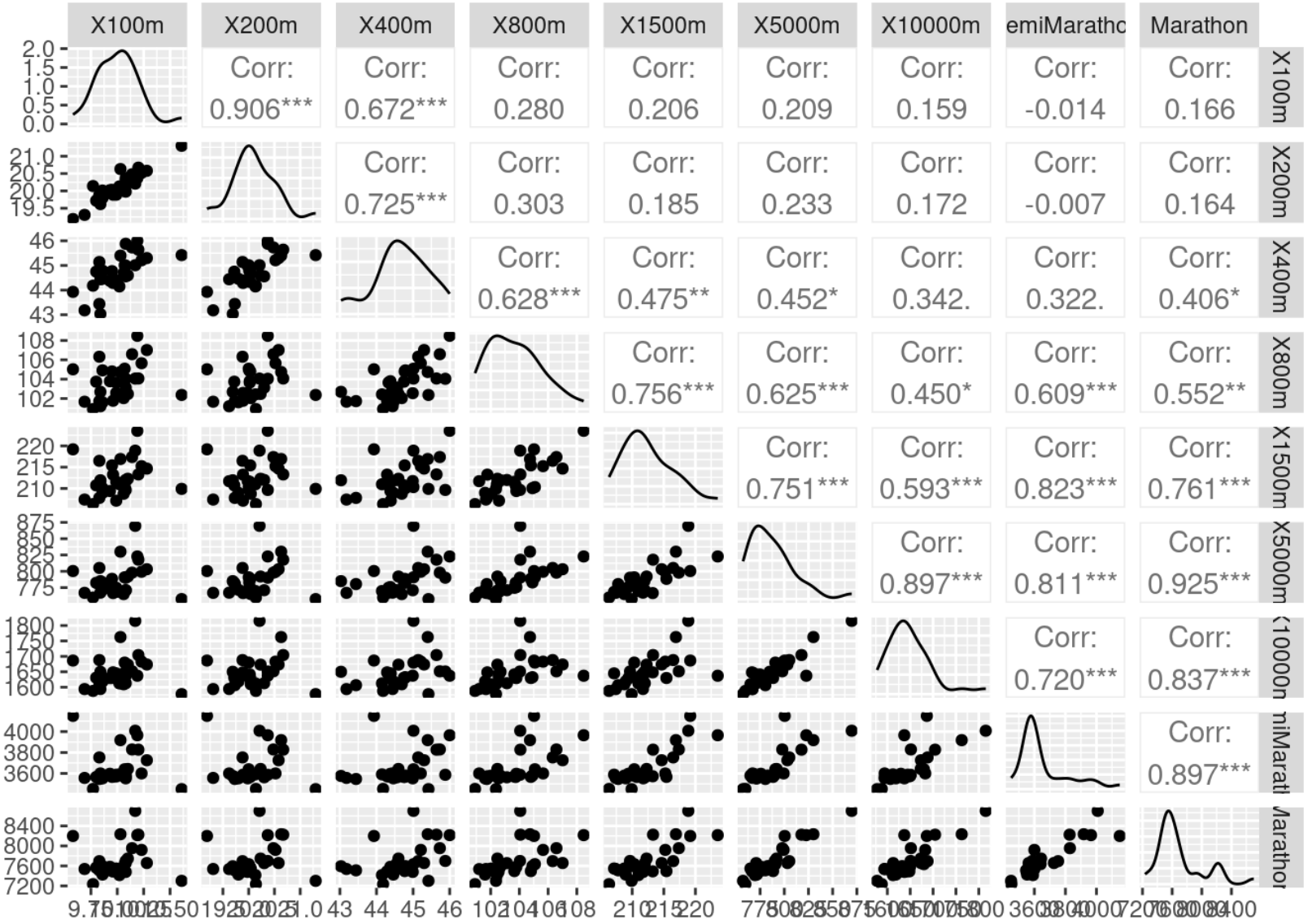
Exemple : les données decathlon

Durées (en s) transformées en vitesse (m/s) pour les épreuves chronométrées

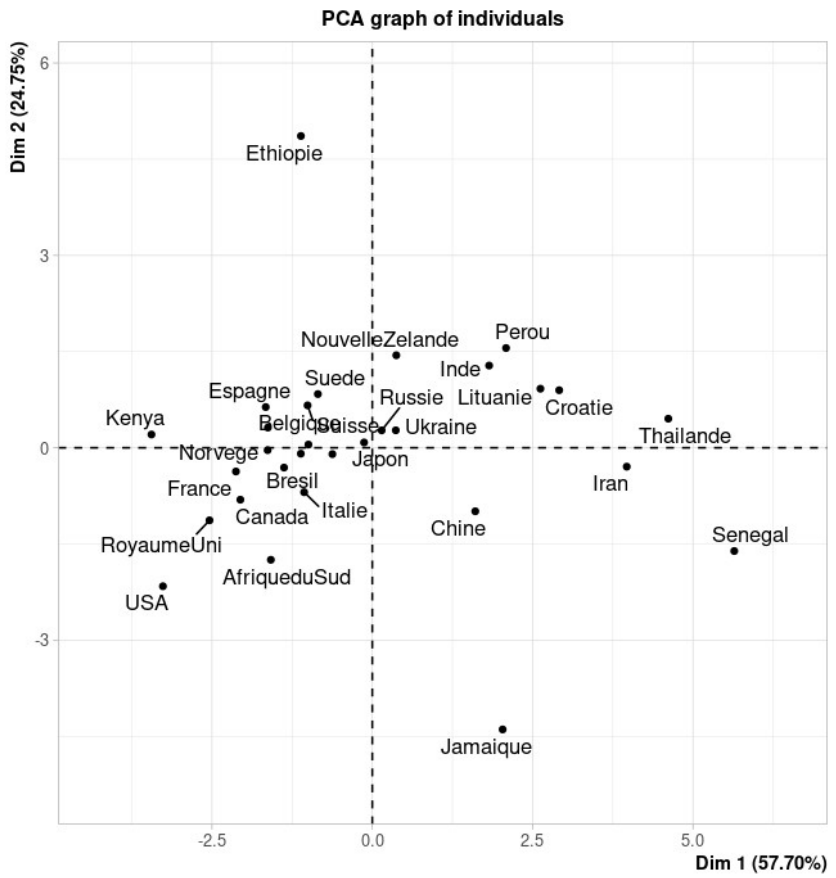


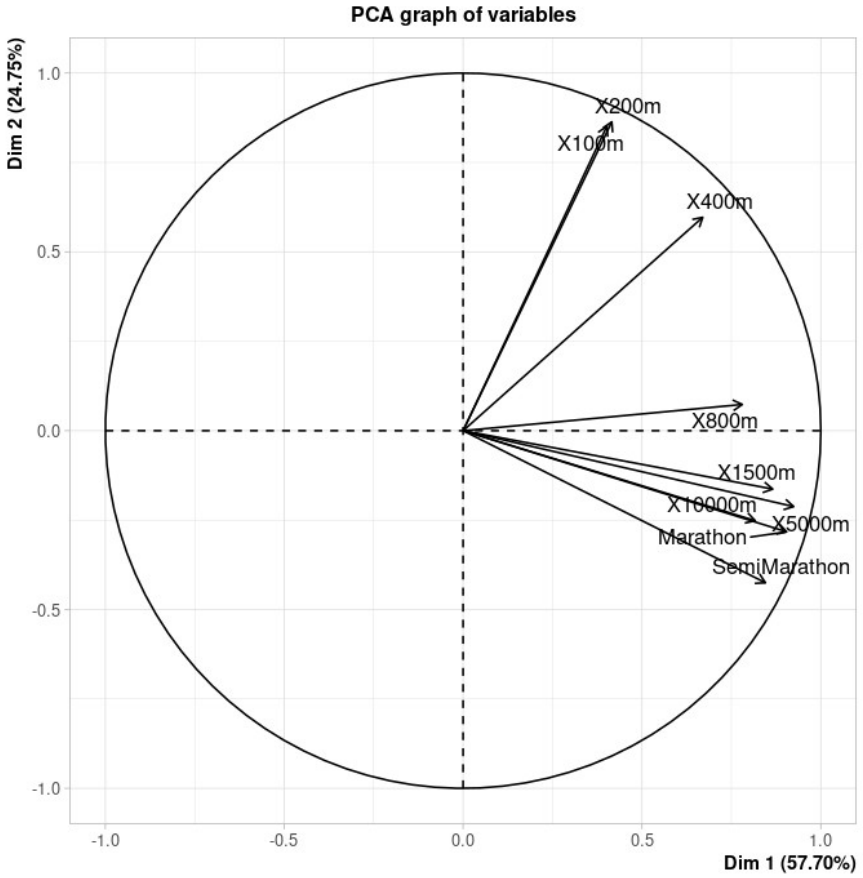


Exemple : les données records_athle

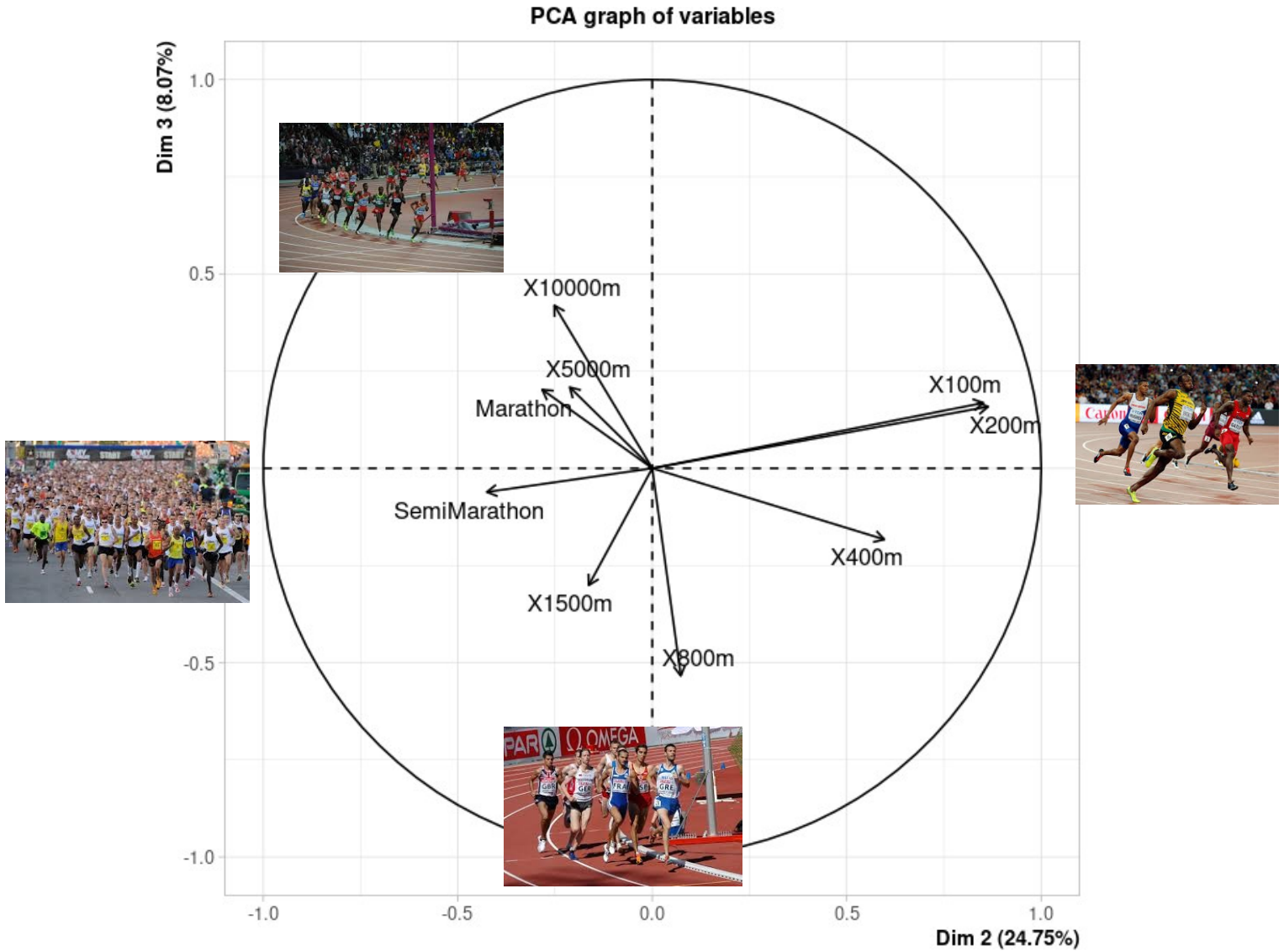


Exemple : les données records_athle



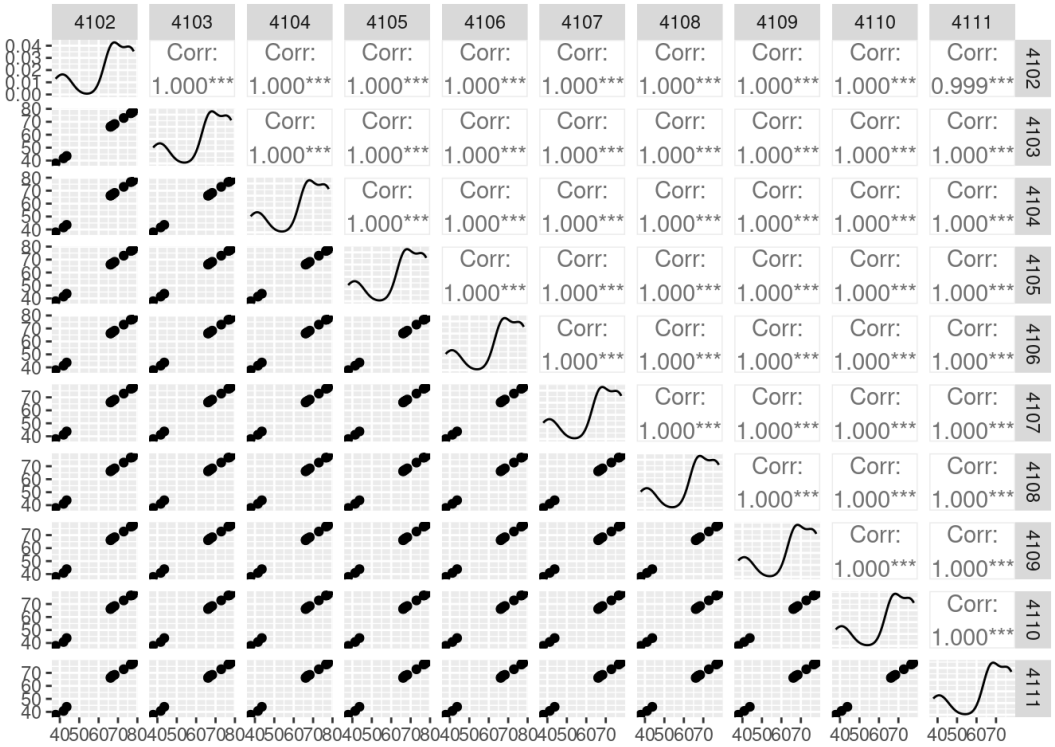


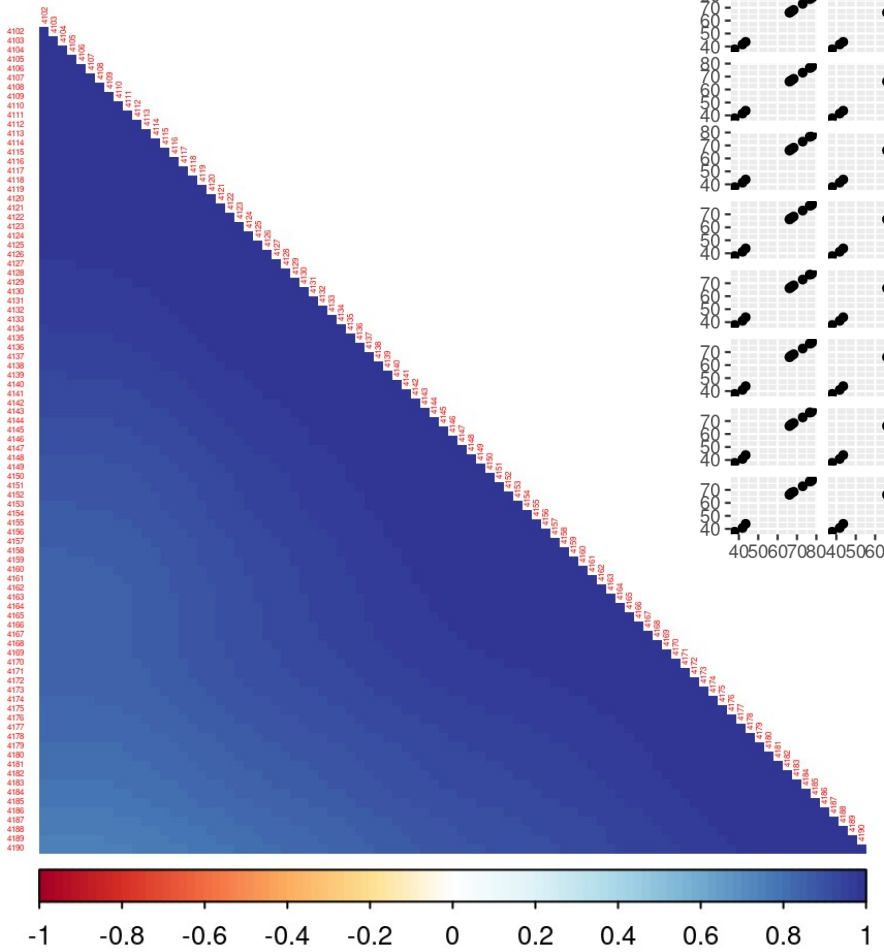
Exemple : les données `records_athle`



Exemple : les données GPS_rugby

Analyse mise en œuvre sur les positions en x seulement (position du joueur dans le sens de la longueur du terrain)



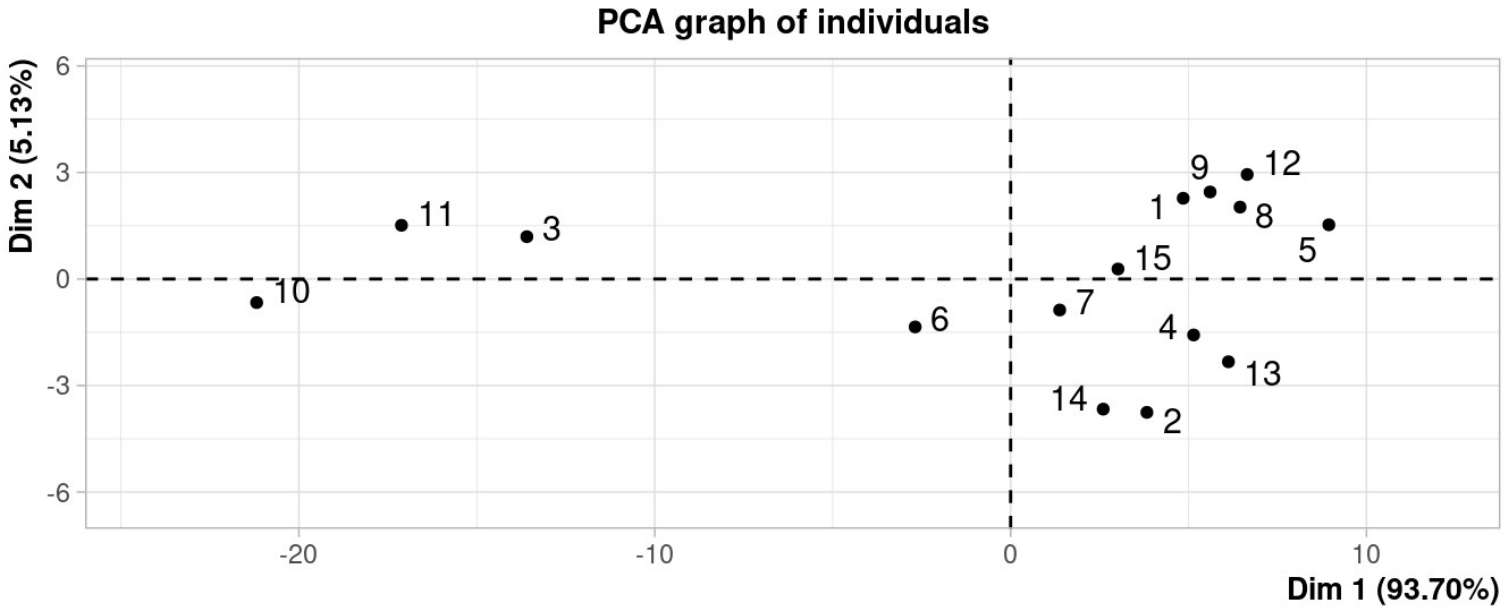


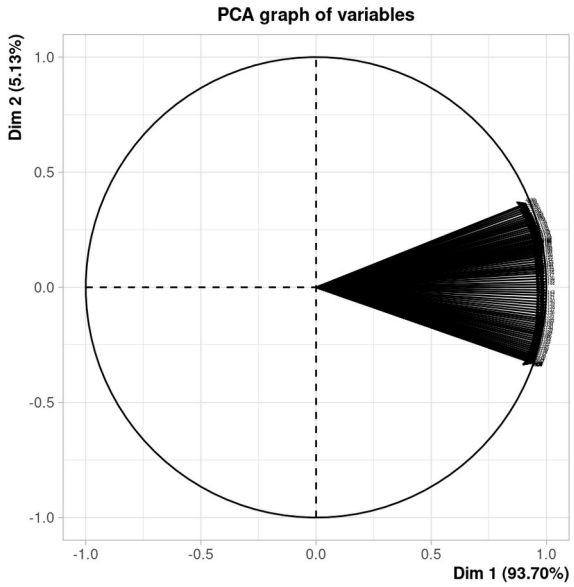
Cascade* réalisée par des professionnels, ne pas reproduire chez vous

