Projet : classification et AFC

Bruno KUBECZKA

07 mai 2023

Abstract

Le projet a pour objet de mettre en oeuvre analyse factorielle des correspondances et classification sur les résultats des élections présidentielles de 2017. En l'occurrence, on orientera l'étude sur l'ancrage territorial des candidats.

Contents

1 Présentation de l'étude

1.1 Introduction

Dans cette étude des résultats de l'élection présidentielle de 2017, on va s'atteler à tirer des tendances sur l'**ancrage des candidats sur le territoire national**.

En d'autres termes, il va s'agir pour chaque candidat (y compris, votes blancs, votes nuls et abstention)

- de rapprocher les candidats au profil similaire,
- de rapprocher ces candidats aux départements votant plus ou moins pour eux

La méthodologie proposée est la suivante

- On procédera à une Analyse des Correspondances (AFC) pour amener les modalités Candidats et Départements dans un espace géométrique commun
- On interprétera les résultats pour en extraire des premières tendances de rapprochements entre modalités
- On procédera à une classification des scores colonnes de l'AFC pour regrouper les candidats les plus proches
- Enfin on caractérisera les groupes de candidats pour identifier les départements les plus marquants pour le groupe.

1.2 Présentation des données

Les données sont fournies sous la forme d'un **tableau de contingence** prenant

- en **ligne**, les **106 départements français**
- en **colonne**, les **11 candidats aux élections présidentielles de 2017**, l'**abstention**, le **vote nul** et le **vote blanc** (soit 14 modalités en tout)

Les valeurs sont les **effectifs (en nombre de votes)** par département pour chacun des candidats.

1.3 Pertinence de l'étude : indépendance des variables d'étude

L'intérêt de l'étude réside dans le fait que lignes (départements) et colonnes (candidats) sont significativement associées.

Pour s'en assurer, on prend le soin de valider la dépendance des 2 variables d'étude à savoir les départements d'un côté et les candidats de l'autre.

On procède à un **test du** χ^2 sur le tableau de contingence tel que

- **H0 hypothèse nulle** : lignes et colonnes du tableau de contingence sont indépendants
- **H1 hypothèse alternative** : lignes et colonnes du tableau de contingence sont dépendantes les unes des autres

Dans ce cas, le nombre de degrés de liberté est (I-1)(J-1) soit (106 - 1) * (14-1) = 1365.

La fonction *chi.test* de R donne le résultat suivant:

```
##
##  Pearson's Chi-squared test
##
## data:  donnees_elections
## X-squared = 3630539, df = 1365, p-value < 2.2e-16
```

En l'occurrence, **la p-valeur de** 2.210^{-6} **est** $< 5\%$, on rejette donc H0: "les variables sont indépendantes".

On confirme la pertinence de l'étude.

2 Analyse factorielle des correspondances (AFC)

2.1 Profils des candidats (colonnes)

A partir du tableau de contingence conjoint des variables Départements et Candidats, on crée un tableau de **profils colonnes** i.e, pour chaque candidat, la **distribution de ses votes répartis sur l'ensemble des départements** ($\sum = 1$).

Pour cela, on calcule

- la **marge ligne** c-a-d la somme des effectifs en colonne soit dans notre cas la somme des votants pour un candidat donné, tout département confondu
- la **marge colonne** c-a-d la somme des effectifs en ligne soit dans notre cas les effectifs par département des inscrits sur les listes électorales
- l'**effectif total** i.e. le nombre d'inscrits sur les listes électorales

Les **profils des colonnes (candidats)** se calculent alors de la façon suivante :

$$profils.colonnes = \frac{colonnes}{marge.ligne}$$

Et le **profil colonne moyen** devient alors

$$profil.colonne.moyen = \frac{marge.colonne}{effectif.total}$$

```
# marge ligne : Somme des effectifs par colonne
marge.ligne <- apply(donnees_elections, MARGIN=2, FUN=sum)

# marge colonne : Somme des effectifs par ligne
marge.colonne <- apply(donnees_elections, MARGIN=1, FUN=sum)

# effectif total : somme des effectifs par département
effectif.total <- sum(marge.ligne)
```

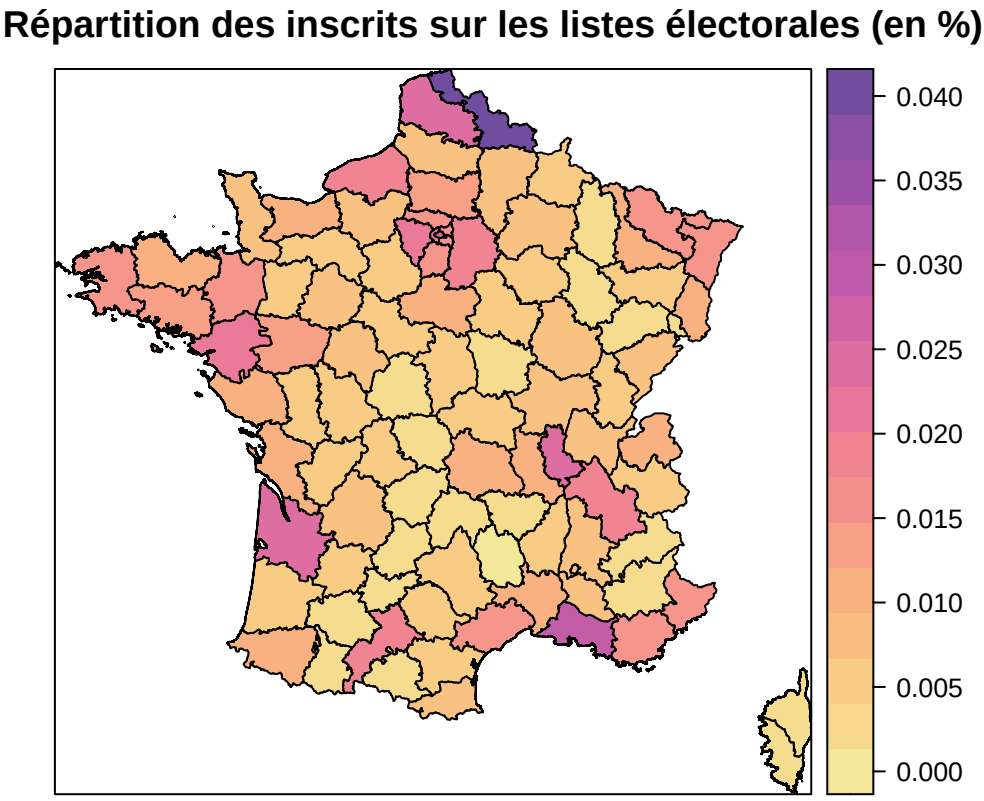
```
# Calcul des profils colonnes
profils.colonnes <- t(t(donnees_elections) / marge.ligne)

# calcul du profil colonne moyen
profil.colonne.moyen = marge.colonne / effectif.total
```

Ainsi calculés

- Un **profil colonne (candidat)** contient la **répartition par département des effectifs ayant voté pour le candidat**.
- Le **profil colonne moyen** contient la **répartition par département des effectifs inscrits sur les listes électorales**. C'est **la situation d'indépendance** pour laquelle la répartition des votants pour un candidat calquerait la répartition des inscrits par département.

On peut ainsi visualiser le profil colonne moyen sur une projection géographique qui fait apparaître la distribution des inscrits sur le territoire national.



Les zones les plus denses sont logiquement localisées autour des grandes métropoles ; on note aussi quelques autres pôles comme le bord de la méditerranée, la Bretagne, la frontière franco-allemande et la Normandie.

2.2 Analyse Factorielle des Correspondances (AFC)

L'analyse factorielle des correspondances va permettre de rapprocher l'ensemble des modalités (candidats et départements) dans un même espace de dimensions restreintes.

Dans l'étude que nous menons sur les ancrages territoriaux, il s'agira d'interpréter les projections des modalités sur des plans pertinents afin

- d'identifier les candidats proches les uns des autres (pour en déduire les prémices d'une classification)
- d'identifier les départements proches de ces candidats (pour en déduire des potentiels départements votant davantage pour ces candidats)

Techniquement, on va procéder à une **analyse factorielle des correspondances** entre les modalités "départements" et les modalités "candidats" en utilisant la fonction *CA* de la librairie *FactoMineR*.

On note que l'AFC a cette caractéristique de construire un espace de dimension *min(I-1, J-1)* (au-delà l'inertie portée par les axes est nulle). Dans notre cas où nous avons en lignes 106 départements (*I vaut 106*), et en colonnes 14 candidats (*J vaut 14*), l'AFC créera un espace à 13 dimensions maximum.

Par ailleurs, on note les points suivants

- le nombre de modalités Départements et Candidats est de taille raisonnable et il n'est pas nécessaire de les diminuer en amont par une pré-classification de type k-moyennes.
- l'espace créé par l'AFC disposera d'un maximum de 13 dimensions ; c'est cette limite dimensionnelle que nous préciserons dans l'appel à la fonction.

d'où l'appel à la fonction suivante:

```
res.CA <- CA(donnees_elections, ncp=13, graph=FALSE)
```

2.3 Interprétation des axes de l'AFC

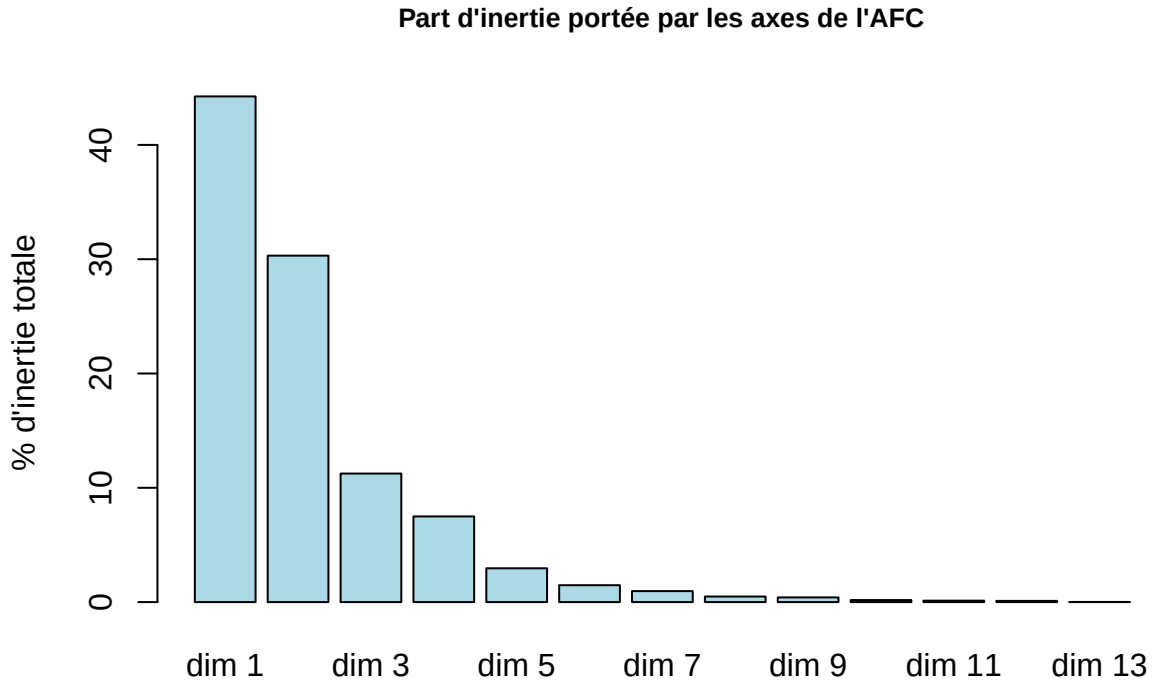
Ce chapitre a pour vocation d'interpréter les axes en mettant en avant, plan par plan,

- les modalités "candidats" les plus contributives à la construction des axes
- les modalités "candidats" les mieux représentées (cos^2) des axes
- les 15 modalités "départements" les plus contributives

2.3.1 Inertie portée par les axes

Avant d'étudier les différents plans, on examine la part d'inertie totale portée par les 13 axes de l'AFC (en % d'inertie totale):

```
##  dim 1  dim 2  dim 3  dim 4  dim 5  dim 6  dim 7  dim 8  dim 9 dim 10 dim 11
##  44.24  30.31  11.25   7.49   2.96   1.48   0.96   0.49   0.41   0.18   0.12
## dim 12 dim 13
##   0.10   0.01
```



Les 4 premiers axes représentent un total de **93,29% de l'inertie totale**.

