TP Données Industrielles

XU Ziqi

Nettoyage et traitement

Pour les '?' dans les quatre variables qualitatives, j'ai supprimé ces individus. Pour les zéros dans les variables quantitatives, je les ai remplacés avec la moyenne de cette variable.

Etape 1

Pour été/hiver, j'ai obtenu la moyenne de chaque composé pour ces deux saisons. La ligne bleu est 'été' et verte pour 'hiver'. Nous pouvons constater qu'en été, la concentration des composés sont plus que celle en hiver, **mais il nous manque les données du BF1**, donc nous ne pouvons pas conclure qu'en été, la concentration est plus que celle en hiver.

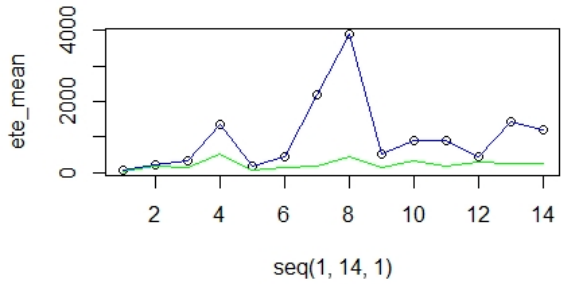

Figure 1 Concentration été/hiver

Pour les six campagnes, nous avons obtenu le résultat ci-dessous. En général la concentration des composés après l'installation est supérieure à celle avant la mise en place du site.

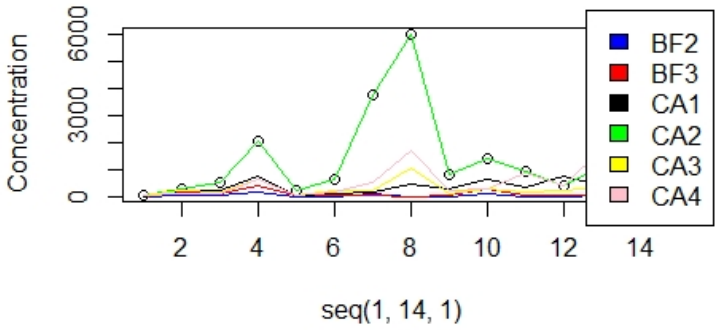

Figure 2 Concentration 6 campagnes

Pour les deux campagnes (avant et après), nous pouvons obtenir le même résultat que le graph 2, la concentration d'après (bleue) est beaucoup plus que celle d'avant(verte).

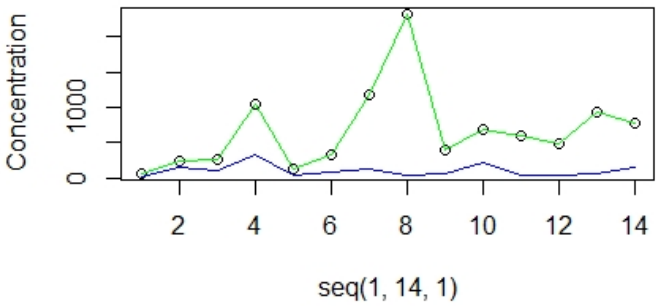

Figure 3 Concentration BF et CA

Etape 2

Nous avons utilisé `prcomp()` pour l'ACP, puis le package `factoextra` pour visualiser les résultats. Voici les valeurs propres : Les deux premiers composants expliquent 70 pourcents de la variance, donc la qualité de la réduction est 70% pour les deux premiers plans.

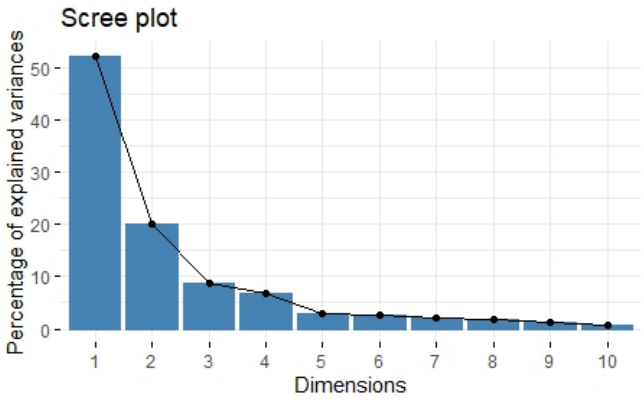

Figure 4 Valeurs propres de l'ACP

Pour les 14 variables, **9_ane** contribue le plus au dimension 1 alors que **B** et **BTM** contribue le plus au dimension 2. Evidemment il y a une forte corrélation entre les différentes variables. Voir Figure 5 pour les détails.