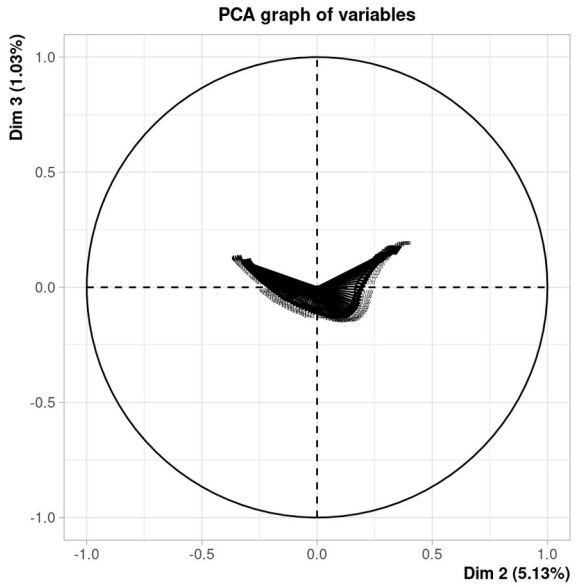
(*) Cascade :
traiter des données fonctionnelles comme si elles étaient multivariées

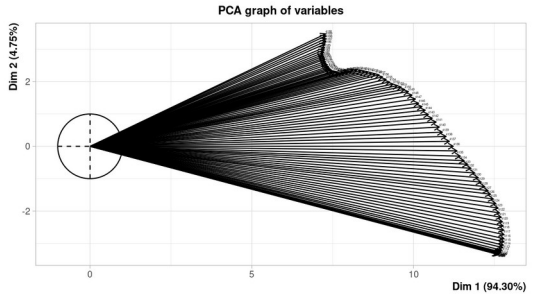
```
R> corrplot(matcor_pearson_GPS_rugby,
            method = "shade",
            type = "lower",
            tl.cex = 0.25)
```

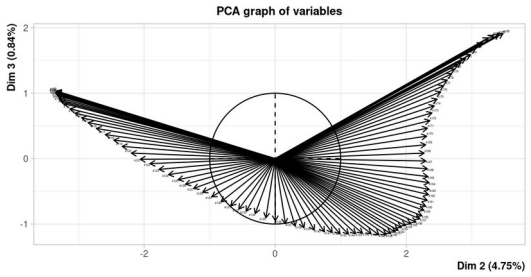
Exemple : les données GPS_rugby





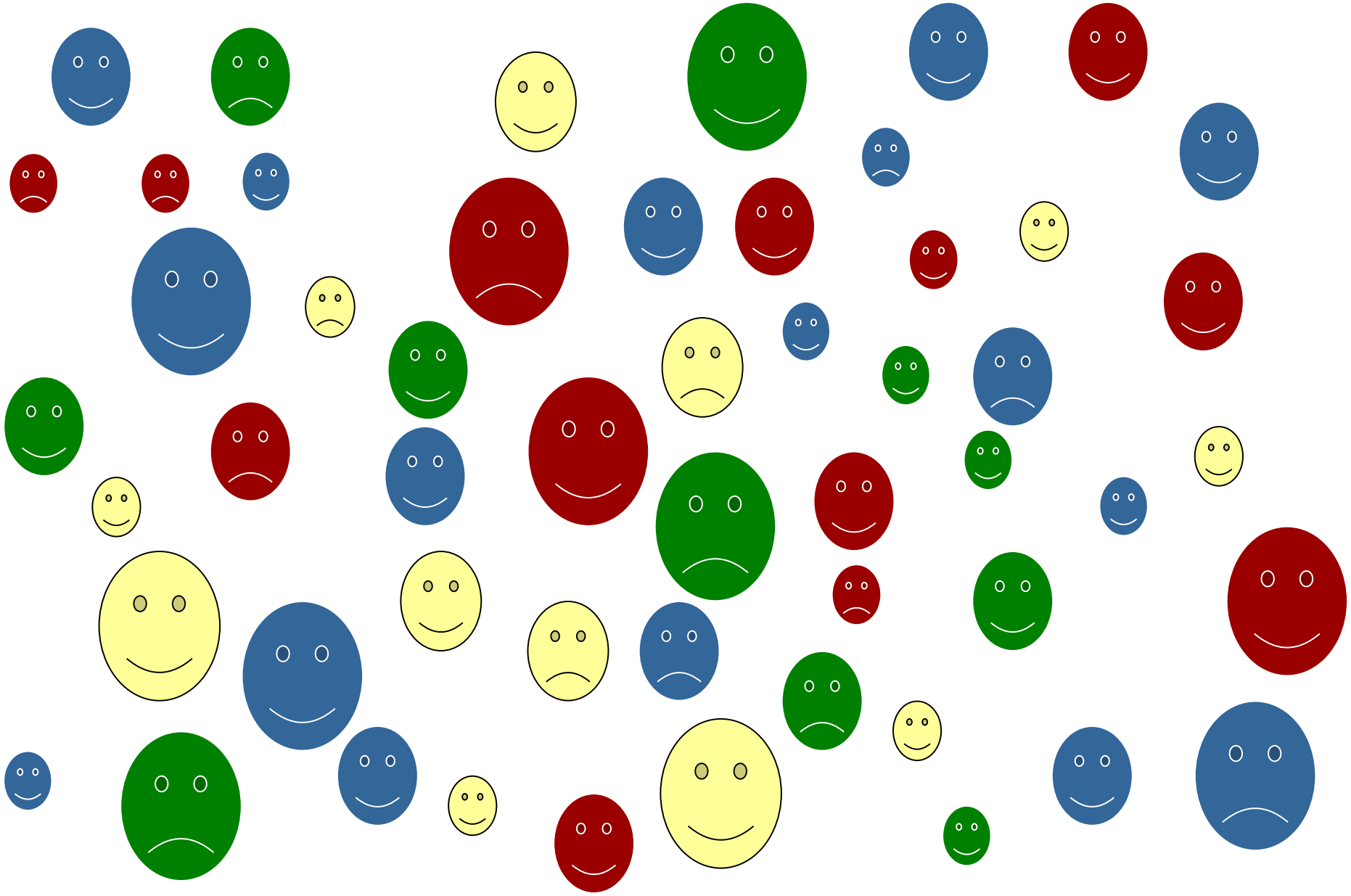






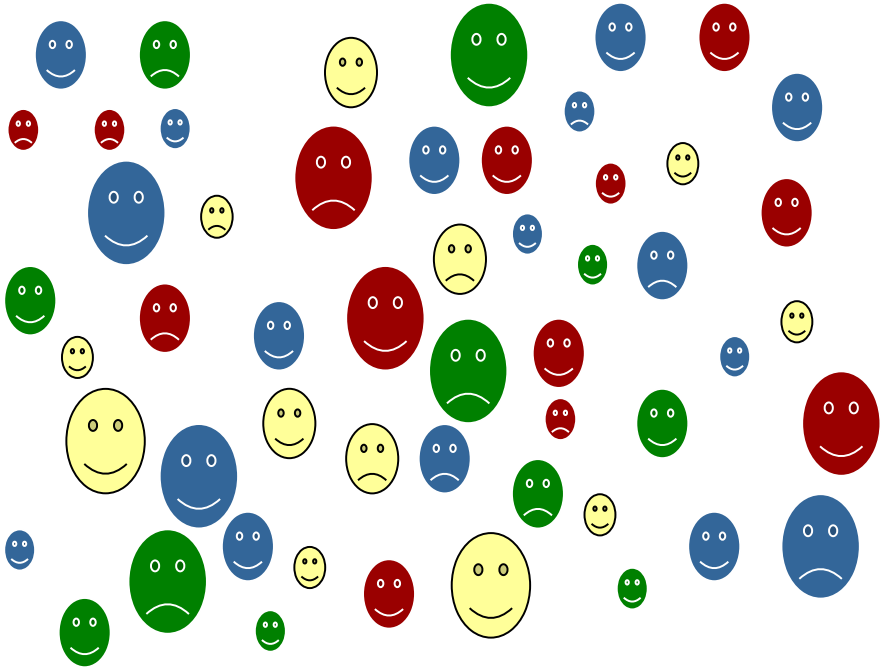
Classification non supervisée

Principe : regrouper des objets qui se ressemblent

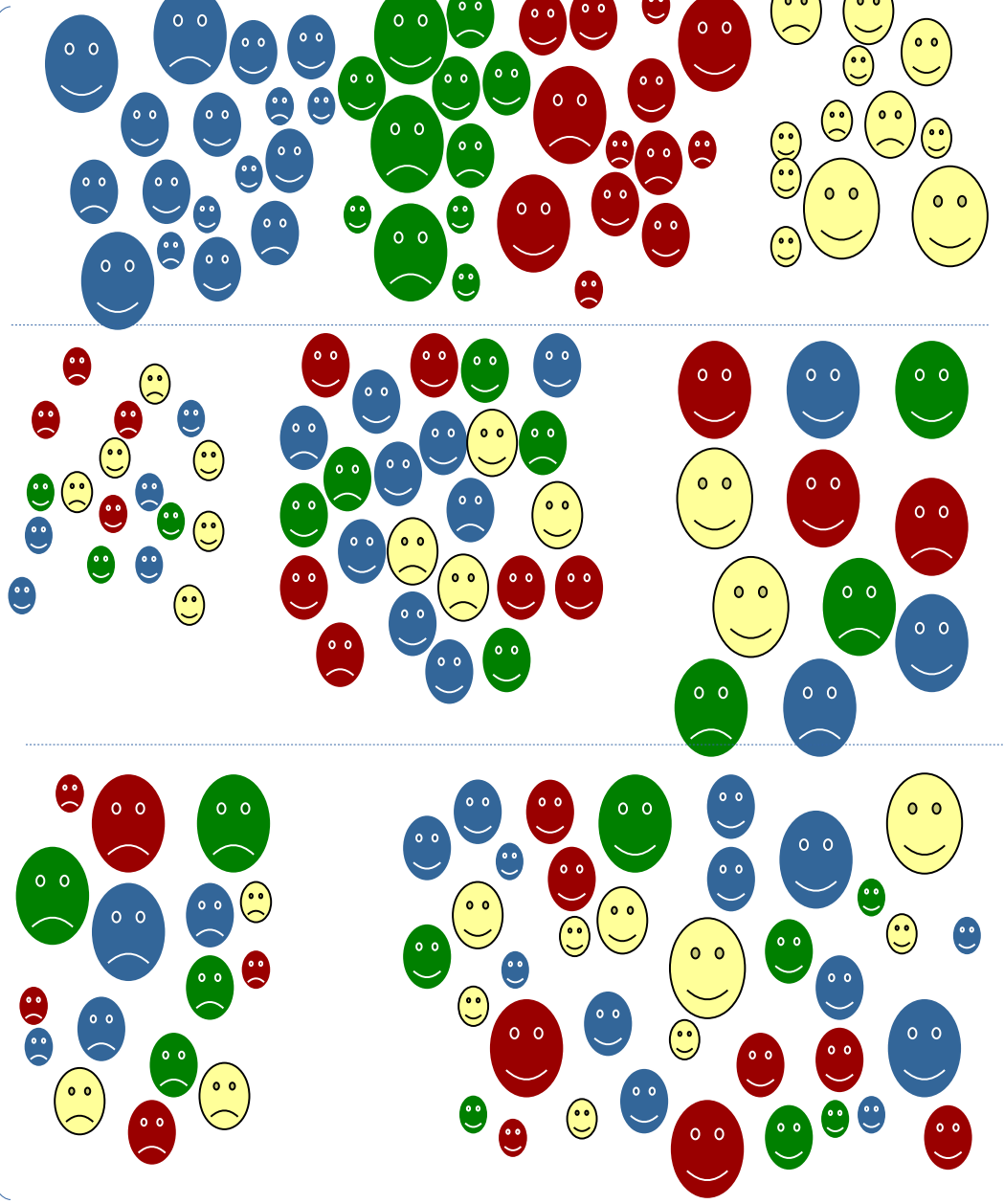


Principe : regrouper des objets qui se ressemblent

Quels sont les groupes homogènes dans mes individus ?



→ ça dépend !



Classification ascendante hiérarchique

- Préalable : choisir
  1) une **distance inter-individus**
  2) un **critère d'agglomération** (distance entre groupes d'individus)

- Procédure itérative
  - **Début** : chaque individu est un groupe
  - **Déroulement** : regroupement des 2 objets les plus proches
  - **Fin** : une classe regroupe tous les individus

- Résultat
  - construction d'un arbre de classification (dendrogramme)

CAH : distance inter-individus

- Distance euclidienne $d_2(x,y)=\sqrt{\sum_{i=1}^{N}(x_i-y_i)^2}$
- Distance Manhattan $d_1(x,y)=\sum_{i=1}^{N}|x_i-y_i|$
- …

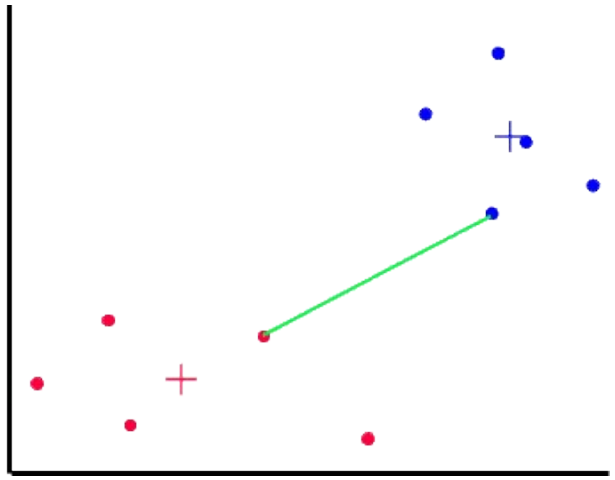
Petit exemple

	tep	tpo	tta	mas	tai		
I1	106.20	89.50	71.50	65.60	174.00		
I2	110.50	97.00	79.00	71.80	175.30		
						Somme	Racine
Diff	-4.30	-7.50	-7.50	-6.20	-1.30		
Diff2	18.49	56.25	56.25	38.44	1.69	171.12	**13.08** = d2(I1,I2)
Absdiff	4.30	7.50	7.50	6.20	1.30	**26.80**	= d1(I1,I2)

CAH : critère d'agglomération (*linkage*)

Minimum

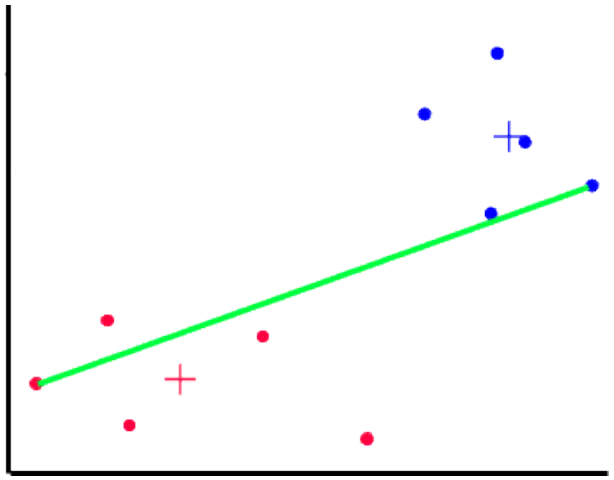
(*single*)



La distance entre 2 groupes est la distance entre les 2 points les plus proches.

Maximum

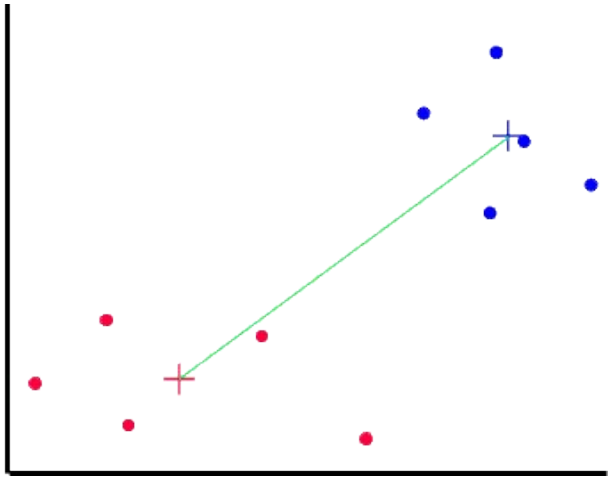
(*complete*)



La distance entre 2 groupes est la distance entre les 2 points les plus éloignés.

Moyen

(*average*)

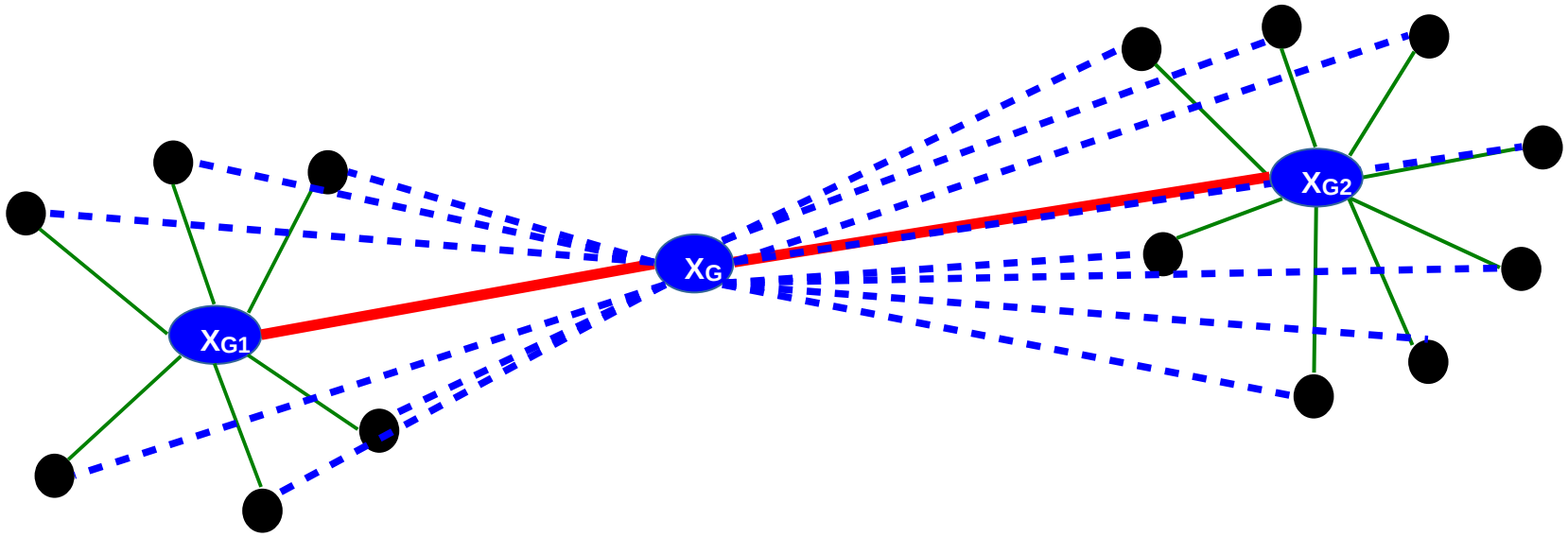


La distance entre 2 groupes est la distance entre les 2 barycentres de chaque groupe.