Pour simplifier le travail sur les candidats et les départements, on utilisera autant que faire se peut ces 4 axes. Néanmoins, on ne s'interdit pas d'utiliser les axes au-delà sur lesquels de l'information relative à des candidats "plus modestes" pourrait apparaître.

2.3.2 Pertinence des axes par candidat

Afin de visualiser les candidats sur les plans les plus pertinents, on étudie 2 données les concernant:

- leur **contribution** à la construction des axes de l'AFC (*col$contrib*), par ordre décroissant
- leur **qualité de représentation** sur les axes de l'AFC (*col$cos2*), par ordre décroissant

Le tableau ci-dessous résume pour chaque candidat (modalité en colonne du résultat de l'AFC) les meilleurs axes selon les critères suivants:

- axes représentant a minima 10% de la contribution de la modalité colonne (*col$contrib>10%*)
- axes sur lesquels la qualité de représentation cos^2 de la modalité colonne vaut a minima 0,25 (*col$cos2 > 0.25*)

	Contribution	Qualité de représentation
Abstentions	1, 2	1, 2

	Contribution	Qualité de représentation
Blancs	10, 8, 11, 9	5
LEPEN	2, 6	2
MELENCHON	3, 5, 4, 6	3, 1
MACRON	8, 1, 9	1, 2
FILLON	3, 4	3, 1
LASSALLE	4, 3	4, 3
DUPONT-AIGNAN	6, 5	2, 5
HAMON	9, 6, 8	1, 2
ASSELINEAU	10, 9, 11, 6	6, 10, 9
POUTOU	12, 8	5, 3
ARTHAUD	11, 12, 5	5
CHEMINADE	13	13, 5
Nuls	7, 9, 8	1, 7

Au vu de ce tableau récapitulatif, on appliquera la **politique de visualisation** suivante :

- On visualisera les 6 premiers axes qui permettent de couvrir l'ensemble des candidats (soit en contributeurs, soit en meilleures représentations)
- Pour les candidats contribuant aux axes jusque 6 (portant 97.72% de l'inertie totale du jeu de données), on les fera apparaître sur les plans auxquels ils contribuent.
- Pour les candidats dont la contribution est concentrée sur les axes au-delà de 6, on prendra le parti de les visualiser sur les axes où ils sont le plus visibles.

  par ex. **POUTOU**, contributeur des "petits axes" 8 et 12, sera volontairement visualisé sur les axes 3 et 5 qui portent la représentation la plus importante.

2.3.3 Interprétation des axes 1 et 2

Candidats les plus contributeurs des axes (*col$contrib[j]>10%*)

Axe 1

```
## Abstentions      MACRON
##       59.19       19.13
```

Axe 2

```
##      LE PEN Abstentions
##       67.26       13.92
```

Candidats les mieux représentés sur les axes (*col$cos2[j]>0.25*)

Axe 1

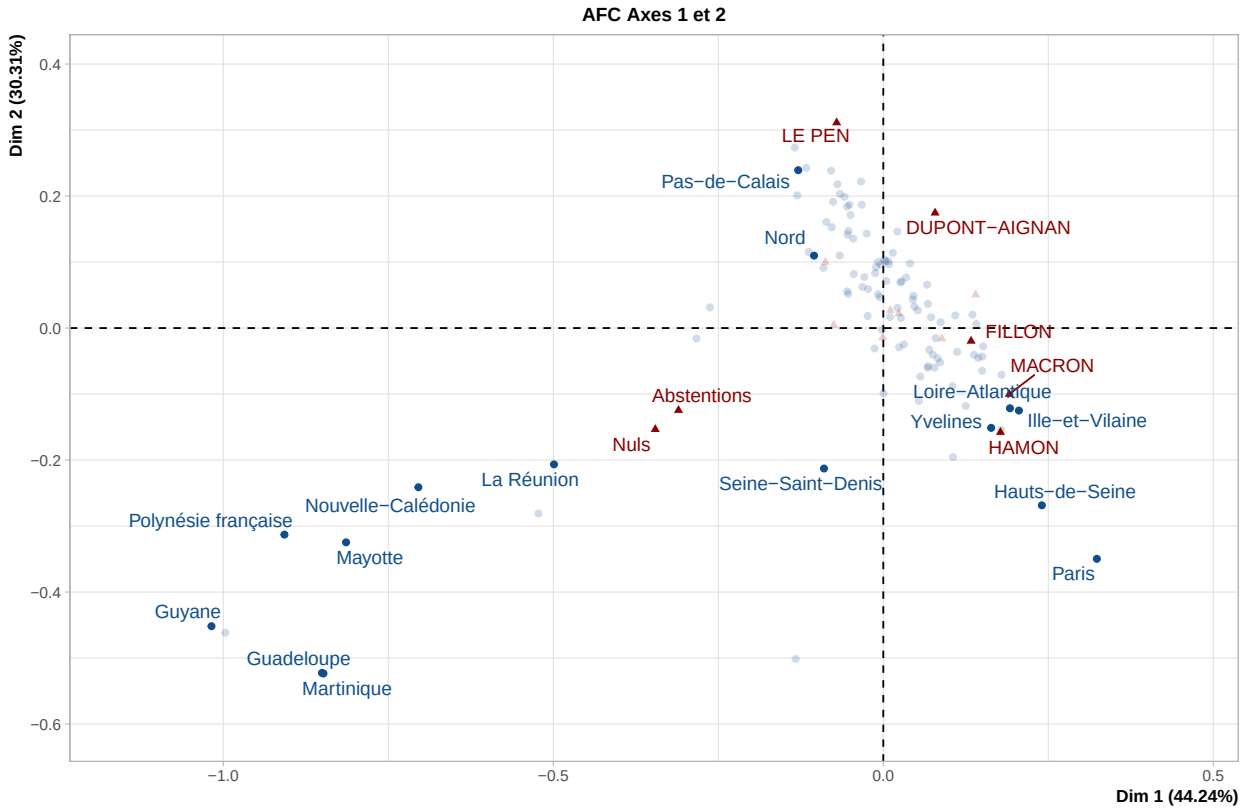
```
## Abstentions      MACRON        Nuls       HAMON      FILLON
##        0.86        0.75        0.45        0.42        0.36
```

Axe 2

```
##        LE PEN DUPONT-AIGNAN         HAMON
##          0.94          0.43          0.33
```

Sur le **plan 1:2**,

- on affichera donc les candidats les plus contributeurs à savoir **Abstentions**, **MACRON**, **LE PEN**
- on complétera avec les candidats non contributeurs les mieux représentés ($cos^2 > 0.25$) : **FILLON**, **HAMON**, **Nuls**, **DUPONT-AIGNAN**



Interprétation

Le plan oppose

- sur l'axe 1 Abstentions et Nuls d'un côté aux candidats **HAMON**, **MACRON**, **FILLON**.
- sur l'axe 2 DUPONT-AIGNAN et LE PEN aux candidats **MACRON** et **HAMON** ainsi que **Blancs** et **Nuls**.

On note par ailleurs

- Le rapprochement des modalités **Nuls** et **Abstentions**.

  Ce sont des modalités qui cotoient des **départements d'Outre-Mer**, parmi les plus contributeurs de la construction des axes 1 et 2.

- Le rapprochement des candidats **DUPONT-AIGNAN** et **LE PEN**.

  Ce sont des modalités proches des départements du nord de la france : **Pas de Calais**, **Nord**

- Le rapprochement des candidats **FILLON**, **MACRON** et **HAMON**.

  Ils côtoient des départements contributeurs de l'île de France (**Paris**, **Hauts-de-Seine**, **Yvelines**) et de l'ouest de la France (**Loire-Atlantique** et **Ille-et-Vilaine**).

2.3.4 Interprétation des axes 3 et 4

Candidats les plus contributeurs des axes (*col$contrib[j]>10%*)

Axe 3

```
##    FILLON  LASSALLE MELENCHON
##     43.30     25.93     24.55
```

Axe 4

```
##  LASSALLE MELENCHON    FILLON
##     70.08     16.82     10.18
```

Candidats les mieux représentés sur les axes (*col$cos2[j]>0.25*)

Axe 3

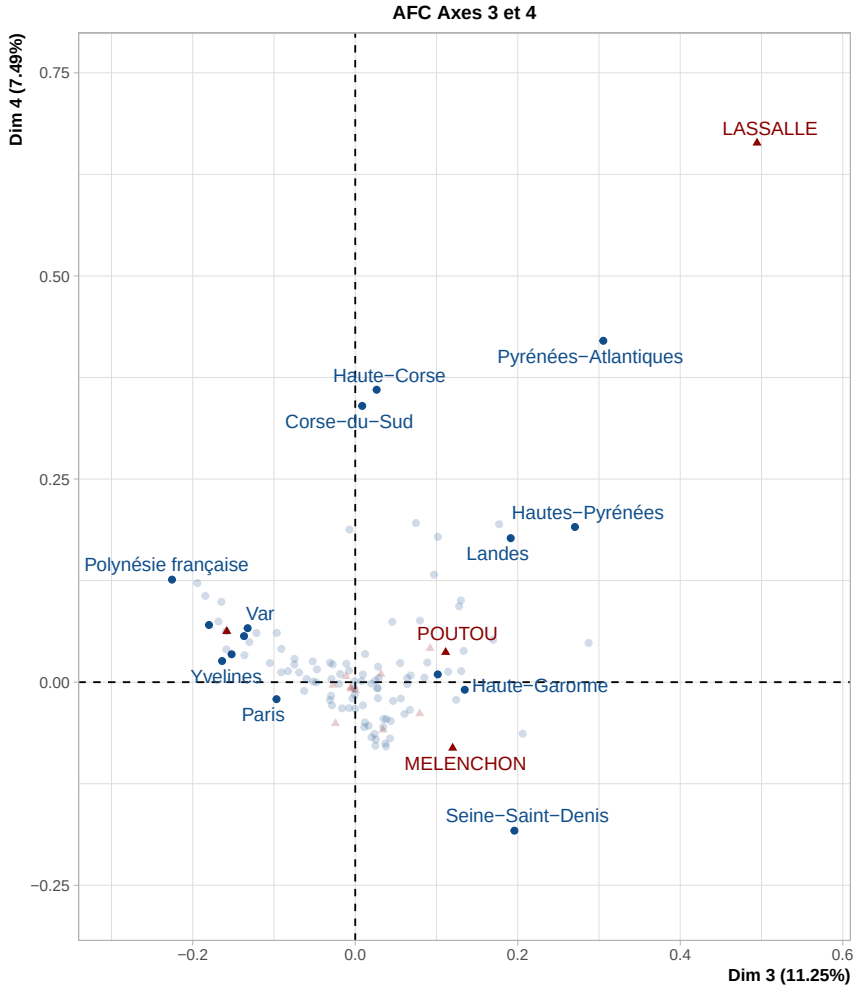
```
##    FILLON MELENCHON  LASSALLE
##      0.51      0.43      0.34
```

Axe 4

```
## LASSALLE
##     0.62
```

Sur le **plan 3:4**,

- on affichera les candidats les plus contributeurs et les mieux représentés à savoir **FILLON**, **LASSALLE**, **MELENCHON**
- on complétera par le candidat **POUTOU** (non contributeur) dont on a vu que l'axe 3 est la meilleure représentation.



Interprétation

Le plan met en avant la particularité du candidat **LASSALLE** très éloigné des autres candidats sur les axes 3 et 4.

