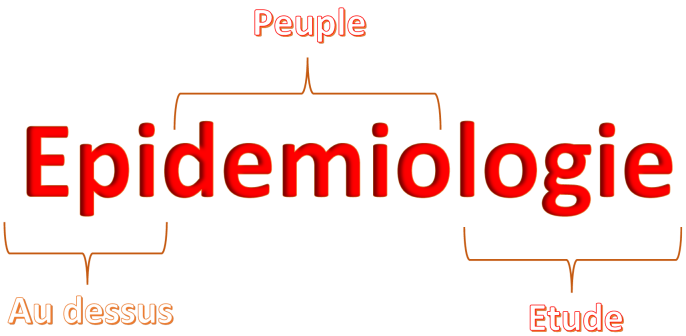
Rapport de stage M1

INSERM U970 PARCC (Paris Centre Cardiovascualaire)

E04 : Integrative Epidemiology of Cardiovascular Deseases

01/04/2019 - 31/08/2019



Odélia Guedj odelia-g@hotmail.fr
M1 Mathématiques en Interaction
Université Evry Val d'Essonne

Dr Jean Philippe Empana
Research Director, MD, PhD
INSERM U970 PARCC

Table des matières

Remerciements

Citations

Introduction

1 Contexte

1.1 L'Inserm U970 Equipe 4

1.2 L'Etude Parisienne Prospective 3 (EPP3)

EPP3 est une étude prospective de cohorte en population générale comptant n = 10157 sujets.
Ces dernier ont été recrutés dans des centres IPC (centres d'examens de santé conventionnés par l'assurance maladie) entre Juin 2008 et Décembre 2011.
Pour entrer dans l'étude, les sujets doivent avoir entre 50 et 75 ans.

Les donnés sont récoltées via des questionnaires envoyés tous les deux ans qui, à l'exception du questionnaire d'inclusion sont élaborés à l'INSERM U970 équipe 4.
Le questionnaire d'inclusion provient de l'IPC. Il contient une partie socio-admnistrative, des question sur l'environnement professionnel des sujets ainsi que sur leurs habitudes de vie (alimentation, tabac, alcool). Une autre partie du questionnaire traite des antécédents médicaux des sujets, tant familiaux que personnels, de leur état de santé actuel et de leurs prescriptions médicamenteuses. La dernière partie traite du bien-être des sujets : on y trouve des questions sur leur stress perçu, leur équilibre mental ainsi que leur nutrition.

Le but de l'étude est la recherche de facteurs de risque pour les maladies cardiovasculaires.
Dans chacun des questionnaires il est demandé aux sujets de déclarer leurs hospitalisations en détaillant le motif d'hospitalisation, le nom de l'hôpital, du service où ils ont été traités ... Chaque hospitalisation déclarée est appelée évènement.

Pour s'assurer de la validité des déclarations des sujets, un protocole de validation d'événement a été mis en place : régulièrement un certain nombre d'événements sont extraits de la base. Pour chaque évènement, on contacte l'hôpital pour qu'il transmette au responsable de l'étude les comptes-rendus hospitaliers (CRH) de l'évènement en question. Ensuite les CRH sont lus par un médecin qui valide, invalide ou corrige le diagnostique déclaré par le sujet.

L'étude EPP3 est une étude longue, il est prévu que le suivi dure 20 ans (10 questionnaires à raison d'un tous les deux ans).
Elle a déjà donné lieu à la publication de plus d'une vingtaine d'articles dans des revues prestigieuses comme le JACC (Journal of the American College of Cardiology) ou le JAMA (Journal of the American Medical Association).

1.3 Sujet de stage : Définition et limites

2 Bref aperçu de l'épidémiologie

2.1 Histoire et définition

2.2 Vocabulaire indispensable (analogie avec le vocabulaire "orienté mathématiques")

2.3 Principaux biais et indicateurs de l'épidémiologie

3 Description et explication des données

3.1 Variable d'expostion : activpro

Dans le questionnaire d'inclusion, 3 questions traitent de l'activité professionnelle des sujets (Question 1,2,3 du questionnaire IPC en annexe). Ces 3 questions ont été codées en 4 variables catégorielles :

- Adm12 : Êtes-vous
  - 6NNNNN contrat emploi-solidarité, intérim, CDD
  - N5NNNN chômeur depuis + de 6 mois
  - NN4NNN chômeur depuis - de 6 mois
  - NNN3NN à la recherche d'un emploi
  - NNNN2N jeune en cours de formation
  - NNNNN1 étudiant

- Adm12a : Êtes-vous
  - JXXX en formation professionnelle
  - X9XX au foyer
  - XX8X retraité(e)
  - XXX7 pré-retraité(e)

- Adm11 : Depuis quand n'exercez-vous plus d'activité professionnelle ?
  - 0 en activité
  - 1 moins d'un an
  - 2 1 an
  - 3 2 ans
  - 4 3 ans ou +
  - 5 jamais travaillé

- Adm10 : Si vous travaillez quelle est votre profession ?
  La réponse à cette question est du texte libre. Après la récupération des questionnaires par l'IPC, un code à deux chiffres est attribué à chaque grand groupe de profession.

L'encodage de la variable Adm10 introduit une première source potentielle d'erreur du fait de la difficulté d'interpréter du texte libre d'une part et des possibles erreurs de "classification humaine" d'autre part. J'ai été confronté à ces dernières lors du codage de ma variable d'exposition. Nous y reviendrons au paragraphe suivant. Il faut également noter qu'un des sujets a un code travail de 88 et que ce code ne correspond à aucune des professions de la liste IPC (en annexe).
Une autre difficulté est due à la possibilité qu'ont les sujet de cocher plusieurs réponses par question. Il en résulte un grand nombre de classes dans chaque variable ce qui a compliqué le codage de la variable activpro.

3.1.1 Codage manuel

L'objectif du projet étant d'étudier l'impact de l'activité professionnelle sur la santé perçue des sujets, il faut tout d'abord résumer l'information contenue dans les 3 variables citées précédemment en une seule variable.
J'ai donc créer une variable catégorielle nommée activproManuel qui comporte 5 classes :

- R pour retraités
- T pour travailleur
- C pour chômeur
- I pour inactif
- NSP pour ne sait pas (cas "inclassables")

Pour cela j'ai fait le choix de me baser de manière successive sur les variables Adm12a puis Adm11 puis Adm12 puis Adm10. La raison en est simple : c'est la variables Adm12a qui propose la réponse "Retraités", or c'est la catégorie qu'il m'intéresse le plus d'étudier. Ensuite la variable Adm11 nous indique si le sujet travaille encore OU depuis combien de temps il a cessé de travailler. La variable Adm12 discrimine les chômeurs. Enfin, la variable Adm10 est utile pour vérifier la cohérence des différentes réponse des sujets ou bien de trancher dans des cas où les autres variables ne fournissaient pas d'informations suffisantes.
Par exemple, il y a 57 individus pour lesquels les variables Adm12 et Adm12a ne sont pas renseignées et qui déclarent ne plus travailler depuis moins d'un an ou plus (c'est à dire qu'ils ont cochés les cases 1,2,3,4 ou 5 du questionnaire). Il est possible que ces sujets soient au chômage, à la retraite ou qu'ils n'aient jamais travaillé.

