UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

ÉCOLE D'ÉCONOMIE
DE LA SORBONNE

Master 1 Économétrie Statistiques

# Projet : Analyse de données

*Prutki Lucas*
*Barbey Charlotte*
*Domergue Corentin*

1er février 2021

# 1 Enquête budget des familles 1994-1995

## 1.1 Déterminer le meilleur modèle possible pour l'emploi d'un(e) employé(e) de maison

L'objectif est de trouver le modèle qui coïncide au mieux avec les données pour pouvoir déterminer si un ménage emploi ou non un employé de maison. Pour cela on se réfère au critère AIC qui est une mesure de la qualité d'un modèle statistique avec un terme de fit (maximum de vraisemblance) et un terme de pénalité sur le nombre de variables explicatives. On réalise notre étude en 2 étapes : premièrement on prend en compte toutes les variables explicatives données et dans un second temps on fait une analyse en considérant un sous-ensemble de ces variables pour voir si cela permet d'obtenir un meilleur modèle.

### 1.1.1 Modèle complet

Variables explicatives : { TYPMEN2,CC,REVTOT,DIPLOPR,DIPLOCJ }
Premièrement on nettoie la base de données : suppression des variables inutiles, suppression des ménages avec données manquantes et conversion des variables qualitatives en "factors". Suite à cette étape, on remarque qu'il ne reste plus que deux valeurs possibles à la variable à expliquer DOMTRAV : *1-Oui 2-Non*. On peut alors procéder à une régression logistique binaire, si il restait des réponses *8-ne sait pas* ou *9-refus* on aurait dû réaliser une régression logistique multinomiale.
On se doute à priori que l'hypothèse d'indépendance des variables explicatives n'est pas vérifiée, on fait alors plusieurs hypothèses de dépendance entre les variables. On suppose 6 effets d'interactions :

- Degré d'urbanisation et type de ménage
- Diplôme de la personne de référence et diplôme du conjoint
- Revenu total et toutes les autres variables

On fait alors des chisq-test pour les 2 premiers effets et des anova à 1 facteur pour les effets entre la variable REVTOT et les autres (on fait des test de bartlett au préalable pour spécifier l'homogénéité ou non des variances).

```
l'approximation du Chi-2 est peut-ętre incorrecte
        Pearson's Chi-squared test

data:  bdf$CC and bdf$TYPMEN2
X-squared = 62.374, df = 30, p-value = 0.0004705

l'approximation du Chi-2 est peut-ętre incorrecte
        Pearson's Chi-squared test

data:  bdf$DIPLOPR and bdf$DIPLOCJ
X-squared = 846.75, df = 49, p-value < 2.2e-16


        Bartlett test of homogeneity of variances

data:  bdf$REVTOT by bdf$CC
Bartlett's K-squared = 106.78, df = 5, p-value < 2.2e-16


        One-way analysis of means (not assuming equal variances)

data:  bdf$REVTOT and bdf$CC
F = 9.1923, num df = 5.0, denom df = 490.8, p-value = 2.244e-08


        Bartlett test of homogeneity of variances

data:  bdf$REVTOT by bdf$DIPLOPR
Bartlett's K-squared = 560.8, df = 7, p-value < 2.2e-16


        One-way analysis of means (not assuming equal variances)

data:  bdf$REVTOT and bdf$DIPLOPR
F = 47.388, num df = 7.00, denom df = 570.28, p-value < 2.2e-16


        Bartlett test of homogeneity of variances

data:  bdf$REVTOT by bdf$DIPLOCJ
Bartlett's K-squared = 311.03, df = 7, p-value < 2.2e-16


        One-way analysis of means (not assuming equal variances)

data:  bdf$REVTOT and bdf$DIPLOCJ
F = 31.955, num df = 7.00, denom df = 345.95, p-value < 2.2e-16


        Bartlett test of homogeneity of variances

data:  bdf$REVTOT by bdf$TYPMEN2
Bartlett's K-squared = 308.33, df = 6, p-value < 2.2e-16


        One-way analysis of means (not assuming equal variances)

data:  bdf$REVTOT and bdf$TYPMEN2
F = 114.48, num df = 6.00, denom df = 584.04, p-value < 2.2e-16
```

Les tests sont très significatifs donc ces variables sont dépendantes entre elles et on intègre alors ces effets d'interactions à notre modèle.

```
DOMTRAV~CC+DIPLOPR+TYPMEN2+DIPLOCJ+REVTOT+TYPMEN2:CC+CC:REVTOT+DIPLOPR:REVTOT+DIPLOCJ:REVTOT+DIPLOPR:DIPLOCJ
```

On utilise la fonction step sur notre modèle avec toutes les variables qui nous stipule qu'enlever les interactions entre les deux variables de diplômes et celle entre le revenu total et le type de ménage nous permet d'avoir le modèle avec AIC plus faible :

```
DOMTRAV~CC+DIPLOPR+TYPMEN2+DIPLOCJ+REVTOT+TYPMEN2:CC+CC:REVTOT+DIPLOPR:REVTOT+DIPLOCJ:REVTOT
```