Il oppose aussi sur l'axe 3 le candidat **FILLON** des candidats **POUTOU** et **MELENCHON**

Le plan confirme aussi la proximité du candidat **FILLON** des départements **Yvelines** et **Paris** et y ajoute les départements **Var**.

On note par ailleurs

- le rapprochement dans une certaine mesure des candidats **POUTOU** et **MELENCHON**, concomitamment proches de la **Haute-Garonne**,
- la proximité du candidat **MELENCHON** et le département **Seine-Saint-Denis**
- la rapprochement du candidat **LASSALLE** des départements corses (**Haute-Corse** et **Corse-du-Sud**) et du sud-ouest (**Pyrénées-Atlantiques**)

2.3.5 Interprétation des axes 5 et 6

Candidats les plus contributeurs des axes (*col$contrib[j]>10%*)

Axe 5

```
## DUPONT-AIGNAN     MELENCHON       ARTHAUD
##         28.61         18.63         10.51
```

Axe 6

```
## DUPONT-AIGNAN         HAMON        LE PEN     MELENCHON
##         38.18         21.64         10.91         10.53
```

Candidats les mieux représentés sur les axes (*col$cos2[j]>0.25*)

Axe 5

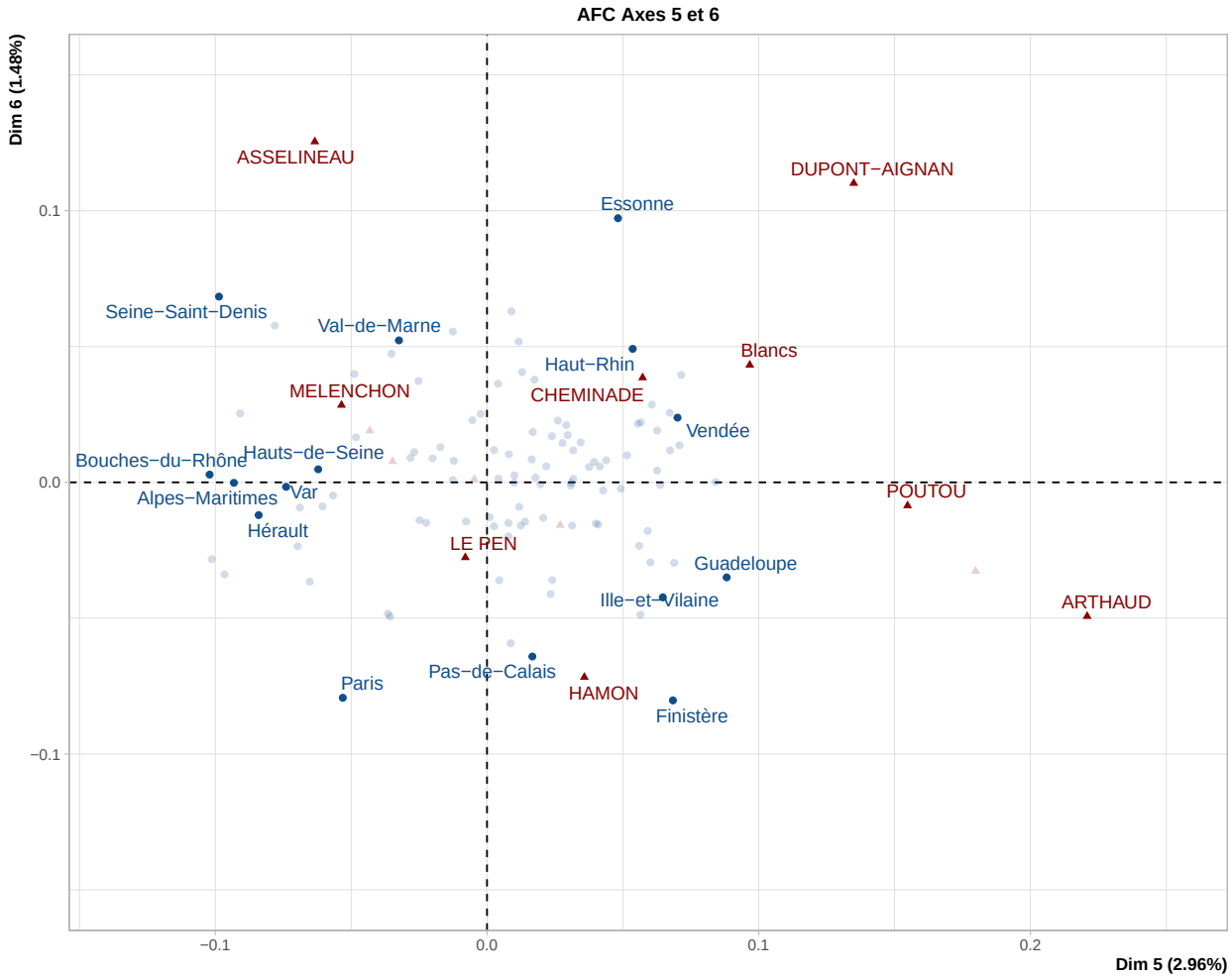
```
##       ARTHAUD        POUTOU        Blancs DUPONT-AIGNAN
##          0.49          0.43          0.27          0.26
```

Axe 6

```
## ASSELINEAU
##       0.35
```

Sur le **plan 5:6**,

- on affichera les candidats les plus contributeurs à savoir **DUPONT-AIGNAN**, **MELENCHON**, **ARTHAUD**, **HAMON**, **LE PEN**
- on complétera avec les candidats non contributeurs les mieux représentés sur le plan à savoir **POUTOU**, **Blancs**, **ASSELINEAU**
- on ajoutera le candidat **CHEMINADE** (non contributeur) dont on a vu que l'axe 5 est la meilleure représentation possible.



Interprétation

Le plan 5:6 (qui a été zoomé pour plus de lisibilité) présente des **candidats très concentrés autour de l'origine**, qui projettent peu de variabilité autour du barycentre vu des axes 5 et 6.

Néanmoins cette concentration autour de l'origine met en avant

- la proximité des candidats **CHEMINADE**, **ASSELINEAU**, **ARTHAUD**, **Blancs** voire **POUTOU**, peu ou pas représentés sur les autres plans.
- un rapprochement des candidats **POUTOU**, **HAMON**, **MELENCHON**, **LE PEN** et **DUPONT-AIGNAN** que l'on considérera de façon mesurée, sachant que ce sont des candidats qui sont par ailleurs mieux représentés sur d'autres plans.
- un candidat **DUPONT-AIGNAN** proche du département **Essonne** (où il est maire et député)
- un candidat **ASSELINEAU** proche de départements d'île de France comme **Seine-Saint-Denis**, **Val-de-Marne** et **Essonne**.
- un candidat **HAMON** proche du **Pas-de-Calais** et du **Finistère**.

2.4 Conclusion de la phase d'interprétation : quelques tendances

L'interprétation des axes a pu mettre en évidence quelques tendances

- relatives au rapprochement des candidats entre eux
- relatives au rapprochement des candidats avec des départements

Ainsi, l'interprétation rapproche les modalités suivantes

- **Abstentions** et **Nuls** sur le plan 1:2 s'approchent des départements de la **France d'Outre-Mer**.
- **LE PEN** et **DUPONT-AIGNAN** sur le plan 1:2, puis sur le plan 5:6, s'approchent des départements **Pas-de-Calais**, **Nord** et **Essonne**.
- **MACRON**, **FILLON**, **HAMON** sur le plan 1:2 s'approchent des départements d'Ile-de-France (**Paris**, **Yvelines**, **Hauts-de-Seine**) et de l'ouest de la France (**Ille-et-Vilaine** et **Loire-Atlantique**)
- **MELENCHON** et **POUTOU** sur les plans 3:4 puis 5:6, sans pour autant mettre en avant un rapprochement de départements en particulier
- **Blancs**, **ARTHAUD**, **POUTOU**, **ASSELINEAU**, **CHEMINADE** sur le plan 5:6
- **LASSALLE** sur le plan 3:4, proches des départements corses **Corse-du-sud** et **Haute-Corse**, et du sud-ouest de la France **Pyrénées-Atlantique**, **Haute-Pyrénées**, **Landes**

3 Rapprochement des candidats (classification)

Nous venons d'effectuer une analyse factorielle des correspondances (AFC) qui a permis de rassembler dans un espace unique à 13 dimensions les modalités Candidats et Départements. L'interprétation "à l'oeil" de leur projection sur différents plans pertinents a fait apparaître des rapprochements "géométriques" de quelques candidats d'une part, et des candidats et des départements d'autre part.

L'objet de ce chapitre est de consolider les **rapprochements des candidats** de façon algorithmique.

Pour ce faire, on va procéder à une **classification des candidats en se basant sur les distances entre leurs scores colonnes respectifs**.

Le nombre de colonnes étant limité (14 modalités), on procédera à une **classification ascendante hiérarchique** (CAH) qu'on consolidera ensuite par une **classification par la méthode des k-moyennes** (les centres initiaux découleront des classes identifiées par la CAH)

3.1 Classification Ascendante Hiérarchique (CAH) sur les scores colonnes (candidats)

L'étude des candidats nous amène à traiter les scores en colonnes.

L'intérêt de procéder à une classification sur les résultats de l'analyse factorielle est que l'analyse factorielle concentre l'information du jeu de données (sa variabilité) sur les 1ères composantes principales, les dernières composantes pouvant être assimilées à du bruit i.e de l'aléatoire. En éliminant les dernières composantes de l'AFC, on stabilisera les résultats de la classification.

Afin d'éliminer ce qu'on peut considérer comme du bruit, on va procéder à une classification basée sur la **distance des colonnes sur les 4 premiers axes** représentant **93.29% de l'inertie totale**.

La classification se base sur le **critère de perte minimale d'inertie inter-classes de Ward sur distance euclidienne** entre les **scores des candidats (en colonnes)**, exprimés pour l'occasion sur **4 dimensions** de l'AFC.