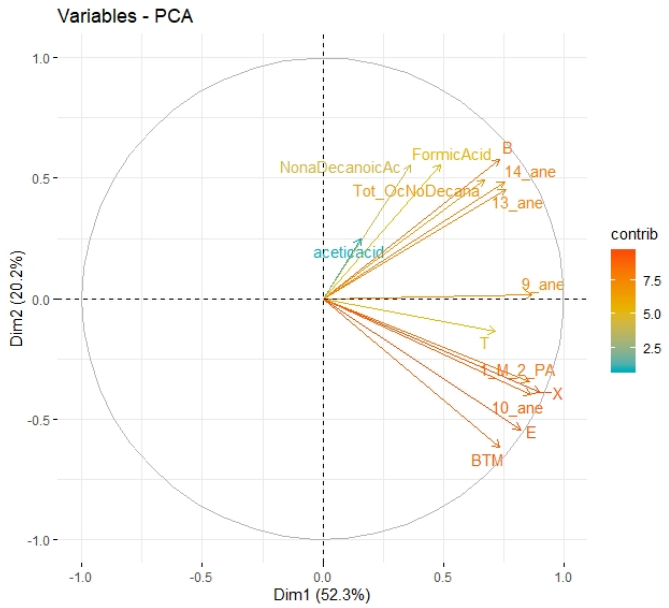

Figure 5 Variables PCA

La qualité des projections est calculée par l'attribut **cos2** du résultat. Voici les cinq premières dimensions des cinq premières variables.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	8.079196e-01	1.184279e-01	1.412230e-02	1.213744e-03	5.072282e-02
2	7.935642e-01	1.268476e-01	1.337137e-02	2.202819e-03	5.559325e-02
3	6.240410e-01	1.917678e-01	6.633620e-02	2.501521e-02	5.352665e-03
4	7.498311e-01	2.001242e-01	2.914387e-02	9.609240e-03	2.494614e-03
5	7.754709e-01	1.761822e-01	2.168760e-02	9.566169e-04	1.978491e-02

Figure 6 Quality Individual PCA

Pour les individus, si l'on considère les périodes été/hiver, nous pouvons constater que les individus 'été' se regroupent dans le premier quadrant, c'est à dire ce groupe est plus lié aux composés. En même temps, le groupe CA2 est bien expliqué par ce deux dimensions (ils sont plus loin du point d'origine) alors que ce n'est pas le cas pour CA4. Ce phénomène est confirmé par la qualité de la projection des individus.

L'individu 76 est anormal, c'est possible de l'enlever dans les calculs suivants. Pour l'instant on le garde.

En conclusion, nous pourrons garder certains composés dans notre mesure (9_ane, 13_ane etc.) afin de réduire le nombre du composés.