Voici le tableau récapitulant comment ont été classés ces sujets en fonction de leur code travail :

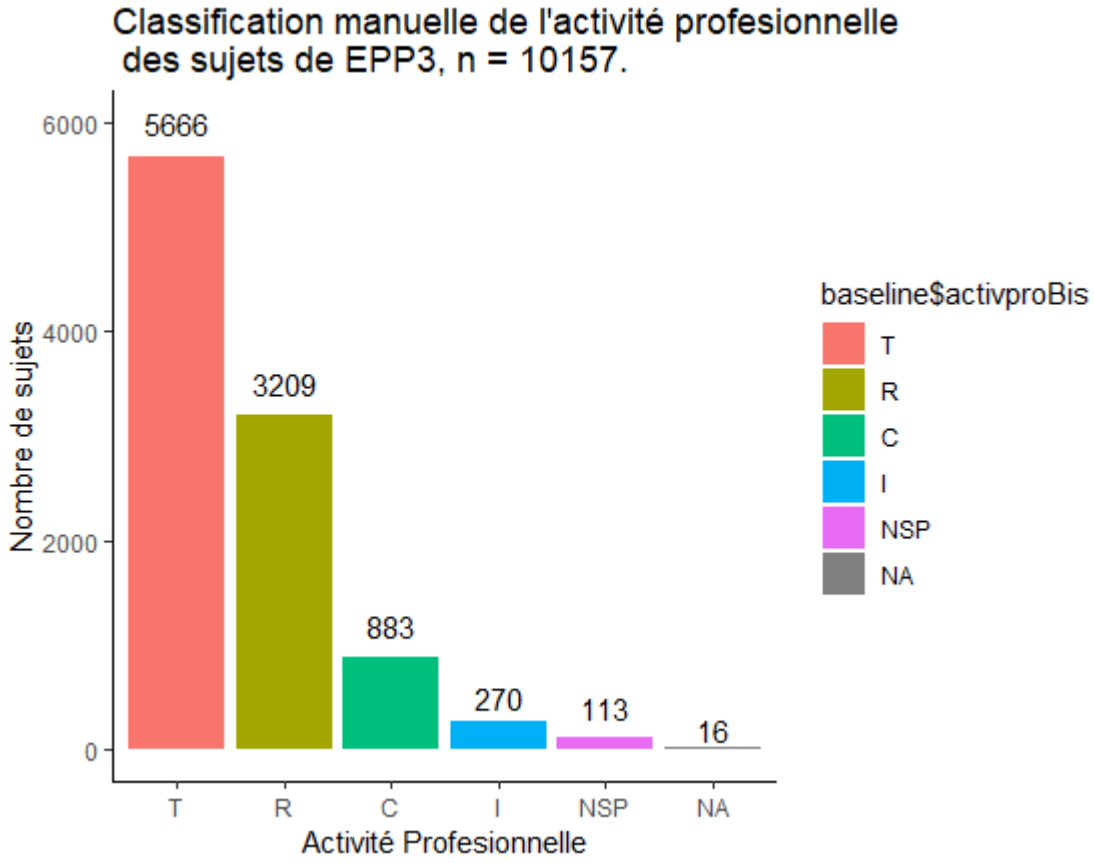
	Codes
Retraités	71 72 73 74 75 76 77 78
Chômeurs	91 92 93 94 95 96
Inactifs	81 82 84 85 86
Ne sait pas	88
Travailleurs	Le reste

Règle de décision : Les sujets indiquant qu'ils sont à la retraite et qu'ils travaillent sont classés en tant que travailleurs et ce, même si leur code travail (c'est à dire la variable Adm10) est de type "Ancien X" (codes 71 à 78 de la liste IPC) car il est possible de percevoir une pension de retraite et de continuer à travailler (par exemple en tant qu'expert). Or ce qui est intéressant pour l'étude est la santé perçue des sujets ne travaillant plus du tout. Il vaut donc mieux classer ceux qui complètent leur retraite avec un emplois en tant que travailleurs qu'en tant que retraités.
Ainsi, sont étiquetés inactifs les sujets n'aillant jamais travaillé, ou déclarant être au foyer ou ayant un code travail parmi les deux suivants : 85/86. Ces codes correspondent à des individus inactifs non retraités respectivement de moins de 60 ans et de plus de 60 ans.
Si les variables Adm12 OU Adm10 indiquent que le sujet est chômeur (en incluant la simple recherche d'emplois), il est classé en chômeur.En effet, on peut raisonnablement faire l'hypothèse que la précarité induite par la recherche d'un emplois occasionnera une plus mauvaise santé perçue.
On classe retraités tous les sujets n'ayant aucun indicateur de chômage ou d'activité professionnelle et n'étant pas inactifs.

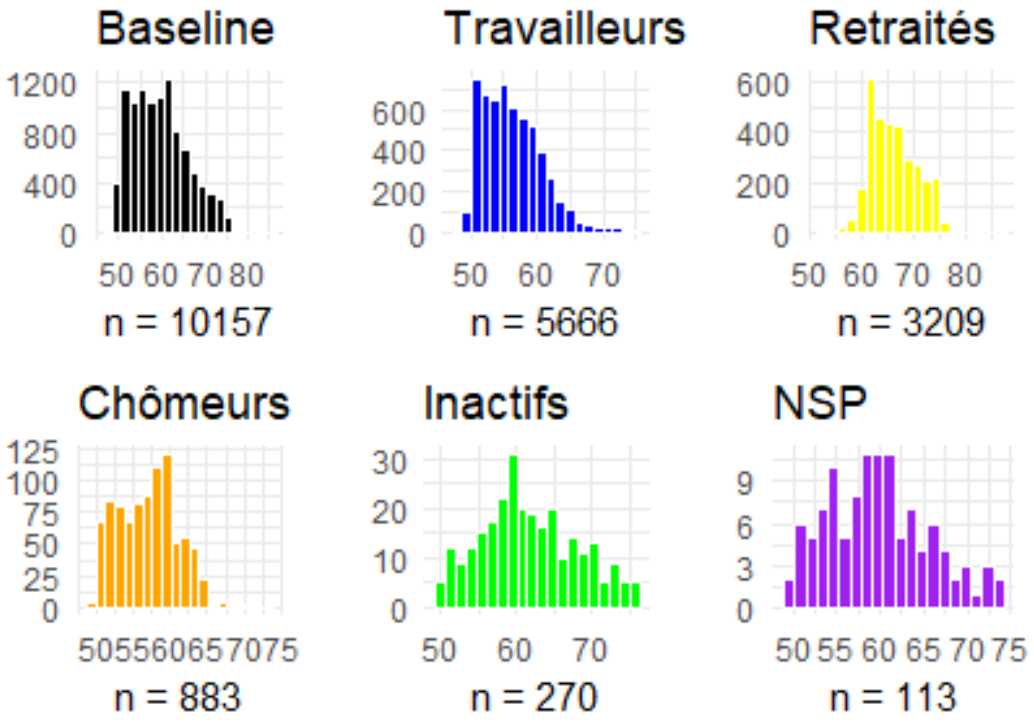
Enfin, le reste est classé comme travailleurs.