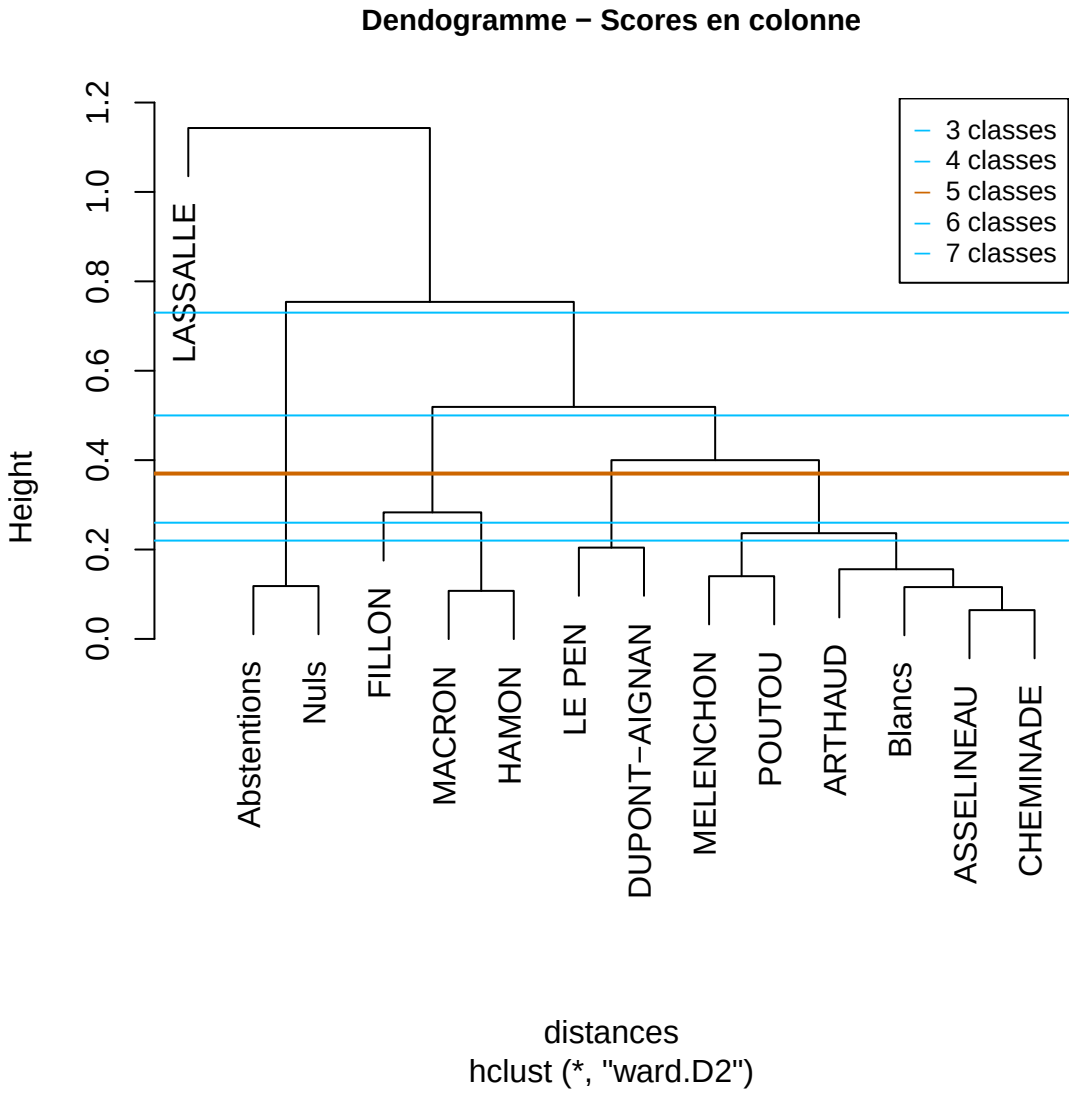
Pour cela,

- on extrait les scores colonnes sur 4 dimensions
- on calcule la distance euclidienne entre les scores colonnes restreints
- on utilise la fonction *stats::hclust* sur les distances et on visualise le **dendogramme** qui en résulte.

```
## Scores des candidats (colonnes)
## sur 4 dimensions
scores.nb.dimensions <- 4
scores.colonnes <- res.CA$col$coord[,1:scores.nb.dimensions]

## Distances euclidiennes entre les scores colonnes
distances = dist(scores.colonnes)

## Classification "Ward" sur distance euclidienne
res.CAH.colonnes = hclust(distances, method="ward.D2")
```



De haut en bas (de 2 classes à n classes), on visualise les **sauts d'inertie intra-classes** au fur-et-à-mesure qu'on augmente le nombre de classes (attribut *height* du résultat de la classification) : plus le saut entre "étages" est grand, plus le "gain" d'inertie intra-classes est grand.

On cherche le nombre de classes tel que les variations de l'inertie intra-classes (représentée par l'ordonnée *height*) restent **petites et stables**.

```
##  [1] 1.14302260 0.75402885 0.51906629 0.39993982 0.28308476 0.23658198
##  [7] 0.20431791 0.15592238 0.14034872 0.11828212 0.11599432 0.10751109
## [13] 0.06443673
```

En l'occurrence, les sauts peuvent être considérés comme importants jusqu'au passage de 5 à 6 classes ; ils deviennent minimes au-delà.

Ainsi une segmentation en 6 classes semble optimale au sens de la minimisation de l'inertie intra-classes. néanmoins elle isole le candidat FILLON qu'on préférera intégré dans un groupe plus large.

On prend donc le parti de conserver **une classification à 5 classes** plus lisible.

3.2 Choix des classes

La méthode *stats::cutree* permet d'extraire les classes du résultat de la CAH. En l'ocurrence, on spécifie une segmentation en **5 classes** (attribut *k=5*)

```
nb_groups = 5
groups <- cutree(res.CAH.colonnes, k=nb_groups)
```

Les **5 groupes de candidats** qui en découlent sont donc :

```
## $groupe1
## [1] "Abstentions" "Nuls"
##
## $groupe2
## [1] "Blancs"     "MELENCHON"  "ASSELINEAU" "POUTOU"     "ARTHAUD"
## [6] "CHEMINADE"
##
## $groupe3
## [1] "LE PEN"        "DUPONT-AIGNAN"
##
## $groupe4
## [1] "MACRON" "FILLON" "HAMON"
##
## $groupe5
## [1] "LASSALLE"
```

3.3 Consolidation des classes par la méthode des k-means / Qualité de la classification

La partition par la méthode de classification hiérarchique peut ne pas être optimale, à cause justement du principe de hiérarchisation.

Afin de stabiliser la classification (et vérifier qu'un candidat n'a pas intérêt à se rapprocher d'un groupe plus proche),

- on va procéder à une **classification des scores colonnes par la méthode des k-means**,
- en spécifiant comme **centres initiaux** les centres des classes identifiées précédemment (on utilise les scores en colonnes sur les 4 premières dimensions)

```
## calcul des coordonnées des centres initiaux
scores.colonnes.centers <- colMeans(scores.colonnes[c("Abstentions", "Nuls"),])

scores.colonnes.centers <- rbind(scores.colonnes.centers,
                                 colMeans(scores.colonnes[c("Blancs", "MELENCHON",
                                                            "ASSELINEAU", "POUTOU",
                                                            "ARTHAUD", "CHEMINADE"),]))
scores.colonnes.centers <- rbind(scores.colonnes.centers,
                                 colMeans(scores.colonnes[c("LE PEN", "DUPONT-AIGNAN"),]))

scores.colonnes.centers <- rbind(scores.colonnes.centers,
                                 colMeans(scores.colonnes[c("MACRON", "FILLON", "HAMON"),]))

scores.colonnes.centers <- rbind(scores.colonnes.centers,
                                 scores.colonnes["LASSALLE",])
```

```
res.CA.kmeans <- kmeans(scores.colonnes, centers=scores.colonnes.centers)
```

```
## K-means clustering with 5 clusters of sizes 2, 6, 2, 3, 1
##
## Cluster means:
##          Dim 1       Dim 2       Dim 3        Dim 4
## 1 -0.327912835 -0.13901152  0.04030236  0.024368049
## 2 -0.006730801  0.02037293  0.04443506 -0.025025087
## 3  0.003909502  0.24282999 -0.01368960 -0.006628047
```

```
## 4  0.167153251 -0.09258325 -0.02799653  0.005230090
## 5  0.139885004  0.05042179  0.49447993  0.663691562
##
## Clustering vector:
##   Abstentions        Blancs          Nuls        LE PEN     MELENCHON
##             1             2             1             3             2
##        MACRON        FILLON      LASSALLE DUPONT-AIGNAN         HAMON
##             4             4             5             3             4
##    ASSELINEAU        POUTOU       ARTHAUD     CHEMINADE
##             2             2             2             2
##
## Within cluster sum of squares by cluster:
## [1] 0.00699533 0.05879368 0.02087290 0.04584781 0.00000000
##  (between_SS / total_SS =  89.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

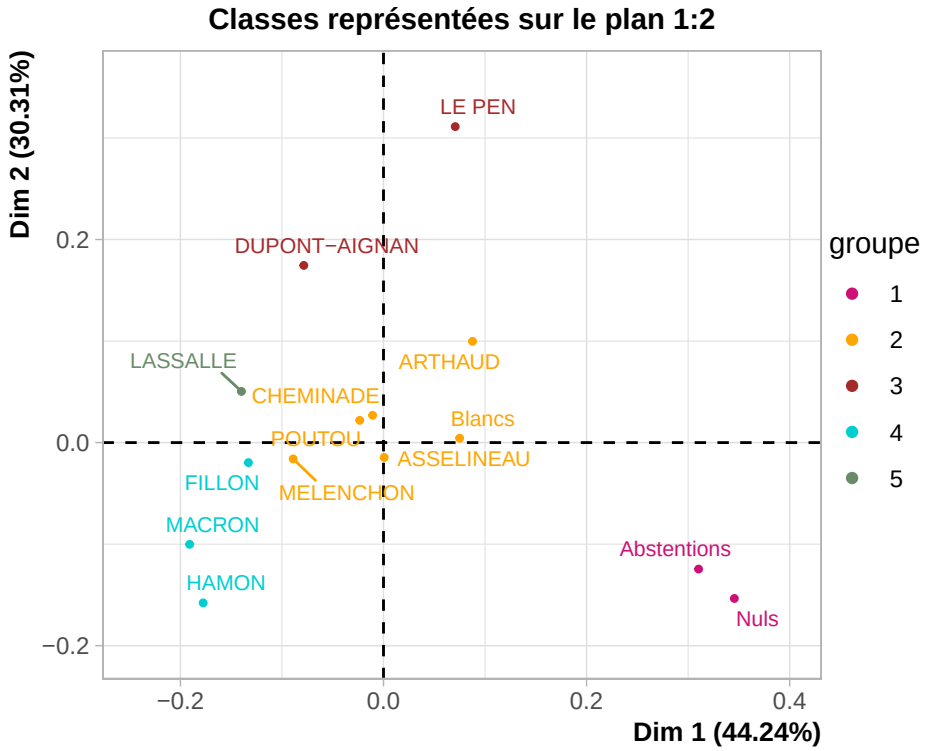
La **consolidation par k-means confirme la classification** construite par la CAH. Aucun candidat n'a changé de classes.

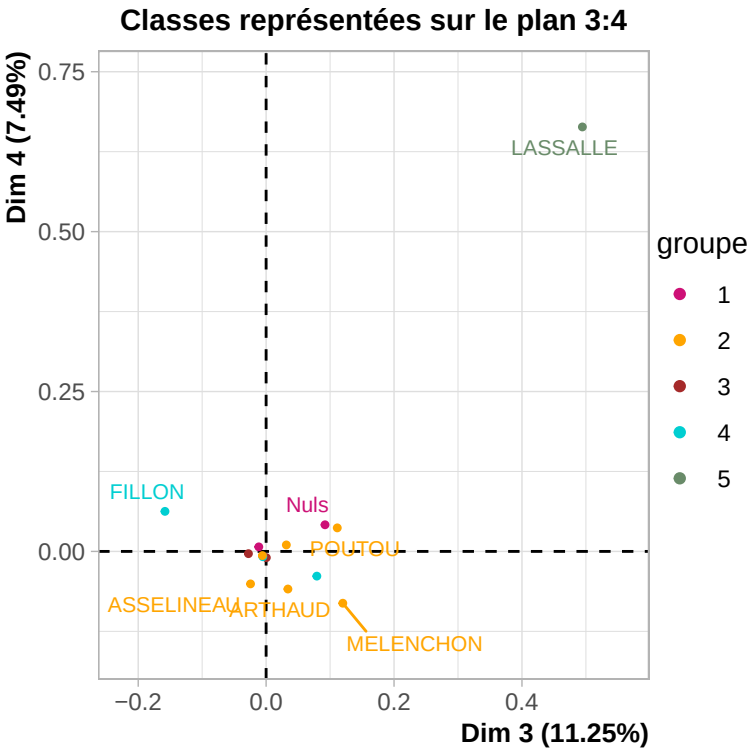
On note par ailleurs la **bonne qualité de classification** avec un R^2 **de l'ordre de 89.7%** :

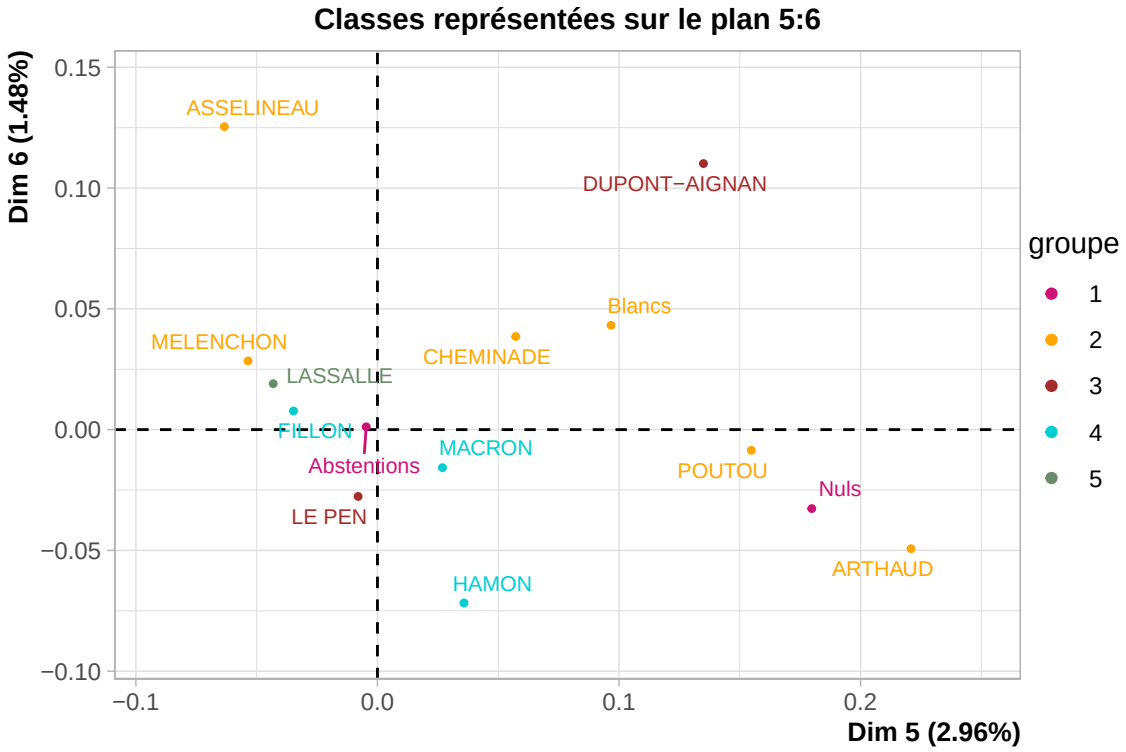
- les groupes sont concentrés et présentent de **faibles inerties intra-classes**,
- l'**inertie totale** est **portée majoritairement par l'inertie inter-classes**.

3.4 Visualisation des groupes

On visualise plan par plan les candidats colorés de leur groupe, pour identifier le plan où les groupes sont le plus distinctement visibles.







Le plan 1:2 est le meilleur pour **visualiser simultanément et de façon dissociée** les 5 groupes de candidats.

4 Caractérisation des groupes

Les groupes constitués par la CAH sont donc

- **Groupe 1** : **Abstentions - Nuls**
- **Groupe 2** : **Blancs - MELENCHON - ASSELINEAU - POUTOU - ARTHAUD - CHEMINADE**
- **Groupe 3** : **LE PEN - DUPONT-AIGNAN**
- **Groupe 4** : **FILLON - MACRON - HAMON**
- **Groupe 5** : **LASSALLE**

On cherche maintenant à **comprendre ce qui rapproche les candidats d'un même groupe**. Autrement dit, on cherche les départements qui votent plus ou moins que la moyenne pour les candidats du groupe.

Pour ce faire, on propose de **caractériser les groupes de candidats** en étudiant les **liens entre les variables quantitatives Départements et une nouvelle variable qualitative Groupe**.

Pour l'exercice, le tableau de jeu de données a été transposé :

- en ligne, on retrouve les 14 candidats
- en colonne, on retrouve les 106 départements
- une nouvelle colonne **Groupe** est ajoutée en parallèle des départements ; elle associe un candidat à sa **modalité Groupe** (1, 2, 3, 4, 5)

Par symétrie des variables Candidats et Départements dans le tableau, la transposition ne modifie pas les résultats de l'AFC incluant la variable qualitative supplémentaire Groupe (*quali.sup=107*). Seules les projections sur les plans peuvent présenter des symétries par rapport aux axes d'étude.

Dans les chapitres suivants, après avoir identifié les plans les plus pertinents pour visualiser les groupes,

- on étudiera la **caractérisation de chaque groupe** par les départements.
- on visualisera les résultats sur un **plan de l'AFC** puis sur une **projection géographique**.

4.1 Dimensions pertinentes

L'AFC intégrant les groupes en variable qualitative supplémentaire nous indique les axes sur lesquels ils sont les mieux représentés (on s'intéresse aux valeurs $cos^2 > 0.25$) à savoir :

