STT7335 - Travail de session

Gabriel Rieger Junqueira

Hiver 2020

Question 1

Les graphiques sont à la fin du document

Prétraitement

Je commence à vérifier s'il y a des valeurs manquantes, aberrantes ou extrêmes. Je vois aussi que les données sont déjà au bon format, des fois il faut les transformer en facteur mais c'est tout. Il n'y a pas de cas Homme/H/h/HOMME pour la variable genre par exemple, ni des problèmes similaires pour les autres variables qualitatives
Je trouve une valeur de poids égal à 575, je pense qu'il s'agit d'une valeur extrême.
Je vérifie que la base n'a pas des doublons. Je prends la décision de remplacer les données manquantes des variables qualitatives par le mot 'inconnu' en creante un nouvelle catégorie qui pourrait nous fournir des informations.
Pour l'ACP la fonction PCA de FactoMineR fait la standardisation et l'imputation.
Pour les autres items j'analyserai si je fais une imputation par modèle ou si j'exclue les données manquantes.
J'exclu la colonne ID parce qu'elle ne sert à rien.

ACP

J'ai décidé faire l'ACP plusieurs fois:
i) avec les quatre variables qualitatives et imputation automatique de la fonction PCA

ii) avec les quatre variables qualitatives et imputation par prédiction de la fonction mice

iii) sans la variable dette_2013 avec imputation de la fonction PCA

iv) sans la variable salaire_2013 avec imputation de la fonction PCA

Avec l'ACP j'ai pu observer que les variables salaire_2013 et age sont très corrélées, on pourrait n'utiliser qu'une. La même chose on voit avec les variables emprunt et dette_2013.
Si on travaille avec moins de variables on réduit notre risque au surapprentissage lors de l'étape des modèles. J'ai trouvé intéressant que la variable age change des sens à mesure qu'on ne considère pas la dette_2013. J'ai ne trouvé pas un bonne idée imputer les variables qui seront prédites dans les prochaines items, mais j'ai décidé regarder chaque item séparément et exploiter au max les possibilités, ça reflète un difficulté d'avoir une vue d'ensemble de ma part.

On voit que les variables emprunt et dette_2013 sont positivement bien corrélées, elles ont beaucoup d'influence sur la première composante. Les variables age et salaire_2013 sont positivement bien corrélées aussi et ont beaucoup d'influence sur la deuxième composante.

Deux à deux emprunt et dette_2013 ne sont presque pas corrélées à age ni salaire_2013.

Pour les observations, avec 10342 individus il est difficile de voir quelque chose cependant si on essaie de superposer les graphiques des variables par dessus du graphique des individus

on voit que environ 2/3 des observations sont beaucoup impactées pour les 4 variables numériques.

Une approche comme sélection de variables serait de choisir seulement emprunt ou dette_2013, vue quelles sont dans la même direction, un autre choix seraiy possible entre age et salaire_2013.

Analyse de correspondances multiples

Dans ce point j'ai eu beaucoup de difficulté de voir quelque chose interessante.

En général les observations sont homogènes et il y a beacoup des variables avec influence sparses et quelques unes avec un peu plus d'influence.

Pour mieux segmenter je fais les images par trois groupes des variables différentes, soit:

i)Variables de nature sociodémographique:

ii)Variables concernant le diplôme obtenu en 2010 et les emprunts qui y sont liés:

iii) Variables concernant le statut du diplômé en 2013:

Regroupement

Dans ce point j'utilise l'approche hiérarchique car je ne sais pas combiens de groupes seront fournis, mais j'ai pensé qu'on aura entre 2 et 3 groupes.

Après regarder qu'il y a une très grande perte d'inertie entre les groupes 1 et 2, 2 et 3, 3 et 4, mais pas beaucoup après le groupe 4 j'ai choisi 4 comme nombre de groupes.

Pour simplifier, mon regroupement à été fait à partir d'un autre ACP (pas exactement le même de la partie ACP) avec 3 composantes à considérer. Et avant de faire le regroupement hiérarchique l'algorithme fait des tests avec k-moyens avec k=100)

Explication cluster 1: plus grand pouvoir discriminant de la variable age et scolarite_avant=scolarite_avant_<= DES

Explication cluster 2: plus grand pouvoir discriminant salaire_2013

Explication cluster 3: plus grand pouvoir discriminant scolarite_avant=scolarite_avant_<= DES

et salaire_2013

Explication cluster 4: dette_2013 et emprunt dans un sens et salaire_2013 dans le sens contraire.

Question 2

Pour ce point j'ai décidé de faire une forêt aléatoire avec la librairie et fonctio ranger par la simplicité et pour voir l'importance de chaque variables dans la prédiction.

J'ai divisé le fichier valid_entrain en 80% pour entrainement et validation et 20% pour test

J'ai exécuté le code plusieurs fois, mon exactitude étaient bonne mais pas la précision, pour améliorer ça j'ai testé quelques chiffres pour le *mtry* et pour la taille des noeuds terminaux et même le nombre d'arbres. Cependant j'ai continué avec une mauvaise précision, parce que les données ne sont pas bien distribuées selon le niveau de satisfaction.

Quelques chiffres:

echantillon de test

bonne exactitude mais précision entre 16% et 25% avec mtry entre 10 ou 18, 500 arbres.