Malgré toutes les étapes de la classification et la minutie des vérifications, il existe un certain nombre de cas "impossibles" à classer car les informations d'un sujet pour les différentes variables sont contradictoires Ces sujets sont étiquetés NSP.

Finalement on obtient la répartition suivante :



Pour vérifier la cohérence des classes obtenues on peut tracer les distributions de l'âge dans chaque classe.



On remarque que 1116 sujets classés Travailleurs ont moins de 60 ans (on peut raisonnablement supposer que l'âge de la retraire est 60 ans car les sujets de la cohorte ont, à l'inclusion, un âge compris entre 50 et 75 ans). De même, 134 sujets sont classés en tant que Retraités et ont moins de 60 ans.

Ceci peut s'expliquer soit par une erreur de classification : ayant effectué cette dernière à la main il y a un risque non négligeable que je n'ai pas appliqué exactement le même critère de jugement pour chacun des cas. Il peut aussi ne pas s'agir d'une erreur dans ce cas les données sont ainsi et on veillera simplement à garder cela à l'esprit lorsque nous interpréterons les résultats des tests statistiques ultérieurs.

3.1.2 ACM et CAH

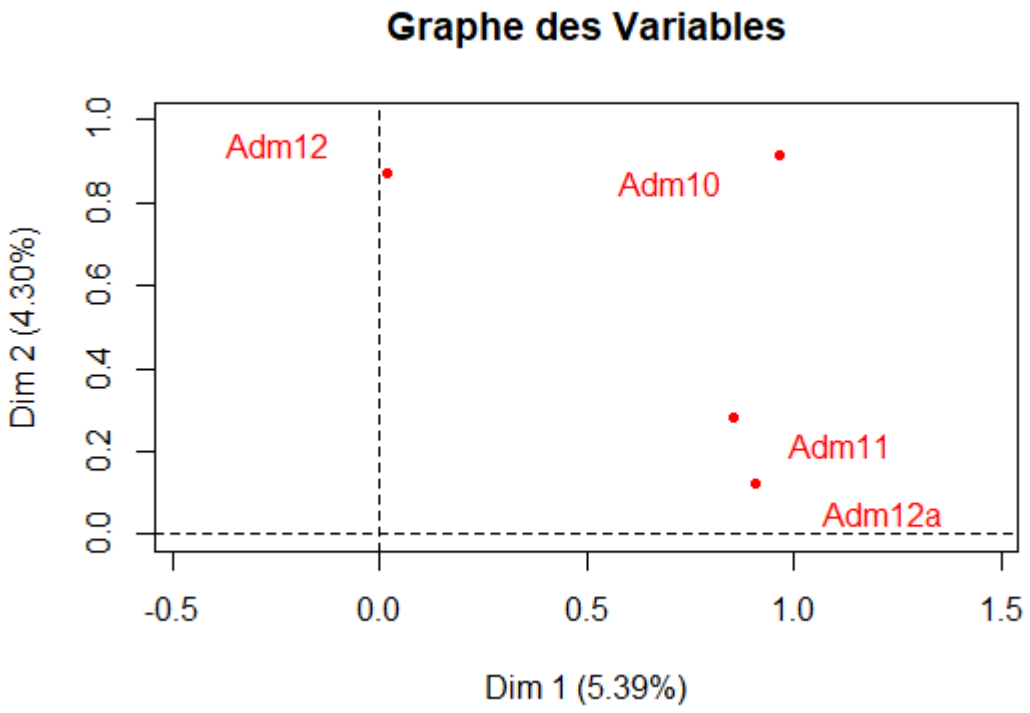
Pour vérifier la cohérence des résultats présentés au paragraphe précédent, j'ai effectué une Classification Ascendante Hiérarchique sur une base de données constituée des quatre variables utilisées pour la classification "manuelle" : Adm12a, Adm11, Adm12 et Adm10.
J'ai également veillé à supprimer les individus (n = 29) ayant une valeur manquante pour l'une de ces quatre variables. Les résultats de l'ACM sont rassurants, dans le sens où l'emplacement des quatre variables qui nous intéressent dans le plan factoriel correspond à l'intuition qu'on en avait à savoir :

— Les variables Adm12a et Adm11 apportent globalement la même information, cohérent puisque Adm12a et Adm11 discriminent les retraités et les travailleurs.

— Adm12 et Adm10 sont sur le même plan horizontal : elles discriminent les chômeurs (avec les codes travail 91 à 96).