On retient alors ce modèle auquel on stipule le paramétrage $weights = COEF$ pour tenir compte du coefficient de pondération et on obtient un modèle avec un AIC de 552,5.
Commande utilisée :

```
model <- glm( DOMTRAV~CC+DIPLOPR+TYPMEN2+DIPLOCJ+REVTOT+TYPMEN2:CC+CC:REVTOT+DIPLOPR:REVTOT+DIPLOCJ:REVTOT, data =
bdf, family = binomial(logit), weights = COEF)
```

### 1.1.2 Modèle réduit

Il y a 2 variables qui peuvent potentiellement poser des problèmes dans notre régression : DIPLOCJ car elle dispose de beaucoup de valeurs manquantes et réduit beaucoup notre nombre d'observations et TYPMEN2 car une fois la base nettoyée elle ne dispose que de 5 modalités sur les 7 de base : *2 - couple sans enfant, 3 - couple avec 1 enfant, 4 - couple avec 2 enfants, 5 - couple avec 3 enfants ou plus, 7 - autres cas*. On réitère la méthode pour le modèle complet sur les bases de données réduites (une sans DIPLOCJ l'autre sans TYPMEN2) et on remarque assez rapidement que les modèles sans DIPLOCJ ont un AIC bien plus forts que le modèle complet donc on décide de garder cette variable. Pour le modèle réduit sans TYPMEN2, on arrive à un modèle avec AIC de 572, ce qui est encore plus fort que ce qu'on a sur le modèle complet.
Modèle sans TYPMEN2 avec AIC le plus faible (572) :

```
DOMTRAV~DIPLOCJ+REVTOT+CC:REVTOT+DIPLOCJ:REVTOT
```

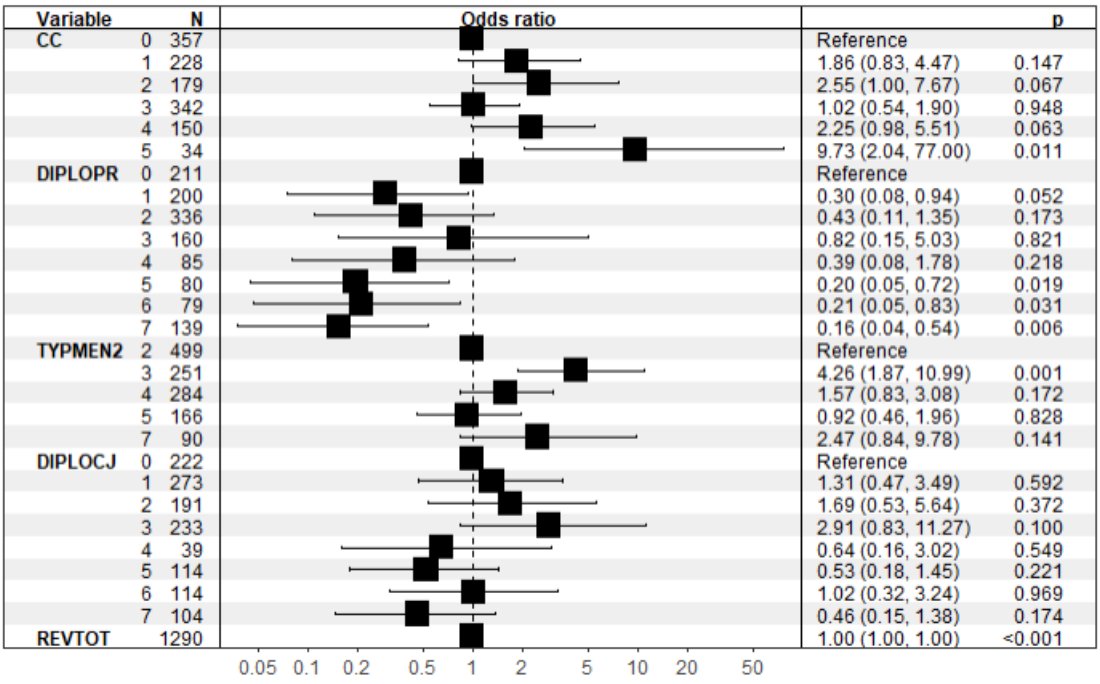
Ce modèle semble tout de même incomplet, on décide alors de garder notre modèle complet qui semble être le meilleur (AIC le plus faible).