CAH : critère d'agglomération (*linkage*)

$$\sum_{j=1}^{p}\sum_{i=1}^{n}\left(X_{ij}-X_G\right)^2 = \sum_{j=1}^{p}\left(X_{Gj}-X_G\right)^2 + \sum_{j=1}^{p}\sum_{i=1}^{n}\left(X_{ij}-X_{Gj}\right)^2$$

Critère de Ward (décomposition de l'inertie totale)



L'affectation d'un point à une classe est effectuée selon la règle :

Minimiser l'inertie intra-classe <=> Maximiser l'inertie inter-classes

CAH : construction d'un dendrogramme

Données
5 individus
3 variables

	V1	V2	V3
I1	1	2	3
I2	4	2	5
I3	4	3	7
I4	8	9	6
I5	4	2	3

Effectuer la classification hiérarchique des individus avec la distance euclidienne et le critère de saut minimum.

Étape 0

Calcul des distances inter-individus

$$d(I1,I2)=\sqrt{(1-4)^2}+(2-2)^2+(3-5)^2=\sqrt{(3^2+0^2+2^2)}=\sqrt{13}$$

	I1	I2	I3	I4
I2	3.61			
I3	5.10	2.24		
I4	10.34	8.12	7.28	
I5	3.00	2.00	4.12	8.60

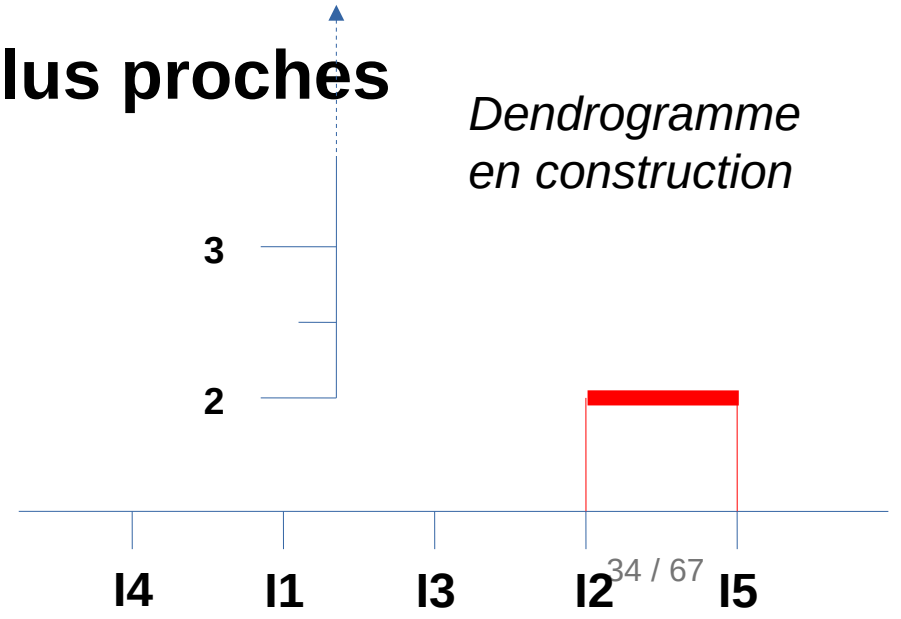
À chaque étape, regroupement des 2 objets les plus proches

Étape 1

- Regroupement des individus I2 et I5 pour former le nœud N1 à la hauteur 2
- Avant de passer à l'étape suivante, il faut calculer les distances entre les nouveaux objets (I1, I3, I4 et **N1**) grâce au critère d'agglomération.

```
d(I1,N1) = min{d(I1,I2),d(I1,I5)} = min(3.61,3) = 3
d(I3,N1) = min{d(I3,I2),d(I3,I5)} = min(2.24,4.12) = 2.24
d(I4,N1) = min{d(I4,I2),d(I4,I5)} = min(8.12,8.60) = 8.12
```

Dendrogramme en construction



CAH : construction d'un dendrogramme

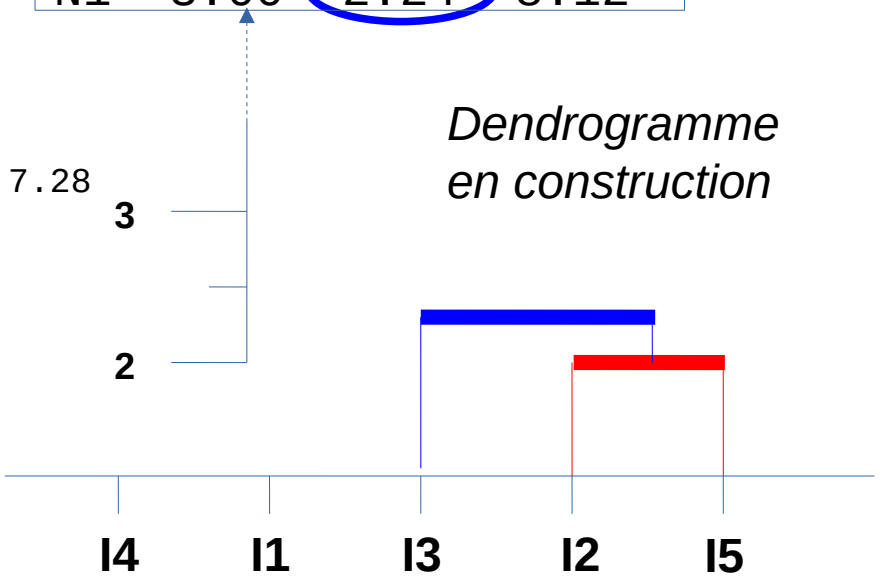
Étape 2

- Regroupement des objets I3 et N1 pour former le nœud N2 à la hauteur 2.24
- Comme précédemment, on calcule les distances entre I1, I4 et N2.

```
d(I1,N2) = min{d(I1,I3),d(I1,N1)} = min(5.10,3) = 3
d(I4,N2) = min{d(I4,I3),d(I4,N1)} = min(7.28,8.12) = 7.28
```

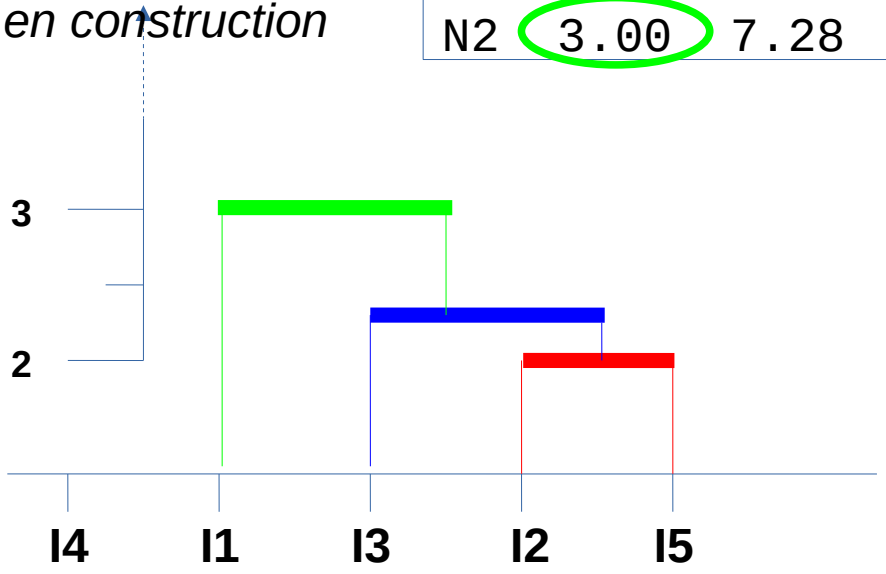
	I1	I3	I4
I3	5.10		
I4	10.34	7.28	
N1	3.00	2.24	8.12

Dendrogramme en construction



	I1	I4
I4	10.34	
N2	3.00	7.28

Dendrogramme en construction



Étape 3

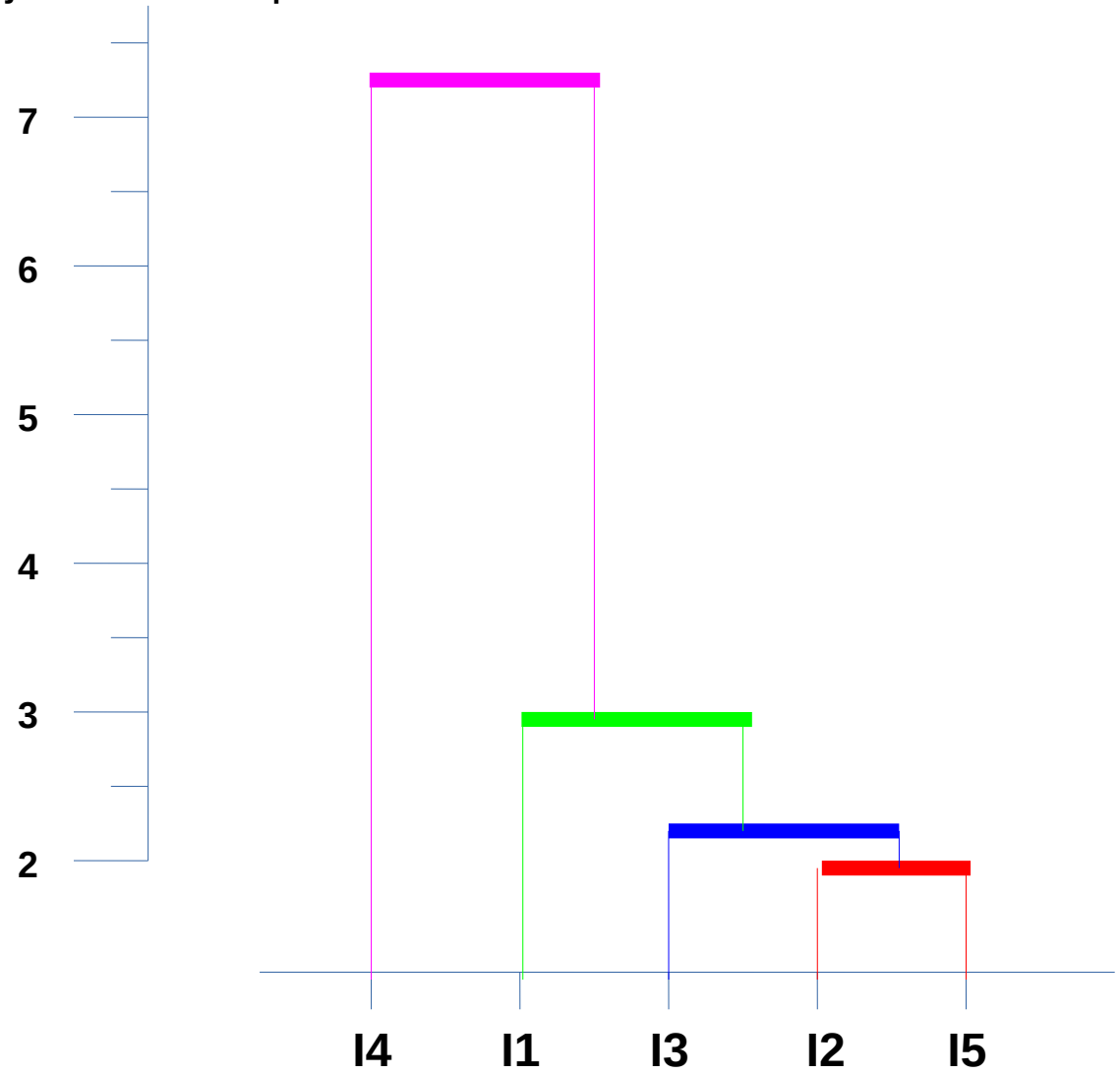
- Regroupement des objets I1 et N2 pour former le nœud N3 à la hauteur 3.
- Il reste à calculer la distance entre I4 et N3.

```
d(I4,N3) = min{d(I4,I1),d(I4,N2)} = min(10.34,7.28) = 7.28
```

CAH : construction d’un dendrogramme

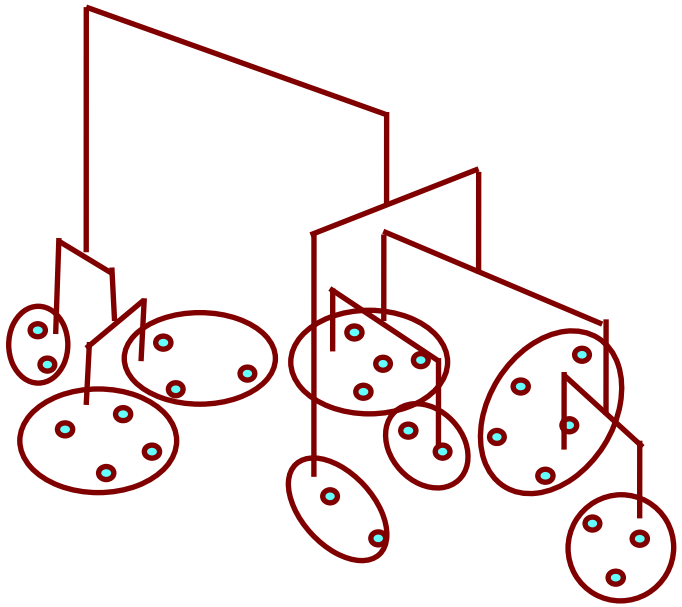
Étape 4 (fin de l'algorithme)

- Regroupement des objets I4 et N3 pour former le nœud terminal N4 à la hauteur 7.28

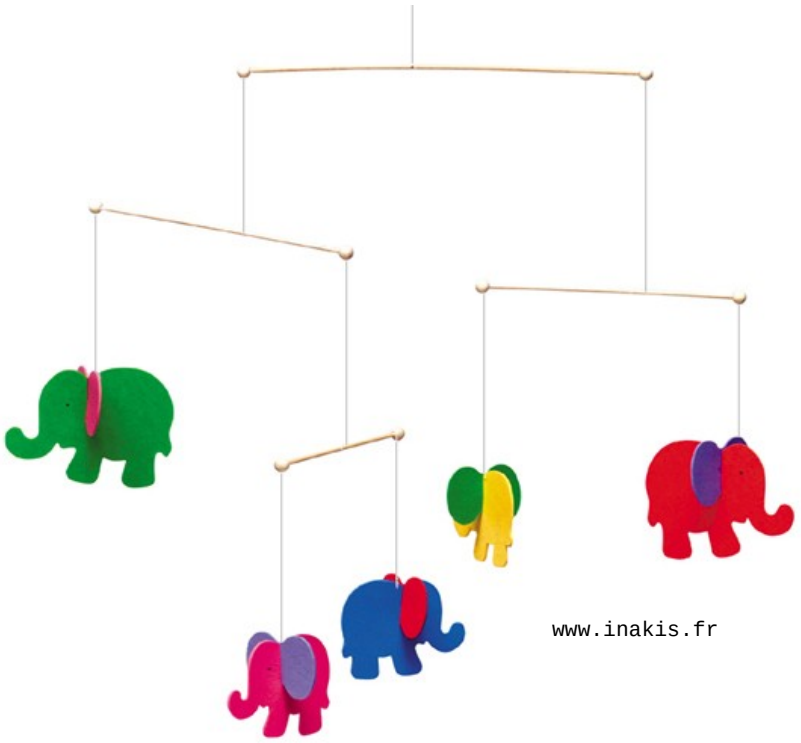


	V1	V2	V3
I1	1	2	3
I2	4	2	5
I3	4	3	7
I4	8	9	6
I5	4	2	3

CAH : interprétation d’un dendrogramme

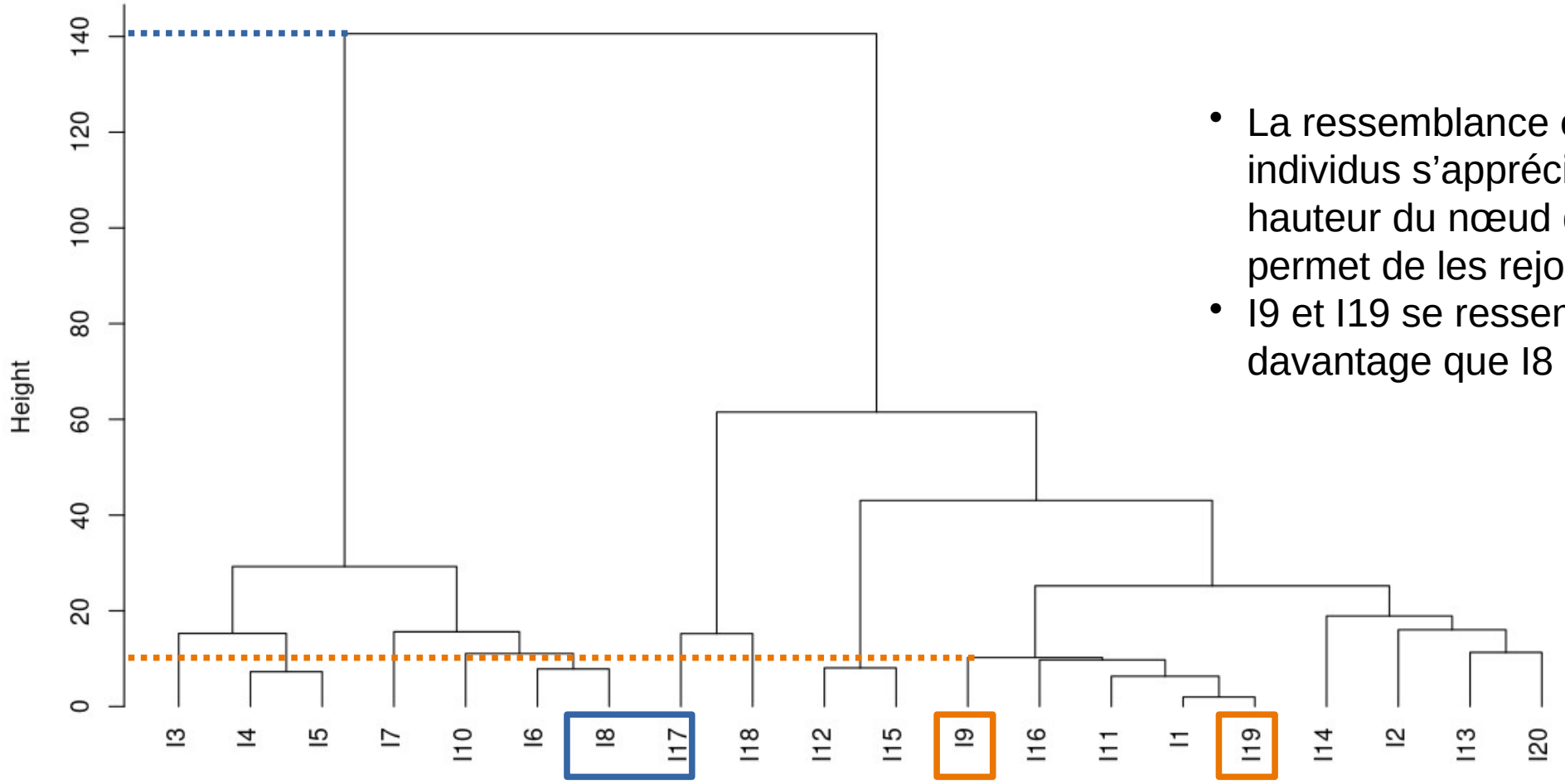


Un dendrogramme doit être vu comme un mobile pour bébé (le truc que l’on pend au dessus d’un berceau et dont les branches peuvent tourner). Ainsi, les « distances » entre les individus ne sont valables qu’en remontant les branches pas selon leur proximité « visuelle ».



