AUBRY Maxime, NORROY Tom, PIGE Loïc

Projet statistique M1

Analyse factorielle des correspondances

I. Introduction

L'analyse factorielle des correspondances (AFC) est une méthode statistique largement utilisée pour examiner les relations entre les variables catégorielles dans un tableau de contingence. Elle permet de mettre en évidence des structures sous-jacentes dans les données en réduisant leur dimensionnalité tout en préservant au mieux les informations importantes.

Dans le cadre de notre étude, nous avons appliqué une AFC aux résultats des dernières élections présidentielles en France (2022) par département. Pour ce faire, nous avons organisé les données avec les départements en ligne et les candidats au 1er et 2nd tour en colonnes. De plus, nous avons inclus différentes variables démographiques et socio-économiques importantes telles que le « taux d'immigration », le « niveau de vie médian » et le « taux de pauvreté », afin d'explorer les relations potentielles entre ces facteurs et les résultats aux élections.

En utilisant l'AFC, nous avons cherché à identifier des tendances, des regroupements ou des associations entre les variables catégorielles examinées. Cette approche nous permet de visualiser graphiquement les relations complexes entre les départements, les candidats et les variables socio-économiques, et d'identifier les facteurs qui pourraient influencer les résultats aux élections.

Notre problématique est la suivante : ***Pouvons-nous expliquer le choix des candidats votés par département en fonction de différentes données démographiques et socio-économiques ?***

II. Les variables actives et supplémentaires

Nous avons choisi de mettre en variables actives les variables qualitatives correspondant aux candidats présents au 1er tour. Nous avons 13 candidats en incluant le vote blanc, soit 13 variables actives. Les départements correspondent aux individus, ils sont au nombre de 99 en incluant les DROM (sans Mayotte pour cause de données manquantes). La variable « abstention » n'est pas prise en compte dans l'AFC car elle ne permet pas de répondre à notre problématique, qui s'intéresse uniquement aux votes.

De plus, nous avons mis en colonnes supplémentaires les candidats présents au 2nd tour afin d'observer les tendances des départements et des votants face à un choix restreint de candidats.

Enfin, nous avons mis en variables explicatives les variables quantitatives supplémentaires suivantes :

- Taux d'immigration par département (%)
- Niveau de vie médian par département (€)
- Taux de pauvreté par département (%)

- Part des différentes catégories socio-professionnelles par département (%) : agriculteurs/exploitants ; artisans/commerçants/chefs d'entreprise ; cadres/professions intellectuelles supérieures ; professions intermédiaires ; employés ; ouvriers
- Nombre de personnes des différentes classes d'âges par département : 0-19 ans ; 20-39 ans ; 40-59 ans ; 59-74 ans ; 75 et +

Ces variables sont cruciales car elles nous permettent d'explorer comment les caractéristiques démographiques et socio-économiques des départements influencent les résultats aux élections. Par exemple, les parts des « agriculteurs/exploitants » et des « artisans/commerçants/chefs d'entreprise » peuvent être associées à des zones rurales ou à des régions agricoles, ce qui peut avoir un impact sur les préférences politiques locales. De même, la répartition des différentes classes socio-professionnelles peut influencer les dynamiques politiques régionales, avec des tendances électorales différentes observées dans les régions à forte concentration de « cadres/professions intellectuelles supérieures » par rapport à celles dominées par les « ouvriers » ou les « employés ».

Ensuite, les variables socio-économiques telles que le « taux d'immigration », le « niveau de vie médian » et le « taux de pauvreté » sont également cruciales, elles peuvent jouer un rôle significatif dans la formation des opinions politiques des électeurs. Par exemple, des départements avec un niveau de vie plus élevé pourraient avoir des préférences politiques différentes de ceux avec un haut niveau de pauvreté. De même, la composition démographique, comprenant les différentes « classes d'âge » et le « taux d'immigration », peut influencer les attitudes politiques et les choix électoraux.