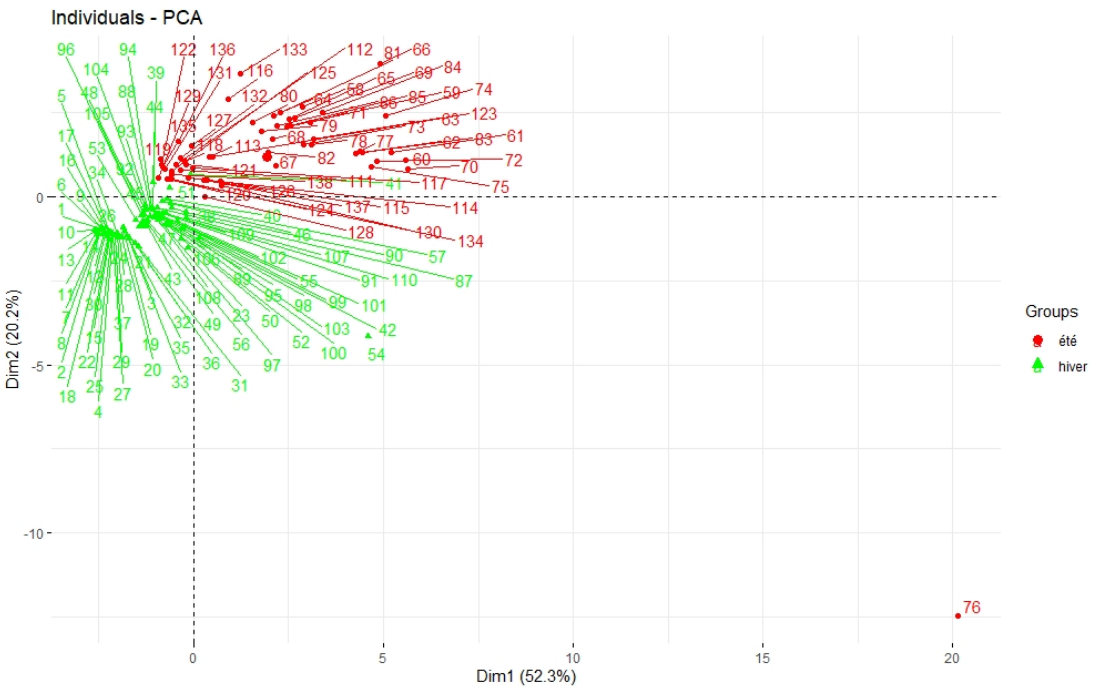

Figure 7 Individus été/hiver

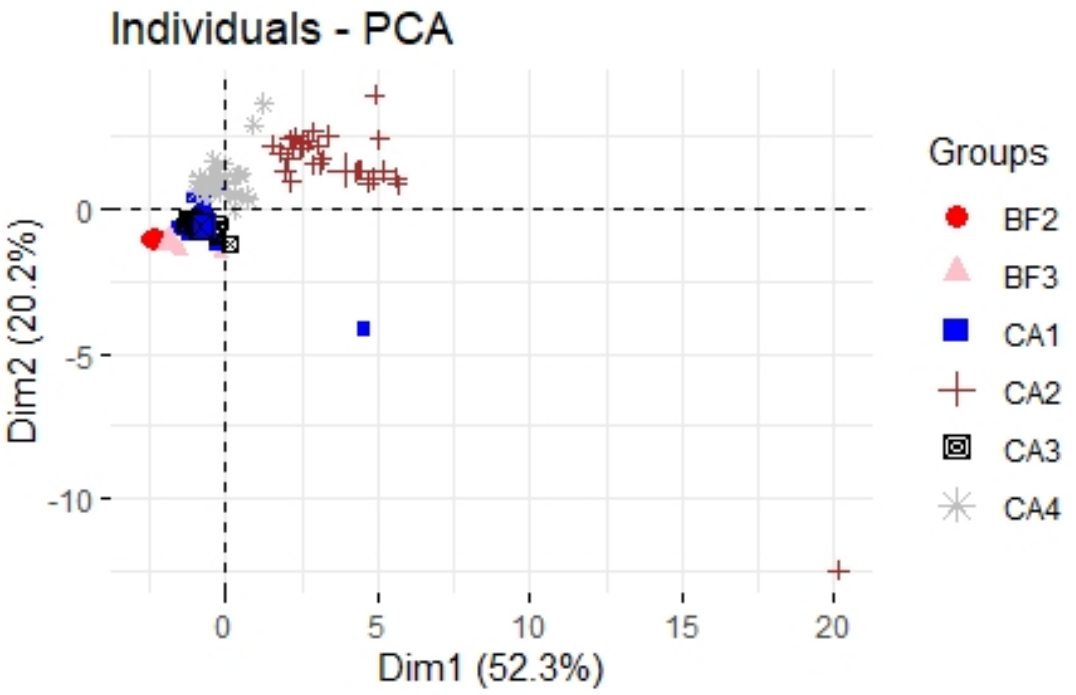

Figure 8 Individus campagnes

Etape 3

Si l'on ne considère que les quatre campagnes CA, la concentration du CA2 (été) est beaucoup plus que les autres campagnes, alors que la différence entre CA4 (été) et CA1 ou CA3 (hiver) n'est pas assez grande (Figure 2).

Le résultat de l'AFD sur le groupe été/hiver est ci-dessous. Les deux groupes sont bien distingués selon le graph. Compte tenu qu'il y a deux modalités, nous obtenons seulement une dimension après l'AFD, donc il n'y a pas de variance expliquée.