```
## [1] "groupe.1"
## Dim 1
##  0.86
## [1] "groupe.2"
## Dim 3
##   0.5
## [1] "groupe.3"
## Dim 2
##  0.97
## [1] "groupe.4"
## Dim 1
##  0.75
## [1] "groupe.5"
## Dim 4 Dim 3
##  0.62  0.34
```

Ainsi :

- Le **plan 1:2** donnera une bonne visibilité des **groupes 1, 3 et 4**
- Le **plan 3:4** donnera une bonne visibilité des **groupes 2 et 5**

4.2 Départements caractérisant le mieux les groupes de candidats

4.2.1 Notion de caractérisation

La **caractérisation** consiste à mesurer l'**intensité de la liaison** des **variables quantitatives Départements** (en colonne pour l'occasion) et la **variable qualitative Groupe** dont les modalités sont associées aux candidats (en ligne pour l'occasion).

On utilisera une fonction **catdes** donnant, **groupe par groupe** :

- les départements ayant un lien significatif avec le groupe
- pour les départements significatifs,
  - la moyenne et l'écart-type des effectifs du département pour les candidats du groupe (attributs *Mean in category* et *sd in category*)
  - la moyenne et l'écart-type des effectifs du département, indépendamment du groupe (attribut *Overall Means* et *Overall sd*)
  - la valeur **v.test**, une valeur de test telle que

$$v.test = \frac{\bar{x_j} - \bar{x}}{\sqrt{\frac{s^2}{I_j}(\frac{I-I_j}{I-1})}}$$

  où

  - $\bar{x_j}$ est la moyenne des votants du département pour les candidats du groupe j
  - $\bar{x}$ est la moyenne des votants du département, indépendamment du groupe de candidats
  - le dénominateur est une version corrigée de l'écart-type des effectifs du département prenant en compte I_j, le nombre de candidats du groupe j, et **I**, le nombre total de candidats
  - enfin, la **probabilité critique** (p-valeur) correspondante selon un loi normale

Quelques précisions quant à la valeur **v.test** et la **p-valeur** :

- Si **v.test est positif**, le département caractérise positivement le groupe i.e. **le département vote plutôt plus pour les candidats du groupe**
- Si **v.test est négatif**, le département caractérise négativement le groupe i.e. **le département vote plutôt moins pour les candidats du groupe**
- Dans l'hypothèse où le département ne caractérise par le groupe de candidats, alors v.test suit une loi normale.

  Ainsi si **v.test** va au-delà de la valeur critique fixée par le seuil critique (par ex. v.test < -1.96 ou v.test > 1.96 pour un seuil à 5%), on considère que v.test ne provient pas d'une loi normale et que le département caractérise le groupe de candidats considéré.

- Par défaut, un **seuil de probabilité critique de 5%** permet d'identifier les départements les plus caractérisant selon leur **p-valeur**.

  Néanmoins, pour ce seuil à 5%, il se peut qu'aucun département ne soit identifié comme ayant un lien avec une modalité de la variable Groupe pour le seuil en question (la liste des départements pour le groupe en question vaut alors *NULL*).

  On pourra dans ce cas, augmenter le seuil pour faire apparaître les 1ers départements liés au groupe concerné.

Une première approche de la caractérisation des groupes

```
res.catdes <- catdes.w(donnees_elections_w_groups, num.var=107, proba=0.05)
```

```
##
## Description of each cluster by quantitative variables
## ======================================================
## $`1`
##                                 v.test Mean in category Overall mean sd in category Overall sd    p.value
## Guadeloupe                    2.375204          99102.0   22596.0714        90592.0 47412.2178 0.01753927
## Guyane                        2.367766          30727.0    6573.9286        29383.0 15015.1517 0.01789587
## Martinique                    2.349315          96600.0   22200.8571        90255.0 46614.6836 0.01880799
## Saint-Martin/Saint-Barthélemy 2.320840           8654.5    1813.2143         8547.5  4338.9897 0.02029549
## Mayotte                       2.283983          24308.0    5931.7857        22564.0 11842.9317 0.02237253
## Polynésie.française           2.265082          63276.5   14566.8571        61238.5 31653.8900 0.02350765
## Nouvelle-Calédonie            2.136209          49784.0   13534.5000        48473.0 24977.8121 0.03266241
## La.Réunion                    2.026033         138767.0   45677.2143       125462.0 67631.8986 0.04276141
## Saint-Pierre-et-Miquelon      1.973391           1132.0     354.7143         1106.0   579.7804 0.04845101
##
## $`2`
## NULL
##
## $`3`
## NULL
##
## $`4`
##                   v.test Mean in category Overall mean sd in category Overall sd    p.value
## Paris            2.545350        256433.33     92974.07      110205.39  120919.96 0.01091682
## Hauts-de-Seine   2.364386        181784.67     70480.71       88750.63   88639.66 0.01805996
## Yvelines         2.182514        159486.67     67939.14       75770.32   78981.54 0.02907164
## Ille-et-Vilaine  2.098345        116275.00     52153.00       52261.44   57539.50 0.03587466
## Mayenne          2.098270         35319.00     15880.36       17744.38   17443.78 0.03588131
## Finistère        2.080523        108280.00     49326.43       42585.20   53354.80 0.03747756
## Vendée           1.995033         79383.00     36747.21       41051.88   40240.20 0.04603929
## Maine-et-Loire   1.980489         86708.67     40673.21       40751.43   43767.93 0.04764865
## Morbihan         1.969841         87635.67     41215.14       39967.02   44372.59 0.04885659
## Loire-Atlantique 1.961502        152481.67     70967.00       68556.50   78249.74 0.04982047
##
## $`5`
## NULL
```

En maintenant le seuil critique à 5%, l'analyse retourne

- le fait que le groupe 1 (Abstentions / Nuls) est caractérisé significativement par les départements de la France d'Outre-Mer (résultat observé précédemment)
- le fait que le groupe 4 (**FILLON**, **MACRON** et **HAMON**) est caractérisé significativement par les départements d'Île-de-France et de l'Ouest de la France (résultat observé précédemment)
- le fait que les groupes 2, 3 et 5 ne sont pas fortement caractérisés (au seuil de 5%)

Analysons plus finement les groupes dans les chapitres suivants :

- on identifiera les départements "influents" parmi la liste exhaustive des départements et leur *v.test*
- on projettera les départements en question sur un plan de l'AFC,
- on projettera les valeurs *v.test* des départements sur une réprésentation géographique du territoire,
- enfin, on conclura sur les affinités territoriales des groupes de candidats

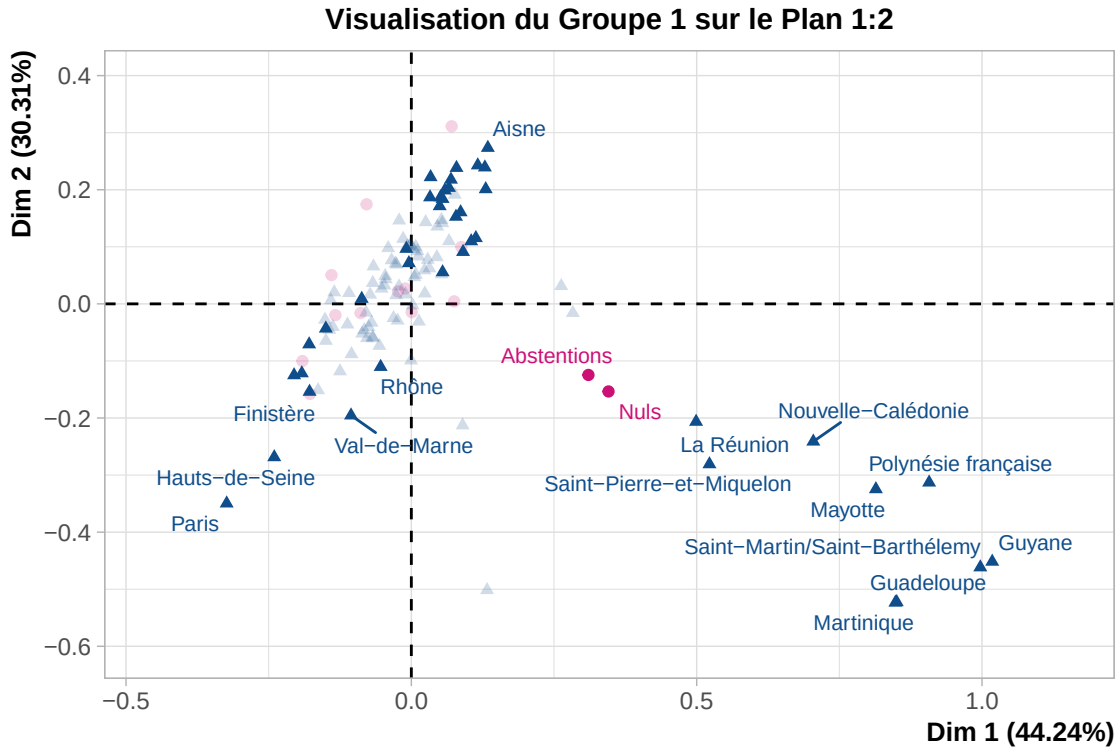
4.2.2 Groupe 1 : Nuls - Abstentions

Caractérisation du groupe (seuil de significativité de 5%)

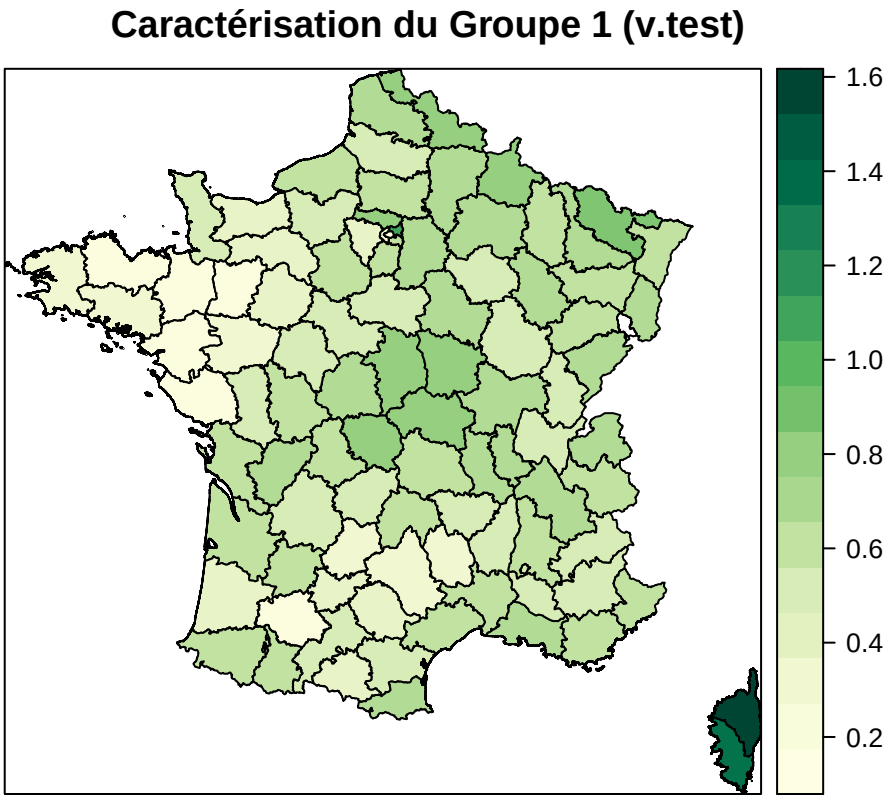
```
res.catdes.groupe.1 <- catdes.w(donnees_elections_w_groups, num.var=107, proba=0.05)
res.catdes.groupe.1$quanti$`1`
```