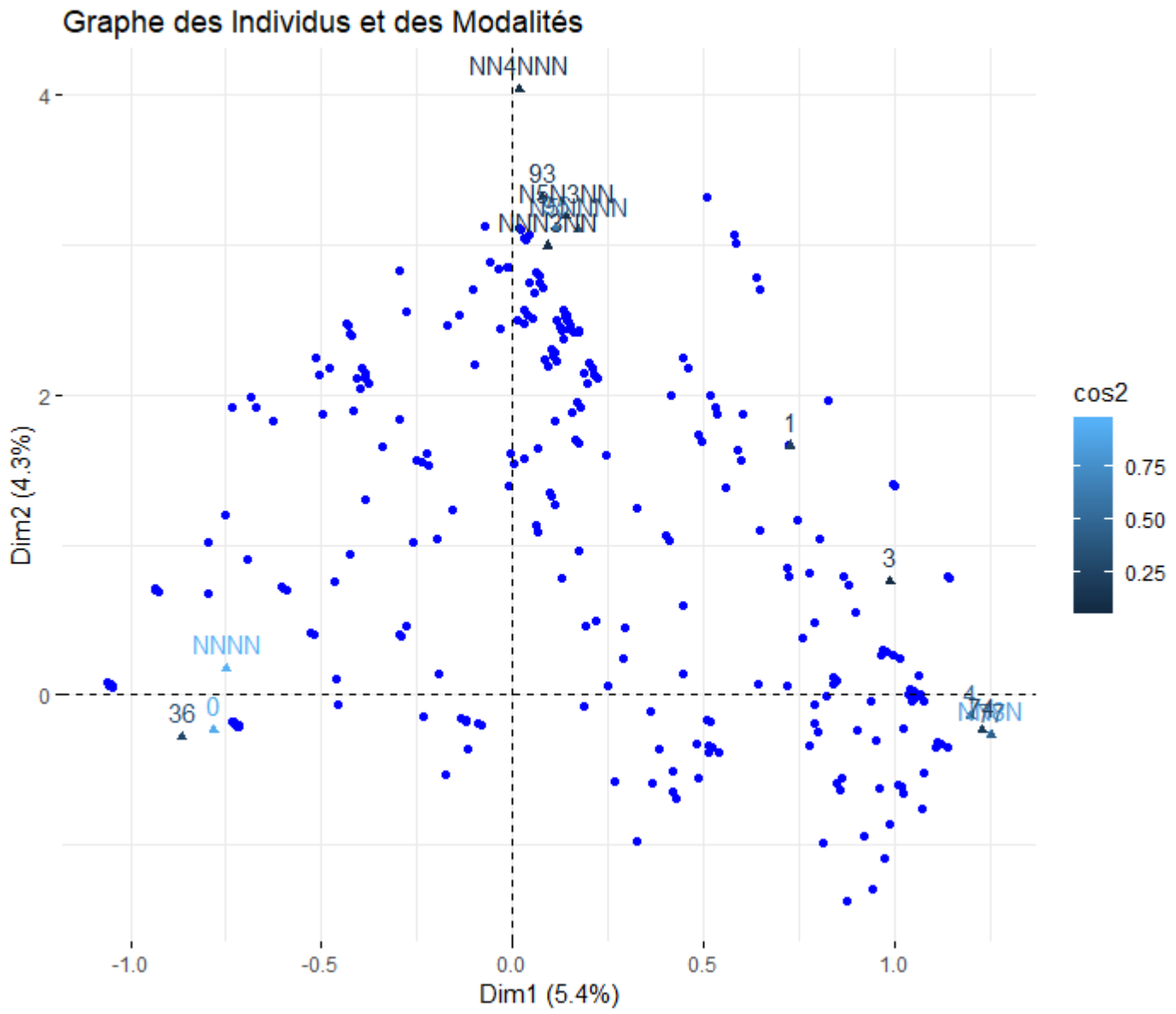
— Adm12a, Adm10 sont dans le même plan vertical : elles discriminent les inactifs.

On s'attend donc à ce que la CAH effectuée sur les résultats de l'ACM produise une classification proche de celle effectuée à la main.



NB : Les faibles pourcentages d'inertie sont tout à fait normaux dans le cadre d'une ACM. En effet, comme l'explique Jérôme Pagès dans son livre "Analyse factorielle multiple avec R", si les variables étaient toutes identiques dans le cas d'une ACP la première dimension aurait 100% d'inertie alors que dans une ACM la première dimension aurait au maximum $\frac{100}{(nb_modalités-1)}$ % d'inertie.

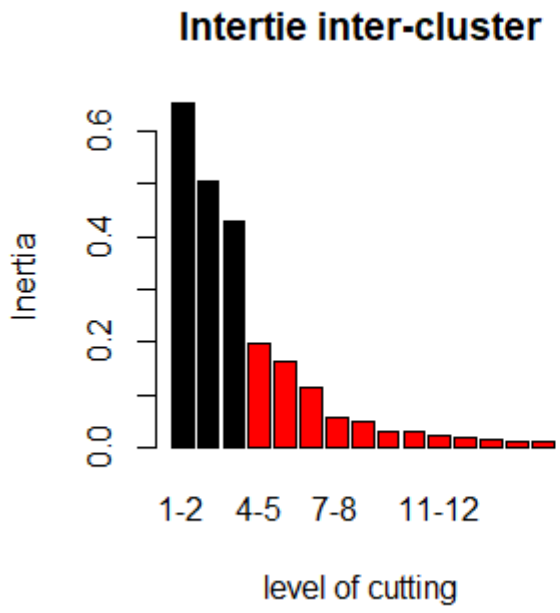
Sur le graphe suivant on affiche les individus ainsi que les 15 modalités ayant la plus grande contribution. Ces 15 modalités sont affichées selon un gradient de couleur en fonction de leur cos2.

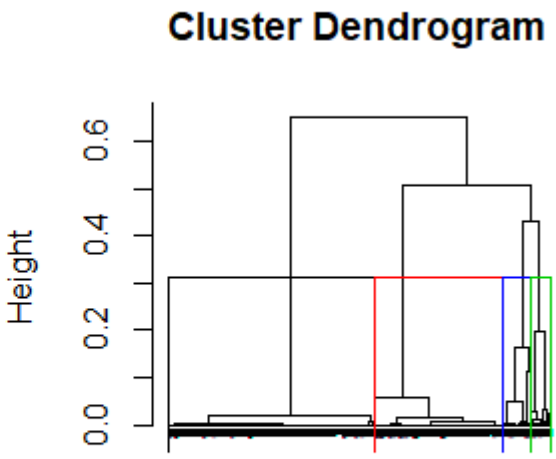


On voit de façon assez claire trois groupes distincts : le premier en bas à droite est certainement celui des sujets retraités, on y retrouve les modalités 1,3,4 de Adm11 (ne travaille plus depuis resp, moin de 1 an, deux ans, trois ans et plus), 74 de Adm 10 (Ancien cadre), NN8N de Adm12a (Retraité).
Le groupe en haut correspond sans doute aux chômeurs puisque les modalités NNN3NN, NN4NNN et N5N3NN de Adm12 (resp. A la recherche d'un emploi , Chômeur depuis moins de 6 mois et A la recherche d'un empois + Chômeur depuis 6 mois) et 93 de Adm10 (cadres et professions intellectuelles chômeurs).

