Examen d'analyse de données

Fabrice Rossi

5 février 2013

Il est vivement conseillé de traiter les questions de chaque exercice dans l'ordre, afin de pouvoir éventuellement s'appuyer sur les conclusions d'une analyse pour faciliter les suivantes. Les deux exercices sont totalement indépendants.

J'attends des réponses précises et concises. Les questions qui appellent une réponse binaire (oui/non) demandent bien entendu une réponse argumentée sans laquelle aucun point ne sera compté.

1 Votes des représentants (États-Unis d'Amérique)

1.1 Présentation des données

On considère la base de données des votes effectués par les membres de la Chambre des représentants des EUA en 1984 sur 16 propositions importantes. Chaque individu est un membre de la Chambre décrit par 17 variables nominales. La variable Parti prend les modalités Démocrate et Républicain. Les autres variables, V1 à V16 représentent les votes et prennent les valeurs OUI, NON et NSP (pour une absence de vote). Il y a 267 représentants démocrates et 168 représentants républicains.

1.2 Visualisation

Question 1 *La figure 1 représente le vote effectué à la première proposition en fonction du parti d'appartenance du représentant. La proposition 1 est elle l'objet d'un consensus entre les deux parti ? Peut-on déterminer simplement à partir de la figure 1 si la proposition est adoptée à la majorité (relative) ?*

Question 2 *La figure 2 représente le pourcentage de réponses OUI en fonction de la question. D'après cette figure, quelle est la proposition pour laquelle le vote est le plus indécis ? Si on fait l'hypothèse que les votes les plus francs sont ceux qui opposent le plus les républicains aux démocrates, sur quelles propositions pourra-t-on constater ce clivage ? Comment déterminer de façon simple la proposition qui oppose le plus républicains et démocrates (de façon exacte, c'est-à-dire sans faire l'hypothèse simplificatrice précédente) ?*

Question 3 *En utilisant un critère bien choisi, l'analyste décide d'étudier les propositions représentées par la figure 3. Qu'est-ce qui oppose ces deux figures ? D'après ces figures, peut-on espérer retrouver l'orientation politique d'un représentant en fonction de ses votes ? Parmi les trois propositions représentées graphiquement dans les figures 1 et 3, quel vote semble le plus adapté pour une telle tâche ?*