```
##                                 v.test Mean in category Overall mean sd in category Overall sd    p.value
## Guadeloupe                    2.375204          99102.0   22596.0714        90592.0 47412.2178 0.01753927
## Guyane                        2.367766          30727.0    6573.9286        29383.0 15015.1517 0.01789587
## Martinique                    2.349315          96600.0   22200.8571        90255.0 46614.6836 0.01880799
## Saint-Martin/Saint-Barthélemy 2.320840           8654.5    1813.2143         8547.5  4338.9897 0.02029549
## Mayotte                       2.283983          24308.0    5931.7857        22564.0 11842.9317 0.02237253
## Polynésie.française           2.265082          63276.5   14566.8571        61238.5 31653.8900 0.02350765
## Nouvelle-Calédonie            2.136209          49784.0   13534.5000        48473.0 24977.8121 0.03266241
## La.Réunion                    2.026033         138767.0   45677.2143       125462.0 67631.8986 0.04276141
## Saint-Pierre-et-Miquelon      1.973391           1132.0     354.7143         1106.0   579.7804 0.04845101
```

Projection du groupe sur le plan 1:2



Projection géographique des *v.test* (métropole + Corse)



Analyse

Les valeurs v.tests sont toutes positives (cf. l'échelle de la projection géographique) :

- Globalement, tous les départements ont une tendance à l'abstention ou au vote nul plus forte que la moyenne par modalité.
- Les départements "plus foncés" sont ceux qui présentent la tendance la plus forte à l'abstention et au vote nul

Ainsi, on note les particularités géographiques suivantes :

- La **France d'Outre-Mer** (non représentée) et la **Corse** sont les territoires à la **tendance aux votes Nuls et abstentions la plus forte**
- La tendance à l'abstention et au vote nul est homogènement répartie sur le reste du territoire.

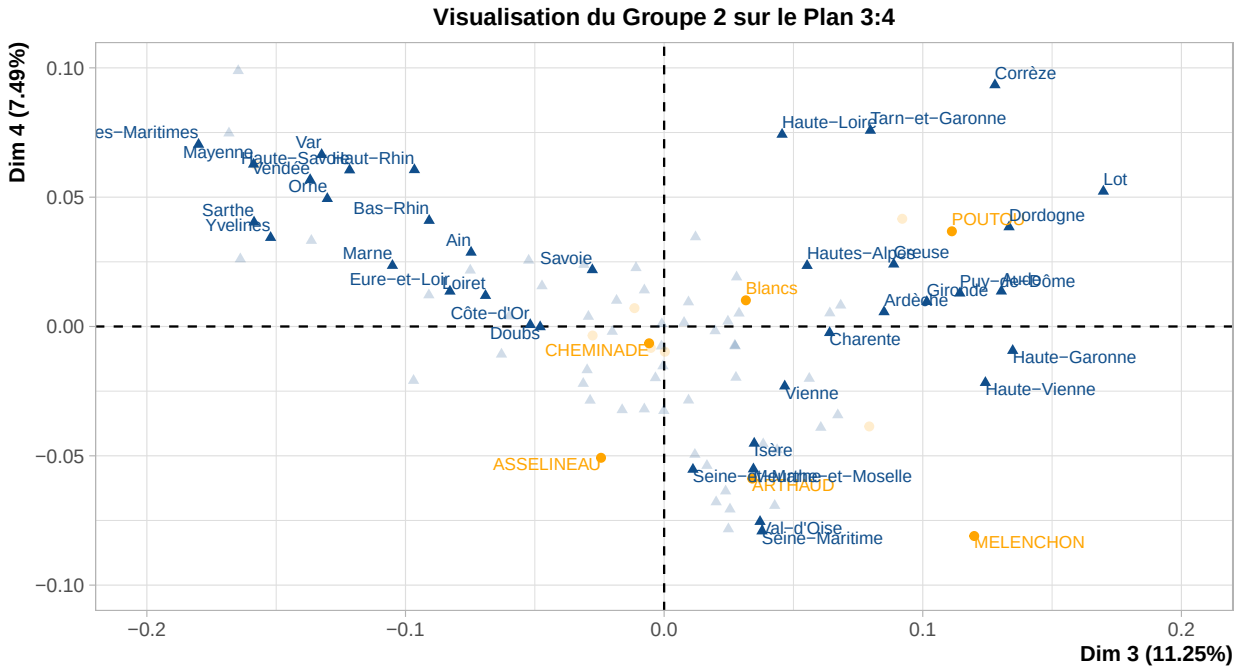
4.2.3 Groupe 2 : Blancs - ARTHAUD - POUTOU - CHEMINADE

Caractérisation du groupe (seuil de significativité de 8%)

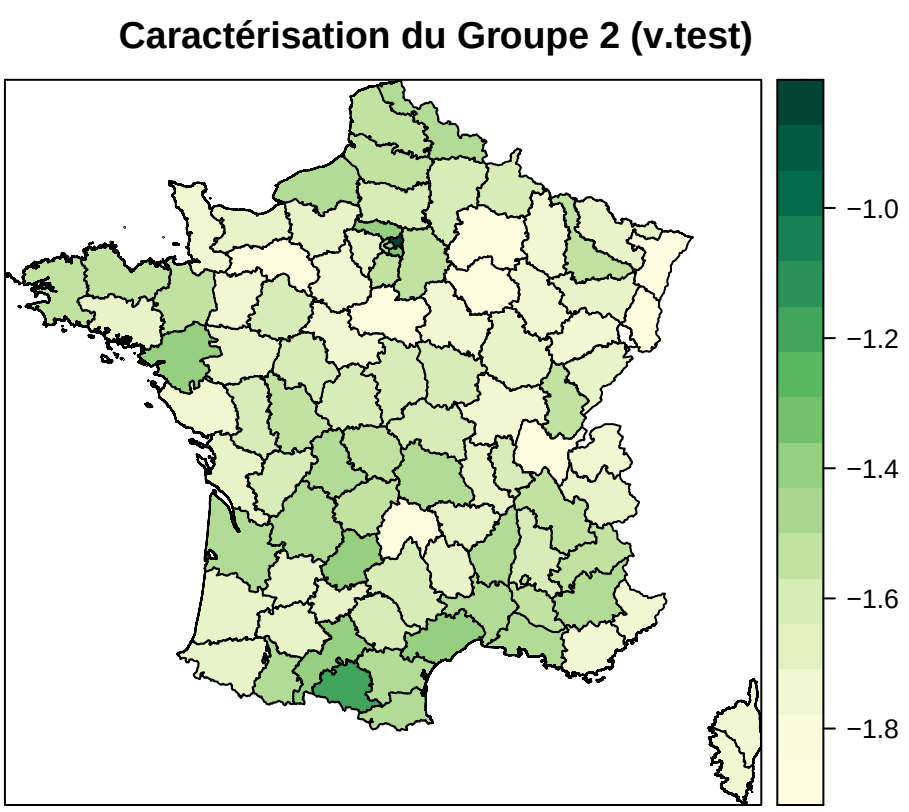
```
##                  v.test Mean in category Overall mean sd in category Overall sd    p.value
## Haute-Corse      -1.756444         2390.667    8918.7143     3927.25437 11605.1681 0.07901271
## Mayenne          -1.759997         6048.167   15880.3571     9351.74015 17443.7784 0.07840825
## Haute-Savoie     -1.761519        14472.167   38573.6429    23629.44299 42722.6859 0.07815055
## Eure-et-Loir     -1.762196         8291.833   21533.2857    13356.24828 23463.0032 0.07803611
## Loir-et-Cher     -1.772325         6821.333   17522.9286    11114.26673 18854.1672 0.07634055
## Vendée           -1.773524        13891.500   36747.2143    22155.35913 40240.2026 0.07614200
```