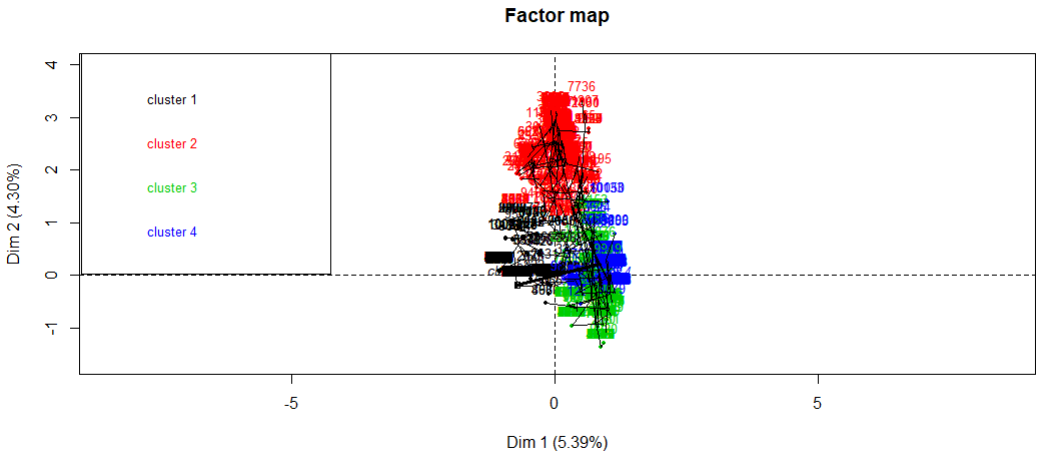
Enfin le dernier groupe est celui en bas à gauche, il représente surement les travailleurs puisque la modalité 0 de Adm11 s'y trouve (Travaille) ainsi que 36 de Adm10 (Cadres) et NNNN de Adm12a. La présence de cette dernière modalité fait sens puisqu'elle représente les individus n'ayant rien coché dans la variable Adm12a doit les choix étaient pré-retraités, retraités, personne au foyer, en formation professionnelle. Or si ces sujets travaillent, aucune de ces catégories ne les concerne.
Il reste deux choses qu'il me semble pertinent de relever : l'ACM semble particulièrement discriminer les cadres, on les retrouve dans les chômeurs, les retraités et les travailleurs. Ceci s'explique par le fait que les modalités 36, 74 et 93 (resp Cadres d'entreprises, Anciens cadres et Cadres et professions intellectuelles chômeurs) représentent à elles seules 3848 sujets (sur 10157 sujets dans la base globale et 10128 dans la base ayant permis l'ACM).
La deuxième chose à noté est l'absence d'un quatrième groupe qui aurait représenté les inactifs. Une explication possible est la faible proportion de ces sujets dans la base. En effet lors de la classification manuelle on n'avait étiqueté "que" 270 sujets comme inactifs.

La question que je me suis ensuite posée est celle de l'allure qu'aurait ma classification si elle était effectuée par un algorithme. De plus, il serait intéressant d'analyser la classification des individus que je n'ai pas su classer. Pour cela j'ai effectué une Classification Ascendante Hiérarchique avec le package FactomineR de R en choisissant la distance du chi-deux et le critère d'agrégation de Ward.





Le graphe d'inertie semble indiquer que le nombre idéal de clusters est 4. On peut voir sur le dendrogramme que c'est effectivement le nombre de clusters retenu ce qui est bon signe puisque l'idée était de repartir tous les sujets en 4 groupes : travailleurs, retraités, chômeurs et inactifs.



Pour pouvoir interpréter les cluster obtenus par la CAH il est indispensable de comprendre quelles sont les modalités qui contribuent le plus à chaque cluster.
Pour cela on cherche à déterminer si la forte présence d'une modalité dans un Cluster est due au hasard. Dans le cas contraire cela signifie que la modalité est représentative du Cluster.

On effectue le test suivant pour chaque modalité dans chaque cluster : H_0 : La proportion de la modalité m dans le cluster c est due au hasard H_1 : La proportion de la modalité m est anormalement élevée/basse dans le cluster c.

Ainsi sous H_0 on a : $\frac{n_{mc}}{n_c} = \frac{n_m}{n}$ où :
n_{mc} est le nombre de sujet du cluster c présentant la modalité m,
n_c est le nombre de sujets dans le cluster c,
n_m est le nombre de sujets présentant la modalité m,
n est le nombre total d'individus.

Sous H_0 : $n_{mc} = \frac{n_m.n_c}{n}$
Or le rapport $\frac{n_m.n_c}{n}$ suit une loi hypergéométrique.
Soit N_{mc} la variable aléatoire représentant l nombre d'individus de la modalité m dans le Cluster c.

$$N_{mc} \underset{H0}{\sim} \mathcal{H}(n, n_c, \frac{n_m}{m}) \tag{1}$$

La fonction catdes() du package factoMineR nous résume les résultats de ce test pour chaque modalité dans chacun des clusters en classant les modalité dans l'ordre décroissant de leur contribution au cluster. Seuls les test significatifs sont affichés.

Dans les résultats affichés juste aprés, la première colonne correspond au nom de la modalité, la deuxième calcule le rapport $\frac{n_{mc}}{n_m}$, la troisième le rapport $\frac{n_m}{n_c}$, la quatrième le rapport $\frac{n_{mc}}{n}$.
La cinquième colonne est le calcul de la p-valeur :

$$\mathbb{P}_{H0}(N_{mc} \geqslant n_{mc,obs}) = \mathbb{P}_{\mathcal{H}(n,n_c,\frac{n_m}{n})}(N_{mc} \geqslant n_{mc,obs}) \tag{2}$$

Enfin, la dernière colonne est la valeur de la statistique de test calculée. Lorsqu'elle est positive la modalité à laquelle elle est associée est sur représentée dans le Cluster et lorsqu'elle est négative le modalité en question est sous représentée dans le Cluster.