III. Interprétation

```
Call:
"res.CA<-CA(data.election,col.sup=c(14,15),quanti.sup=c(16,17,18,19,20,21,22,23,24,25,26,27,28,29),graph=FALSE)"

The chi square of independence between the two variables is equal to 2612187 (p-value =  0 ).

Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10  Dim.11
Variance               0.035   0.016   0.011   0.007   0.002   0.002   0.001   0.000   0.000   0.000   0.000
% of var.             47.372  21.505  15.297   8.822   2.476   2.197   0.923   0.602   0.435   0.206   0.121
Cumulative % of var.  47.372  68.877  84.174  92.996  95.472  97.669  98.591  99.193  99.628  99.833  99.955
                      Dim.12
Variance               0.000
% of var.              0.045
Cumulative % of var. 100.000
```

Figure 1 : Test du khi² et variabilité expliquée par l'AFC

Le test du khi² (cf figure 1) montre qu'il n'y a pas indépendance entre les départements et les candidats votés, la p-value est quasiment égale à 0. Les trois premières dimensions expliquent à elles seules 84% de l'information contenue dans le jeu de données.

Dans un premier temps, nous avons regardé les contributions à la construction des axes. Les lignes qui contribuent le plus à la construction des axes 1 et 2 sont (dans l'ordre décroissant de contribution): Paris, La Seine-Saint-Denis, les Hauts-de-Seine, le Pas-de-Calais, la Guadeloupe, la Réunion et le Val de Marne. Les colonnes qui contribuent le plus à la construction des axes 1 et 2 sont: Mme. Marine Le Pen, M. Jean-Luc Mélenchon, M. Emmanuel Macron et M. Yannick Jadot. Enfin, M. Jean Lassalle contribue à lui seul à 76% de la construction de la troisième dimension.

Par ailleurs, les Pyrénées-Atlantiques sont très bien projetées et contribuent à plus de 20% à la construction de la troisième dimension.

Dans un second temps, nous avons regardé les qualités de représentation. En affichant les départements avec un cos^2 supérieur à 0,5, on obtient plus de la moitié des départements sur le graphique. Dans l'ensemble, les départements sont bien projetés sur les 2 premières dimensions. Pour les candidats, il nous faut un cos^2 supérieur à 0,2 pour en représenter la moitié, ils sont donc moins bien projetés sur les 2 premières dimensions.

Sur le graphique suivant, on représente les lignes (départements) dont le cos^2 est supérieur à 0,7 par soucis de lisibilité ainsi que les colonnes (candidats) dont le cos^2 est supérieur à 0,2 :