Variance_totale	9.197
Variance_été	1.544
Variance_hiver	0.619
Variance_interclasse	7.034
η	76.5%

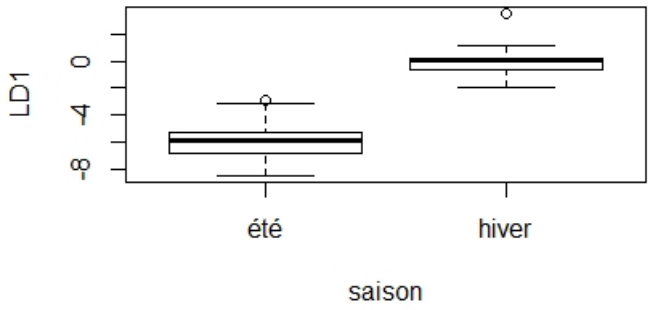

Figure 9 AFD Saison

Le résultat de l'AFD sur le groupe BF/CA est ci-dessous. Les deux groupes sont moins distingués selon le graph par rapport au groupe été/hiver. La variance de la modalité CA est beaucoup plus grande que celle de BF, donc la qualité globale de la réduction est affectée.

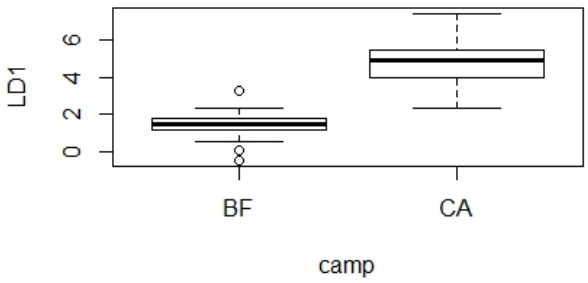

Figure 10 AFD BF/CA

Variance_totale	3.259
Variance_BF	0.383
Variance_CA	1.222
Variance_interclasse	1.654
η	50.8%

La fonction linaire discriminante est ci-dessous.

	LD1
B	6.985888e-02
T	2.331274e-03
E	2.674020e-05
X	-5.301742e-04
`9_ane`	6.170527e-03
`10_ane`	-1.385907e-05
`13_ane`	-1.097567e-03
`14_ane`	2.065326e-04
`1_M_2_PA`	2.260878e-04
BTM	7.074986e-05
FormicAcid	-3.030703e-04
aceticacid	5.366296e-04
NonaDecanoicAc	2.419678e-04
Tot_OcNoDecana	-3.862969e-04

Figure 11 Fonction discriminante

En conclusion, d'après l'ACP, la mise en place du site a un grand effet sur l'environnement (voir Figure 8), car les points du CA se sont situés au premier et deuxième quadrant, mais en hiver cette différence est moins significative. Pour l'AFD, il sert à distinguer les différentes modalités, donc l'écart (variance interclasse) est plus grand que l'ACP (Figure 12 vs Figure 9).

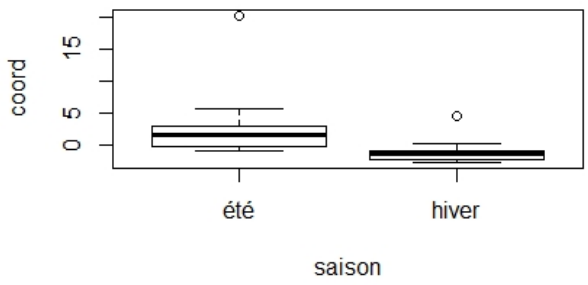

Figure 12 ACP Saison Premier Plan

Etape 4

Le résultat de l'AFMD n'est pas assez satisfaisant, car les cinq premiers plans ne représentent que 37% d'inertie (figure 13). Si l'on regard la contribution des variables quantitatives entre l'ACP et l'AFMD (Figure 14 et Figure 5), les figures sont similaires, '9_ane' est toujours la contribution la plus importante au premier plan. Pour les variables quantitatives (Figure 15), il existe quatre types de variables. 'TYPE', les trois modalités 'compostage', 'urbain' et 'rural'