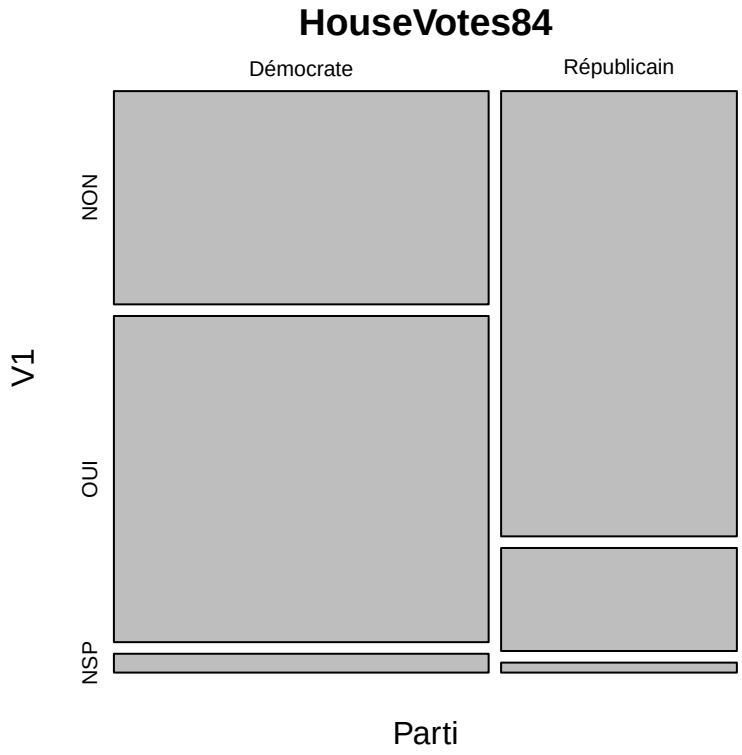

FIGURE 1 – Diagramme mosaïque pour la première proposition

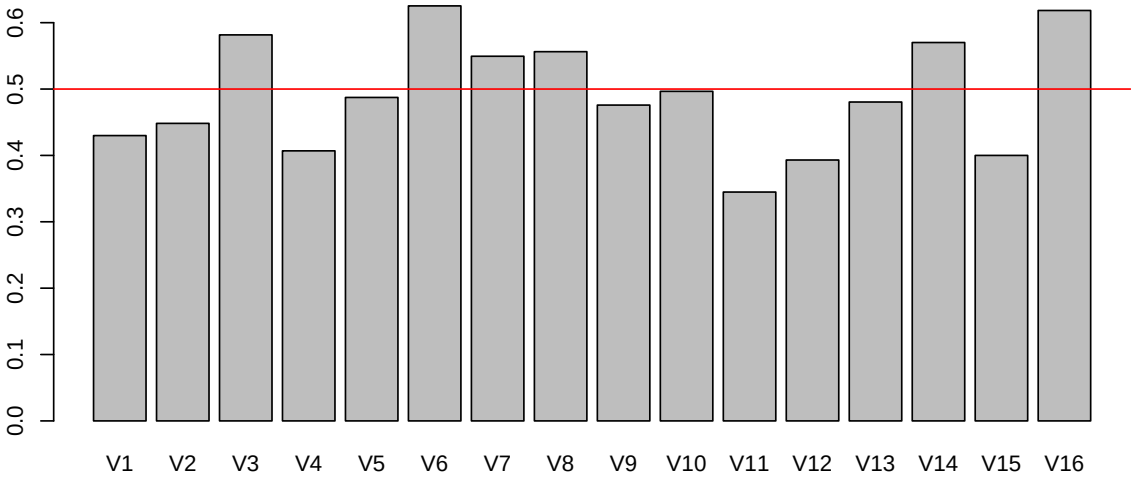

FIGURE 2 – Pourcentage de vote OUI à chaque proposition

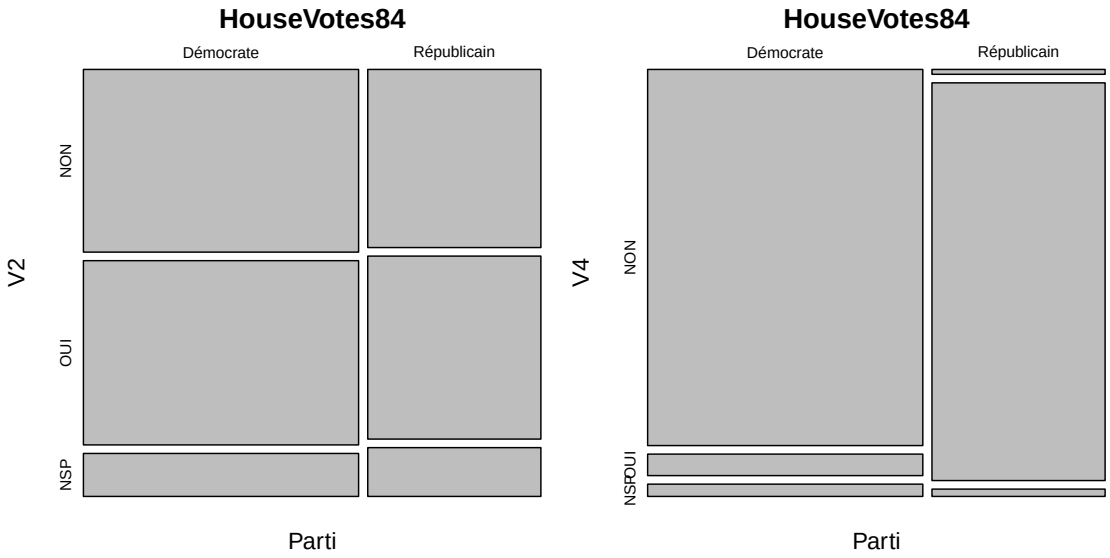

FIGURE 3 – Diagrammes mosaïques pour deux propositions

1.3 Analyse supervisée

Dans cette partie, on cherche à prédire l'appartenance au parti démocrate en fonction des votes effectués par un représentant en utilisant un classifieur bayésien naïf.

Question 4 *Rappelez l'hypothèse d'indépendance conditionnelle associée au classifieur bayésien naïf (dans le contexte des votes). Cette hypothèse semble-t-elle* a priori *réaliste pour les données étudiées ?*