CAH : interprétation d'un dendrogramme

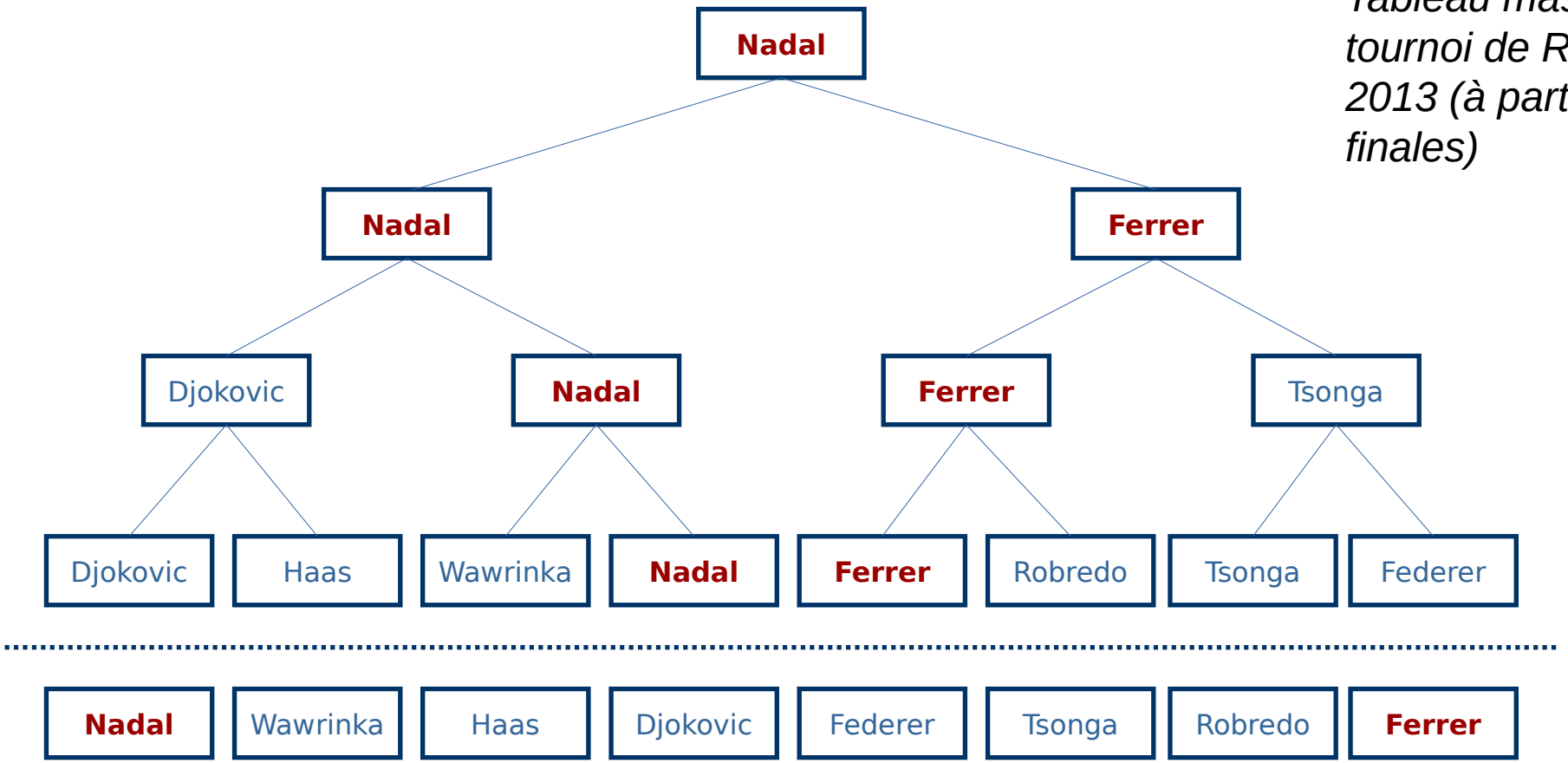
Classification des individus : distance euclidienne, critère de Ward



- La ressemblance entre 2 individus s'apprécient à la hauteur du nœud qui permet de les rejoindre
- I9 et I19 se ressemblent davantage que I8 et I17

CAH : interprétation d'un dendrogramme

Tableau masculin du tournoi de Roland Garros 2013 (à partir des ¼ de finales)

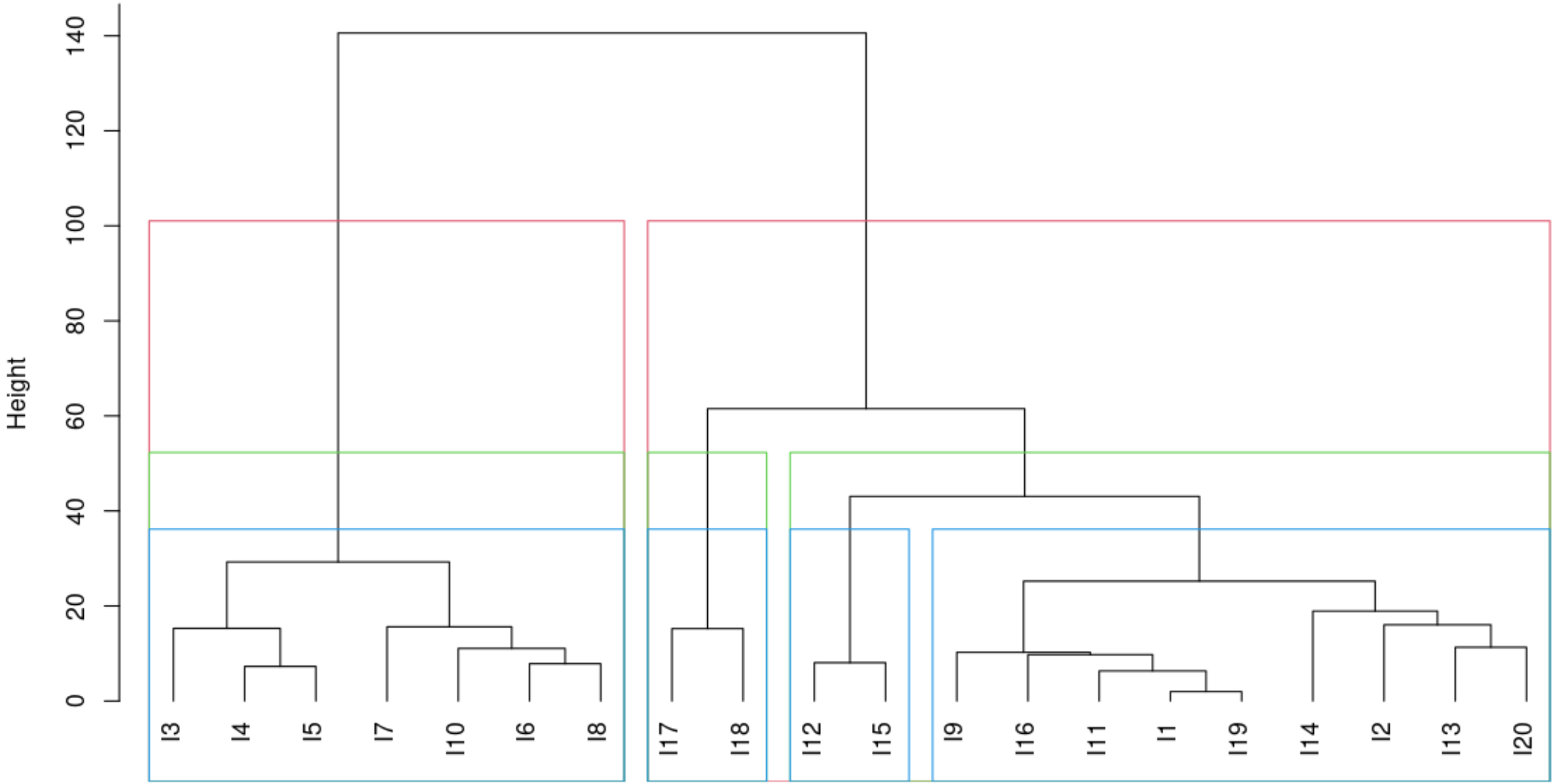


Ferrer et Nadal ne pouvaient se rencontrer qu'en finale bien que leur position « visuelle » dans le tableau initial puissent être différentes, soit proches soit à l'opposé.

CAH : interprétation d’un dendrogramme

- Un dendrogramme fournit une classification des éléments lorsque l'on se donne une « hauteur de coupe » de l'arbre.
- Plus l'arbre est coupé « bas » (proche des éléments initiaux) plus la classification obtenue est fine.
- Une hauteur de coupe est pertinente si elle se trouve entre 2 noeuds dont les hauteurs sont « relativement » éloignées.

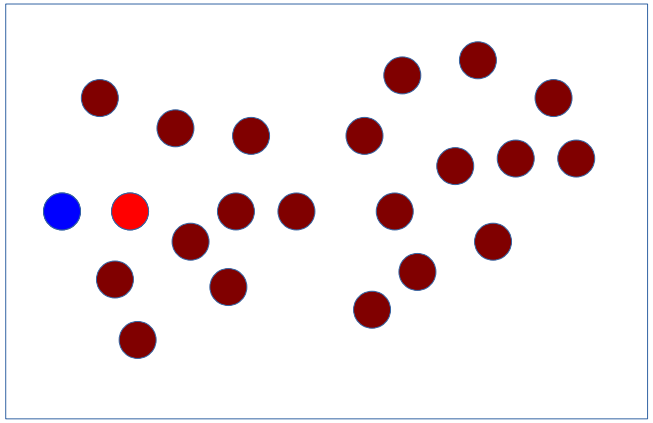
CAH : interprétation d’un dendrogramme



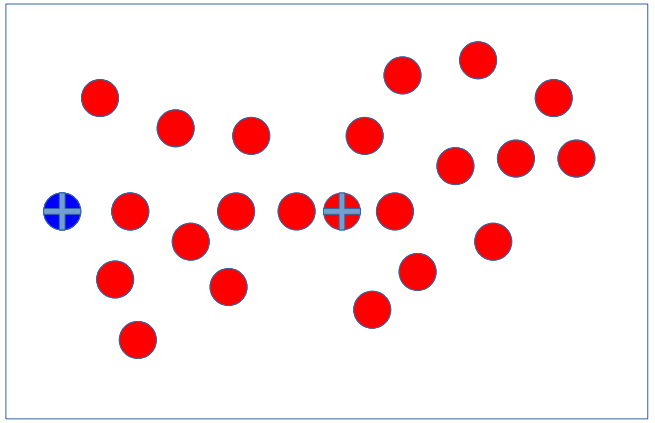
Agrégation autour de centres mobiles (*k-means*)

- Préalable : déterminer le nombre de groupes (k) soit par une connaissance a priori du phénomène étudié, soit par une autre méthode (classification hiérarchique par exemple)
- Procédure itérative
  - **Début** : k centres (tirés aléatoirement ou imposés)
  - **Déroulement** :
    - tous les individus sont affectés au centre le plus proche
    - les centres de chaque groupe sont recalculés
  - **Fin** : les individus ne changent pas de groupe entre 2 étapes successives
- Résultat
  - répartition des individus en k groupes

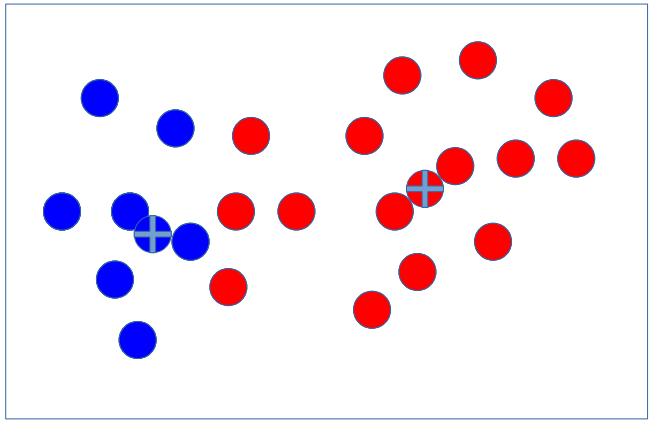
Agrégation autour de centres mobiles



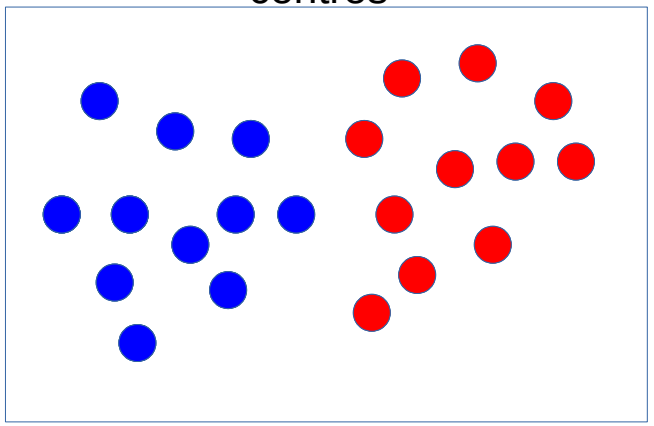
Tirage aléatoire de 2 centres initiaux



Affectation de chaque point au centre le plus proche et calcul des nouveaux centres



Affectation et calcul des nouveaux centres



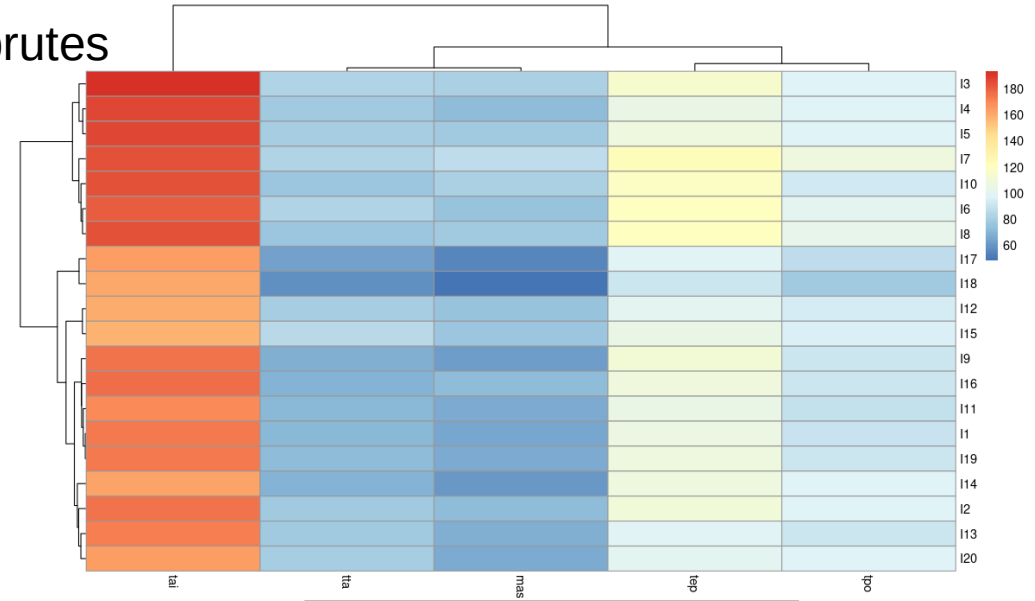
Dernière affectation

Représentation *heatmap*

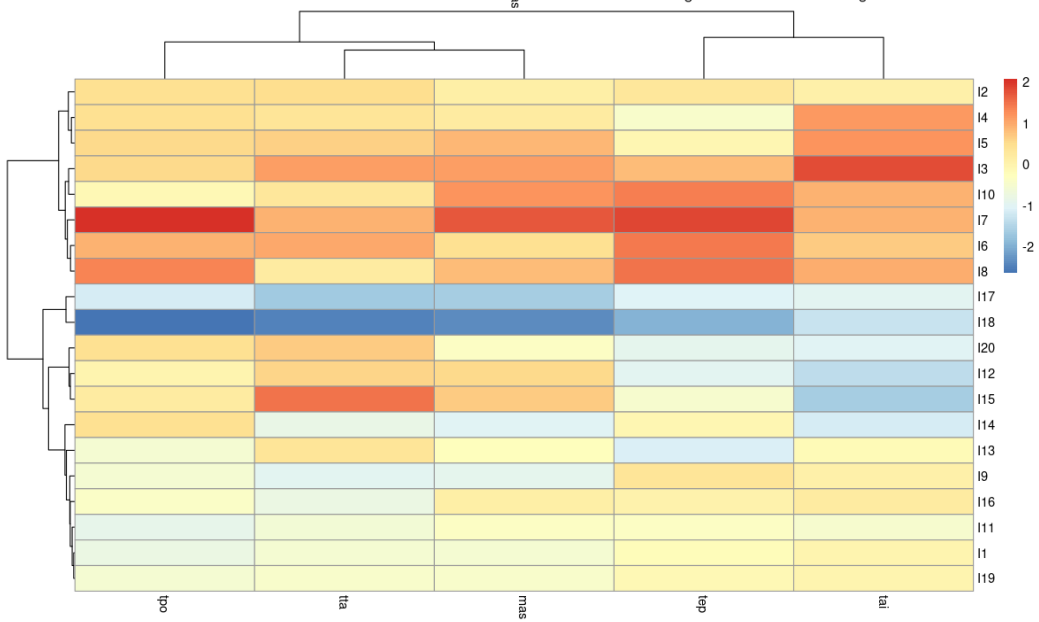
```
R> library(pheatmap)
R> pheatmap(donnees,
          clustering_method = "ward.D")
```

- Classification des individus et des variables
- Représentation *heatmap* : les dendrogrammes sont représentés autour d'une image de la matrice des données
- L'interprétation du graphique obtenu vise à mettre en relation la classification des individus et des variables