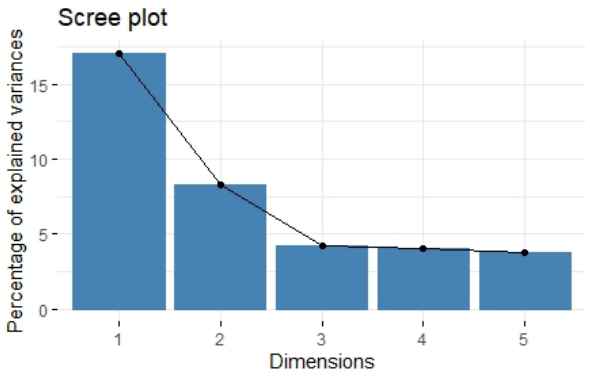

Figure 13 Valeurs propres AFMD

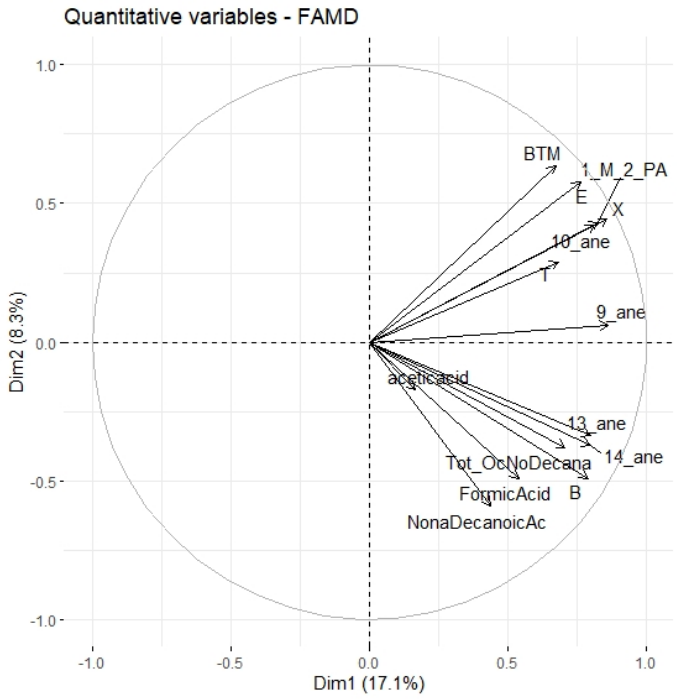

Figure 14 Variables quantitatives AFMD

sont proches du point d'origine, alors que 'source industrie' est mieux distingué, c'est-à-dire les composés sont plutôt liées à la source industrie. 'Localisation', certains points de mesure (P33, P10, P32, etc.) ont contribué plus à la première dimension, donc il y a plus de pollution à ces points-là. 'Campagne', il n'y a pas de grande différence entre 'avant le site' et 'après', car les points BF2, BF3 et CA1, CA3 sont proches dans le graph, mais il y a plutôt une différence entre 'hiver' et 'été'.