Avec 100 arbres j'ai 33% de précision

Avec 1000 arbres précession de 16%

Avec min.node.size de 81 et 500 arbres et mtry 8

donne precision de 33%.

Si min.size.node = 1 donne précision de 50%

J'ai fait un modèle avec toutes les variables et un autre seulement avec les variables concernant le statut en 2013.

J'ai obtenu des résultats similaire pour exactitude et précision, si je reduit le nombre d'arbres cela s'ameliore.

une aute apporche comme selection de variabels serait de choisir seulement emprunt ou dtte, vue quelles sont dans la meme direction ou dans un cas pareil agee et salaire

On perds environ 20% des observations en excluant les valeurs manquantes liées à satisfaction_emploi

J'ai fait des tests sans etude_temps_plein=oui et inconnu, et résultats restent similaires.

Importance des variables (modèle avec toutes les variables)

1) emprunt
2) age
3) dette_2013
4) diplome_secteur_2010
5) scolarite_avant
6) genre
7) scol_mere

Importance des variables (modèle avec les variables concernant le statut en 2013)

1) lien_emploi_etude
2) emprunt
3) etude_temps_plein
4) secteur_emploi_2013
5) age

Question 3

Pour utiliser une autre méthode et comme il s'agit de la prédiction d'une variable quantitative j'ai choisi faire une régression linéaire.

J'ai divisé le fichier valid_entrain en 80% pour entrainement et validation et 20% pour test et après j'ai fait la prédiction sur le fichier test. Je me suis orienté seulement avec les variables de le fichier test_numero_3, car sinon ne sera pas possible faire un prédiction sur ce fichier.

Evidemment j'ai enlevé les valeurs manquantes du fichier test_numero_3.

Difficultés rencontrées avec la variable salaire_2013 qui a beaucoup des lignes à zéro ou à faible revenue, ce fait à affecte énormément l'erreur de mon modèle.
J'ai calculé l'EQM et l'EAM.

Même en sélectionnant seules les revenues plus grandes que zéro l'erreur n'est pas acceptable.
J'aurai du penser à un autre algorithme.

Une chose qui aiderait sera avoir l'information sur etude_temps_plein temps dans le fichier test pour pouvoir l'utiliser et même filtrer les étudiants à temps plein qu'en général n'ont pas de revenue, ou leur venue est très faible.

Question 4

J’ai créé une variable appelé moitie_emprunt, et je l’a comparé avec dette_2013, si dette_2013 est plus petit que moitie_emprunt j’ai marqué OUI dans une deuxième nouvelle variable appelée moitie_paye et NON dans le cas contraire.

Par la suite j’ai fait mon modèle, une forêt aléatoire, pour prédire la variable moitie_paye

Les imputations au long de ce TP affaiblissent la performance de modèle, dans ce item j’ai exclut aussi toutes les lignes avec emprunt = valeur manquante

Avant de soumettre la prédiction sur le fichier test_numero_4 j’ai divisé le fichier valid_entrain en 80% pour entrainement et validation et 20% pour test pour avoir une idée.

Je pense qu’il faut faire ça sinon je n’aurai aucune information sur la performance de mon modèle sur d’autres échantillons.

Pour évaluer le modèle j’ai utilisé l'exactitude, la précision, la sensibilité et la spécificité.

Les trois premières sont acceptables, seulement la spécificité n’est pas bonne

C’est un modèle simple mais performant.

Evidemment j’ai enlevé les valeurs manquantes du fichier test_numero_4.