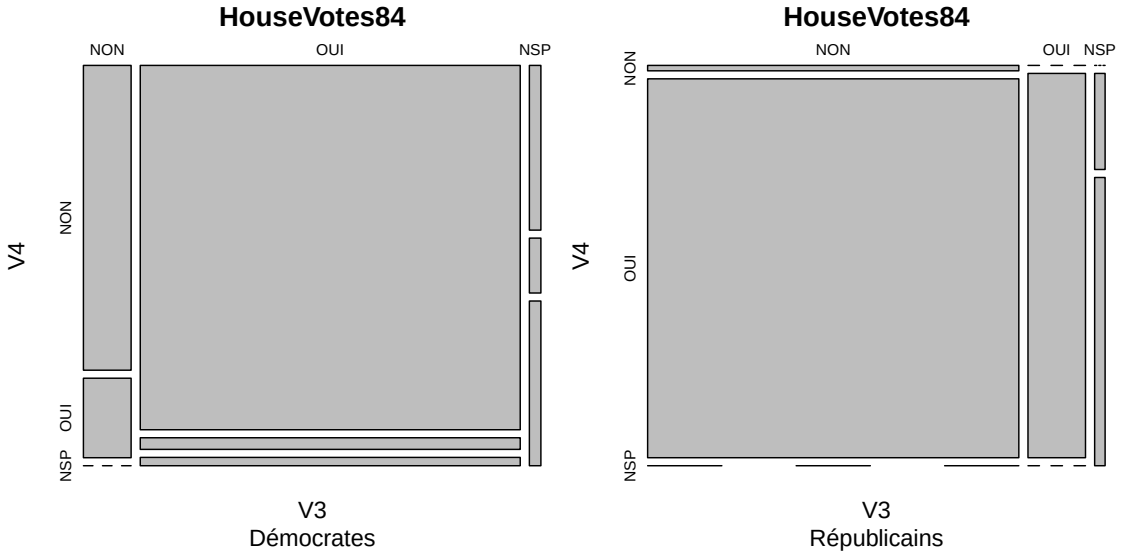

FIGURE 4 – Diagrammes mosaïques pour deux propositions selon le parti d'appartenance

Question 5 *La figure 4 représente les votes à la proposition 4 en fonction des votes à proposition 3 (parti par parti). Discutez la pertinence de l'hypothèse du classifieur bayésien naïf en vous appuyant sur cette figure.*

Républicains

	NON	NSP	OUI
V1	134	3	31
V2	73	20	75
V3	142	4	22
V4	2	3	163
V5	8	3	157
V6	17	2	149
V7	123	6	39
V8	133	11	24
V9	146	3	19
V10	73	3	92
V11	138	9	21
V12	20	13	135
V13	22	10	136
V14	3	7	158
V15	142	12	14
V16	50	22	96

Démocrates

	NON	NSP	OUI
V1	102	9	156
V2	119	28	120
V3	29	7	231
V4	245	8	14
V5	200	12	55
V6	135	9	123
V7	59	8	200
V8	45	4	218
V9	60	19	188
V10	139	4	124
V11	126	12	129
V12	213	18	36
V13	179	15	73
V14	167	10	90
V15	91	16	160
V16	12	82	173

TABLE 1 – Table des votes de chaque parti aux 16 propositions

Question 6 *La table 1 représente les votes de chaque parti aux 16 propositions. Quelles grandeurs nécessaires à la mise en œuvre d'un classifieur bayésien naïf peuvent être évaluées grâce à cette table ?*

Question 7 *Soit un représentant ayant voté selon le vecteur de réponse V suivant :*

V1	*V2*	*V3*	*V4*	*V5*	*V6*	*V7*	*V8*
OUI	*NON*	*NSP*	*OUI*	*NON*	*OUI*	*OUI*	*OUI*

V9	*V10*	*V11*	*V12*	*V13*	*V14*	*V15*	*V16*
NON	*NON*	*OUI*	*NON*	*NON*	*NON*	*NON*	*OUI*

Donner le rapport de probabilités $\frac{P(Démocrate \mid V)}{P(Républicain \mid V)}$ *tel qu'estimé par le classifieur bayésien naïf. On rappelle qu'il y 267 représentants démocrates et 168 représentants républicains. On se contentera de donner une formule numérique sans chercher à simplifier la fraction obtenue (ni à évaluer les produits).*

	Démocrate	Républicain
Démocrate	239	28
Républicain	14	154

TABLE 2 – Matrice de confusion du classifier bayésien : chaque ligne correspond au parti réel des représentants, chaque colonne au parti prédit par le classifieur.

Question 8 *En utilisant le classifieur bayésien naïf, on obtient la matrice de confusion de la table 2 (sur l'ensemble des données). Commentez les résultats. Peut-on en particulier estimer de façon satisfaisante la qualité du modèle à partir de cette table ?*

Question 9 *L'analyste s'intéresse aux représentants les plus difficiles à classer pour le classifieur bayésien naïf (c'est-à-dire les plus ambigus). Comment déterminer ces représentants ?*

La table 3 donne les votes des 6 représentants les plus difficiles à classer. Expliquez pourquoi les deux premiers représentants de cette table sont effectivement difficile à classer (en vous appuyant notamment sur la table 1).