## 1.2 En déduire une description des ménages qui emploient un ou une employé(e) de maison

On a remarqué que sur la régression de notre modèle retenu, le coefficient pour la modalité 5 de la variable CC (Degré d'urbanisation) était complètement absurde car de l'ordre de $4.408e+02$ et qui donc est encore plus absurde lorsqu'on y applique l'exponentielle. Après plusieurs essais, on a remarqué que le problème venait de l'intéraction entre CC (Degré d'urbanisation) et TYPMEN2 (Type de ménage), on a donc décidé d'enlever cette intéraction, cependant on a ensuite remarqué que suite à ce retrait les effets d'interactions avec la variable REVTOT permettent un gain d'AIC négligeable et perturbent l'interprétation donc on a décidé de faire notre interprétation sur le modèle sans effets d'intéractions qui a un AIC de 578.
Le modèle :

```
model <- glm(DOMTRAV~CC+DIPLOPR+TYPMEN2+DIPLOCJ+REVTOT, data = bdf,
family = binomial(), weights = COEF)
```

Odds ratios associés :




On remarque alors que très peu de coefficients sont significatifs car ces coefficients ont de très fortes variances, ce qui peut s'expliquer par une grande disparité dans les comportements humains et parce que notre modèle a, en réalité, quelques défauts : variables omises, multicolinéarité, biais dans les données... Cependant, on observe quand même des tendances qui nous permettent de dresser un profil des ménages qui emploient un ou une employé(e) de maison.
Dans ce modèle, la référence pour la variable à expliquer DOMTRAV est la valeur 1 donc "Oui", ainsi des odds ratios inférieurs à 1 indiquent une réponse vers ce "Oui" et des odds ratios supérieurs à 1 indiquent une réponse vers "Non".
Interprétation :

— CC (Degré d'urbanisation) : On a pour référence la modalité 0 qui correspond aux zones rurales. On remarque donc globalement que plus on habite en zone urbaine, moins on aura tendance à employer du personnel de maison (logements plus petits, plus chers) exceptées les

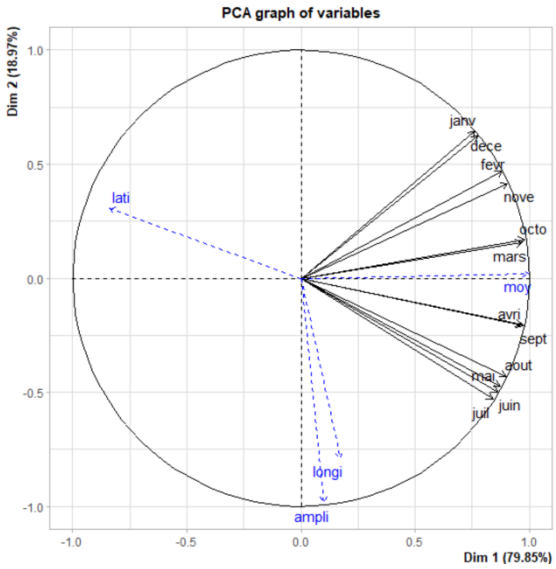
grandes villes hors Paris, qui suivent plutôt la tendance des zones rurales.

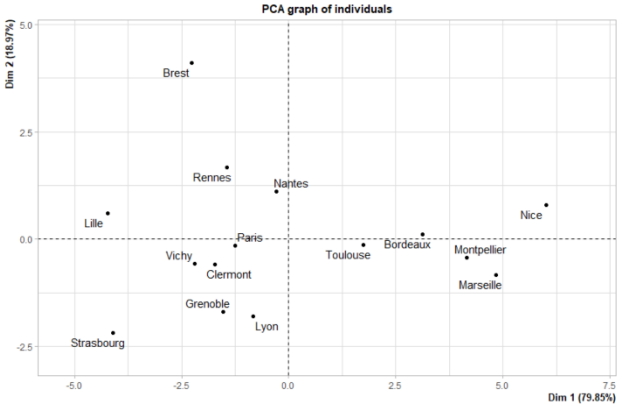
— DIPLOPR (Diplôme de la personne de référence) : Ici, les personnes sans diplôme constituent la référence. Globalement, on remarque que plus le diplôme est élevé, plus on est susceptible d'employer du personnel de maison.
— TYPMEN2 (Type de ménage) : Ici, la référence est un type de ménage où il y a un couple sans enfant. Globalement, par rapport aux couples sans enfant, la tendance sera plutôt de ne pas employer de personnels pour les couples avec 1 ou 2 enfants et pour les familles monoparentales. On remarque aussi que plus il y a d'enfants, plus le coefficient diminue, et donc moins on est enclin à employer du personnel de maison.
— DIPLOCJ (Diplôme du conjoint) : De même que pour DIPLOPR, les personnes sans diplôme constituent la référence. L'interprétation est semblable à celle pour DIPLOPR, à la différence que les ménages où les conjoints ont des diplômes en dessous du BAC auront plutôt tendance à ne pas employer du personnel de maison. On remarque que sur cette variable aucun coefficient n'est significatif.
— REVTOT (Revenu total) : Paradoxalement, le coefficient n'est pas significatif, on s'attendrait plutôt à un coefficient (odd ratio) inférieur à 1. En effet, plus on a de revenu, plus on peut se permettre d'employer des personnes pour les tâches de maison. Cela s'explique certainement par la colinéarité de REVTOT avec les autres variables.