Données brutes



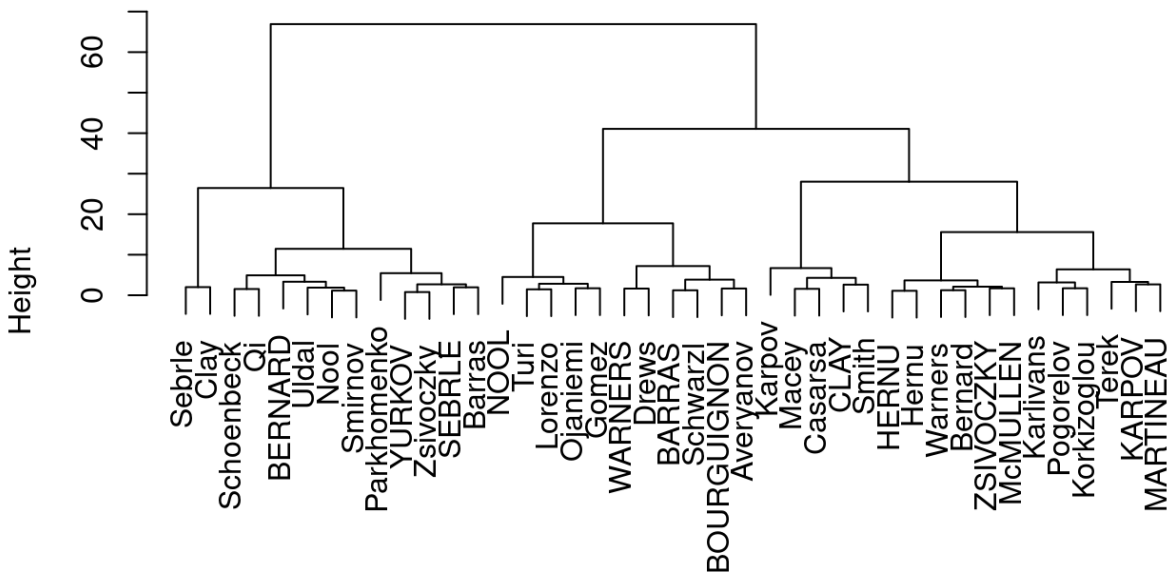
Données centrées-réduites



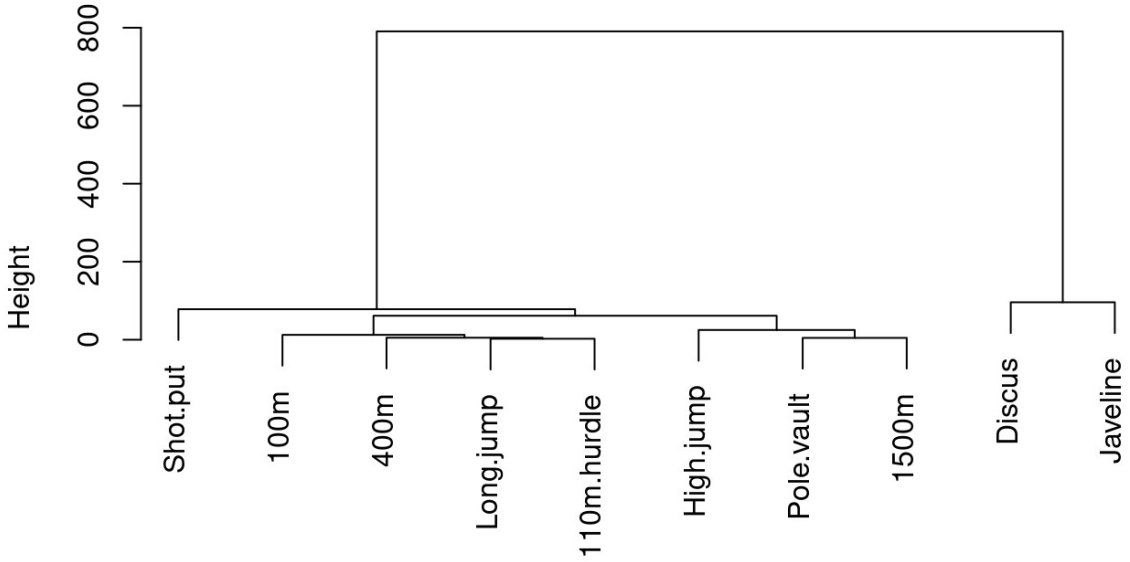
CAH : les données decathlon

Durées (en s) transformées en vitesse (m/s) pour les épreuves chronométrées

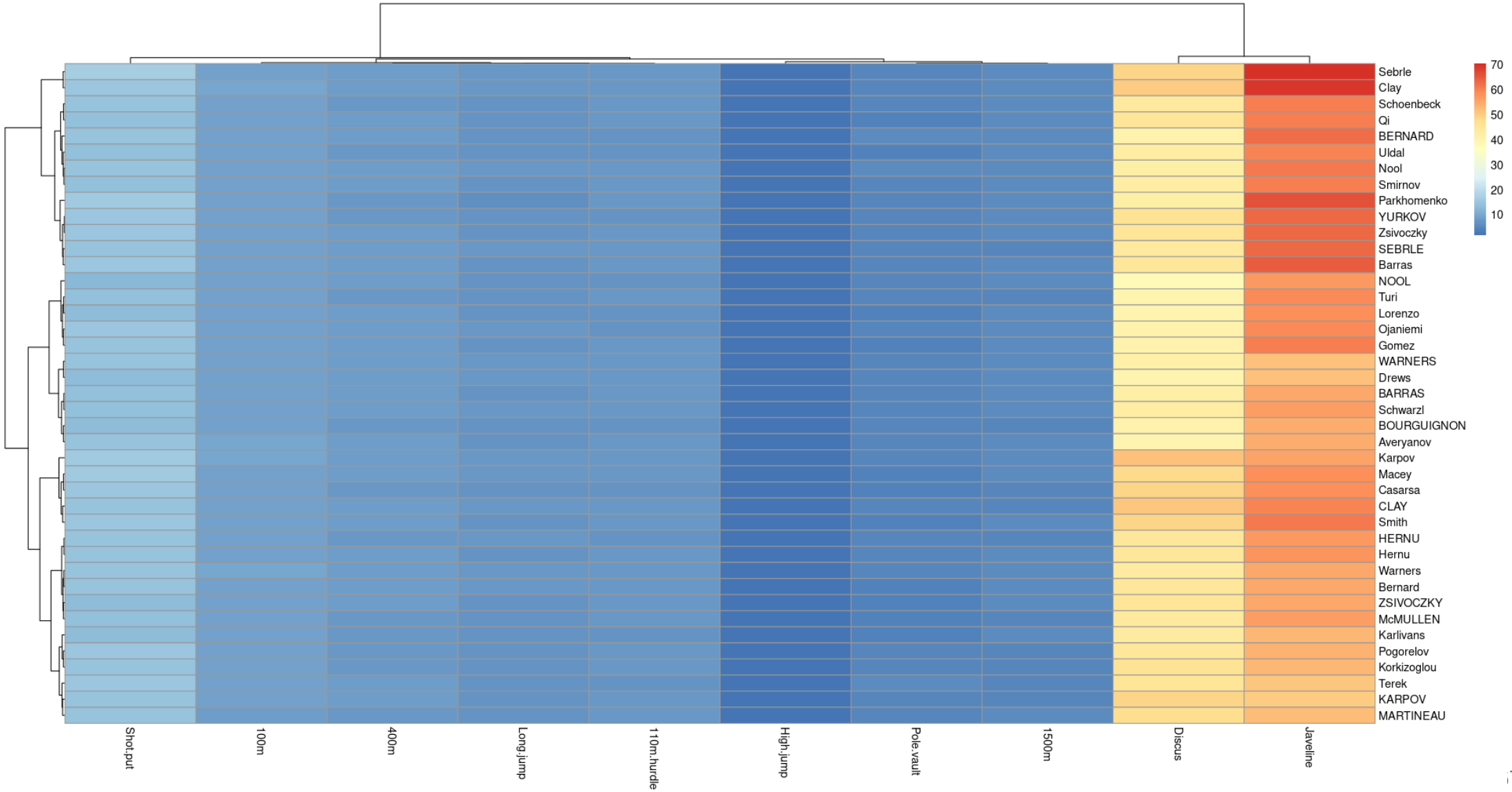
Classification des athlètes



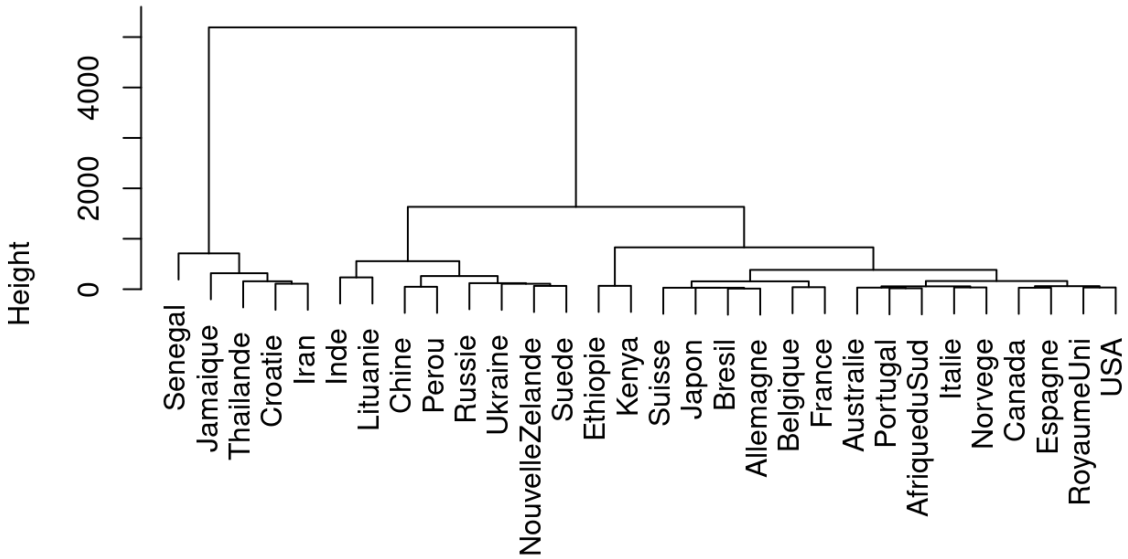
Classification des épreuves

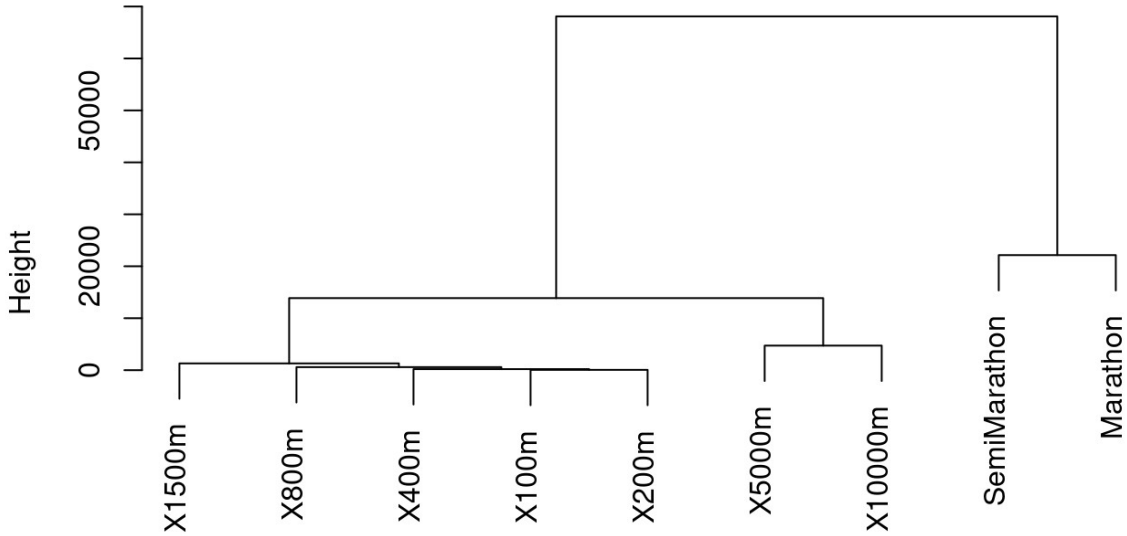


CAH : les données decathlon

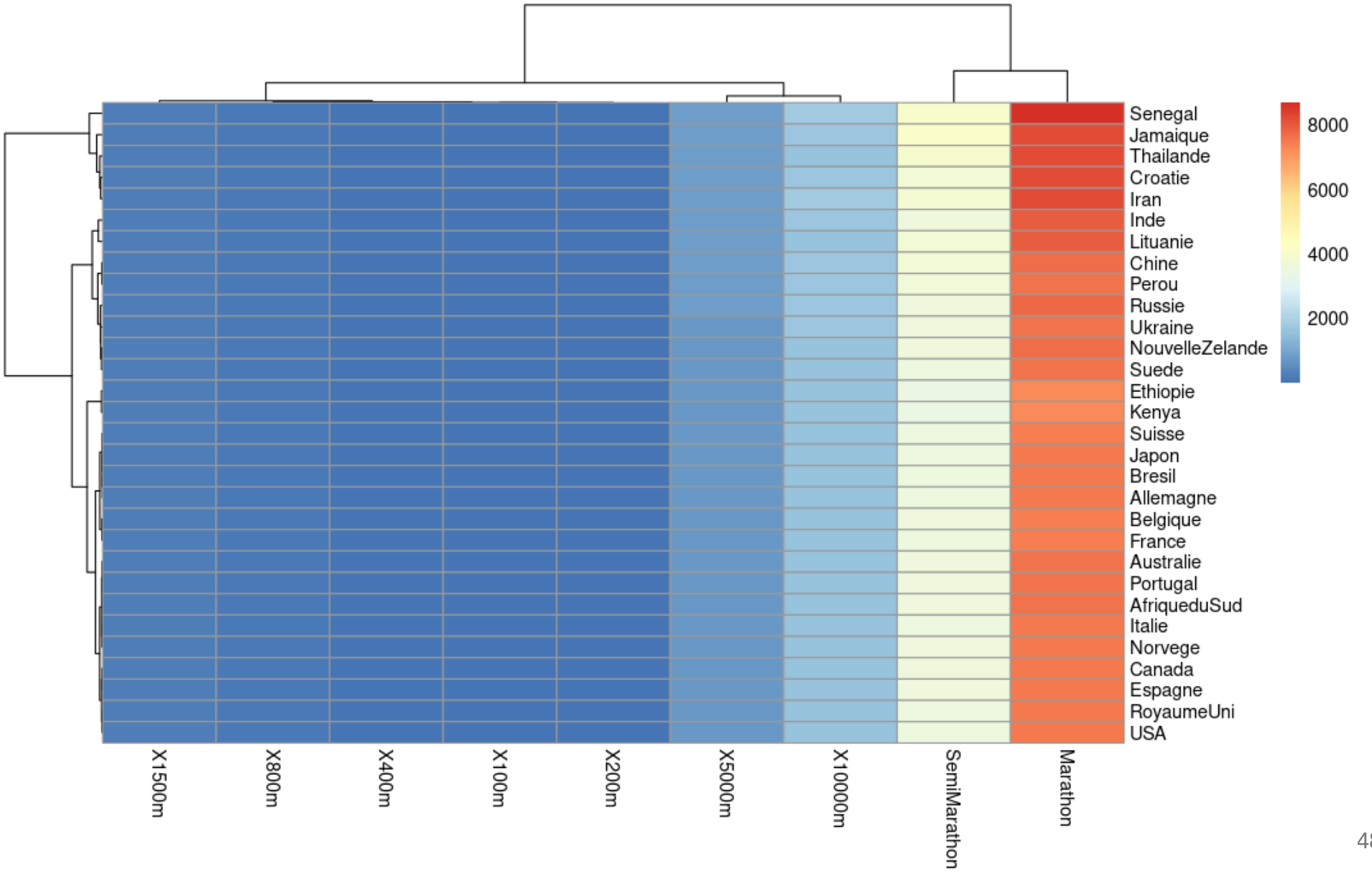


CAH : les données records_athle



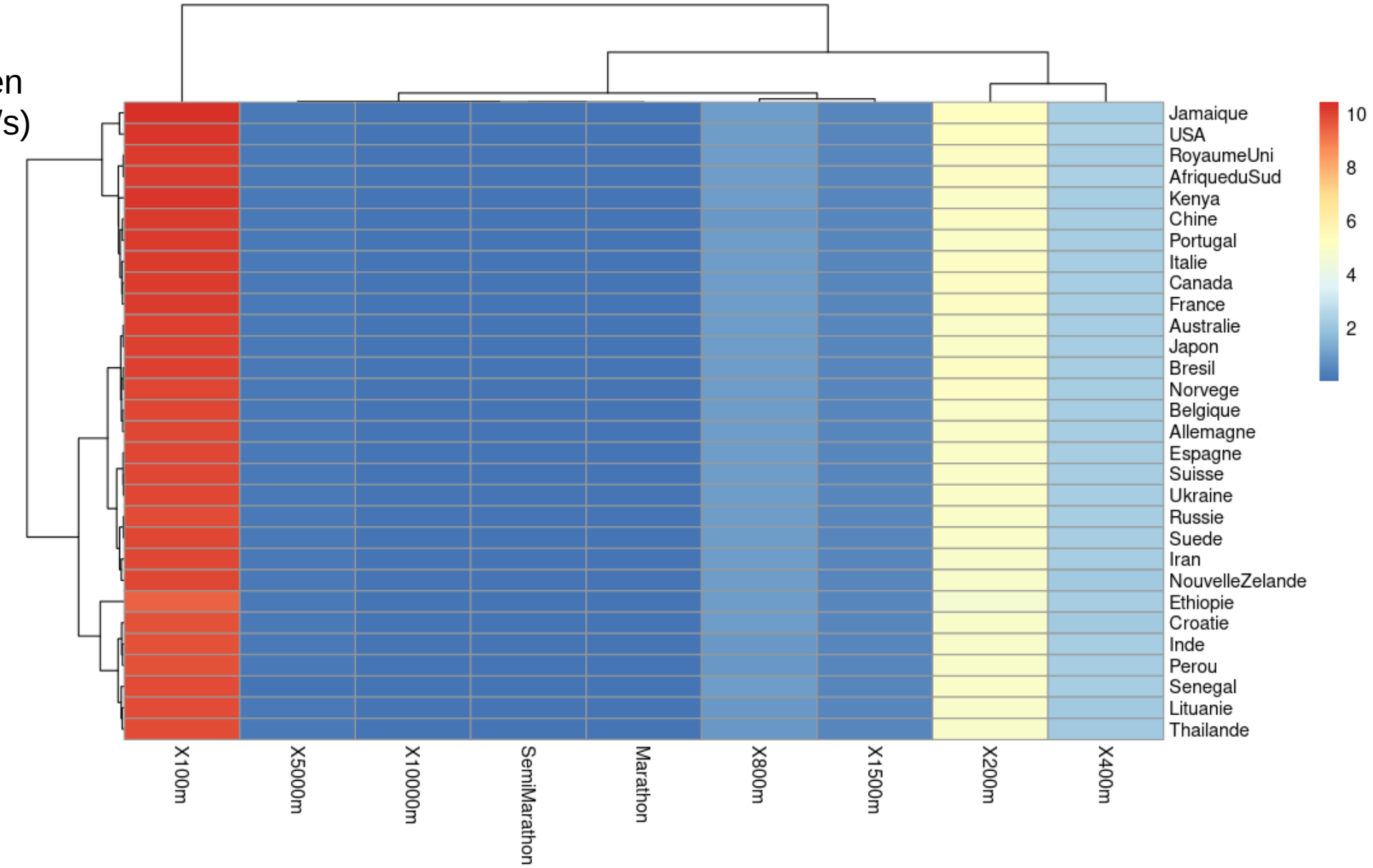


CAH : les données records_athle



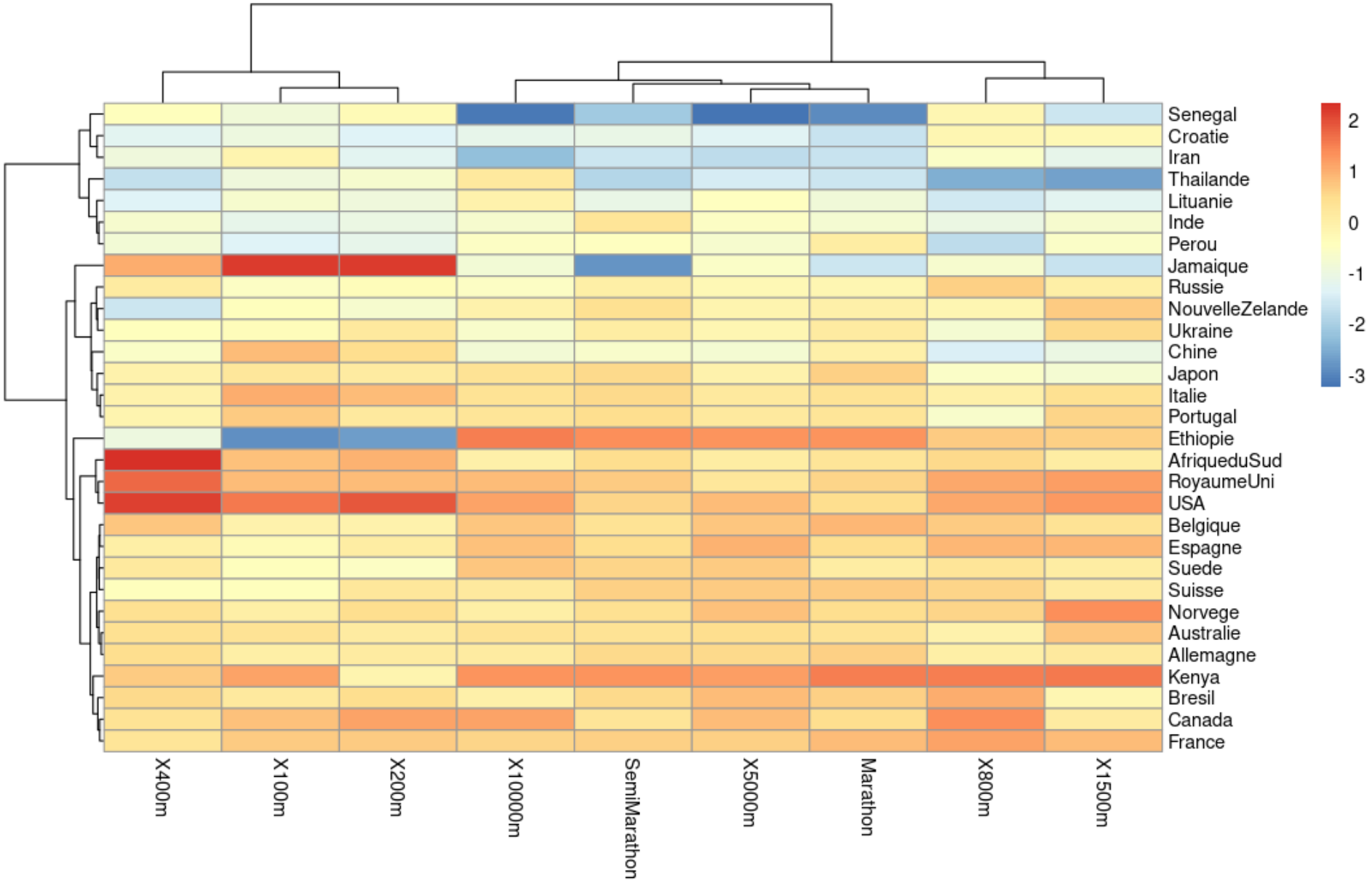
CAH : les données records_athle

Durées (s) converties en vitesses (m/s)