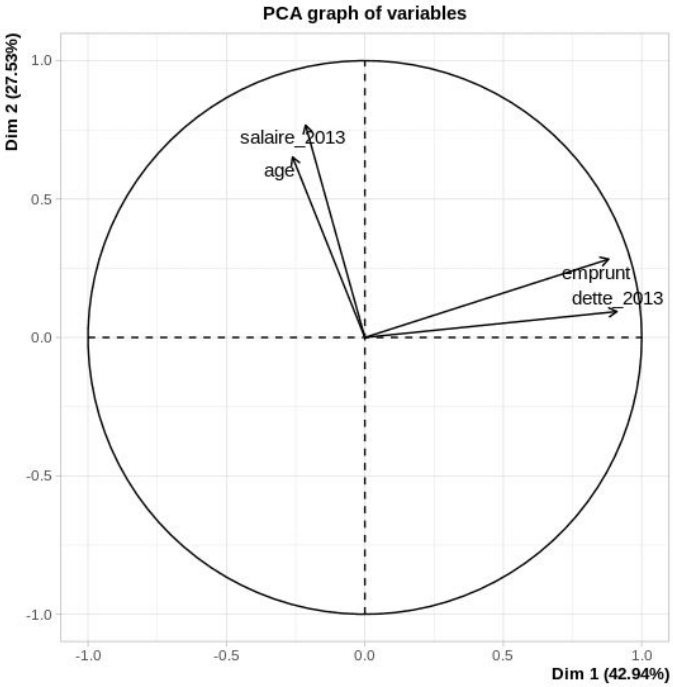
Importance des variables

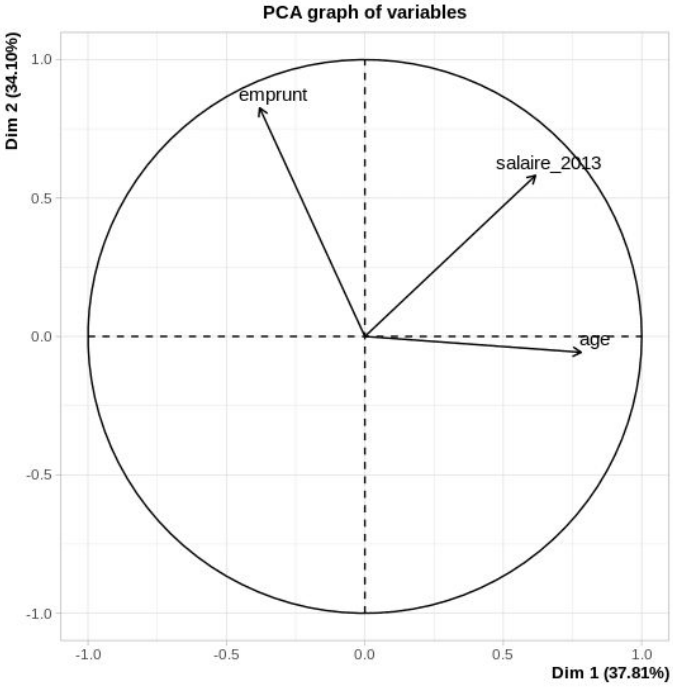
1) emprunt
2) emprunt_gouv
3) minorite_visible
4) diplome_sectuer_2010
5) age
6) langue

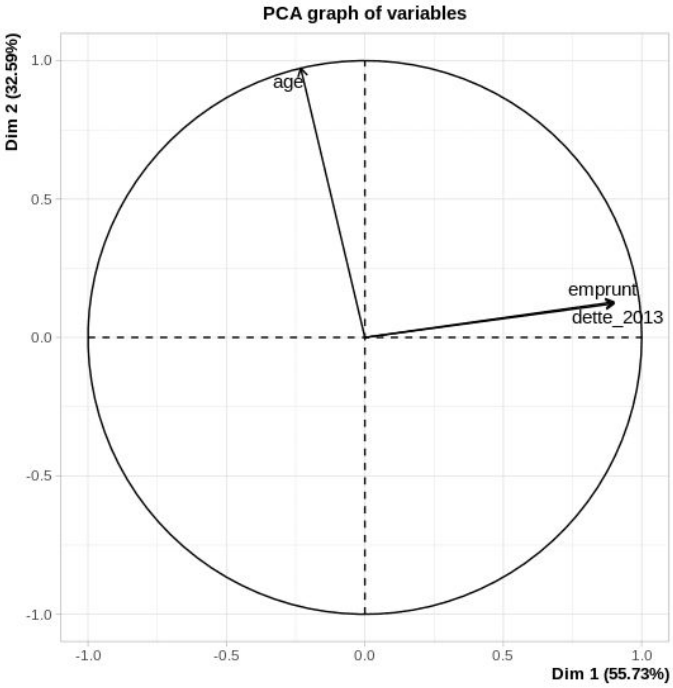
Pour évaluer sur l'égard de variables sensibles j’ai pensé en tester sur les variables genre, minorite_visible, scolarite_avant et langue concernant l'équité sur l’exactitude, taux de faux positifs et taux de faux négatifs.