# 2 Températures et précipitations en France

## 2.1 Analyse en composantes principales

L'ACP est réalisée grâce à la commande suivante :
> acp=PCA(villes, graph = T, quanti.sup = c(13 :16))




### 2.1.1 Quel est le pourcentage d'inertie expliquée par le premier plan factoriel ?

On obtient les pourcentages d'inertie expliquée :

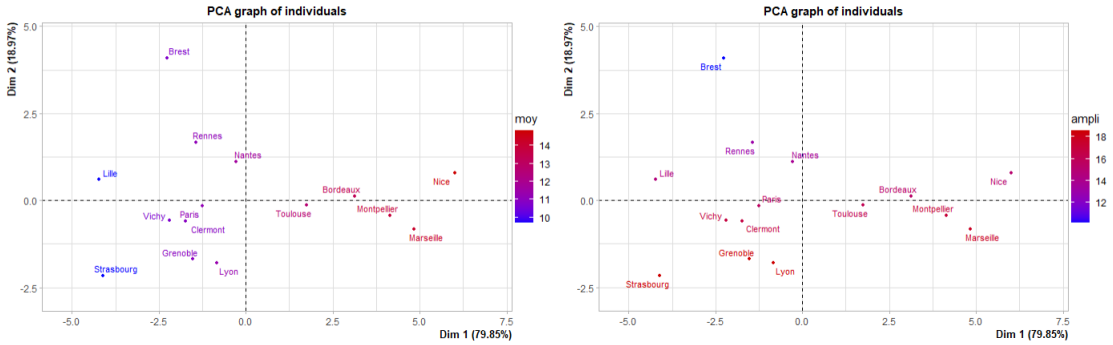
| | eigenvalue | variance.percent | cumulative.variance.percent |
|---|---|---|---|
| Dim.1 | 9.5817795809 | 7.984816e+01 | 79.84816 |
| Dim.2 | 2.2764183987 | 1.897015e+01 | 98.81832 |

Le pourcentage d'inertie expliqué par le premier plan factoriel est la somme de la variance expliquée par les deux premiers facteurs soit environ 98,8 % .

### 2.1.2 Interpréter les 2 axes à l'aide du cercle des corrélations et de la représentation des villes sur le premier plan factoriel

Le cercle des corrélations nous indique un effet taille sur le premier axe ainsi qu'un effet forme sur le second. L'intégralité des vecteurs représentant les mois sont situés à droite du premier axe et sont bien représentés car les flèches arrivent en bordure du cercle.

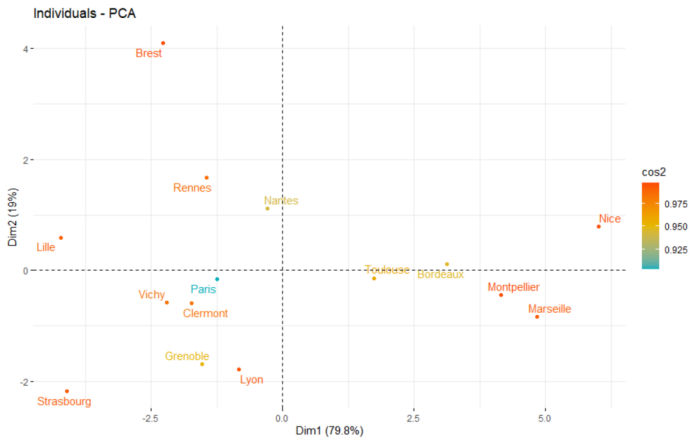
L'effet taille sur le premier axe oppose alors les villes les plus froides aux plus chaudes en moyenne peu importe la saison. On remarque grâce aux variables en quantités supplémentaires que le vecteur de la moyenne (très bien représenté) est quasiment confondu au premier axe. On aura alors à droite les villes les plus chaudes de France en moyenne sur l'année et à gauche les villes les plus froides. L'effet forme sur le second axe oppose le semestre le plus chaud en moyenne au semestre le plus froid. On aura alors en haut les villes où les températures sont plus douces en hiver/automne et/ou le plus frais en été/printemps et en bas les villes où les températures sont plus chaudes en été/printemps et/ou plus froides en hiver/automne. On remarque grâce aux variables en quantités supplémentaires que cet effet forme est équivalent à un effet taille sur la variable amplitude (très bien représenté) dont le vecteur est proche de l'axe 2. Plus on est bas dans le graphique, plus on a une forte amplitude (différence entre le mois le plus chaud et le mois le plus froid).




On peut aussi faire une interprétation complémentaire avec les vecteurs longitude et latitude, qui sont plutôt bien représentés. Le vecteur latitude s'oppose comme prévu au vecteur moyenne. Évidemment, les villes les plus au sud sont les villes les plus chaudes en moyenne et ce vecteur se place vers le haut, s'opposant à la période été/printemps où il fait le plus chaud au sud. Le vecteur longitude quant à lui se rapproche du vecteur amplitude, ce qui signifierait que plus on se déplace vers l'est (longitude positive), plus l'amplitude augmente.

### 2.1.3 Les individus "villes" sont-ils bien représentés sur le premier plan factoriel ?

On obtient les $cos^2$ des villes :
À l'aide du graphique ci-dessus, on remarque que l'échelle des $cos^2$ commence à 0.90. Il s'agit de Paris (en turquoise sur le graphique), qui est la ville la




moins bien «représentée». Toutefois, Paris a un $cos^2$ d'environ 0,9 et donc on peut dire que toutes les villes sont très bien représentées.