	398	167	243	74	391	278
Parti	Démocrate	Républicain	Républicain	Républicain	Démocrate	Républicain
V1	OUI	OUI	NON	OUI	NSP	NON
V2	OUI	NON	NON	NON	NSP	NON
V3	NON	OUI	NON	OUI	NON	OUI
V4	NON	OUI	NON	OUI	NON	OUI
V5	OUI	OUI	OUI	OUI	NSP	OUI
V6	NSP	OUI	OUI	NON	OUI	OUI
V7	NON	OUI	OUI	OUI	NSP	OUI
V8	NON	OUI	NON	NON	NON	OUI
V9	NON	NON	NON	OUI	NON	NON
V10	NON	OUI	NON	OUI	NON	OUI
V11	OUI	NON	NON	NON	OUI	NON
V12	NON	OUI	NSP	NON	OUI	NON
V13	OUI	NON	NON	OUI	NON	NON
V14	OUI	OUI	OUI	OUI	OUI	OUI
V15	NON	OUI	OUI	NON	NON	NON
V16	OUI	OUI	OUI	OUI	NSP	OUI

TABLE 3 – Représentants difficiles à classer pour le classifieur bayésien naïf

Question 10 *L'analyse s'intéresse ensuite aux représentants les plus mal classés par le classifieur bayésien naïf. Comment déterminer ces représentants ?*

La table 4 donne les votes des 6 représentants les plus mal classés. Expliquez pourquoi les deux premiers représentants de cette table sont mal classés (en vous appuyant notamment sur la table 1). Peut-on espérer bien classer ces représentants avec une autre méthode que le classifieur bayésien naïf ?

Question 11 *En utilisant une validation croisée à 3 blocs, l'analyste estime la matrice de confusion de la méthode du classifieur bayésien naïf donnée par la table 5. Commentez les résultats, notamment en comparant avec la table 2.*

2 Élections au *Bundestag* (Allemagne)

2.1 Présentation des données

On étudie les élections du Parlement allemand (le *Bundestag*) en 2009. L'Allemagne est constituée de 299 circonscriptions (zones électorales). Chaque circonscription (un individu dans les données) est décrite par 16 variables numériques et une variable nominale. La variable nominale `state` précise l'état de la circonscription (l'Allemange est découpée en 16 états). Les 16 autres variables sont à valeurs entières et désignent des votes, exceptées la variable `eligible` qui donne le nombre d'électeurs de la circonscription. Chaque citoyen donne 2 votes qui peuvent être soit valide, soit invalide (`valid1` et `valid2` comptent les votes valides, `invalid1` et `invalid2` comptent les votes invalides). Les 10 autres variables restantes donnent le nombre de premier et second votes obtenus par les cinq partis allemands : les sociaux démocrates `SPD1` et `SPD2`, les conservateurs chrétiens `UNION1` et `UNION2`, les verts `GRUENE1` et `GRUENE2`, les libéraux `FDP1` et `FPD2`, et enfin la gauche `LINKE1` et `LINKE2`.

	408	268	376	72	389	177
Parti	Démocrate	Républicain	Démocrate	Républicain	Démocrate	Républicain
V1	NON	OUI	NON	OUI	NON	NON
V2	NON	NON	OUI	OUI	OUI	NON
V3	NON	NON	NON	OUI	OUI	OUI
V4	OUI	NON	OUI	OUI	OUI	OUI
V5	OUI	NON	OUI	NON	OUI	NON
V6	OUI	NON	OUI	NON	OUI	NON
V7	NON	OUI	NON	OUI	NON	OUI
V8	NON	OUI	NON	OUI	NON	OUI
V9	NON	OUI	NON	OUI	NON	OUI
V10	NON	OUI	NON	OUI	NON	OUI
V11	OUI	NON	OUI	OUI	NON	NON
V12	OUI	NON	OUI	NON	OUI	NON
V13	OUI	NON	NON	NON	OUI	NON
V14	OUI	OUI	OUI	OUI	OUI	OUI
V15	NON	NON	NON	NON	NON	OUI
V16	NON	OUI	NON	OUI	NSP	OUI

TABLE 4 – Représentants mal classés par le classifieur bayésien naïf

	Démocrate	Républicain
Démocrate	239	28
Républicain	15	153

TABLE 5 – Matrice de confusion du classifier bayésien estimée par validation croisée : chaque ligne correspond au parti réel des représentants, chaque colonne au parti prédit par le classifieur.

Question 12 *D'après la description des données, quelles relations de dépendance entre certaines variables peut-on attendre ?*

2.2 Nombre de votants

Question 13 *La figure 5 donne une estimation de la distribution des tailles des circonscriptions en nombre d'électeurs (chaque tiret vertical sous la courbe représente une circonscription). Sachant que le nombre moyen d'électeurs par circonscription est de 207921 et que l'écart-type de cette grandeur est de 22936, que dire de l'homogénéité des circonscriptions en terme de nombre d'électeurs ?*

Question 14 *La figure 6 représente le nombre de votants en fonction du nombre d'électeurs, avec en superposition les prévisions d'un modèle linéaire reliant les deux variables. Commentez cette figure.*

2.3 Orientation politique des circonscriptions

On s'intéresse dans cette section aux votes aux différents partis. On exclue donc de l'analyse la variable indiquant l'état, et les variables donnant respectivement le nombre d'électeurs, le nombre de votes et le nombre de votes valides. Les votes sont ensuite transformés en pourcentage des votes exprimés (ce qui revient, par exemple, à diviser la variable `SPD1` par `valid1` et la variable `SPD2` par `valid2`).