Par souci de lisibilité nous n'afficheront que les modalités surreprésentée dans chaque cluster :

Cluster 1	Cla/Mod	Mod/Cla	Global	p.value	v.test
Adm12a=NNNN	87.25	99.91	61.73	0	Inf
Adm11=0	92.41	99.29	57.92	0	Inf
Adm10=36	99.88	46.21	24.94	0	Inf
Adm10=54	99.62	14.41	7.80	3.80e-217	31.45
Adm12=NNNNNN	57.94	97.60	90.82	7.30e-155	26.51
Adm10=51	100	9.03	4.87	1.12e-137	24.98
Adm10=32	99.78	8.19	4.42	3.78e-122	23.50
Adm10=61	100	6.01	3.24	9.09e-91	20.20
Adm10=56	100	3.86	2.08	3.58e-58	16.08
Adm10=66	100	3.59	1.94	4.94e-54	15.48
Adm10=47	99.28	2.53	1.37	2.75e-36	12.58
Adm10=48	100	2.20	1.18	3.43e-33	12.00
Adm10=55	100	1.41	0.76	1.70e-21	9.52
Adm12=6NNNNN	80.52	2.27	1.52	4.19e-12	6.93
Adm10=41	100	0.77	0.41	4.99e-12	6.91
Adm10=46	100	0.51	0.28	2.97e-08	5.54

Les modalités de la variable Adm10 sur représentées dans le **Cluster 1** correspondent toutes à des codes travail de sujet qui travaillent. La modalité la plus présente dans ce Cluster est Adm12a = NNNN, c'est à dire les sujets n'ayant coché aucune modalité de cette variable. Cela fait sens puisqu'aucune des modalités possibles ne traitaient d'une activité professionnelle. Enfin les modalités de la variable Adm12 représentées dans le Cluster 1 sont NNNNNN c'est à dire ceux n'ayant rien coché et 6NNNNN : en contrat emplois-solidarité / intérim / CDD.

Cluster 2	Cla/Mod	Mod/Cla	Global	p.value	v.test
Adm12=N5NNNN	99.74	45.49	3.79	0	Inf
Adm10=95	98.52	71.14	6.00	0	Inf
Adm10=93	100	17.93	1.49	1.42e-169	27.76
Adm12=N5N3NN	100	17.81	1.48	2.05e-168	27.66
Adm12a=NNNN	12.59	93.47	61.73	1.89e-108	22.12
Adm11=1	37.58	28.38	6.28	1.39e-103	21.61
Adm12=NNN3NN	90.65	11.52	1.06	1.38e-94	20.63
Adm12=NN4NNN	100	7.13	0.59	2.13e-66	17.21
Adm10=96	82.95	8.6698337	0.86887836	7.223367e-65	17.007518
Adm12=NN43NN	100	4.9881235	0.41469194	1.646295e-46	14.319786
Adm12a=NNN7	55.56	5.9382423	0.88862559	6.429777e-31	11.561864
Adm11=2	26.68	13.1828979	4.10742496	3.196727e-30	11.423349
Adm11=3	24.17	14.6080760	5.02567141	5.036790e-29	11.181263
Adm10=94	100	1.3064133	0.10860979	1.234652e-12	7.101438
Adm12=65NNNN	78.57	1.3064133	0.13823065	3.632761e-10	6.269042
Adm12=6NN3NN	70	0.8313539	0.09873618	2.657340e-06	4.695670
Adm11=4	10.34	32.5415677	26.15521327	1.612450e-05	4.312738
Adm12=65N3NN	100	0.4750594	0.03949447	4.745827e-05	4.067804
Adm12=6NNNNN	17.53	3.2066508	1.52053712	2.141441e-04	3.701718
Adm12=6N43NN	100	0.3562945	0.02962085	5.727234e-04	3.444213
Adm12a=JNNN	66.67	0.2375297	0.02962085	2.013941e-02	2.323740

Dans le **Cluster 2**, les codes travail présents sont 94/93/94 : ils correspondent aux sujets chômeurs. De plus la modalité la pus représentée dans ce cluster est Adm12 = N5NNNN : sujes chômeurs depuis plus de 6 mois. Toutes les autres modalités de Adm12 résentent dans le Clusteur concernent également des chômeurs. Concernant la variable Adm11, les modalités présentes sont 1,2,3,4 c'est à dire les sujets ne travaillant plus depuis au moins moins d'un an. On remarque aussi que certains sujets en formation professionnelle (Adm12a = JNNN) sont inclus dans le Cluster.

Cluster 3	Cla/Mod	Mod/Cla	Global	p.value	v.test
Adm12a=N9NN	99.41348974	81.4903846	3.36690363	0.000000e+00	Inf
Adm10=85	98.25174825	67.5480769	2.82385466	0.000000e+00	Inf
Adm10=86	85.00000000	28.6057692	1.38230648	3.761709e-149	26.010557
Adm11=5	94.23076923	11.7788462	0.51342812	1.364674e-65	17.104886
Adm11=4	8.72027180	55.5288462	26.15521327	2.065844e-38	12.959875
Adm12a=NNN7	36.66666667	7.9326923	0.88862559	2.674663e-23	9.944143
Adm12=NNNNNN	4.46836269	98.7980769	90.81753555	2.729090e-12	6.991013
Adm12a=N9N7	100.00000000	1.6826923	0.06911532	1.878666e-10	6.370945
Adm12a=N98N	66.66666667	1.4423077	0.08886256	3.567252e-07	5.090710

Dans le **Cluster 3** la modalité la plus significativement représentée est Adm12a == N9NN : les personnes au foyer. On trouve également des sujet se déclarant en pré retraite (Adm12a = NNN7/N9N7 ou à la retraite(Adm12a = N98N). Les seules modalités de la variables Adm10 dans ce cluster sont 85/86 ce qui semble indiquer que le cluster correspind à celui des inactifs. Enfn on

remarque que plus de 94 % des personnes declarant n'avaoir jamais travaillé sont dans e groupe (Adm11 = 5).