```
## Corse-du-Sud      -1.773742         2128.333    7770.0714     3575.74606  9931.7269 0.07610583
## Loiret            -1.779106        12396.167   32329.0714    20525.09028 34984.1270 0.07522236
## Ain               -1.785454        11204.167   29707.7857    18211.83540 32360.1379 0.07418774
## Marne             -1.788163         9515.833   27263.7857    15660.49854 30991.5530 0.07374963
## Cantal            -1.788341         3212.500    8387.5000     5113.67489  9035.7103 0.07372099
## Aube              -1.805468         4890.167   14560.5000     7906.64161 16724.5343 0.07100153
## Orne              -1.810677         5503.167   14909.5714     8557.41042 16221.2767 0.07019090
## Haut-Rhin         -1.830214        13497.333   37584.9286    19986.52021 41095.4584 0.06721791
## Bas-Rhin          -1.849049        19602.167   54753.4286    30863.74411 59360.1203 0.06445072
## Wallis.et.Futuna -1.883099           65.000     604.7714       57.96551   894.7014 0.05968698
```

Projection du groupe sur le plan 3:4



Projection géographique des *v.test* (métropole + Corse)



Analyse

Les valeurs v.tests sont toutes négatives (cf. l'échelle de la projection géographique) :

- Globalement, tous les départements ont une tendance à voter pour le groupe 2 plus faible que la moyenne par candidat.
- Les départements "plus clairs" sont ceux qui présentent la tendance la plus faible à voter pour le groupe 2.
- Les départements "plus foncés" sont ceux qui présentent un vote pour le groupe 2 proche de la valeur moyenne des votes par candidat (v.test=0).

Ainsi, on note les particularités géographiques suivantes :

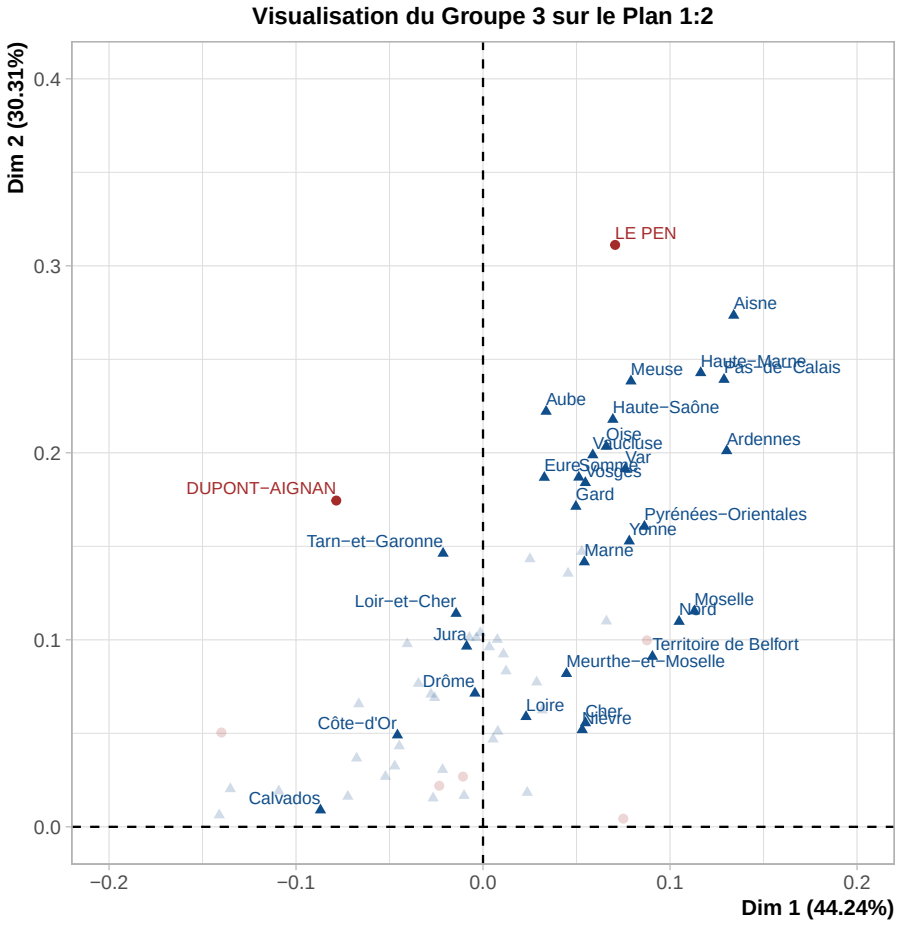
- La tendance à voter pour le groupe 2 est homogènement mesurée sur l'ensemble du territoire.
- Seules la **Seine-Saint-Denis** et l'**Ariège** présentent un vote au groupe 2 proche de la moyenne.

4.2.4 Groupe 3 : LE PEN - DUPONT-AIGNAN

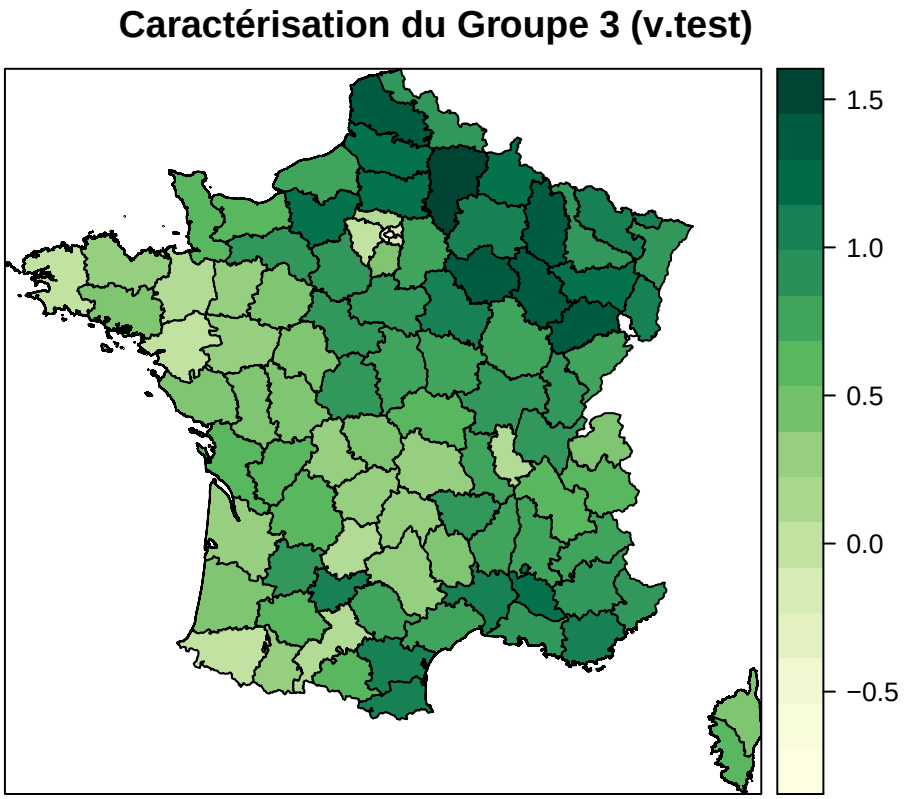
Caractérisation du groupe (seuil de significativité de 20%)

```
##                v.test Mean in category Overall mean sd in category Overall sd   p.value
## Aisne         1.452671          58719.5     26839.43        44067.5   32303.39 0.1463152
## Meuse         1.428037          20702.0      9831.00        13900.0   11205.37 0.1532813
## Haute-Marne   1.409112          20222.0      9582.00        13805.0   11114.56 0.1588022
## Pas-de-Calais 1.371035         163787.0     77883.43       122360.0   92227.14 0.1703641
## Haute-Saône   1.358750          25964.5     12801.00        17788.5   14260.27 0.1742258
## Aube          1.318420          29540.5     14560.50        19305.5   16724.53 0.1873630
```

Projection du groupe sur le plan 1:2



Projection géographique des *v.test* (métropole + Corse)



Analyse

- Les départements “plus foncés” sont ceux qui ont tendance à voter davantage pour les candidats du groupe 3 (v.test>0).
- Les départements “plus clairs” sont ceux qui ont tendance à voter moins pour les candidats du groupe 3 (v.test<0).

Ainsi, on note que le groupe 3 oppose la France en 2 zones Est / Ouest :

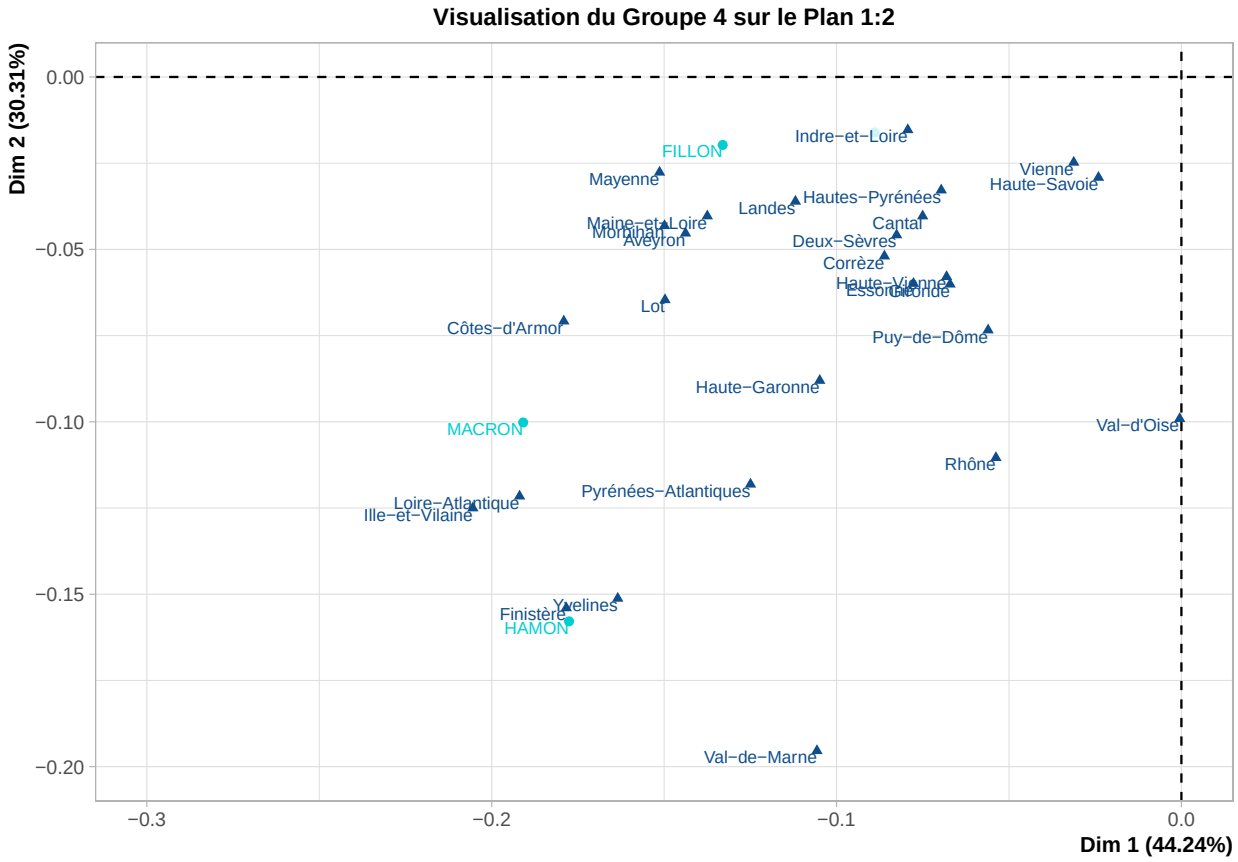
- Les départements de l’Est, du Nord et du Sud-Est de la France ont tendance à plus voter pour le groupe 3.
- Les départements de l’Ouest et du Sud-Ouest de la France ont tendance à moins voter pour le groupe 3.

4.2.5 Groupe 4 : MACRON - FILLON - HAMON

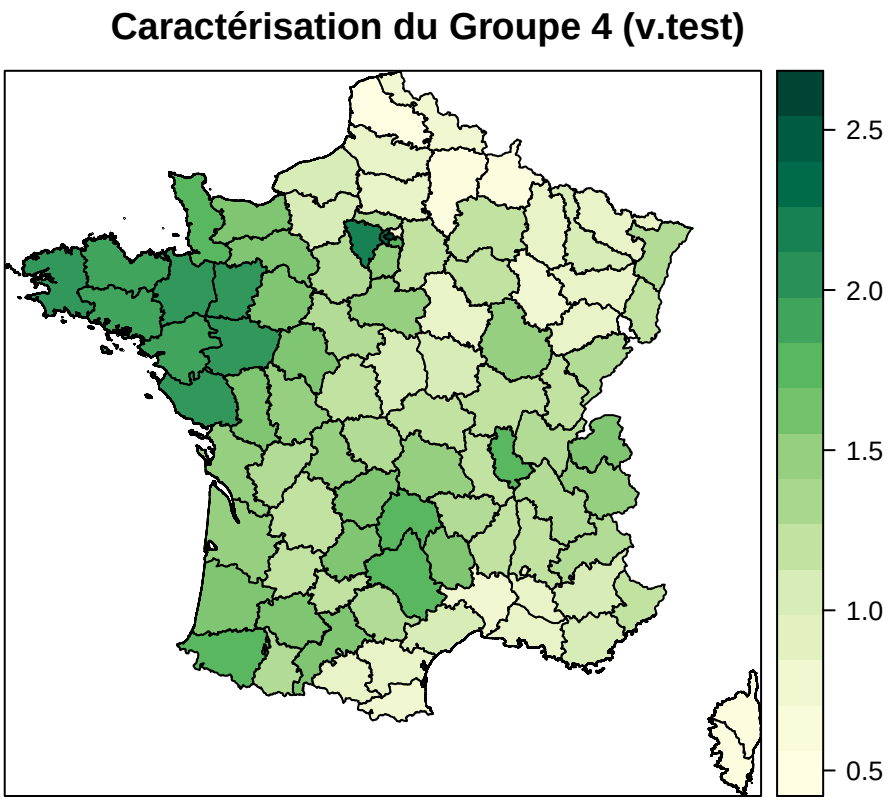
Caractérisation du groupe (seuil de significativité de 5%)