Annexe Images

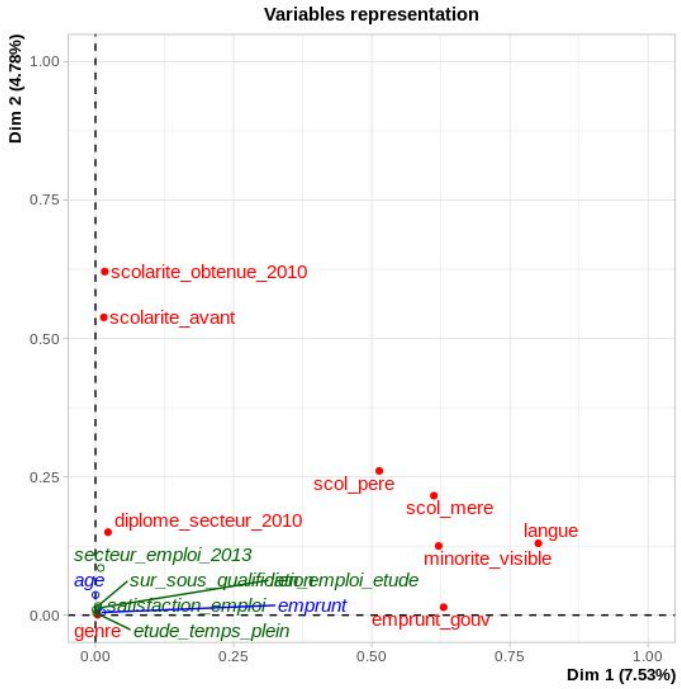
ACP

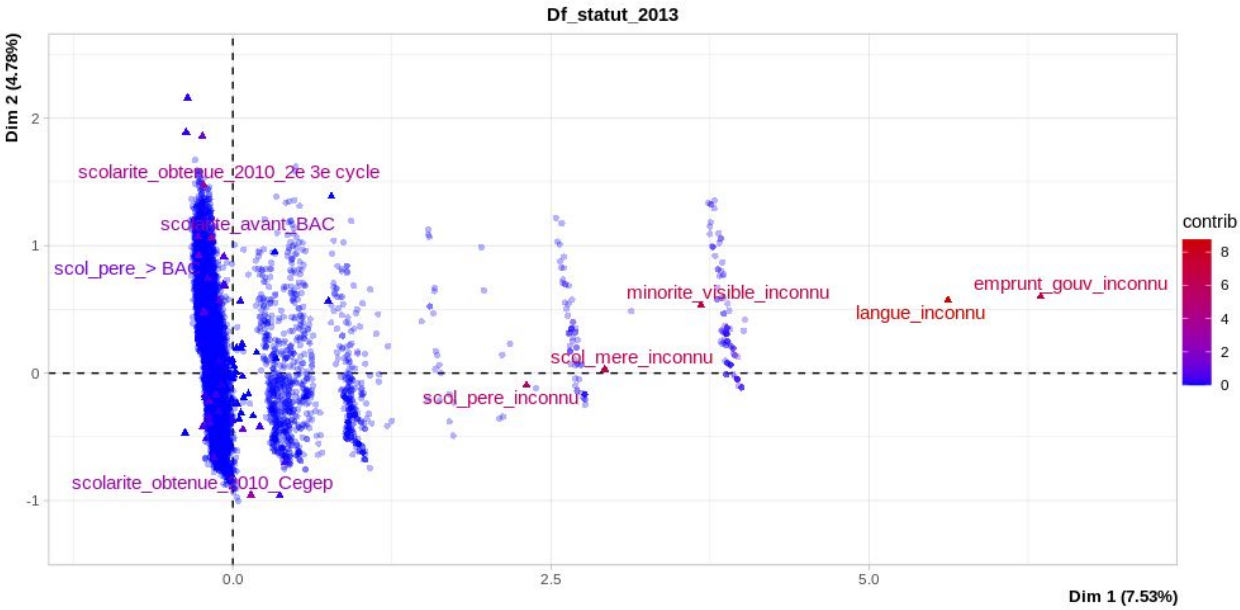

PCA graph of variables
Dim 2 (27.53%)
1.0
0.5
0.0
-0.5
-1.0
salaire_2013
age
emprunt
dette_2013
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (42.94%)

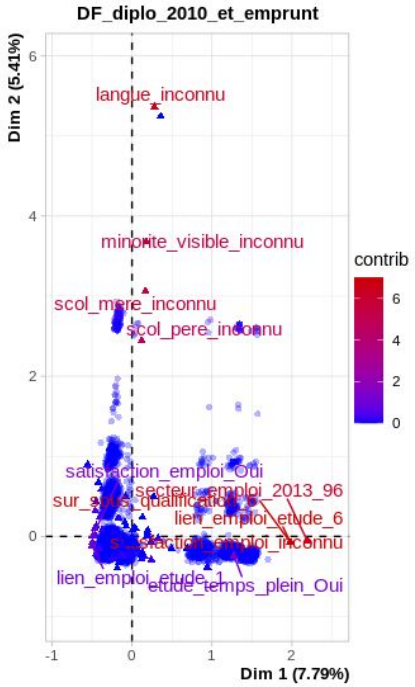

PCA graph of variables
Dim 2 (34.10%)
1.0
0.5
0.0
-0.5
-1.0
emprunt
salaire_2013
age
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (37.81%)

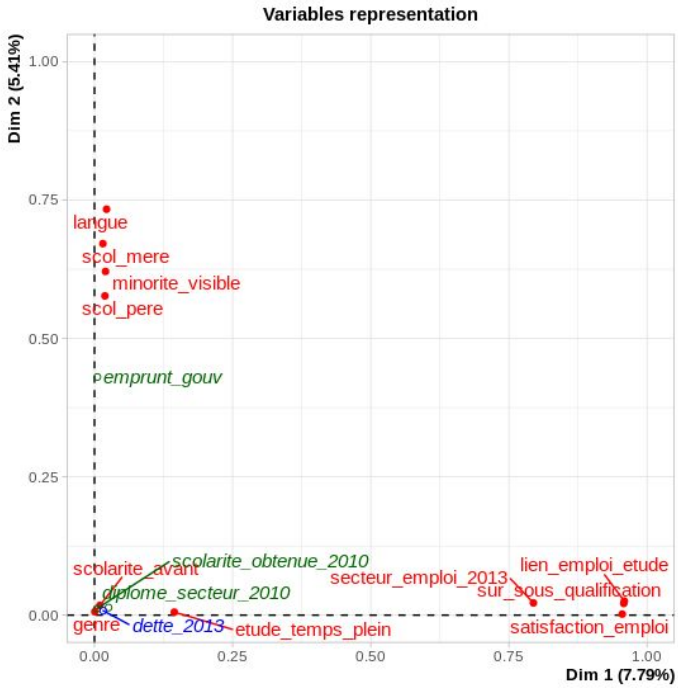

PCA graph of variables
Dim 2 (32.59%)
1.0
0.5
0.0
-0.5
-1.0
age
emprunt
dette_2013
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (55.73%)