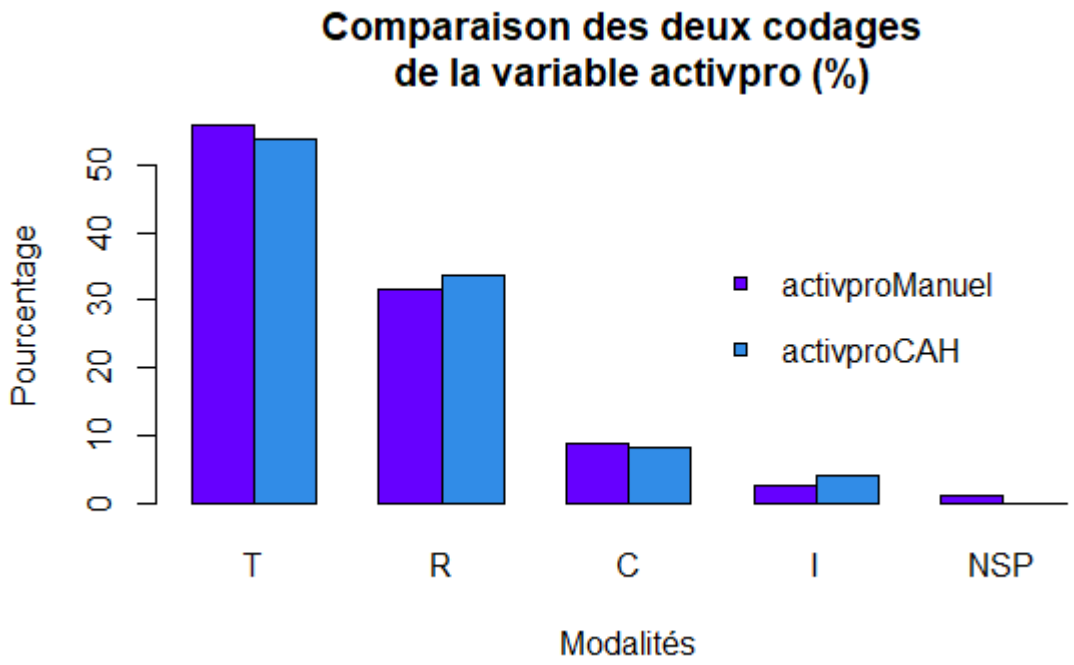
Cluster 4	Cla/Mod	Mod/Cla	Global	p.value	v.test
Adm12a=NN8N	99.09462617	99.50146628	33.80726698	0.000000e+00	Inf
Adm11=4	79.84144960	62.02346041	26.15521327	0.000000e+00	Inf
Adm10=77	98.61687414	62.72727273	21.41587678	0.000000e+00	Inf
Adm10=74	99.57301452	34.19354839	11.56200632	0.000000e+00	Inf
Adm12=NNNNNN	36.98630137	99.76539589	90.81753555	1.996276e-159	26.903543
Adm11=3	72.88801572	10.87976540	5.02567141	7.490306e-77	18.554571
Adm11=2	69.23076923	8.44574780	4.10742496	1.427574e-51	15.108309
Adm11=1	57.70440252	10.76246334	6.27962085	1.321276e-37	12.816748
Adm10=78	100.00000000	2.05278592	0.69115324	5.014862e-34	12.160988
Adm10=75	100.00000000	0.79178886	0.26658768	1.604994e-13	7.378154

Enfin pour le **Cluster 4** on s'attends à trouver des sujets retraités. En effet, la modalité la plus significativement représentée est Adm12a = NN8N : sujets retraités. Les codes travail intervenant (Adm10 sont 77/74/78/75 : uniquement des codes de la forme "Ancien *profession*". En outre on ne trouve dans ce cluster aucun individu déclarant travailler (Adm11 = 0 ou Adm12 = 6NNNNN).

On peut alors conclure que **le Cluster 1 correspond aux travailleurs**, **le Cluster 2 aux chômeurs**, **le Cluster 3 aux inactifs** et **le Cluster 4 aux retraités**.
On remarque alors que l'emplacement des Clusters dans le plan factoriel (cf graphe) correspond à l'emplacement des variables dans ce même plan et que nous avons commenté plus haut.

3.1.3 Comparaisons des deux codages

Étudions las différences entre la classification manuelle et la Classification Ascendante Hiérarchique.
Tout d'abord on classe comme NA tous les sujets qui avaient une valeurs manquante pour au moins une des quatre variables ayant servies à la classification (n = 29).
On a alors :



	T	R	C	I	NSP	NA
activproManuelle	5666	3209	883	270	113	16
activproCAH	5460	3410	842	416	0	29
Différence	206	-201	41	-146	113	-13

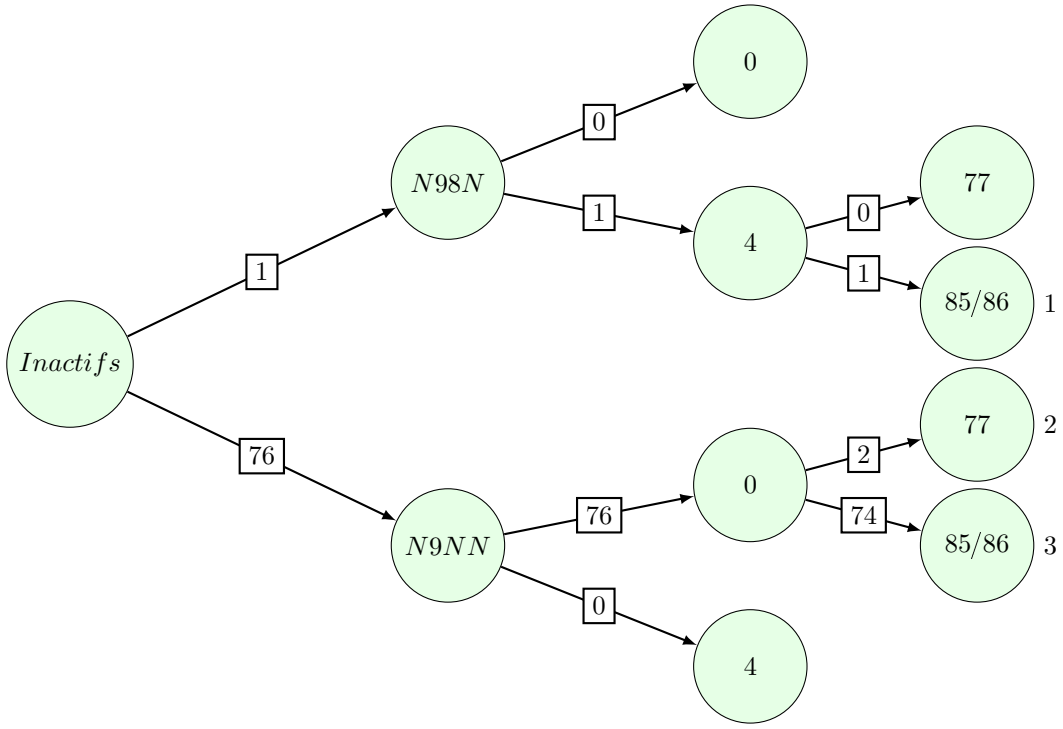
Il y a 509 sujets pour lesquels les deux classifications ne correspondent pas. Sur un total de 10157 cela représente 5,01 % des sujets.
On remarque que la classification Ascendante Hiérarchique a tendance à classer plus facilement les sujets en Retraités et en Inactifs et la classification manuelle en Travailleurs. Une raison simple à cette différence est sans doute la volonté d'être le plus restrictifs possible quant à la classification manuelle des retraités.
Le nombre de valeurs manquantes n'est pas le même pour les deux classifications pour une raison simple : les 16 NA de la classification manuelle sont des sujets pour lesquels TOUTES les variables nécessaires à la classification (Adm12a, Adm11, Adm12 et Am10) étaient manquantes alors que j'ai dû, pour activproCAH, classer en NA les sujets pour lesquels AU MOINS UNE des 4 variables étaient manquantes afin de pouvoir réaliser l'ACM.
Il peut également être intéressant d'étudier la manière dont les sujets classés NSP (n = 113) par la classification manuelle ont été classés par la CAH :