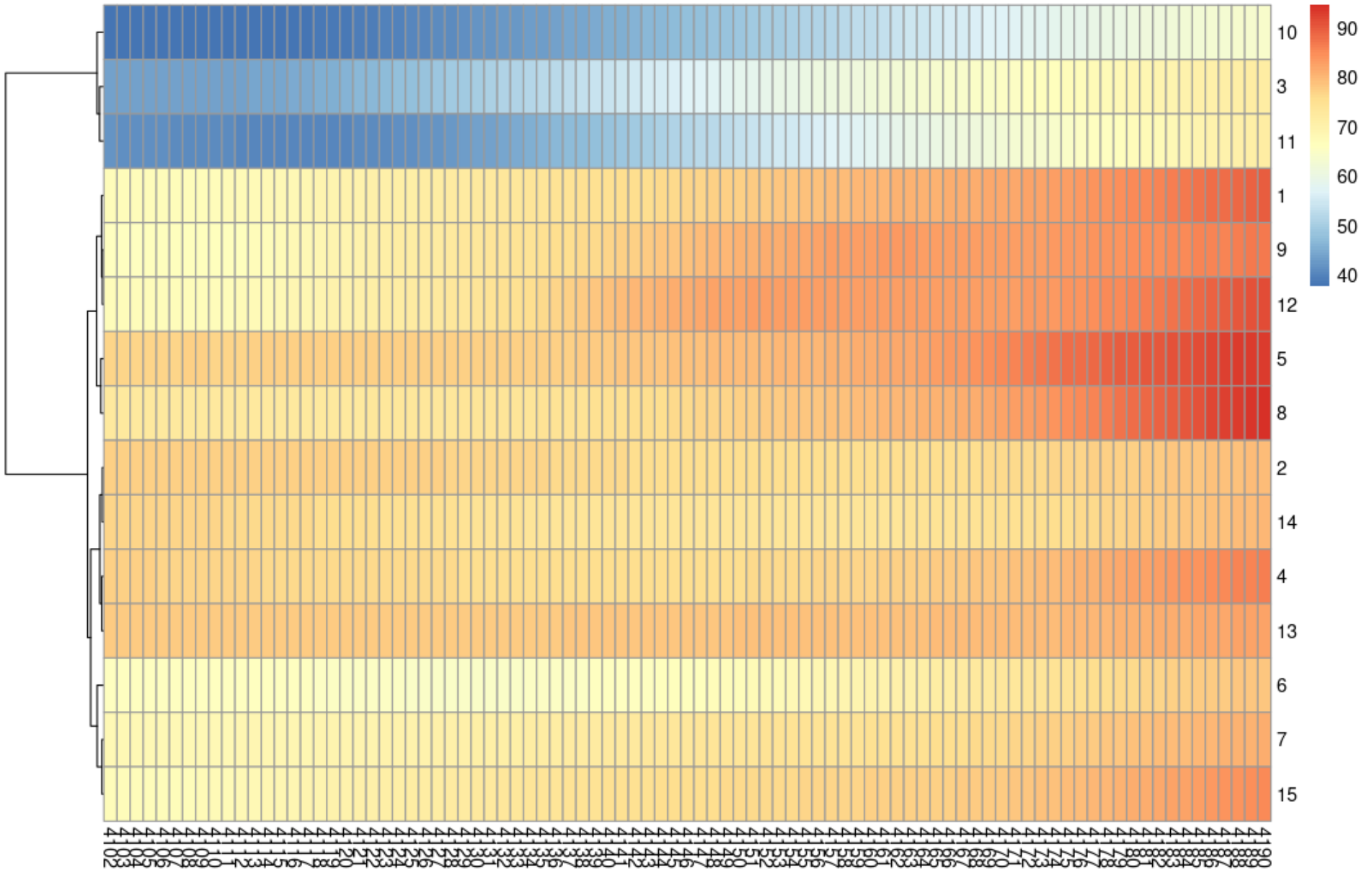


CAH : les données records_athle

- Durées (s) converties en vitesses (m/s)
- Données centrées-réduites



CAH : les données GPS_rugby



Validation et description des groupes

Validation et description des groupes

- Les méthodes de classification fournissent des groupes d'individus ou de variables.

- Comment les caractériser ?
  - Analyse en Composantes Principales, identification des individus selon leur appartenance à un groupe de la classification
  - Description de catégories

Description des groupes

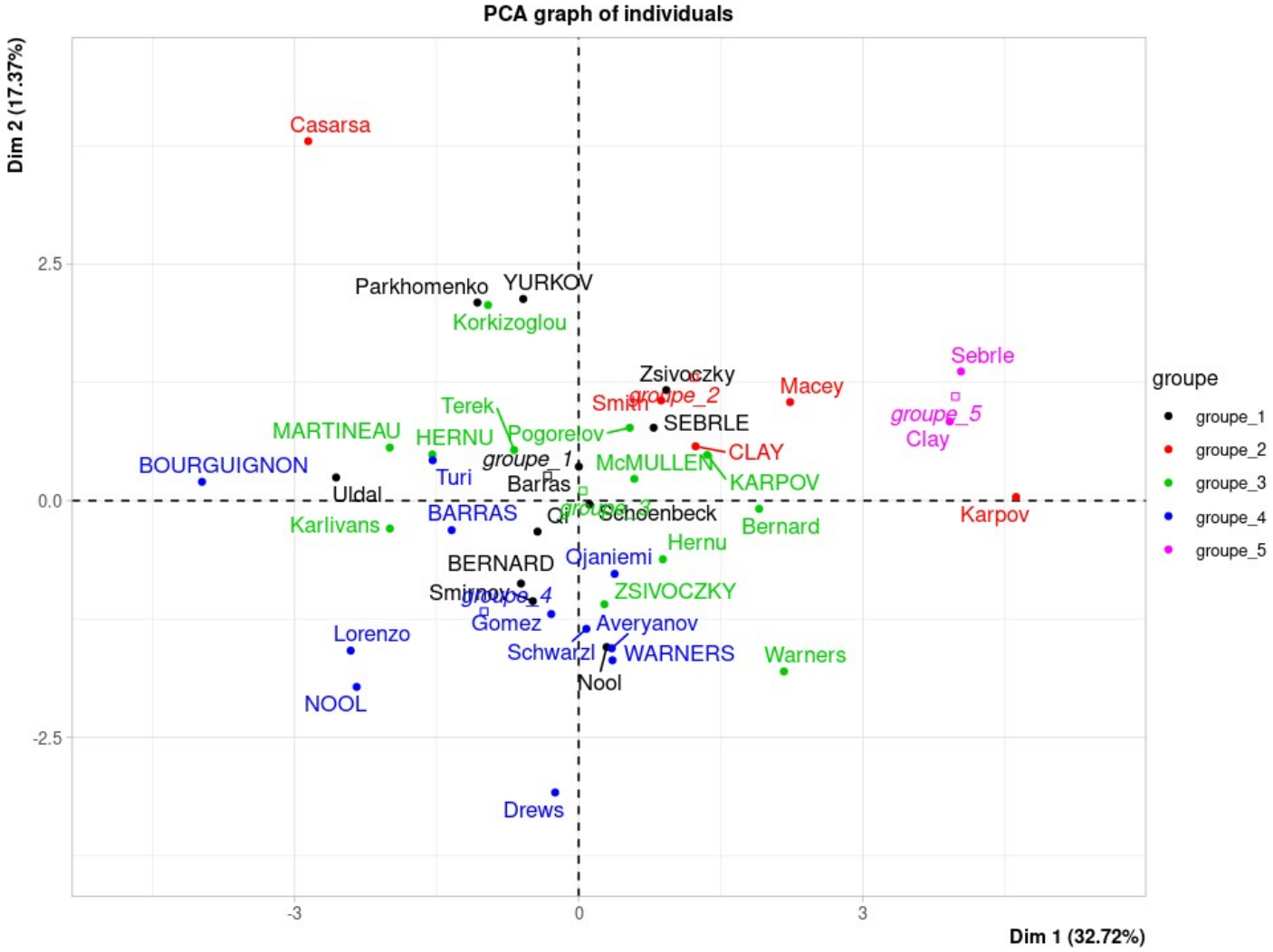
- Sélection des variables les plus caractéristiques de chaque classe
- Valeur-test : Ecart entre valeurs relatives à la classe et valeurs globales

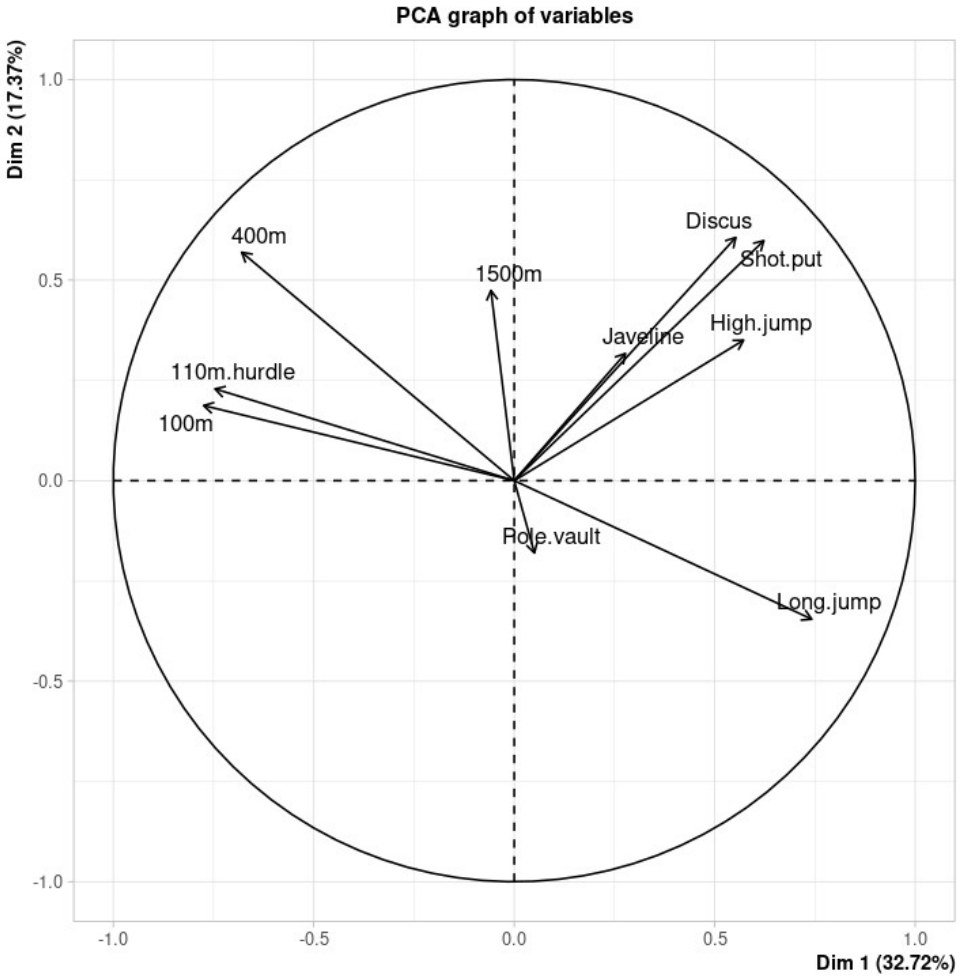
$$t_k(x) = \frac{\bar{X}_k - \bar{X}}{s_k(x)}$$

- Permet de ranger les variables par ordre d'intérêt
- « Profil-type » d'une classe

Ref: Lebart, L., Morineau, A. and Piron, M. (1995)

Les données decathlon

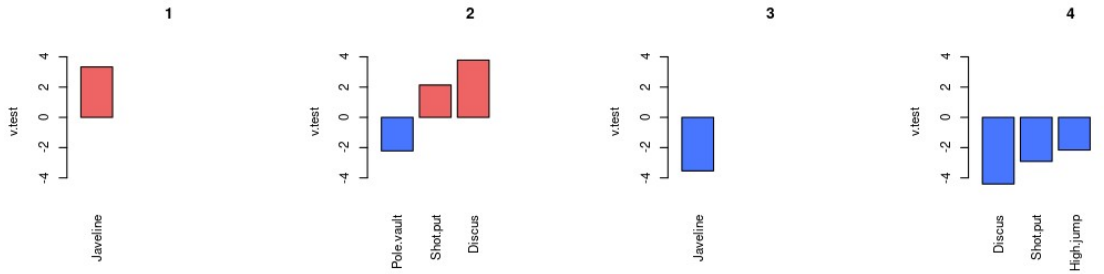




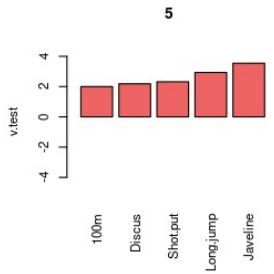
Les données decathlon

```
R> res_catdes <-
      catdes(decathlon_10events_5groupes,
             num.var = 11)
R> plot(res_catdes)
R> plot(res_catdes, barplot = TRUE)
```

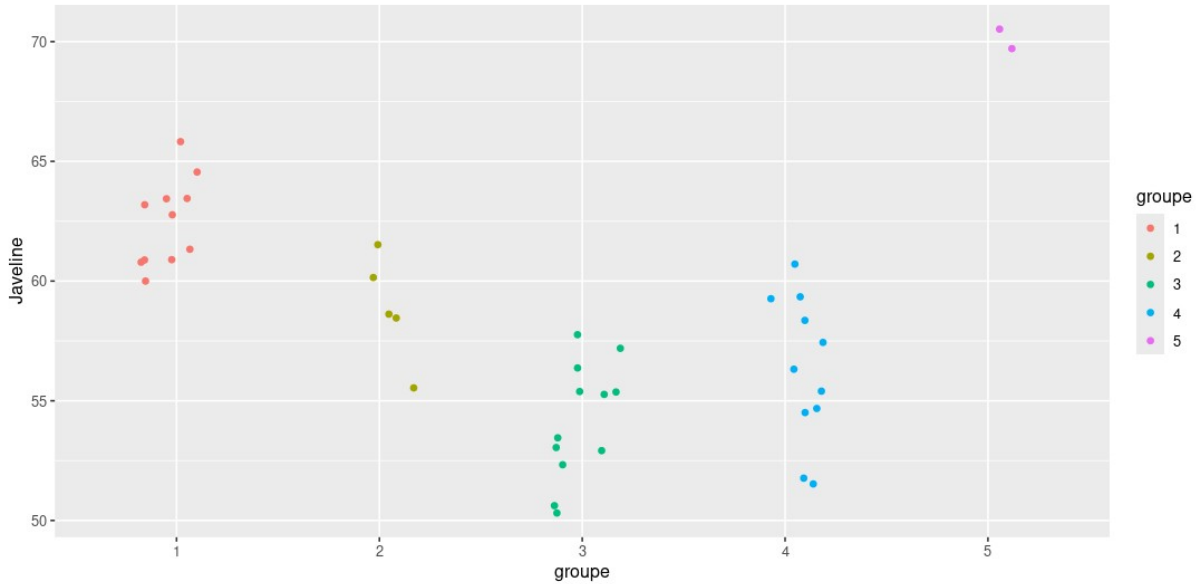

```
Link between the cluster variable
and the quantitative variables
================================
================================
            Eta2  P-value
Discus    0.821 5.39e-13
Javeline  0.772 3.87e-11
Shot.put  0.386 1.20e-03
Long.jump 0.230 4.62e-02
```