On réalise une analyse en composantes principales des données ainsi obtenue en centrant les données sans les normaliser.

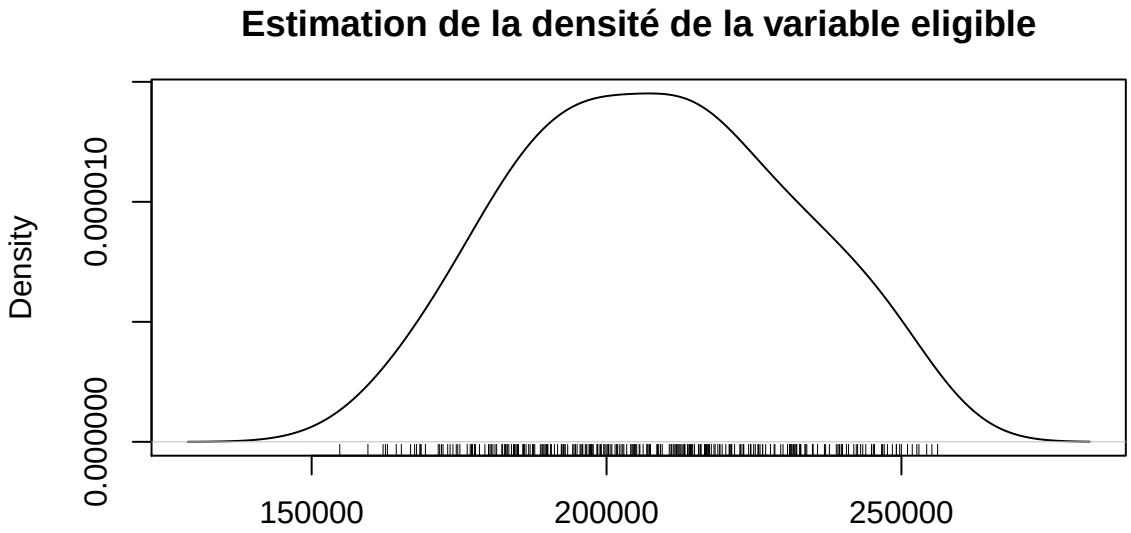

FIGURE 5 – Distribution des tailles des circonscriptions

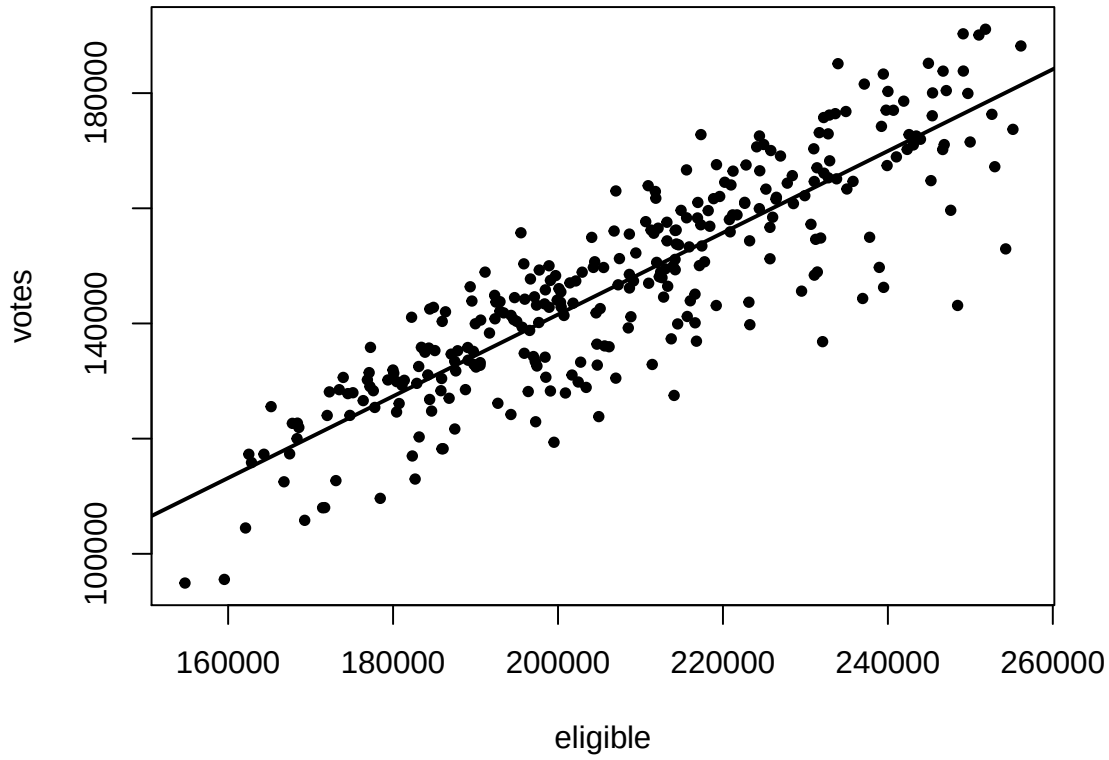

FIGURE 6 – Nombre de votes en fonction du nombre d'électeurs

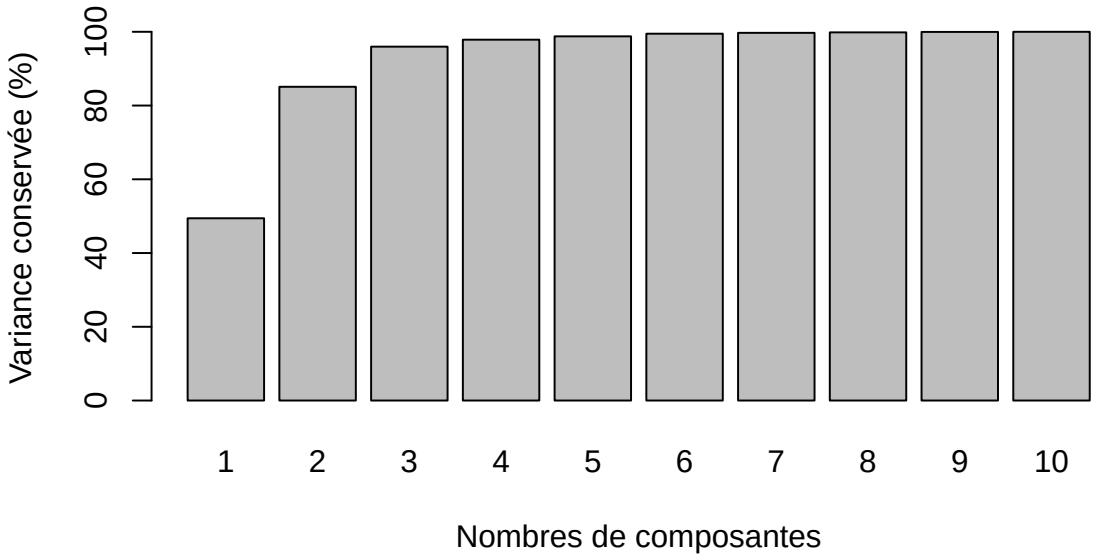

FIGURE 7 – Pourcentage de variance expliquée en fonction du nombre de composantes conservées dans l'ACP

Question 15 *Justifiez brièvement les choix de l'analyste. Quel serait l'effet d'une normalisation dans ce contexte ?*

La figure 7 représente le pourcentage de variance expliquée par l'ACP en fonction du nombre de composantes conservées.

Question 16 *Que peut-on dire des dernières composantes (à partir de la sixième composante) ? Est-ce surprenant ?*

Question 17 *D'après la figure 7, combien de composantes faudrait-il conserver pour ne pas perdre trop d'information ? Que dire d'une représentation sur les deux premiers axes ?*

Question 18 *Commentez la représentation des circonscriptions sur le plan factoriel (deux premiers axes principaux, figure 8), en utilisant en particulier le cercle des corrélations (figure 9) pour interpréter les résultats ainsi que les liens entre les variables.*

Question 19 *La figure 10 représente les circonscriptions sur le plan factoriel état par état (un plan factoriel par état). En utilisant cette représentation et la figure 9, discutez l'homogénéité politique des états et leur orientation (les partis dominants).*

On réalise maintenant une classification hiérarchique des données en utilisant le critère de Ward. Le dendrogramme obtenu est représenté sur la figure 11.

Question 20 *Commentez le dendrogramme en indiquant en particulier quel(s) nombre(s) de classes considérer pour construire une partition des circonscriptions.*

Question 21 *Proposez une technique de visualisation des résultats de la classification permettant d'analyser les liens entre les classes de circonscriptions, les votes et les états. Quels statistiques pourrait-on calculer pour résumer l'orientation politique des circonscriptions dans chaque classe et/ou dans chaque état ?*