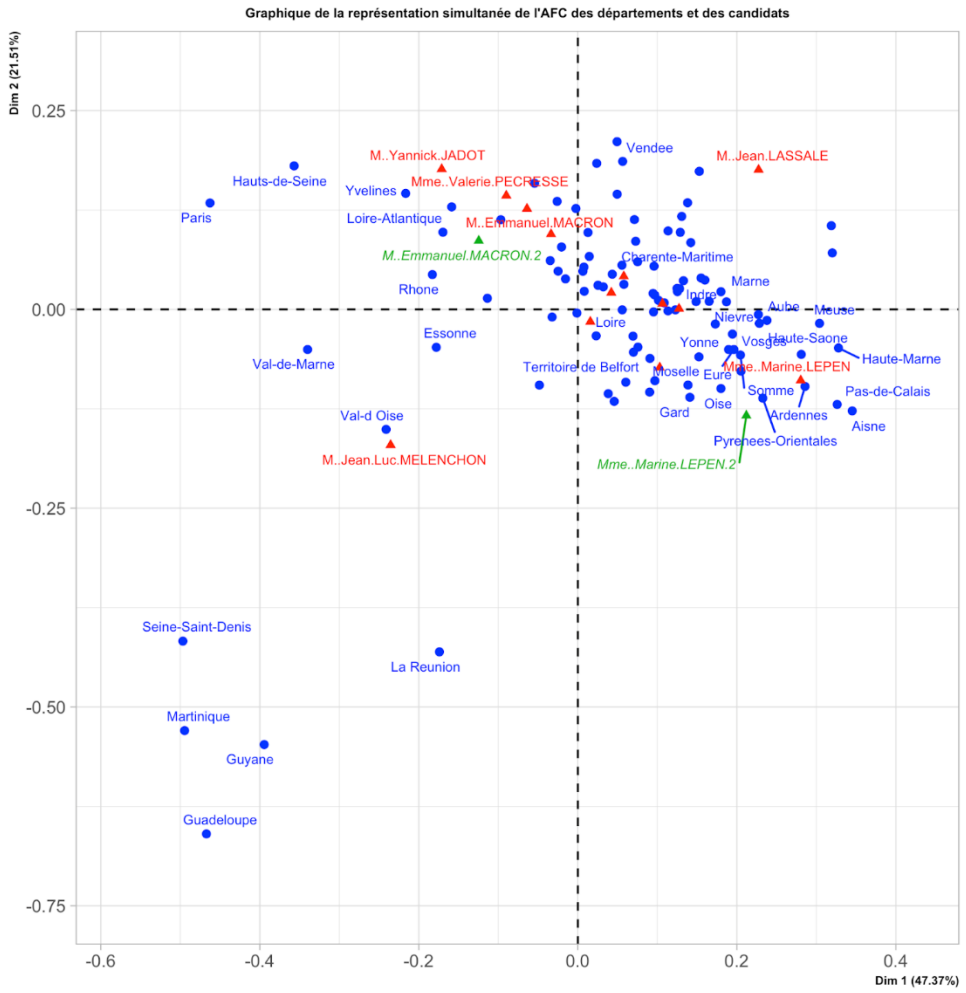

Figure 2 : Graphique de la représentation simultanée de l'AFC des départements et des candidats

Sur ce graphique (cf figure 2), l'axe 1 oppose les candidats plutôt « de gauche » (avec M. Jean-Luc Mélenchon et M. Yannick Jadot notamment), aux candidats plutôt « de droite » (avec Mme. Marine Le Pen et M. Jean Lassalle). M. Emmanuel Macron se situe plutôt au centre et est un candidat centriste. En ce qui concerne l'axe 2, il oppose les candidats « extrêmes » aux candidats plus centraux (de droite ou de gauche).

Lorsque nous regardons les colonnes supplémentaires, soient les candidats présents au 2nd tour, nous voyons que M. Emmanuel Macron s'éloigne du centre et se déplace vers la gauche de l'axe 1, ceci pourrait s'expliquer par le ralliement des électeurs « de gauche », notamment de M. Jean-Luc Mélenchon et M. Yannick Jadot. De plus, Mme. Marine Le Pen se rapproche du centre et se

déplace également vers la gauche de l'axe 1, ceci pourrait s'expliquer par le ralliement des électeurs « de droite ».

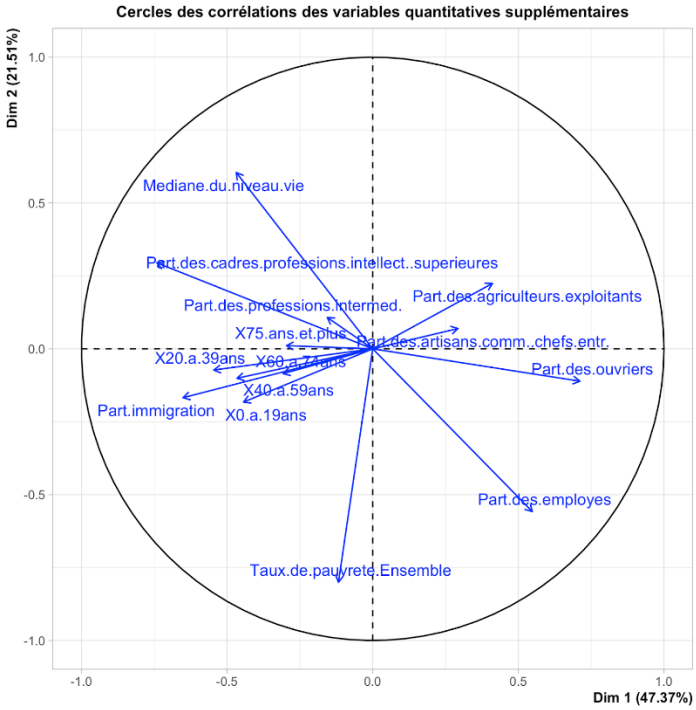

Figure 3 : Cercle des corrélations des variables quantitatives supplémentaires

Sur le cercle des corrélations (cf figure 3), on voit que les classes d'âge sont toutes dans la même direction et mal projetées sur les axes, on ne peut donc pas utiliser ces variables pour expliquer les tendances de votes par département.