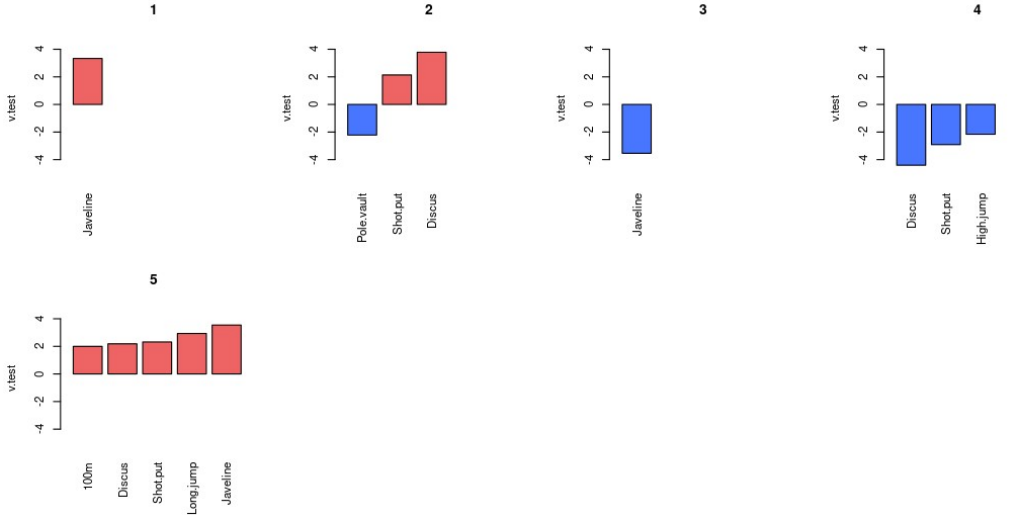


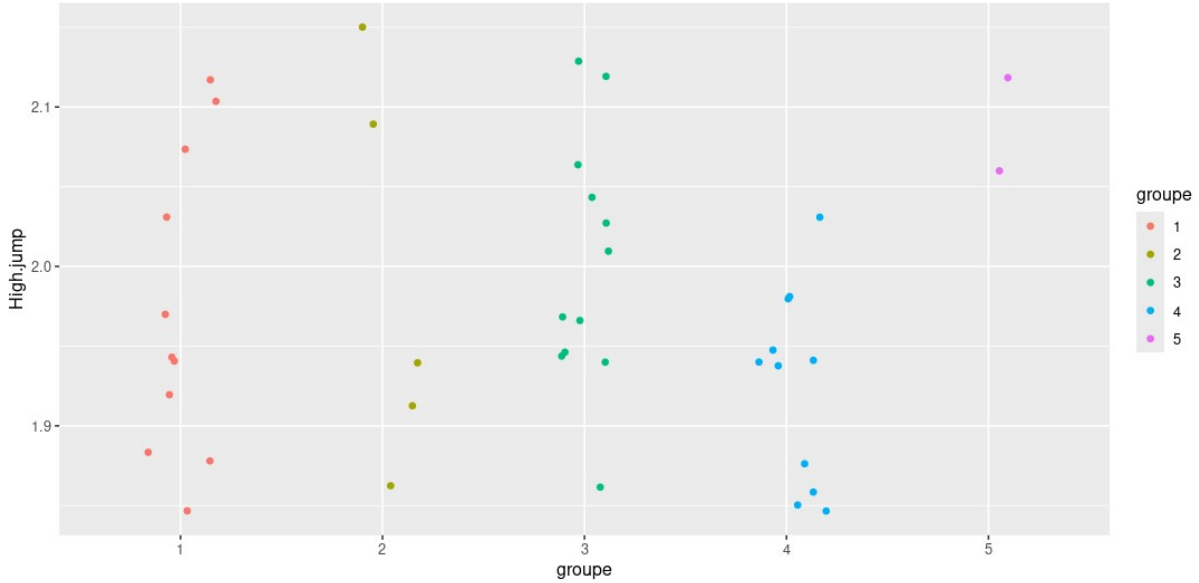

```
Description of each cluster by quantitative variables
=====================================================
$`1`
          v.test Mean_category Overall_mean sd_category Overall_sd  p.value
Javeline    3.33         62.46        58.31        1.74      4.767  0.00086
$`2`
          v.test Mean_category Overall_mean sd_category Overall_sd   p.value
Discus      3.78         49.67        44.32       1.283       3.336  0.00015
Shot.put    2.13         15.21        14.47       0.596       0.814  0.03245
Pole.vault -2.21          4.50         4.76       0.243       0.274  0.02652

[...]
```

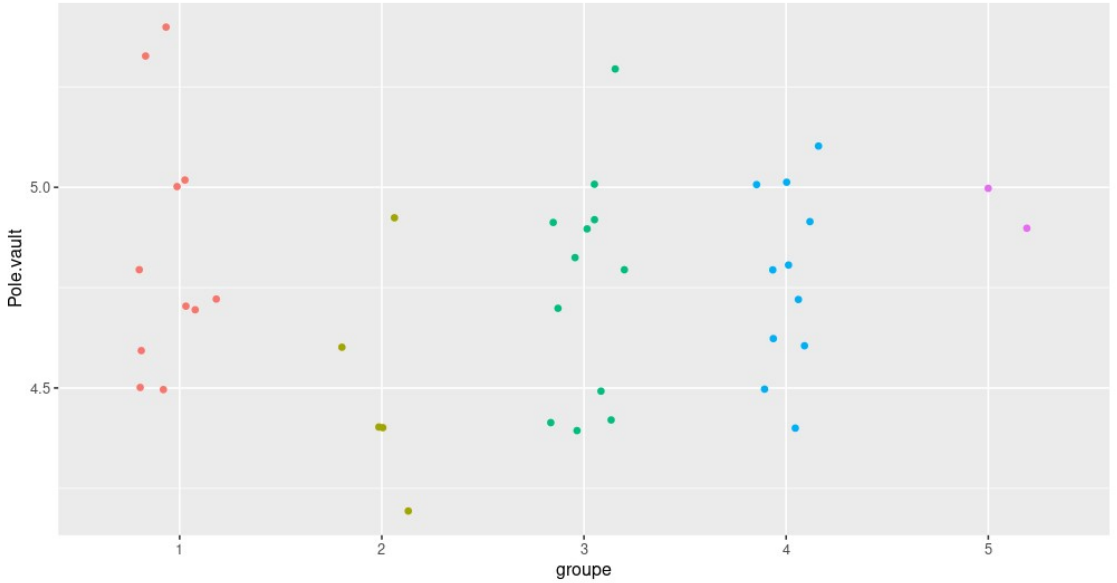


Les données decathlon

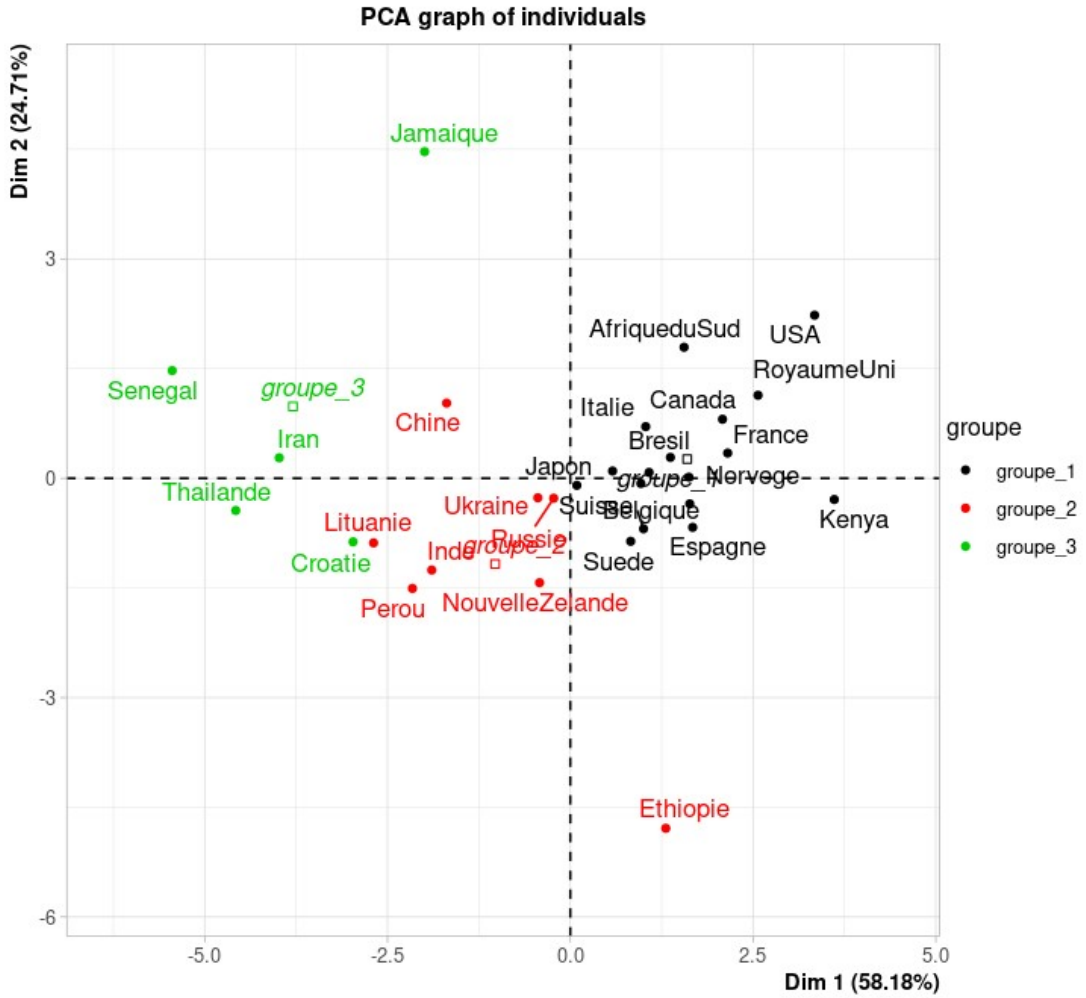


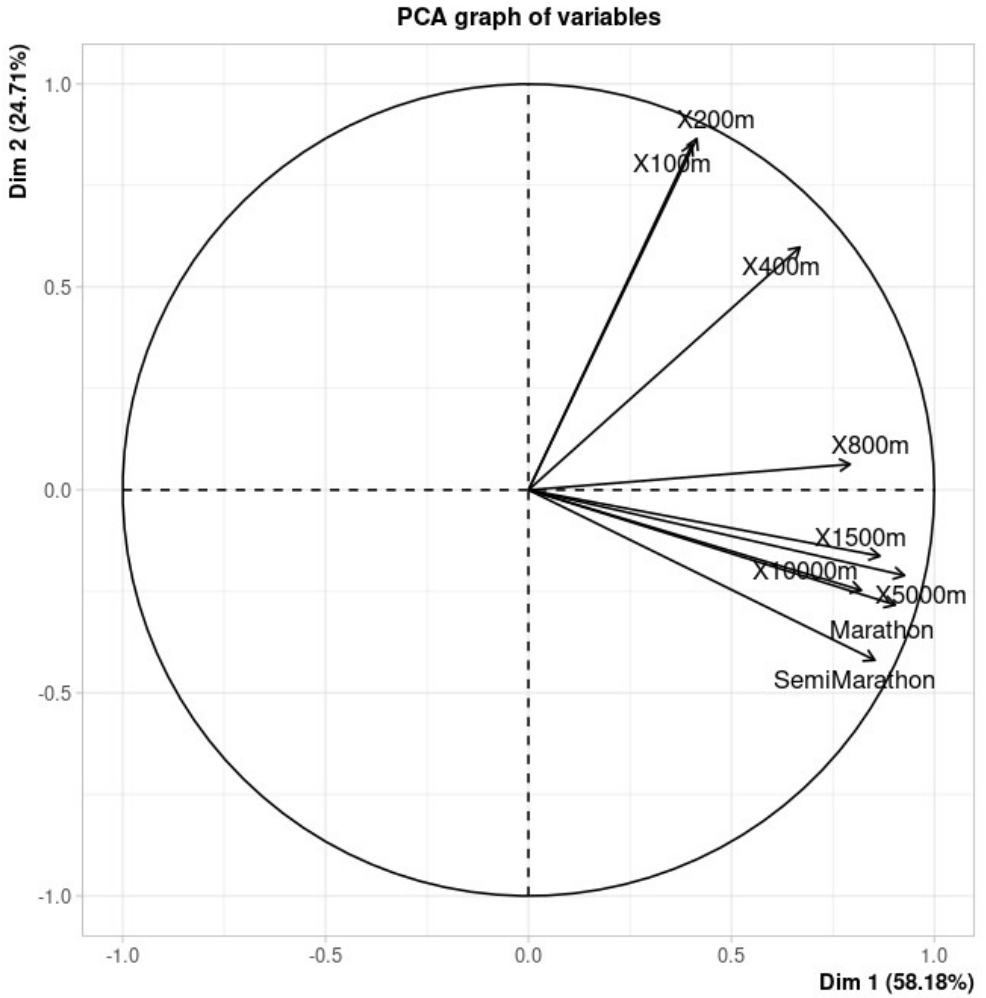






Les données records_athle





Les données records_athle

Link between the cluster variable and the quantitative variables

	Eta2	P-value
Marathon	0.793	5.89e-10
SemiMarathon	0.783	1.13e-09
X5000m	0.695	1.10e-07
X1500m	0.540	2.77e-05
X10000m	0.528	3.93e-05
X400m	0.435	4.51e-04
X800m	0.423	5.96e-04
X100m	0.291	9.62e-03
X200m	0.290	9.79e-03

Description of each cluster by quantitative variables

$`1`

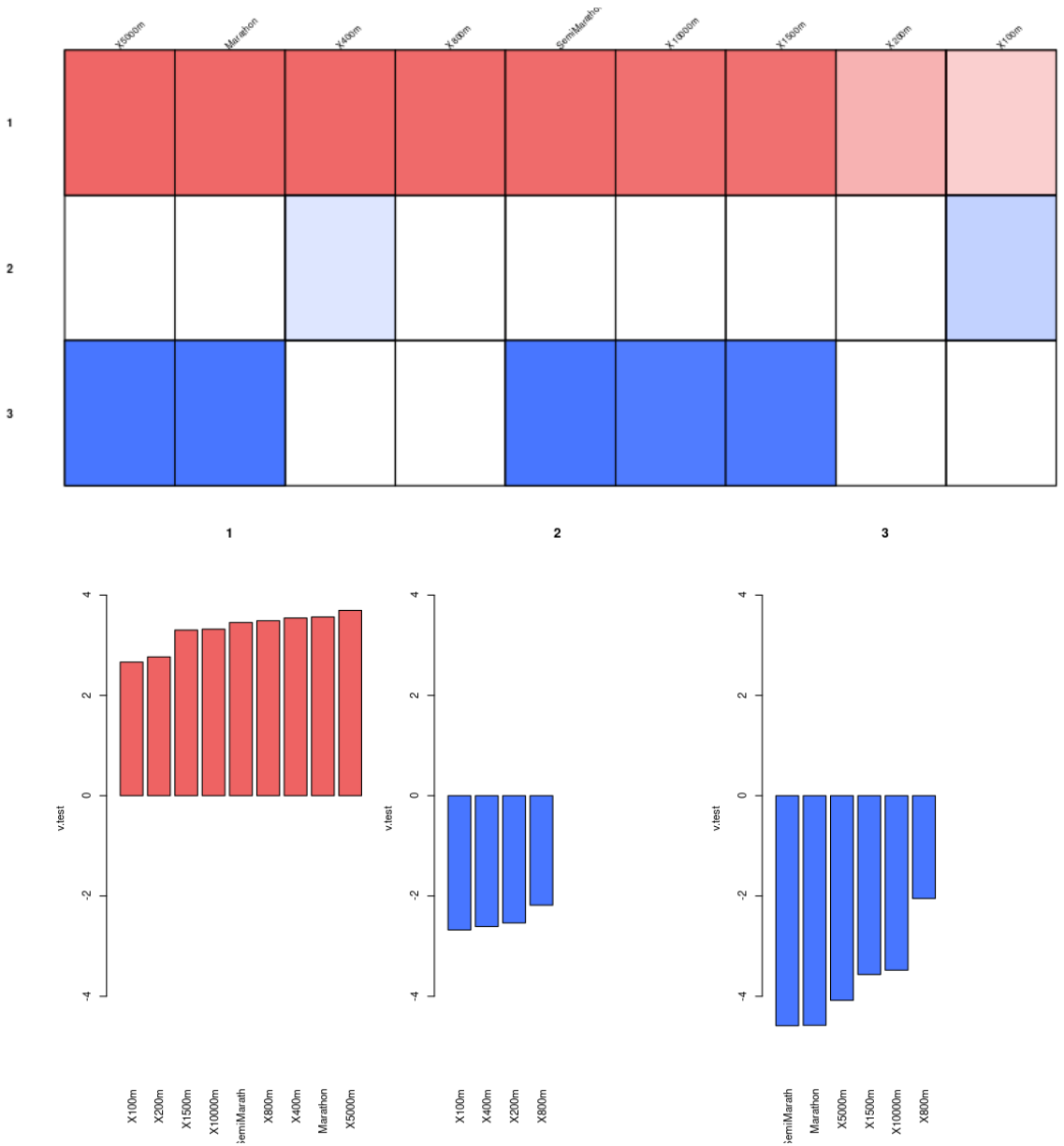
	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
X5000m	3.70	6.47	6.36	0.0645	0.185	0.000219
Marathon	3.56	5.62	5.49	0.0694	0.218	0.000365
X400m	3.54	9.03	8.95	0.1129	0.146	0.000394
X800m	3.49	7.79	7.72	0.0872	0.136	0.000485
SemiMarathon	3.45	5.92	5.78	0.0545	0.245	0.000553
X10000m	3.32	6.16	6.07	0.0729	0.175	0.000898
X1500m	3.30	7.13	7.06	0.0789	0.130	0.000956
X200m	2.77	10.05	9.96	0.1170	0.205	0.005636
X100m	2.66	10.09	10.00	0.1135	0.197	0.007702

$`2`

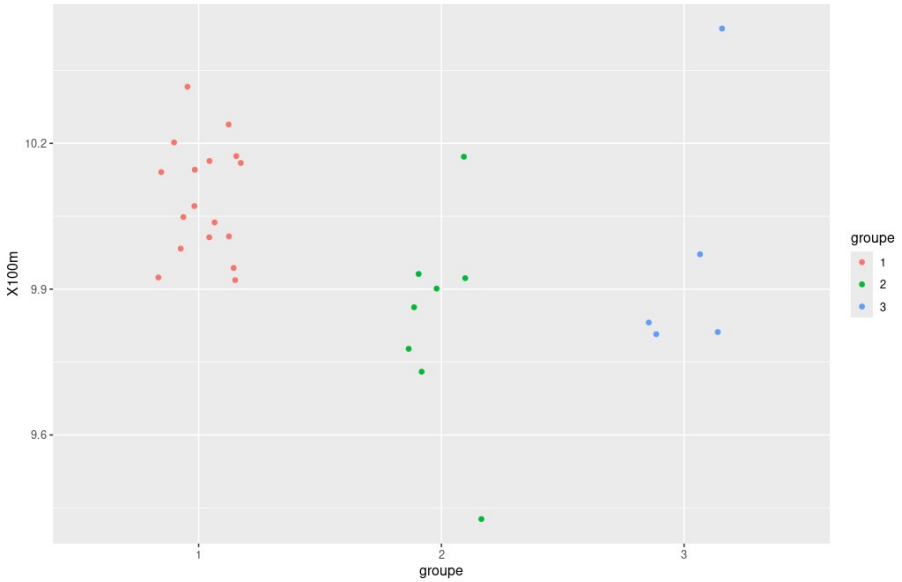
	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
X800m	-2.18	7.63	7.72	0.1239	0.136	0.02910
X200m	-2.54	9.80	9.96	0.1919	0.205	0.01119
X400m	-2.61	8.83	8.95	0.0738	0.146	0.00909
X100m	-2.68	9.84	10.00	0.1996	0.197	0.00745