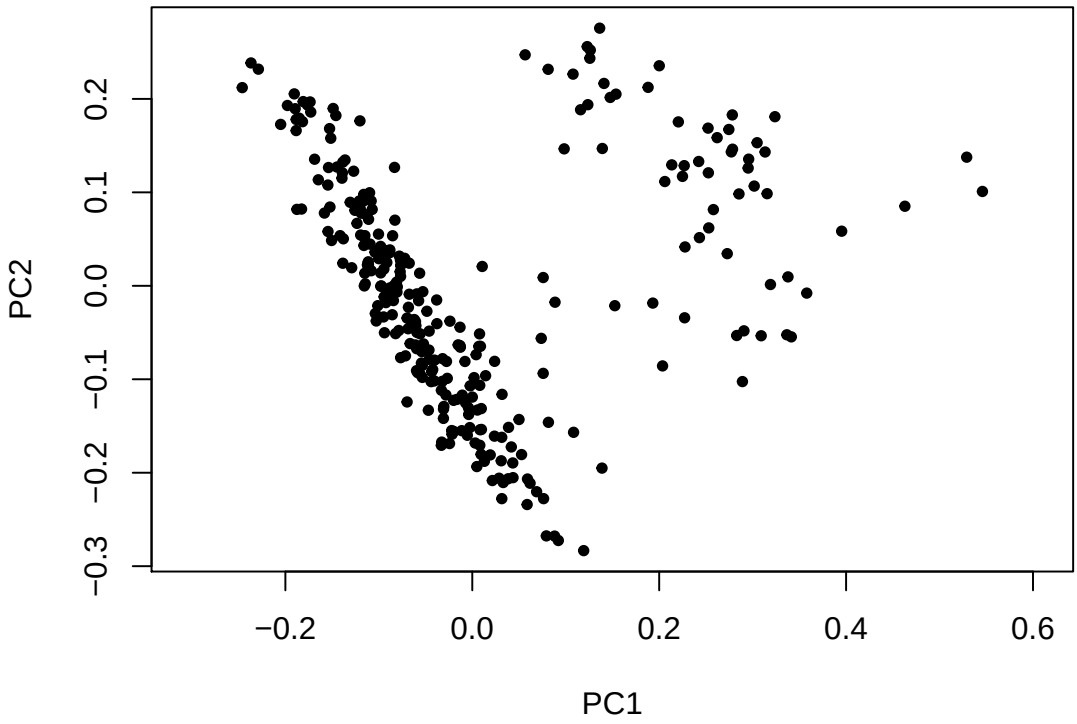

FIGURE 8 – Représentation des circonscriptions sur le premier plan factoriel

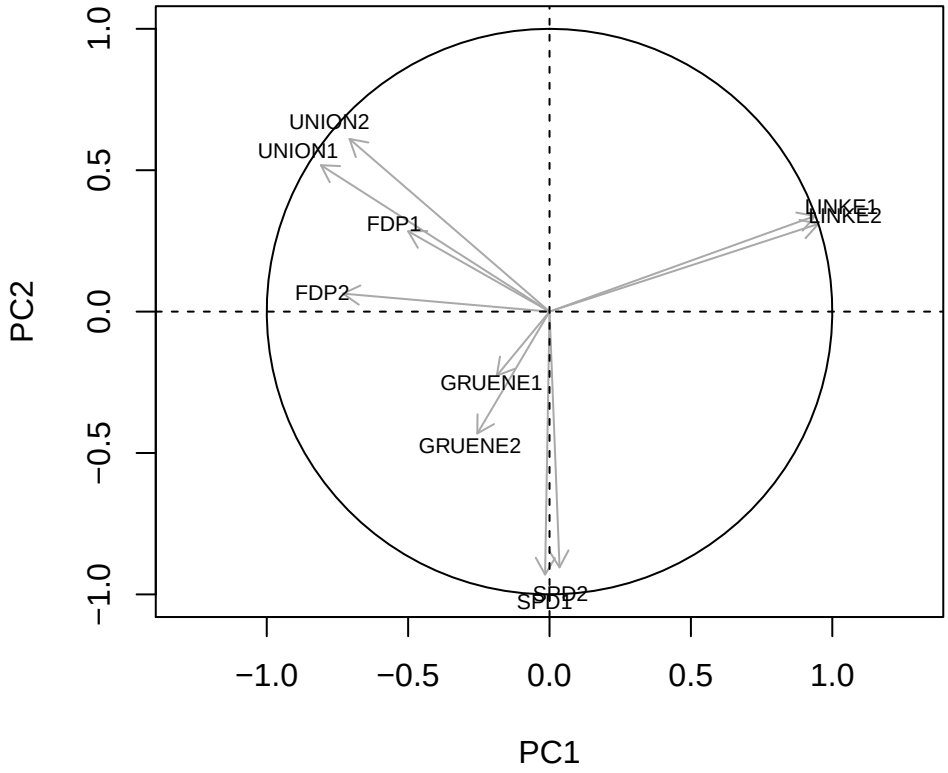

FIGURE 9 – Cercle des corrélations

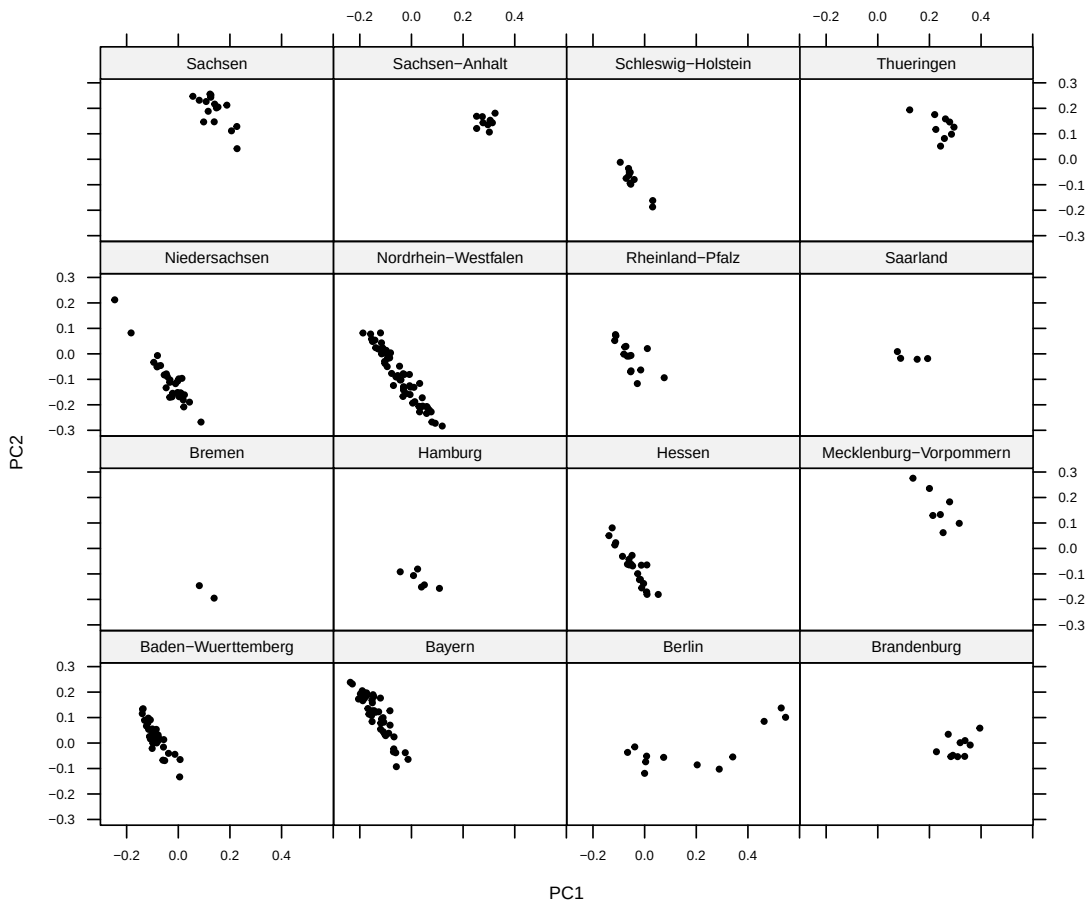