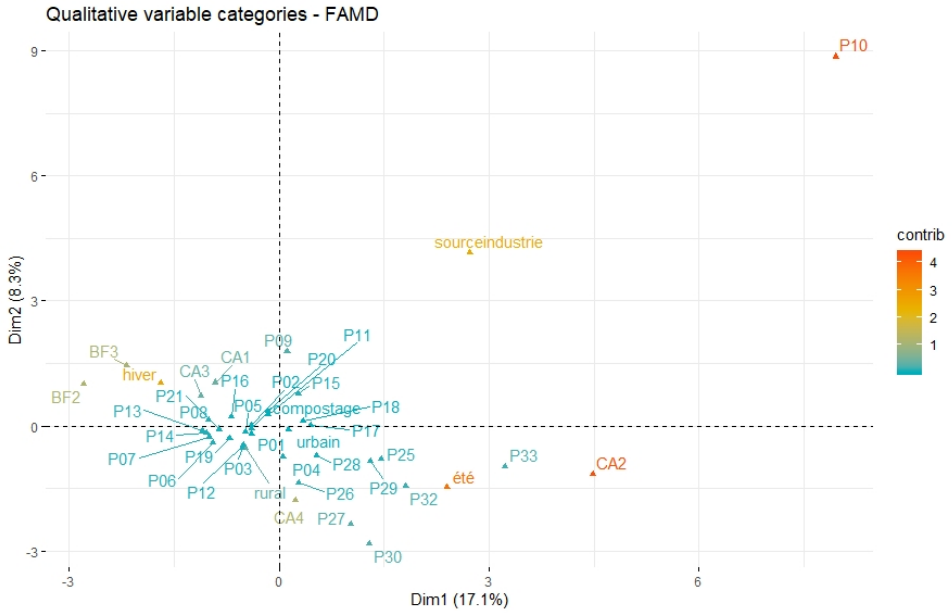

Figure 15 Variables qualitatives

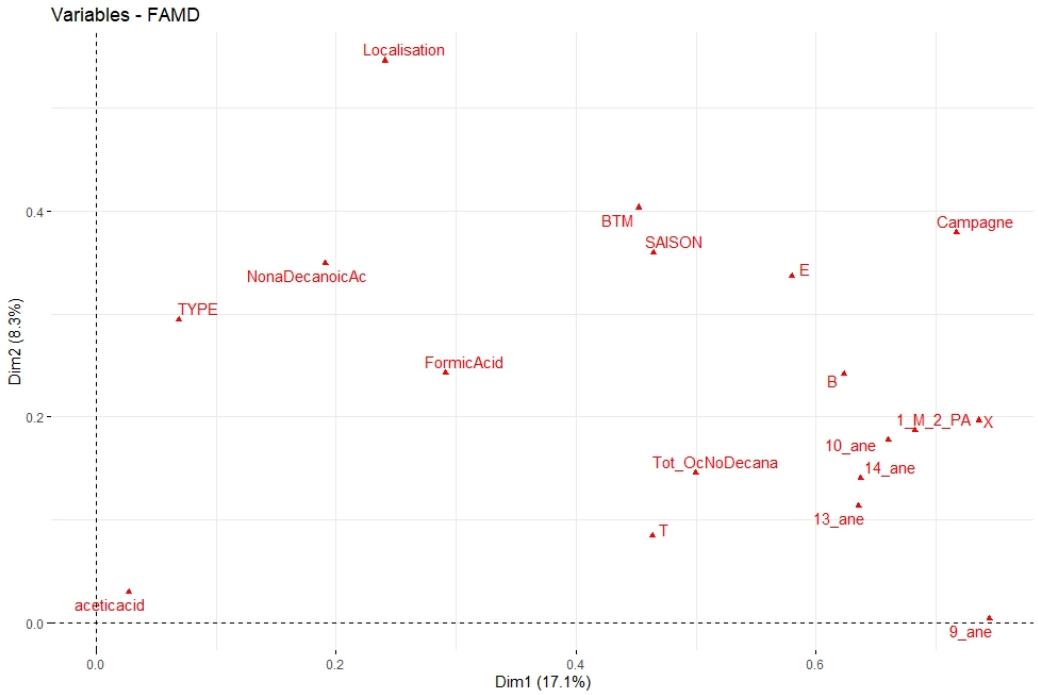

Figure 16 AFMD Variables

Nous pouvons faire une conclusion qu'il n'y a pas de différence après l'installation du site en hiver, mais il faut **les données du BF1** pour vérifier cette conclusion en été. Les points de mesure non significatifs peuvent être enlevés dans le futur.

Comparaison entre l'ACP et l'AFMD : Pour les variables quantitatives, les résultats sont similaires, ils montrent que certains composés sont plus importants ('9_ane' etc.). Pour les variables qualitatives, l'AFMD nous permet d'analyser plusieurs variables en même temps, alors que l'ACP montre au plus une variable qualitative (Figure 7 et 8). Les résultats sont cohérents, par exemple on peut bien distinguer été/hiver mais pas BF/CA à l'hiver.

Questions complémentaires

Le résultat de l'AFD sur 'TYPE' n'est pas satisfaisant même si nous avons obtenu 63.7% de l'inertie expliquée sur le premier plan.

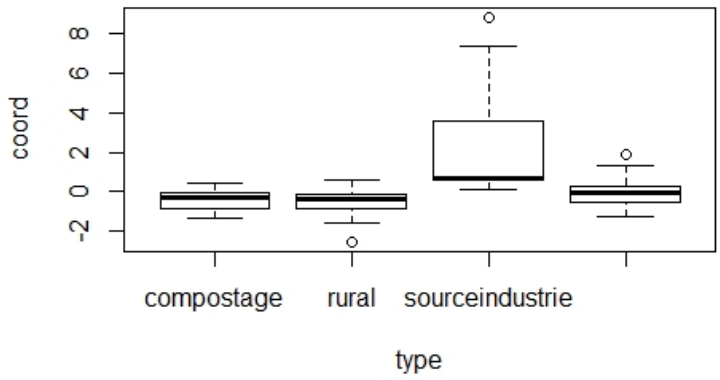

Figure 17 AFD du TYPE premier axe

Il y a une fonction `predict()` qui est capable de prédire l'appartenance à sa classe, mais où sont les nouvelles données ?