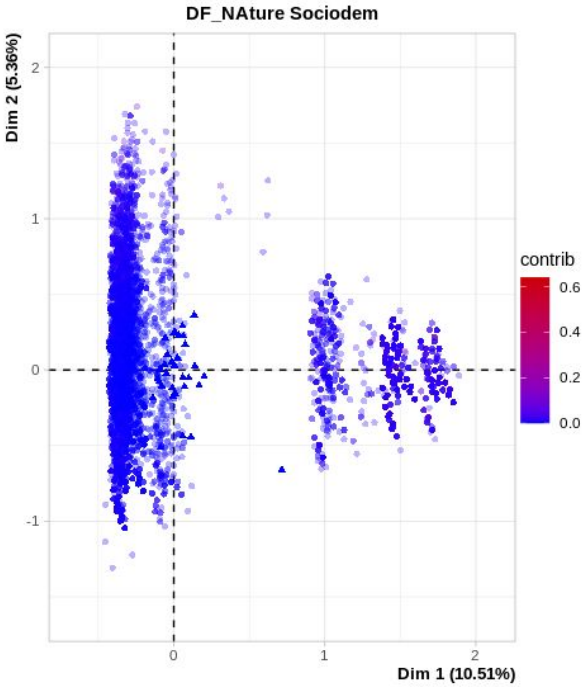
ACM

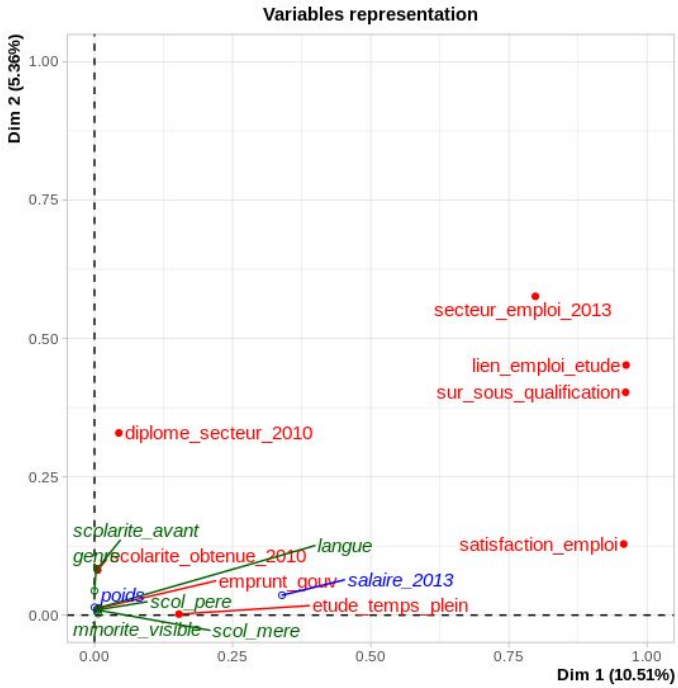

Variables representation
Dim 2 (4.78%)
1.00
0.75
0.50
0.25
0.00
scolarite_obtenue_2010
scolarite_avant
scol_pere
scol_mere
diplome_secteur_2010
langue
secteur_emploi_2013
minorite_visible
age
sur_sous_qualification
emploi_etude
satisfaction_emploi
emprunt
emprunt_gouv
genre
etude_temps_plein
0.00
0.25
0.50
0.75
1.00
Dim 1 (7.53%)


Df_statut_2013
Dim 2 (4.78%)
2
1
0
-1
scolarite_obtenue_2010_2e 3e cycle
scolarite_avant_BAC
scol_pere_> BAC
minorite_visible_inconnu
emprunt_gouv_inconnu
langue_inconnu
scol_mere_inconnu
scol_pere_inconnu
scolarite_obtenue_2010_Cegep
contrib
8
6
4
2
0
0.0
2.5
5.0
Dim 1 (7.53%)

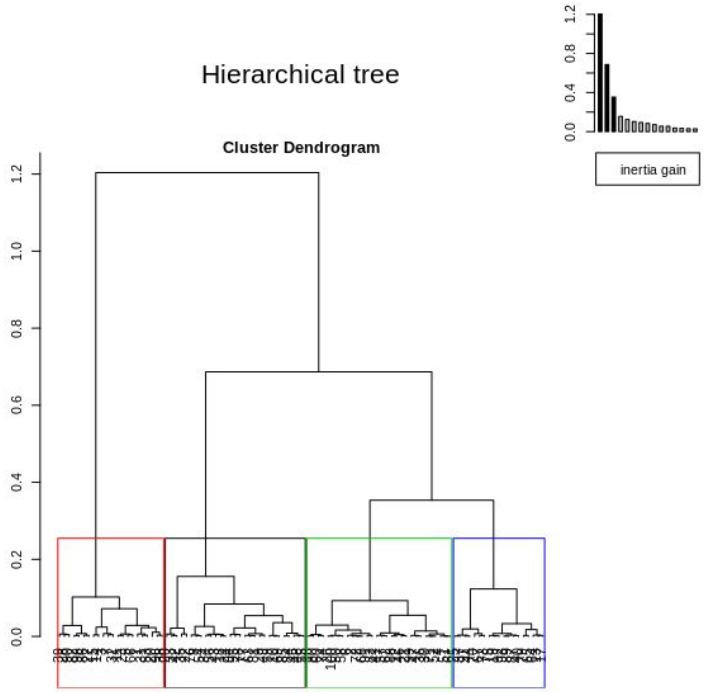

DF_diplo_2010_et_emprunt
Dim 2 (5.41%)
Dim 1 (7.79%)
langue_inconnu
minorite_visible_inconnu
scol_mere_inconnu
scol_pere_inconnu
satisfaction_emploi_Oui
secteur_emploi_2013_96
lien_emploi_etude_6
lien_emploi_etude_1
etude_temps_plein_Oui
contrib