- La première dimension semble séparer les départements ruraux à droite (donc peu peuplés) des départements urbains à gauche (très peuplés) avec une séparation par classe socio-professionnelles notamment. Les « cadres » et « professions intermédiaires » sont du côté des départements urbains, tandis que les « agriculteurs » et les « ouvriers » sont du côté des départements ruraux. En outre, les départements avec un fort « taux d'immigration » sont du côté gauche de la première dimension.

- La deuxième dimension a tendance à séparer les départements riches des départements pauvres. En effet, les départements plutôt pauvres (ex: DROM) sont du côté de la variable « taux de pauvreté » et de la « part des employés », tandis que les départements plutôt riches sont du côté des variables « niveau de vie médian » et « part des cadres/professions intellectuelles supérieures ».

- La troisième dimension non tracée ici sépare plutôt les départements agricoles des non agricoles, avec une contribution importante de M. Jean Lassalle et du département des Pyrénées-Atlantiques, département d'origine du candidat, et une bonne projection de la « part des agriculteurs » et des « artisans ».

Nous obtenons une séparation qui répartit les candidats et les départements de la sorte :

- 1er quart en haut à gauche : départements urbains et riches avec un « niveau de vie médian » élevé et une surreprésentation des « cadres » (et des personnes de plus de 75 ans). Ces types de départements et d'électeurs sont surreprésentés dans les votes des candidats plutôt centraux/non extrêmes, et parallèlement, ces candidats dits centraux sont surreprésentés dans les votes des départements riches et urbains.

- 2ème quart en haut à droite : départements ruraux en majorité avec « niveau de vie médian » élevé ou décent, une surreprésentation des « agriculteurs » et des « artisans et chefs d'entreprise » et un très faible « taux d'immigration » dans ces départements. Dans ce quart, peu de candidats sont surreprésentés et bien projetés, seul M. Jean Lassalle se démarque, il est du côté des départements avec une « part d'agriculteurs » importante.

- 3ème quart en bas à droite : départements ruraux et plutôt pauvres avec un très faible « taux d'immigration » et une forte représentation des « ouvriers » et « employés ». Ces départements ont tendances à voter pour Mme. Marine Le Pen, candidate d'extrême droite, qui est du côté de ce type de département, et assez éloigné du profil moyen.

- 4ème quart en bas à gauche : départements urbains et pauvres, très peuplés avec un « taux de pauvreté » très important et une très forte immigration. Ces départements sont surreprésentés dans les votes de M. Jean Luc Mélenchon, candidat d'extrême gauche. Il y a un cas particulier pour les DROM qui sont présents dans ce quart, à l'extrémité, car ce sont des départements où le « taux de pauvreté » est bien supérieur aux autres départements français : ce ne sont cependant pas des départements très urbains ou peuplés mais les votes pour M. Jean Luc Mélenchon y sont surreprésentés.

IV. Classification

Le jeu de données présente de nombreux individus (100 départements), il est donc envisageable d'effectuer une classification ascendante hiérarchique sur ces données. La classification peut permettre de confirmer les interprétations des votes des départements en fonction des variables explicatives démographiques et socio-économiques. La classification comprend 6 clusters, qui permettent à la fois de minimiser la perte d'inertie et de bien répartir les nombreux départements dans des classes, suivant les variables explicatives.

La classification réalisée sur les individus fait apparaître 6 classes (cf figure 4).