```
##                   v.test Mean in category Overall mean sd in category Overall sd    p.value
## Paris           2.545350        256433.33     92974.07      110205.39  120919.96 0.01091682
## Hauts-de-Seine  2.364386        181784.67     70480.71       88750.63   88639.66 0.01805996
## Yvelines        2.182514        159486.67     67939.14       75770.32   78981.54 0.02907164
## Ille-et-Vilaine 2.098345        116275.00     52153.00       52261.44   57539.50 0.03587466
## Mayenne         2.098270         35319.00     15880.36       17744.38   17443.78 0.03588131
## Finistère       2.080523        108280.00     49326.43       42585.20   53354.80 0.03747756
## Vendée          1.995033         79383.00     36747.21       41051.88   40240.20 0.04603929
## Maine-et-Loire  1.980489         86708.67     40673.21       40751.43   43767.93 0.04764865
## Morbihan        1.969841         87635.67     41215.14       39967.02   44372.59 0.04885659
## Loire-Atlantique 1.961502       152481.67     70967.00       68556.50   78249.74 0.04982047
```

Projection du groupe sur le plan 1:2



Projection géographique des *v.test* (métropole + Corse)



Analyse

- Les départements "plus foncés" sont ceux qui ont tendance à voter davantage pour les candidats du groupe 4 (v.test>0).
- Les départements "plus clairs" sont ceux qui ont tendance à voter moins pour les candidats du groupe 4 (v.test<0).

Ainsi, le groupe 4 vient en opposition au groupe 3 en coupant lui aussi la France en 2 zones Est / Ouest :

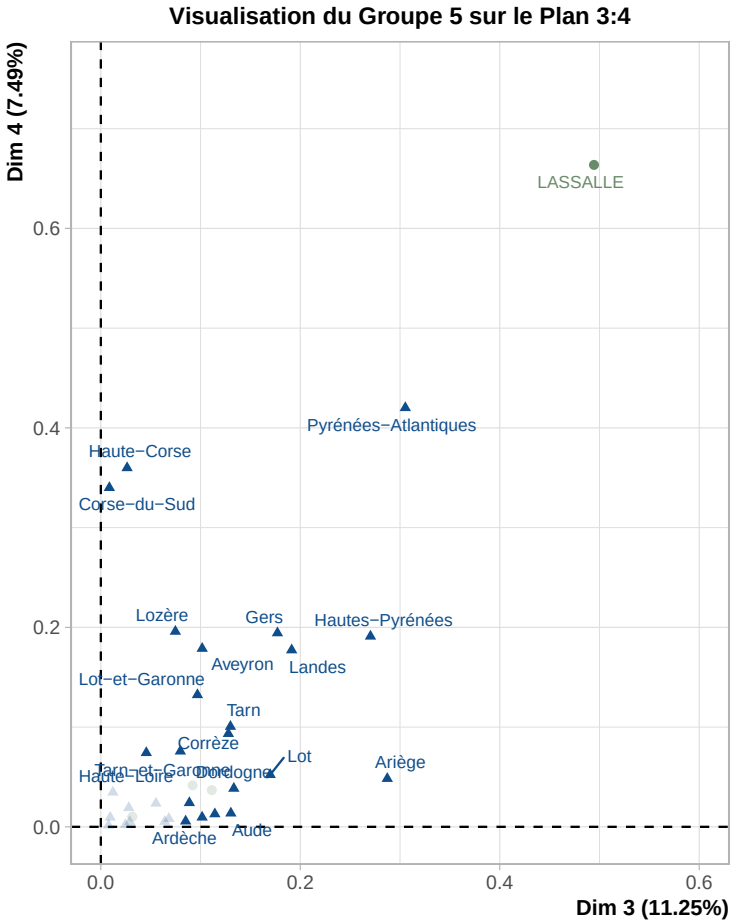
- Les départements de l'Est, du Nord et du Sud-Est de la France ont tendance à moins voter pour le groupe 4.
- Les départements de l'Ouest, du Sud-Ouest et d'île-de-France ont tendance à plus voter pour le groupe 4.

4.2.6 Groupe 5 : LASSALLE

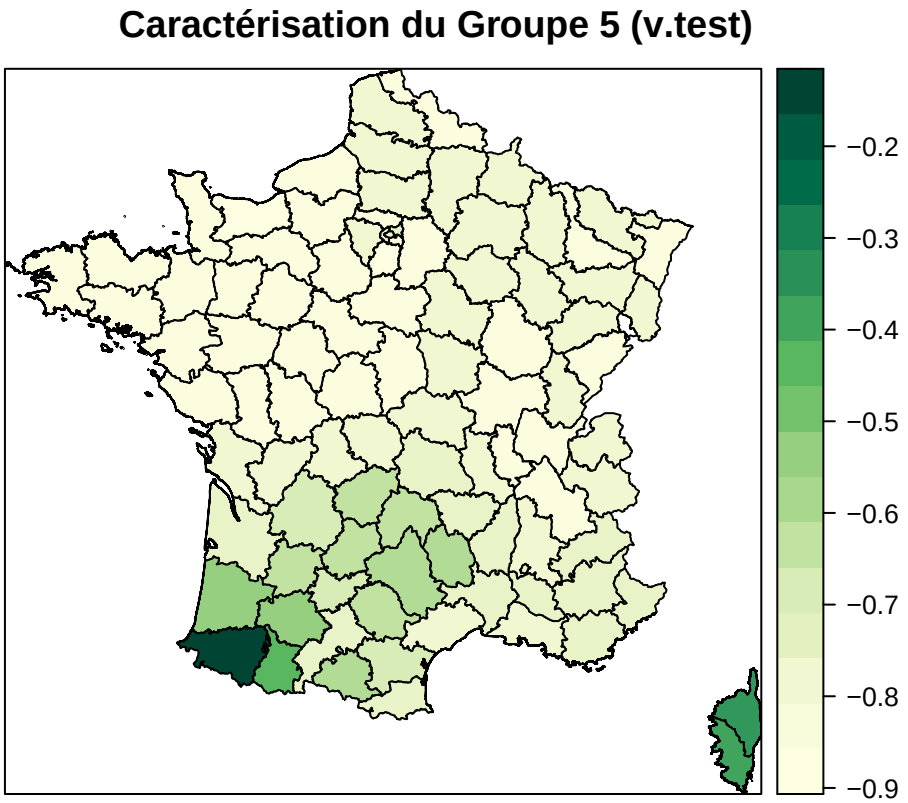
Caractérisation du groupe (seuil de significativité de 40%)

```
##                  v.test Mean in category Overall mean sd in category Overall sd   p.value
## Seine-et-Marne -0.8453311             5185     62968.43             NA   68355.97 0.3979260
## Seine-Maritime -0.8483797             4383     62975.57             NA   69064.09 0.3962266
## Maine-et-Loire -0.8497138             3483     40673.21             NA   43767.93 0.3954842
## Indre-et-Loire -0.8499061             2929     30633.64             NA   32597.30 0.3953773
## Manche         -0.8514537             2524     27078.86             NA   28838.75 0.3945174
## Calvados       -0.8573896             2911     35796.64             NA   38355.54 0.3912296
## Essonne        -0.8580344             4468     56801.93             NA   60992.81 0.3908735
```

Projection du groupe sur le plan 3:4



Projection géographique des *v.test* (métropole + Corse)



Analyse

- Les départements “plus foncés” sont ceux qui présentent un vote pour le candidat LASSALLE proche de la valeur moyenne des votes par candidat (v.test=0).
- Les départements “plus clairs” sont ceux qui ont tendance à voter moins pour le candidat LASSALLE (v.test<0).

Ainsi on note les particularités géographiques suivantes

- on vote globalement peu pour le candidat LASSALLE sur le territoire
- Les départements du Sud-Ouest et la Corse ont tendance à plus voter pour le candidat LASSALLE.

5 Conclusion

Dans cette étude, nous nous sommes intéressés à l'**ancrage territorial des candidats**.

Ainsi, nous avons procédé

- à une **analyse factorielle des correspondances (AFC) entre les modalités des variables Candidats et Départements** pour permettre une représentation des modalités dans un même espace géométrique,
- à une **classification des scores colonnes de l'AFC** pour identifier la proximité des candidats entre eux et en déduire des groupes cohérents,
- à une **caractérisation des groupes de candidats** pour identifier les départements les plus influents sur le groupe.

La classification sur les résultats de l'AFC a permis de faire ressortir **5 groupes**:

- **Groupe 1** : **Abstentions** - **Nuls**
- **Groupe 2** : **Blancs** - **MELENCHON** - **ASSELINEAU** - **POUTOU** - **ARTHAUD** - **CHEMINADE**
- **Groupe 3** : **LE PEN** - **DUPONT-AIGNAN**
- **Groupe 4** : **FILLON** - **MACRON** - **HAMON**
- **Groupe 5** : **LASSALLE**

La caractérisation des groupes a permis de conclure des tendances sur l'**ancrage territorial des candidats**.

On comptera parmi les plus notables

- La Corse et la France d'Outre-Mer ont une forte tendance à l'abstention et au vote nul.
- La France métropolitaine oppose 2 zones Est-Ouest
  - une zone Est votant davantage pour les candidats **LE PEN** et **DUPONT-AIGNAN**
  - une zone Ouest votant davantage pour les candidats **FILLON**, **HAMON**, **MACRON**
- des candidats du groupe 2 sans ancrage territorial particulier
- un candidat **LASSALLE** atypique polarisant ses voix dans les **départements corses et du Sud-Ouest**

6 Références

Site Web de la librairie R FactoMineR :

http://factominer.free.fr/

Caractérisation de variables quantitative sur Chaîne Youtube de François Husson :

https://www.youtube.com/watch?v=JAsvTf-8cXo&t=310s

Code de la fonction catdes adapté (catdes.w) :

https://superstatisticienne.fr/typologie-avec-r/

Représentation géographique (raster / geodata / colorspace) :

https://www.neonscience.org/resources/learning-hub/tutorials/raster-data-r

https://pmbo.pagesperso-orange.fr/STID/Outils_pilotage_2/TD1_cartographie.html

http://colorspace.r-forge.r-project.org/articles/colorspace.html