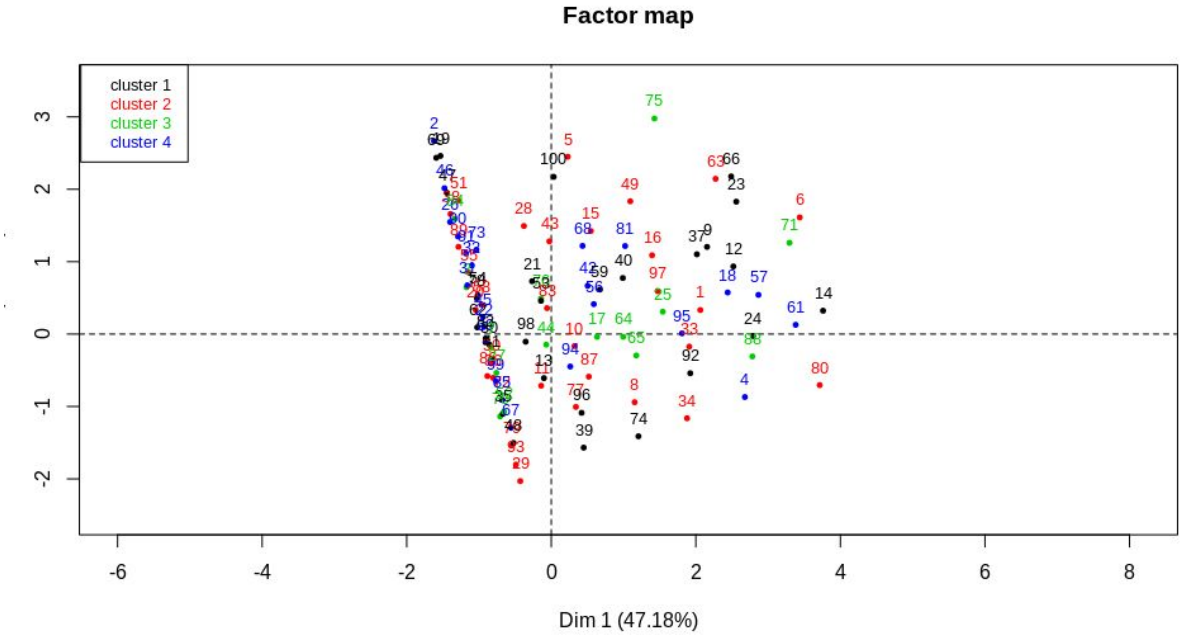

Variables representation
Dim 2 (5.41%)
Dim 1 (7.79%)
langue
scol_mere
minorite_visible
scol_pere
emprunt_gouv
scolarite_avant
scolarite_obtenue_2010
diplome_secteur_2010
genre
dette_2013
etude_temps_plein
secteur_emploi_2013
lien_emploi_etude
sur_sous_qualification
satisfaction_emploi