### 2.1.4 Quelles sont les caractéristiques des villes Lille et Strasbourg ? À quelles villes sont-elles opposées ?

Strasbourg et Lille sont les deux villes les plus à gauche, ce qui signifie que ce sont les villes où les températures sont les plus basses en moyenne sur l'année. La particularité de Strasbourg est qu'elle possède l'amplitude la plus grande, c'est-à-dire que la différence de température entre les saisons est, en moyenne, très grande. Strasbourg s'oppose non seulement aux villes les plus chaudes (Nice, Marseille) sur l'axe 1 mais aussi sur l'axe 2 à Brest, qui est la ville la plus à l'Ouest de notre échantilllon et où les températures sont plus homogènes qu'à Strasbourg, et donc où l'amplitude est la plus faible. Lille est la ville la plus au nord de notre échantillon, c'est celle où les températures sont les plus froides en moyenne. Elle s'oppose alors aux villes où les températures sont les plus chaudes en moyenne : Nice et Marseille.

### 2.1.5 Quelles sont les caractéristiques des villes Brest, Rennes et Nantes ? À quelles villes sont-elles opposées ?

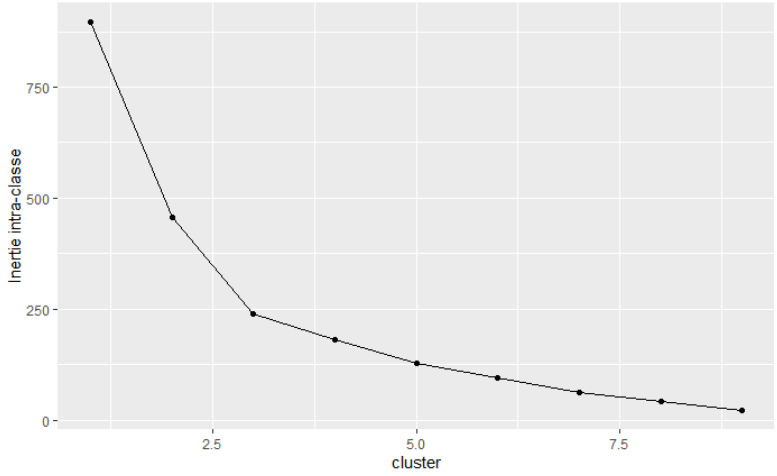
Les villes de Brest, Rennes et Nantes se situent toutes dans le même cadran de l'analyse en composantes principales, celui d'en haut à gauche. Autrement dit, elles se trouvent sur la partie négative de l'axe 1 et sur la partie positive de l'axe 2. Ceci conforte notre interprétation des deux axes : si on part de Rennes et que l'on va vers Nantes ($\simeq$ même longitude), on se dirige vers le sud, et donc vers des températures plus chaudes en moyenne. Cela est cohérent avec l'interprétation de l'axe 1. Maintenant, si on part de Brest vers Rennes ($\simeq$ même latitude), on se dirige vers le méridien de Greenwich, et

donc vers des amplitudes plus fortes. Cela est cohérent avec l'interprétation de l'axe 2. Ces villes s'opposent principalement aux villes du cadran d'en bas à droite (Toulouse, Montpellier, Marseille) mais si on regarde seulement l'amplitude (axe 2), elles s'opposent aussi aux villes les plus à l'est avec une latitude similaire (Strasbourg, Lyon, Grenoble).

## 2.2 Classification

### 2.2.1 Effectuer une classification grâce à la méthode des K-means. Interprétez cette classification. Quel nombre de classes vous semble le plus approprié ?

Le package « factoextra » possède une fonction pour déterminer le nombre de classes optimales selon plusieurs méthodes. Avec la méthode "silhouette" (average silhouette width) et "wss" (total within sum of square), on détermine un nombre de 3 classes optimales.
On peut tracer le graphe de l'inertie intra-classe en fonction du nombre de classes :