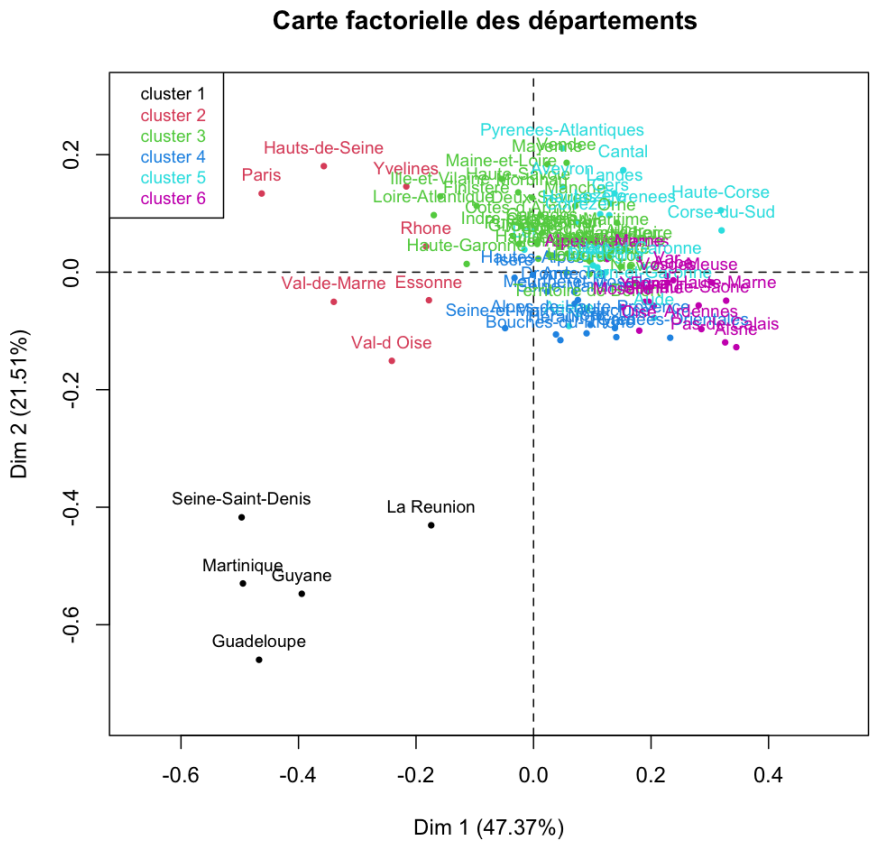

Figure 4 : Classification Ascendante Hiérarchique des départements

Ces différentes classes se distinguent de la façon suivante :

- La classe 1 regroupe des départements tels que la Seine-Saint-Denis, la Martinique, la Guadeloupe, la Guyane et La Réunion. Elle se caractérise par des taux élevés de pauvreté, d'employés, d'immigration, et une population majoritairement jeune, avec une faible présence de cadres et d'ouvriers. Cette classe est proche de M. Jean Luc Mélenchon.

- La classe 2 comprend des départements comme Paris, les Hauts-de-Seine, le Val-de-Marne, le Val-d'Oise, les Yvelines et le Rhône. Ces départements affichent des niveaux élevés de cadres, de revenu médian, d'immigration, et une répartition démographique plus équilibrée. Les taux de pauvreté et la présence d'ouvriers et d'employés y sont relativement faibles. Cette classe est plus proche des candidats M. Yannick Jadot, M. Emmanuel Macron et Mme. Valérie Pécresse.

- La classe 3 rassemble des départements comme la Vendée, caractérisées par une forte proportion d'ouvriers et d'agriculteurs, mais un faible taux de pauvreté, d'immigration et une démographie plus stable. Ces départements sont proches du candidat Emmanuel Macron.

- La classe 4 concerne des départements tels que le Nord, où l'on retrouve une population plus âgée, avec des taux élevés de pauvreté, d'immigration, d'employés, d'ouvriers et de cadres intermédiaires. Ce type de département se rapproche de Mme. Marine Le Pen.

- La classe 5 représente des départements comme les Pyrénées-Atlantiques, où l'agriculture et l'artisanat sont prépondérants, avec une forte présence d'ouvriers et d'employés, mais un faible taux de pauvreté et d'immigration. Cette classe correspond aux départements surreprésentés dans les votes de M. Jean Lassalle.

- Enfin, la classe 6 inclut des départements tels que le Var, le Pas-de-Calais et l'Aisne, caractérisées par des taux élevés d'ouvriers et d'employés, ainsi qu'une concentration de pauvreté, mais une faible présence de cadres et de professions intermédiaires. Cette classe est celle qui se rapproche le plus de Mme. Marine Le Pen.

V. Conclusion

Pour conclure, l'Analyse Factorielle des Correspondances nous a permis de mettre en évidence les tendances politiques des départements français. De plus, grâce aux données démographiques et socio-économiques, utilisées en variables quantitatives supplémentaires, nous avons pu interpréter les choix électoraux des votants de chaque département. Par ailleurs, la classification a confirmé nos interprétations, à savoir, l'existence d'une corrélation entre les départements et les choix des candidats votés en fonction de facteurs socio-économiques.