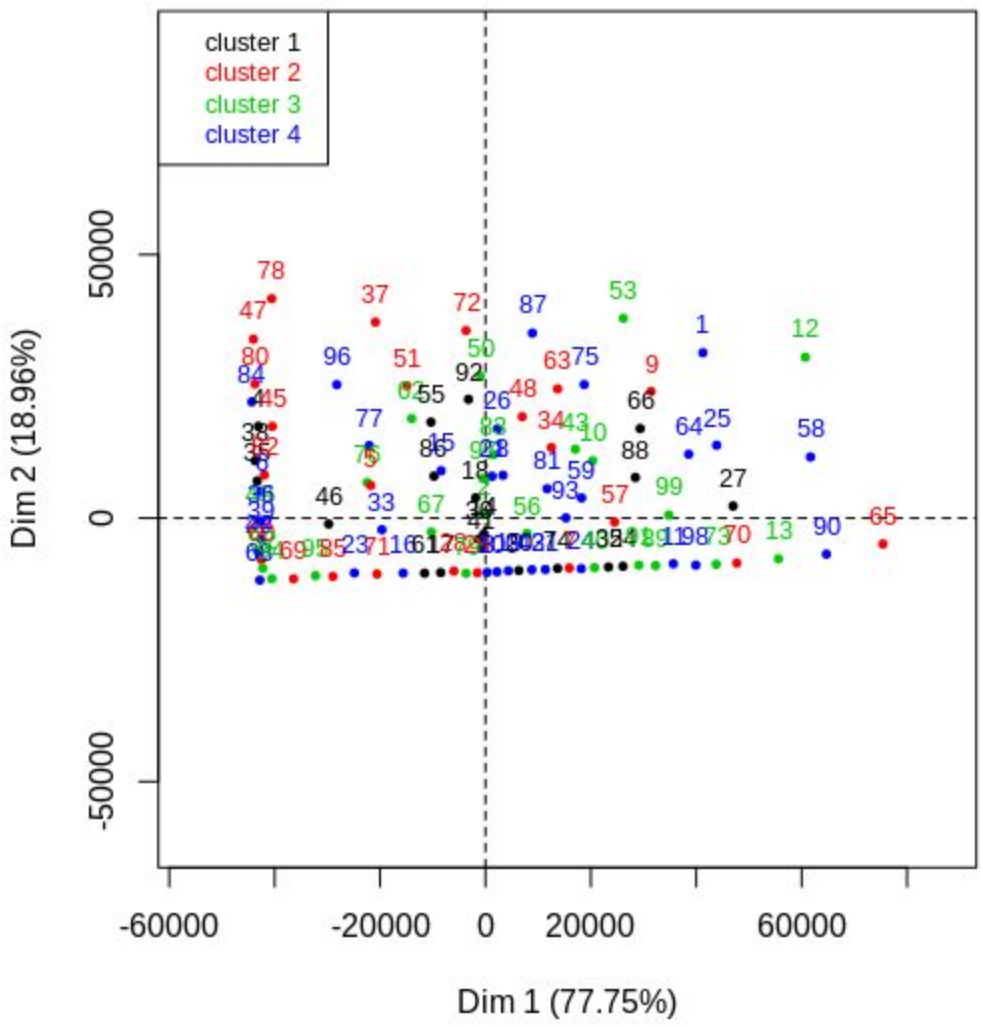

DF_NAture Sociodem
Dim 2 (5.36%)
Dim 1 (10.51%)
contrib
0.6
0.4
0.2
0.0
2
1
0
-1
0
1
2

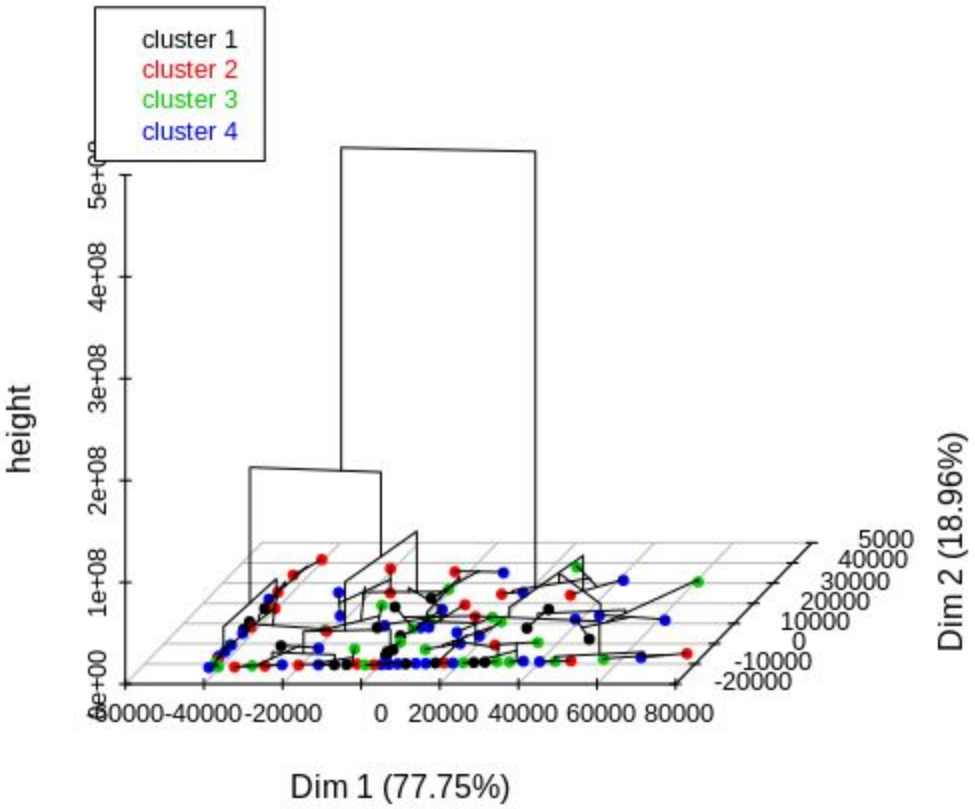

Variables representation
Dim 2 (5.36%)
Dim 1 (10.51%)
1.00
0.75
0.50
0.25
0.00
secteur_emploi_2013
lien_emploi_etude
sur_sous_qualification
diplome_secteur_2010
scplarite_avant
genre
colarite_obtenue_2010
langue
satisfaction_emploi
emprunt_gouv
salaire_2013
poids
scol_pere
etude_temps_plein
mmorite_visible
scol_mere