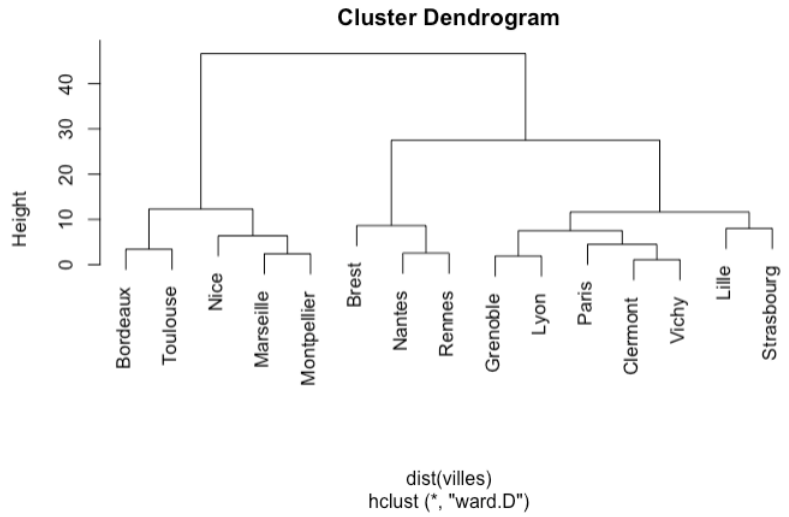

On remarque qu'à partir de 3 classes, le gain en inertie intra-classe n'est pas très important et on remarque aussi que les k-means ne convergent pas toujours vers les mêmes classes lorsqu'on fait des essais avec 4, 5 ou 6 classes. Cela n'est pas très optimal pour l'interprétation.

```
K-means clustering with 3 clusters of sizes 3, 7, 5

Cluster means:
      janv     fevr      mars     avri      mai     juin     juil     aout     sept     octo     nove     dece
1 5.300000 5.466667  8.033333 10.03333 12.86667 15.93333 17.43333 17.46667 15.60000 11.93333 8.333333 5.966667
2 2.114286 3.157143  7.028571 10.15714 13.92857 17.24286 19.24286 18.80000 15.94286 10.90000 6.357143 3.071429
3 5.780000 6.800000 10.040000 12.70000 16.08000 19.80000 22.10000 21.90000 19.28000 14.54000 9.880000 6.660000
      lati     longi      moy    ampli
1 47.80667 -2.343333 11.19667 12.36667
2 47.05000  4.134286 10.66143 17.12857
3 43.56400  3.368000 13.79400 16.34000

Clustering vector:
   Bordeaux       Brest    Clermont    Grenoble       Lille        Lyon   Marseille Montpellier      Nantes
          3           1           2           2           2           2           3           3           1
       Nice       Paris      Rennes  Strasbourg    Toulouse       Vichy
          3           2           1           2           3           2

Within cluster sum of squares by cluster:
[1]  36.49553 116.85443  86.04432
 (between_SS / total_SS =  73.3 %)
```

On remarque alors que la classe 3 se distingue par sa moyenne bien plus élevée que celle des autres classes, ce qui est cohérent avec le fait qu'elle se compose des villes du Sud. Les classes 1 et 2, quant à elles, représentent des villes plutôt au Nord mais elles se différencient par l'amplitude, l'amplitude de la classe 2 est bien plus importante et cela confirme notre interprétation de l'ACP sur le lien longitude/amplitude car la classe 1 regroupe des villes à l'Ouest alors que la classe 2 regroupe des villes plutôt au centre/ à l'Est.

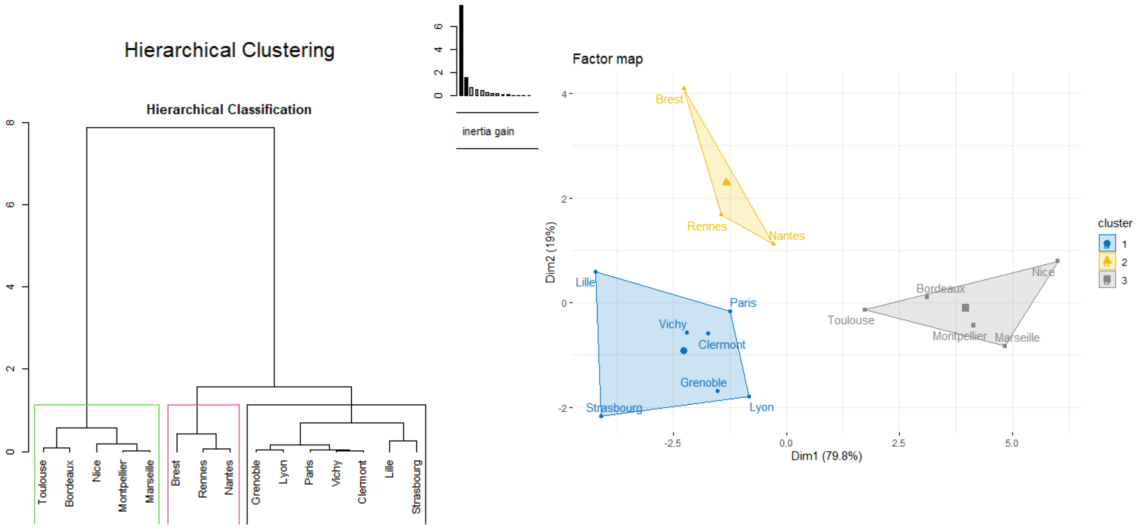
### 2.2.2 Effectuer une classification grâce à une classification hiérarchique. Interprétez cette classification. En combien de classes aurait-on envie de couper le dendrogramme ?