FIGURE 10 – Représentation des circonscriptions sur le premier plan factoriel en fonction des états

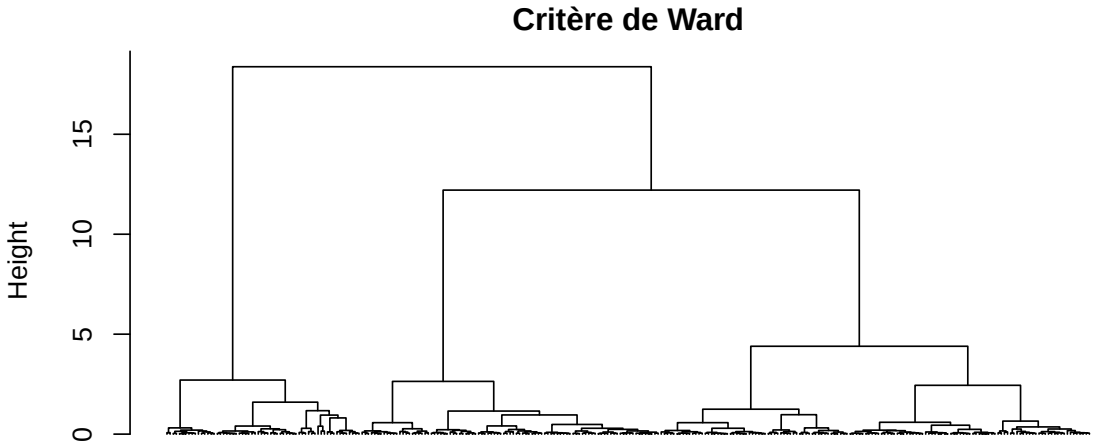

FIGURE 11 – Dendrogramme de la classification pour le critère de Ward