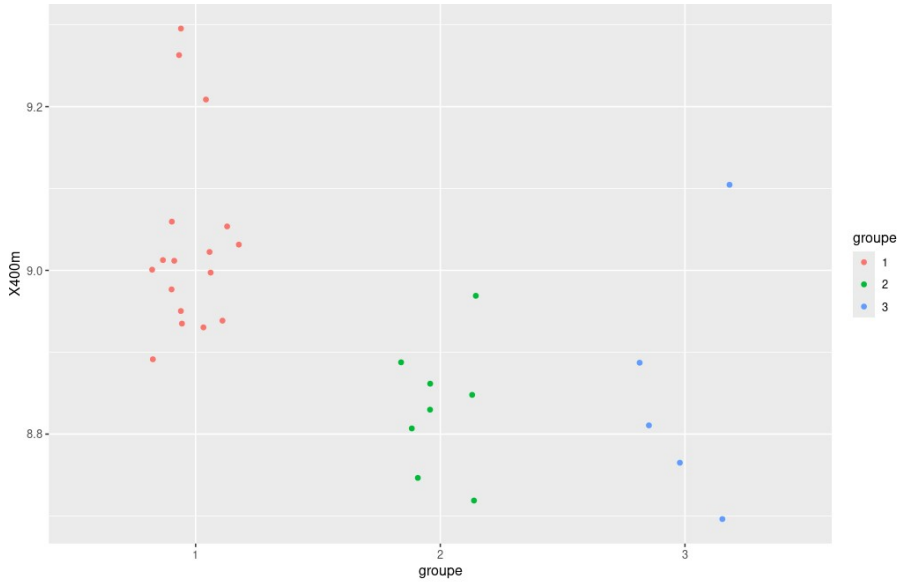
$`3`

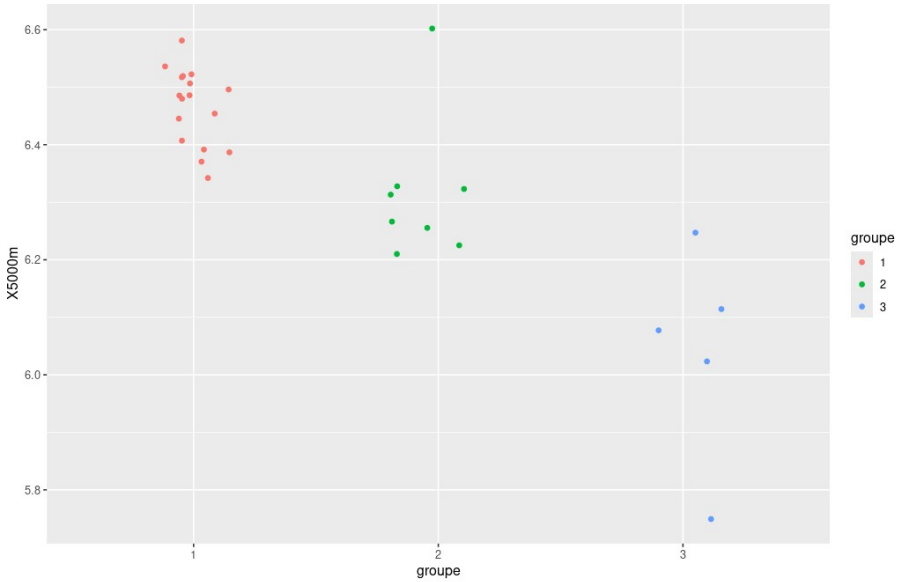
	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
X800m	-2.05	7.60	7.72	0.116	0.136	4.04e-02
X10000m	-3.48	5.82	6.07	0.208	0.175	5.03e-04
X1500m	-3.56	6.87	7.06	0.105	0.130	3.65e-04
X5000m	-4.08	6.04	6.36	0.164	0.185	4.51e-05
Marathon	-4.58	5.08	5.49	0.114	0.218	4.67e-06

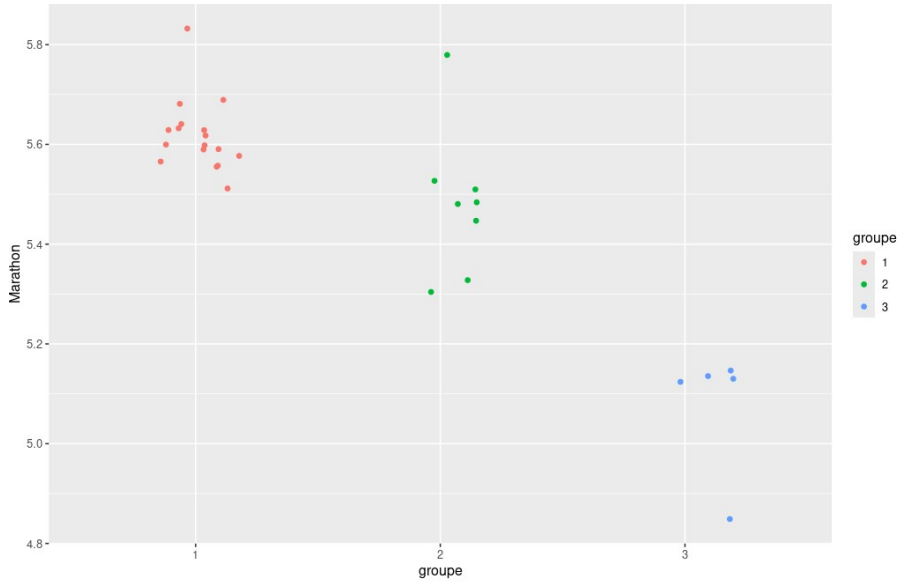


Les données records_athle





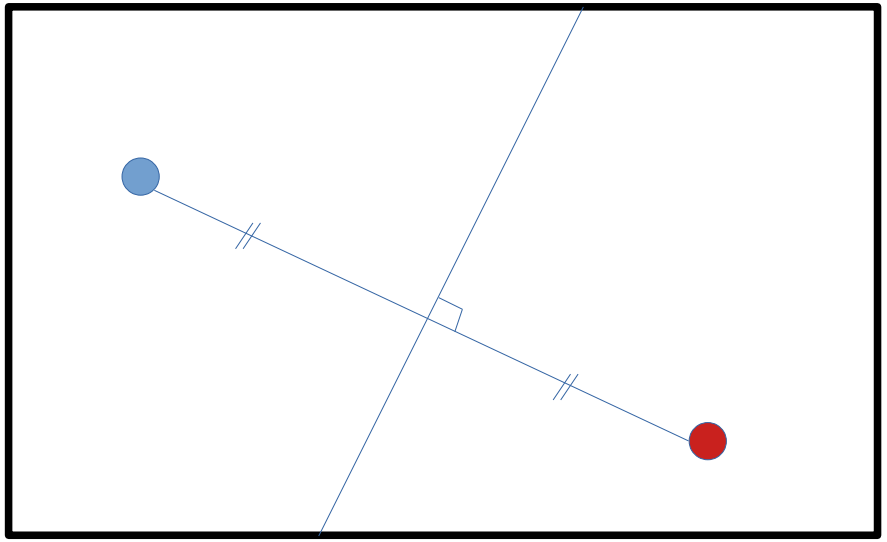


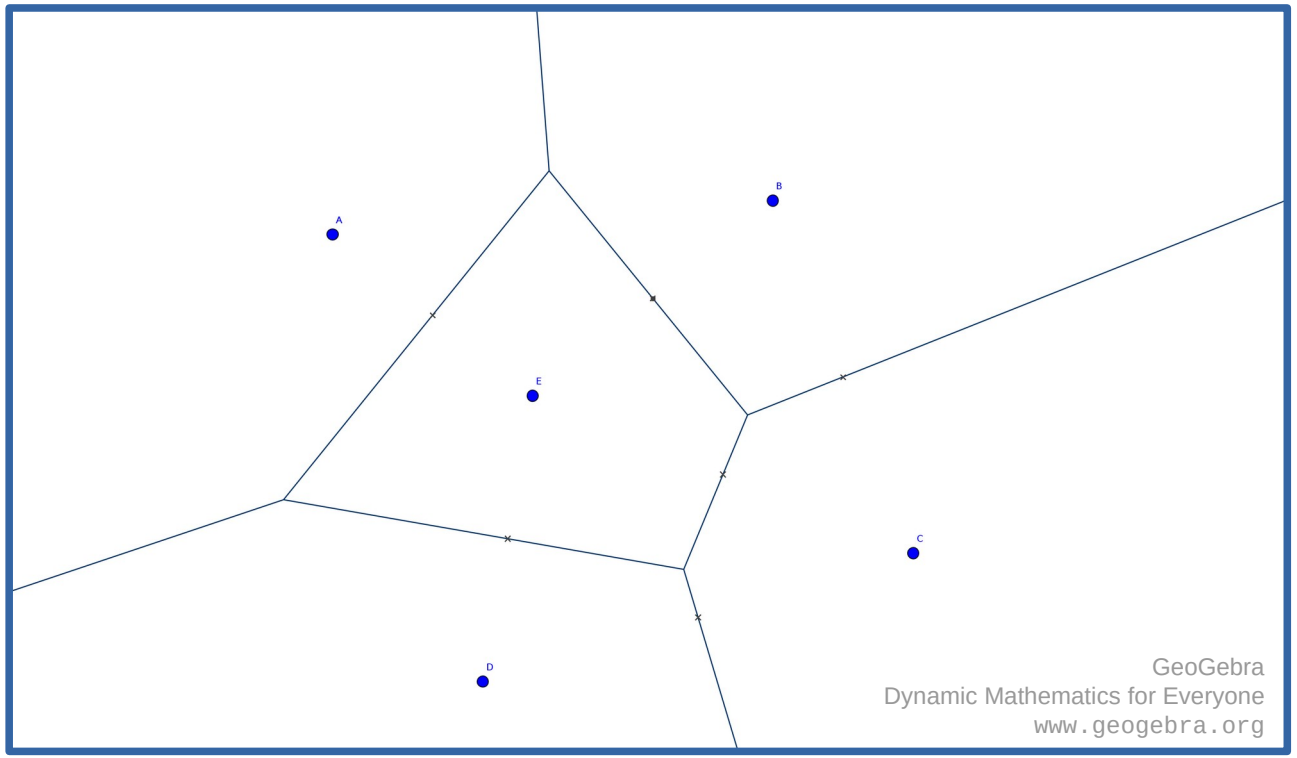


)

Diagramme de Voronoï

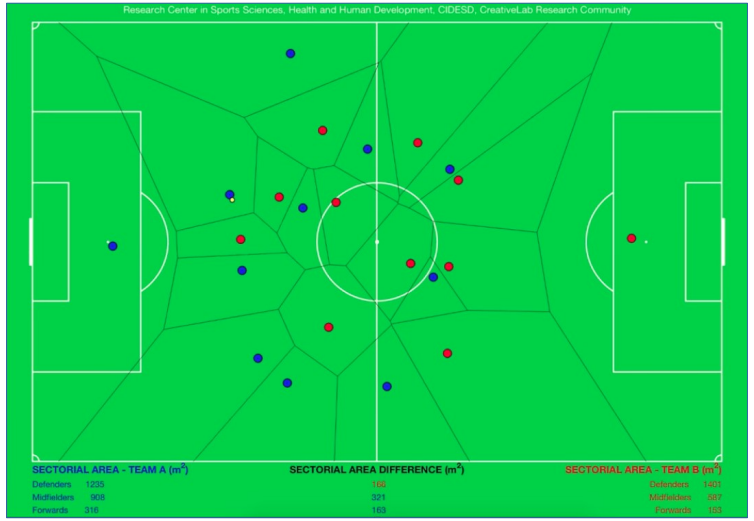
Un outil mathématique simple...





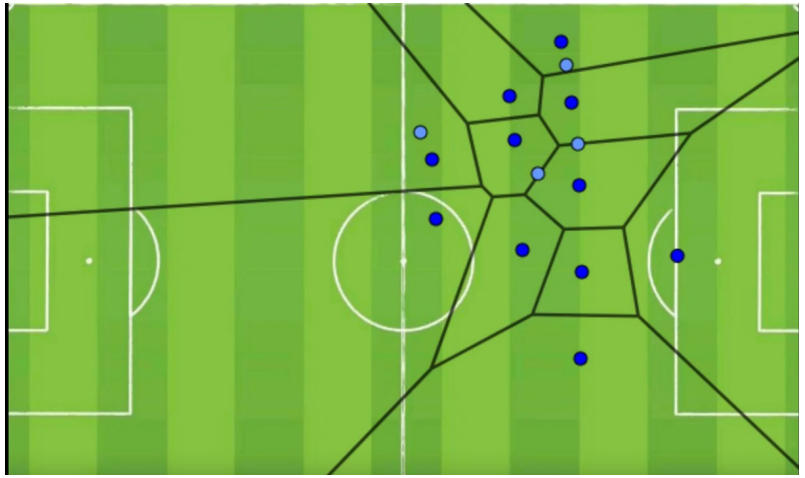
... pour modéliser des formes naturelles...

... mais pas seulement



www.youtube.com/watch?v=ZAz9mDlsWgQ

www.youtube.com/watch?v=UJ0VwY0DcIs



www.youtube.com/watch?v=j2rVU7Fzx3I

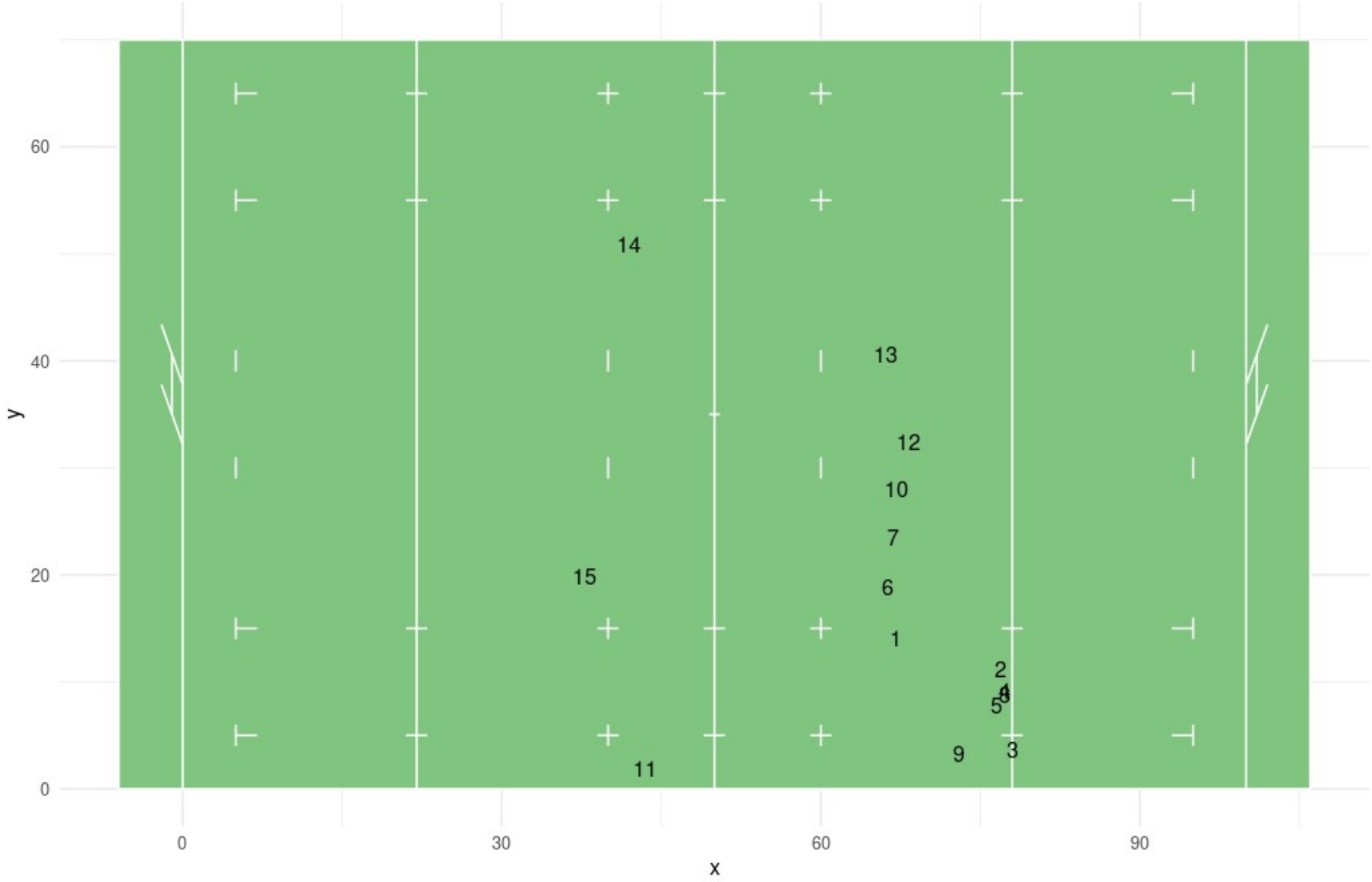
How Manchester City Use Voronoi Diagram?

Autres ressources :

lrhanalytics.co.uk/articles/voronoi-diagrams
medium.com/football-crunching/using-voronoi-diagrams-in-football-ca730ea81c05
...

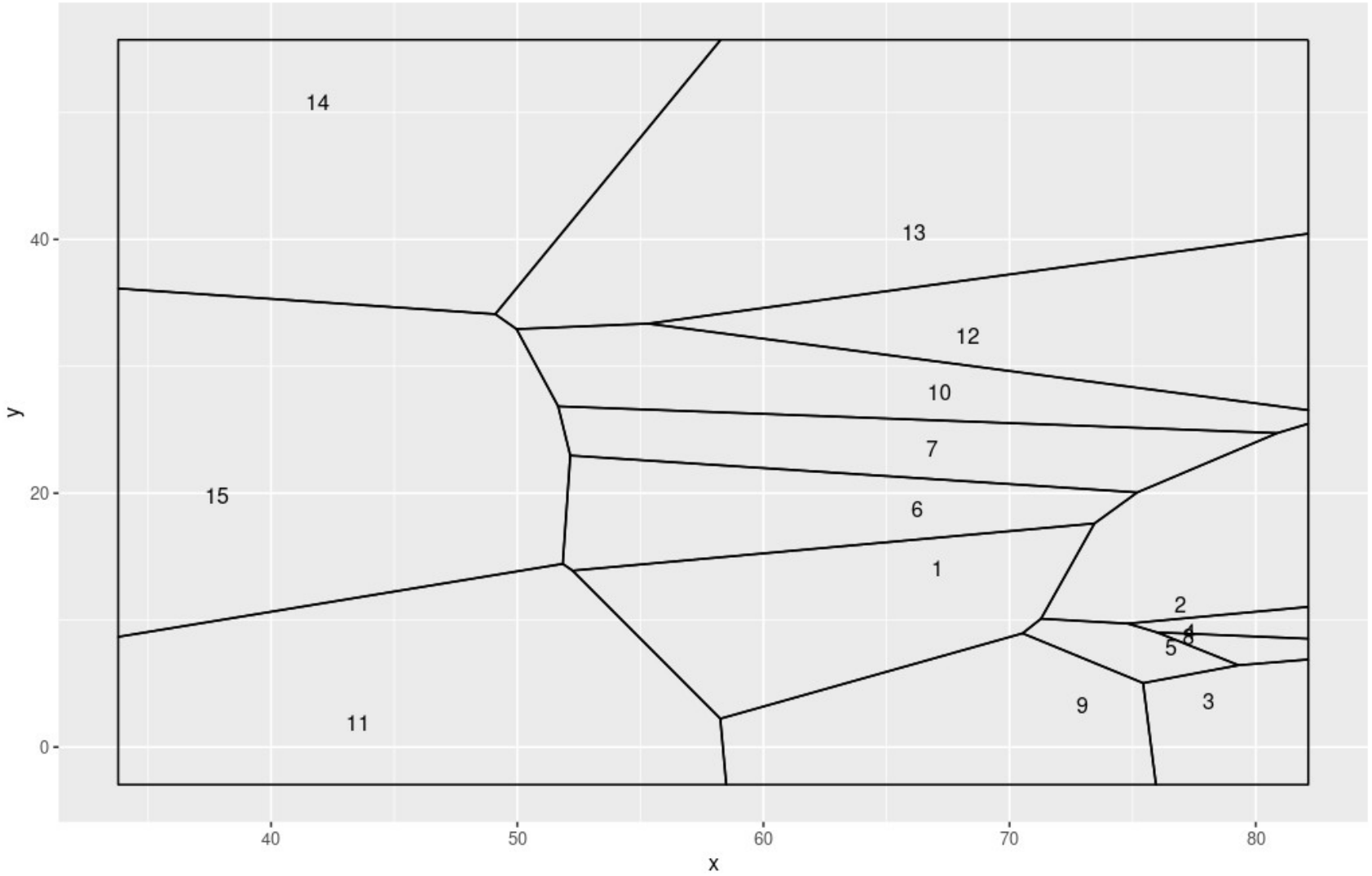
Voronoï : les données GPS_rugby

```
R> library(ggrugby)

R> GPS_rugby %>%
filter(frame_id == 4102) %>%
  ggplot() +
  rugby_pitch() +
  theme_minimal() +
  geom_text(aes(
              x = x,
              y = y,
              label = player))
```



Voronoï : les données GPS_rugby

```
R> GPS_rugby %>%
  filter(frame_id == 4102) %>%
  ggplot(aes(x = x, y = y)) +
    geom_text(
     aes(label = player)) +
    stat_voronoi(geom = "path")
```



Conclusion

- Méthodes pour visualiser des jeux de données sans a priori
- Révéler des informations contenues dans les données
- Ne s’applique pas à tout type de données