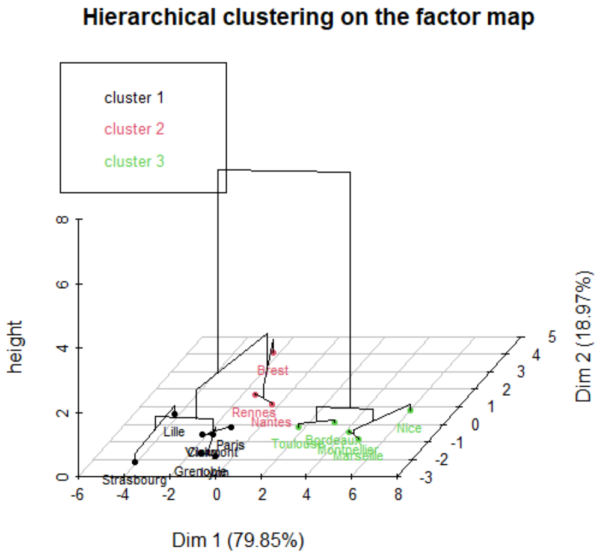
On obtient ce dendrogramme :




En appliquant la classification hiérarchique avec une nombre de classe égal à

3 sur les résultats de l'analyse en composantes principales, on a :

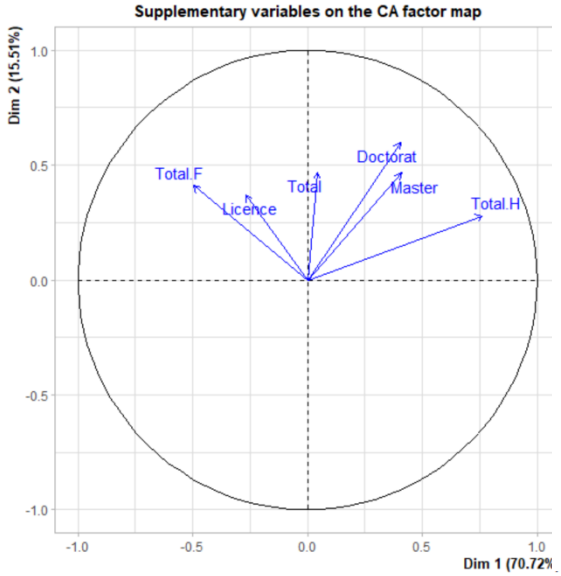


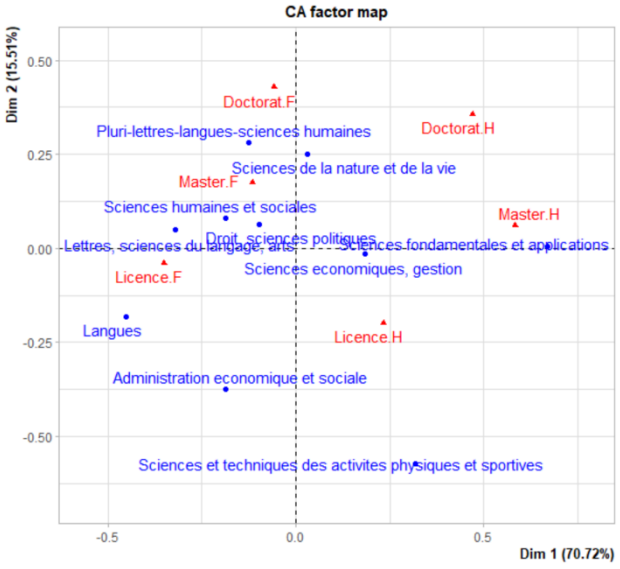




On pourrait couper le dendrogramme en 4, 5 voire même 6 classes car il regroupe évidemment les villes par climat et donc par conséquent de manière géographique. Cependant, on a vu avec les k-means que cela ne semble pas nécessaire de faire autant de classes : les différences de climat entre certaines classes seront minimes, et donc pas forcément intéressantes.

# 3 Universités : Analyse des correspondances

L'analyse des correspondances est réalisée grâce à la commande suivante :
> acf=CA(univ, graph = T, quanti.sup = c(7 :12))




## 3.1 Quel est le pourcentage d'inertie expliquée par le premier plan factoriel ?

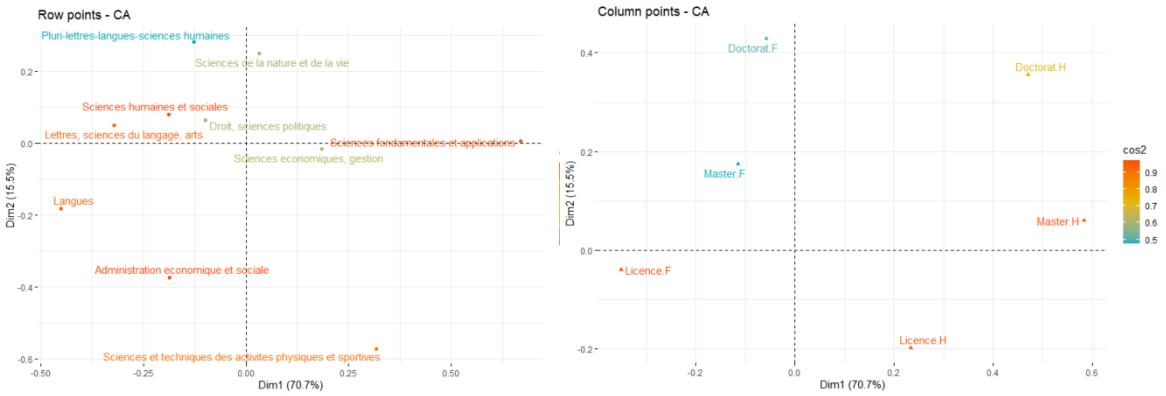
On obtient les pourcentages d'inertie expliquée :

```
       eigenvalue variance.percent cumulative.variance.percent
Dim.1 0.1167660841        70.717688                    70.71769
Dim.2 0.0256061155        15.507973                    86.22566
```

Le premier plan factoriel admet une inertie expliquée d'environ 86.3 %.