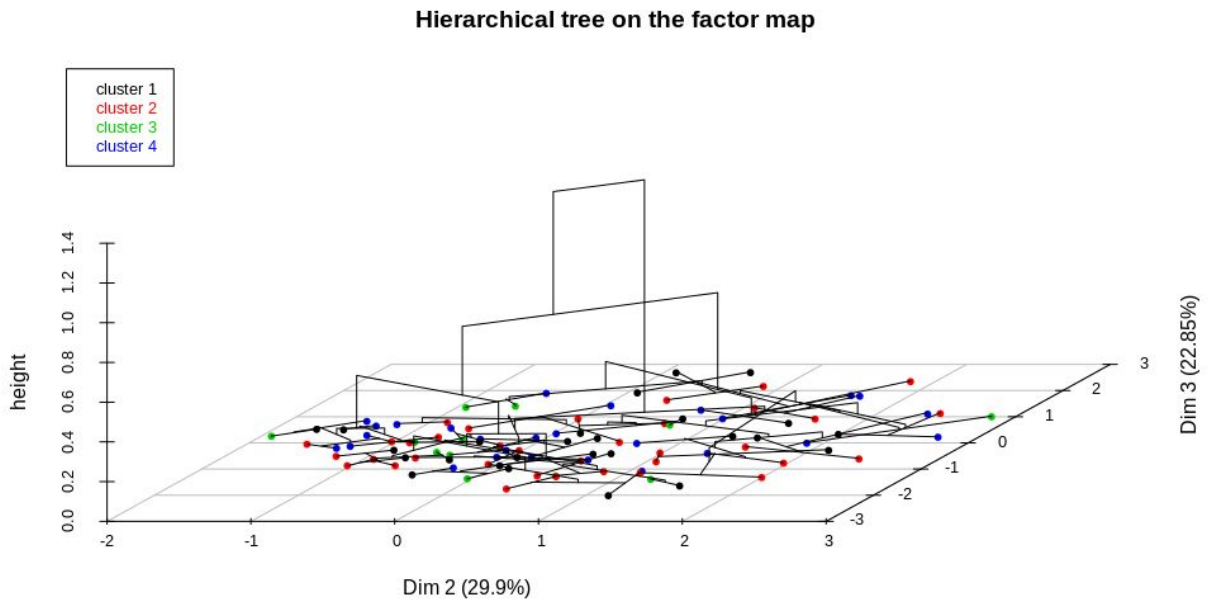
Regroupement






cluster 1
cluster 2
cluster 3
cluster 4
Dim 2 (18.96%)
50000
0
-50000
-60000
-20000
0
20000
60000
Dim 1 (77.75%)


cluster 1
cluster 2
cluster 3
cluster 4
height
0e+00
1e+08
2e+08
3e+08
4e+08
5e+08
-60000 -40000 -20000 0 20000 40000 60000 80000
Dim 1 (77.75%)
5000
40000
30000
20000
10000
0
-10000
-20000
Dim 2 (18.96%)



Annexe

Quelques exemples des observations dans chaque cluster

Le cluster 1 est composé d'individus tels que 13 et 37. Ce groupe est caractérisé par:

• valeurs élevées pour la variable dette_2013.

• des valeurs faibles pour les variables salaire_2013 et age (les variables sont triées des plus faibles).

Le cluster 2 est composé d'individus tels que 75 et 90. Ce groupe est caractérisé par:

• des valeurs faibles pour les variables emprunt, dette_2013, salaire_2013 et age (les variables sont triées des plus faibles).

Le cluster 3 est composé d'individus tels que 53 et 87. Ce groupe est caractérisé par:

• des valeurs élevées pour les variables emprunt et dette_2013 (les variables sont triées des plus fortes).

• des valeurs faibles pour les variables age et salaire_2013 (les variables sont triées des plus faibles).

Le cluster 4 est composé d'individus tels que 1, 9, 47, 68, 70, 72, 73, 80, 92 et 98. Ce groupe est caractérisé par:

• valeurs élevées pour les variables salaire_2013 et age (les variables sont triées des plus fortes).

• des valeurs faibles pour les variables dette_2013, et emprunt (les variables sont triées des plus faibles).

Par les graphiques on peut voir que les axes ne séparent pas bien les groupes

Les variables qui permettent de séparer le mieux les classes

quantitatives

	Eta2	P-value
salaire_2013	0.327	0
dette_2013	0.729	0
age	0.643	0
emprunt	0.539	0

qualitatives

p.value scolarite_avant

scolarite_obtenue_2010

scol_mere

scol_pere

langue

genre

minorite_visible

Individus plus proches du barycentre de chaque cluster

Cluster 1

ID 5745, 3356, 619,2794, 5497, Distances respectives 0.11 0.12 0.13 0.15 0.20

Cluster 2

ID 4919,3062,3022, 42820,2241 Distances respectives 0.08 0.10 0.11 0.11 0.11

Cluster 3

ID 6063. 6388. 9382. 8173. 2590 Distances respectives 0.13 0.17 0.18 0.20 0.21

Cluster 4

ID 3450, 8139, 4958, 4411, 4119 Distances respectives 0.07 0.11 0.20 0.20 0.20

Individus plus loins des centroïdes des autres clusters

Cluster 1

ID 5386 3323 4050 6091 2817 Distances respectives 4.46 4.26 4.15 4.04 4.00

Cluster 2

ID 6230 2900 292 2529 5852 Distances respectives 3.52 3.44 3.40 3.40 3.38

Cluster 3

ID 6266 6532 4161 5030 2866 Distances respectives 2.48 2.45 2.43 2.41 2.40

Cluster 4

ID 9615 6538 3226 7505 3036 Distances respectives 5.29 5.18 5.15 5.14 5.13