## 3.2 Pensez-vous que les disciplines littéraires attirent surtout les femmes ?

On obtient les $cos^2$ :




Dans un premier temps, on regarde les qualités de représentations des niveaux d'études des Femmes. On remarque que Master F et Doctorat F (en bleu turquoise sur le graphique de droite) ne sont pas bien représentés sur le premier plan factoriel. En effet, le $cos^2$ est inférieur ou égal à 0,5. On ne peut ainsi qu'interpréter la proximité de Licence F qui est bien représentée sur le premier plan factoriel avec les disciplines littéraires. Ainsi, en observant la « CA factor map », on remarque que Licence F est proche de Langues et de Lettres, sciences du langage, arts. En outre, on voit que ce sont les Femmes qui se trouvent sur la partie négative de l'axe 1, et les matières plutôt littéraires aussi. L'axe 1 oppose les matières scientifiques aux matières littéraires. Donc les matières littéraires étant sur la partie négative de l'axe 1, au même titre que les Femmes, on peut conclure que les disciplines littéraires attirent surtout les femmes.

## 3.3 Les sciences attirent-t-elles les hommes ?

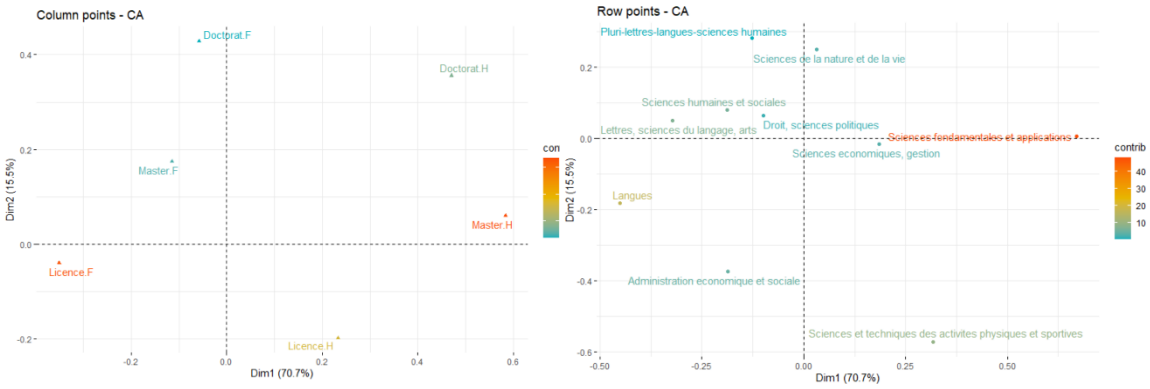
On peut interpréter la proximité entre les sciences et les Hommes car toutes les colonnes ou lignes sont bien représentés dans le premier plan factoriel avec de bon $cos^2$ (cf les deux graphiques illustrant les $cos^2$). À l'inverse de la réponse précédente, les Hommes se trouvent sur la partie positive de l'axe 1 et les matières scientifiques du même côté. Donc on peut dire que les sciences attirent les hommes.

## 3.4 Les études d'AES sont-elles longues ou courtes ?

L'axe 2 oppose quant à lui les études longues (Doctorat) aux études courtes (Licence). Au niveau des $cos^2$, AES est bien représenté sur le premier plan factoriel ainsi que les Licence H et F. Ainsi, on peut interpréter la proximité, et comme AES est proche de Licence H et F, on peut conclure que les études d'AES sont plutôt courtes.

## 3.5 Donner une interprétation des deux premiers axes

Contributions au premier plan factoriel :




L'axe 1 oppose les matières littéraires aux matières scientifiques. Ainsi, l'axe 1 explique le type, la discipline ou le domaine d'études faites par l'individu. L'axe 2, quant à lui, explique la durée des études, il oppose les études longues aux études courtes.

## 3.6 Que pensez-vous de la proximité SVT et Master F ?

La discipline SVT est bien représenté mais le $cos^2$ de Master F étant inférieur à 0,5, on ne peut pas interpréter la proximité de ces deux éléments.

## 3.7 Que pensez-vous de la proximité Lettres, arts et Licence F ?

Les $cos^2$ de Lettres, arts et Licence F sont bons, ainsi on peut affirmer que les études dans le domaine Lettres, arts sont majoritairement féminines et globalement courtes (licence).