R	I	C	T	Total
5	77	0	31	113

Sur quels critères la CAH s'est-elle basée pour classer ces 113 sujets ? Pour le comprendre on s'interesse aux trois modalités dans lesquels elle a répartis les NSP (I,R,T) en construisant des Tableaux de Burt et des arbres de décision.
Par souci de lisibilité j'ai choisis de n'inclure dans les tableaux de Burt que les modalités contenant un nombre de sujets non nul.
Les cases en rouge correspondent simplement aux tableaux de contingences usuels (croisement des variables deux à deux).
Le tableaux est évidement "symétrique", j'ai choisi de n'en remplir que la partie inférieure.
Concernant les arbres , j'ai reproduit les étapes de classification de la CAH (en fixant l'ordre des 4 variables comme étant celui que j'ai moi même appliqué) pour comprendre quelles étaient les modalités discriminant le mieux les classes R,I,T.

INACTIFS

	N98N	N9NN	0	4	NNN3NN	NNNNN	77	85	86
N98N	1	∅							
N9NN	∅	76							
0	0	76	76	∅					
4	1	0	∅	1					
NNN3NN	1	0	0	1	1	∅			
NNNNNN	0	76	76	0	∅	76			
77	0	2	2	0	0	2	2	∅	∅
85/86	1	74	74	1	1	74	∅	58	17



1 : Le sujet déclare être au foyer et à la retraite (Adm12a = "N98N), ne plus tavailler depuis au moins 3ans(Adm11 = 4) et a un code travail correspondant à un inactif non retraité (Adm10 + 85/86). Le classer en Inactifs a donc du sens.

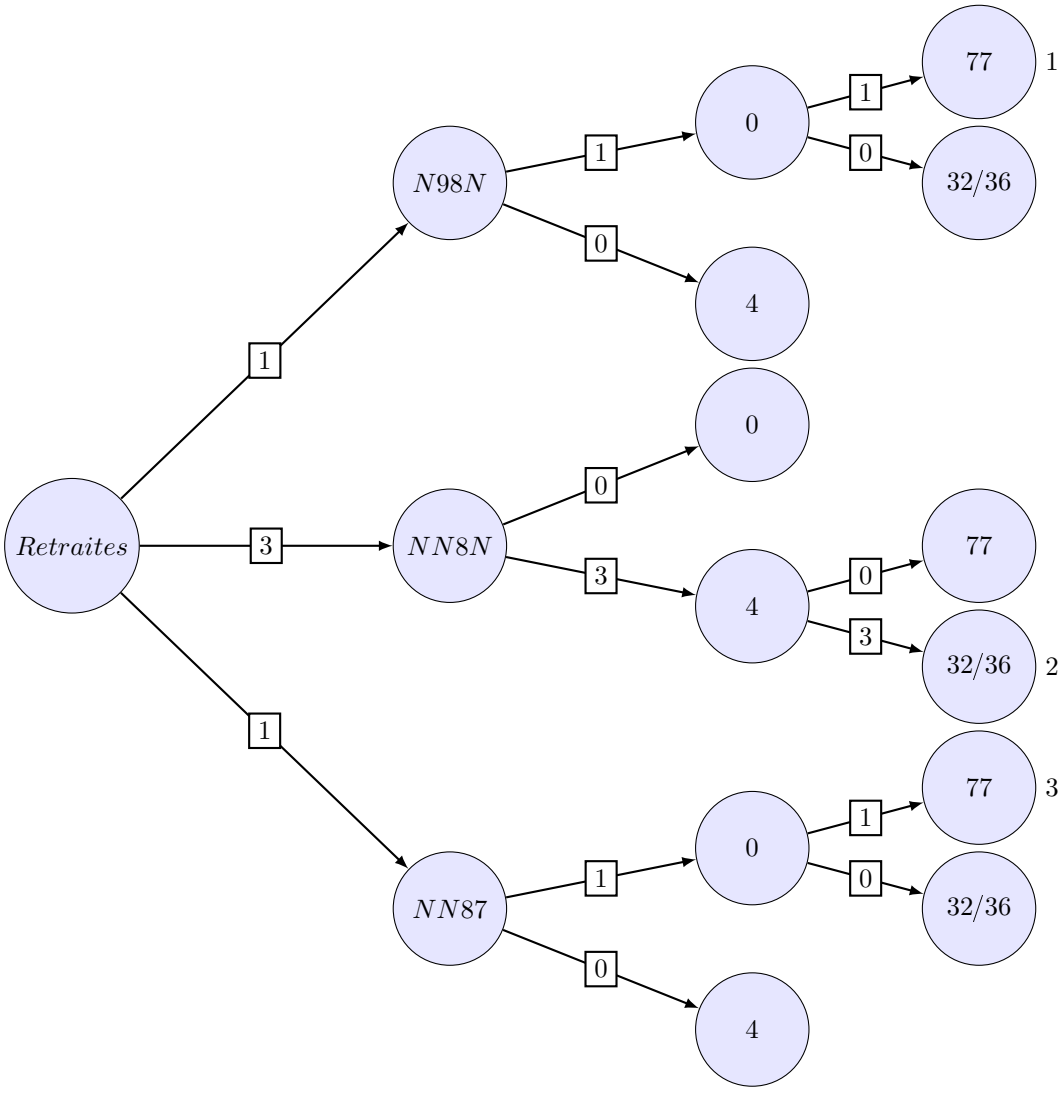
2 : Les 2 sujets déclarent être au foyer (Adm12a = N9NN), travailler (Adm11 = 0) et a un code travail correspondant à un retraité (Adm10 = 77 : Ancien employé). Toutes les variables sont contradictoires. Classer ces suets en Inactifs n'est pas pertinent.

3 : Les 74 sujets presenent les même caractérisqtiques que ceux expliqués précédement à l'exception de la variable Adm10 qui indique que les sujets sont nactifs non retraités. De même, les classer en inactifs n'est pas pertinent.

De manière général on remarque que declarer être "au foyer" est trés discriminant pour la classification en tant qu'nactifs surtout si cette modalité est couplée avec un code travail 85 ou 86.

RETRAITES

	N98N	NN87	NN8N	0	4	NNNNNN	32/36	77
N98N	1	∅	∅					
NN87	∅	1	∅					
NN8N	∅	∅	3					
0	1	1	0	2	∅			
4	0	0	3	∅	3			
NNNNNN	1	1	3	2	3	5		
32/36	0	0	3	0 3	3	1	2	∅
77	1	1	0	2	0	2	∅	2

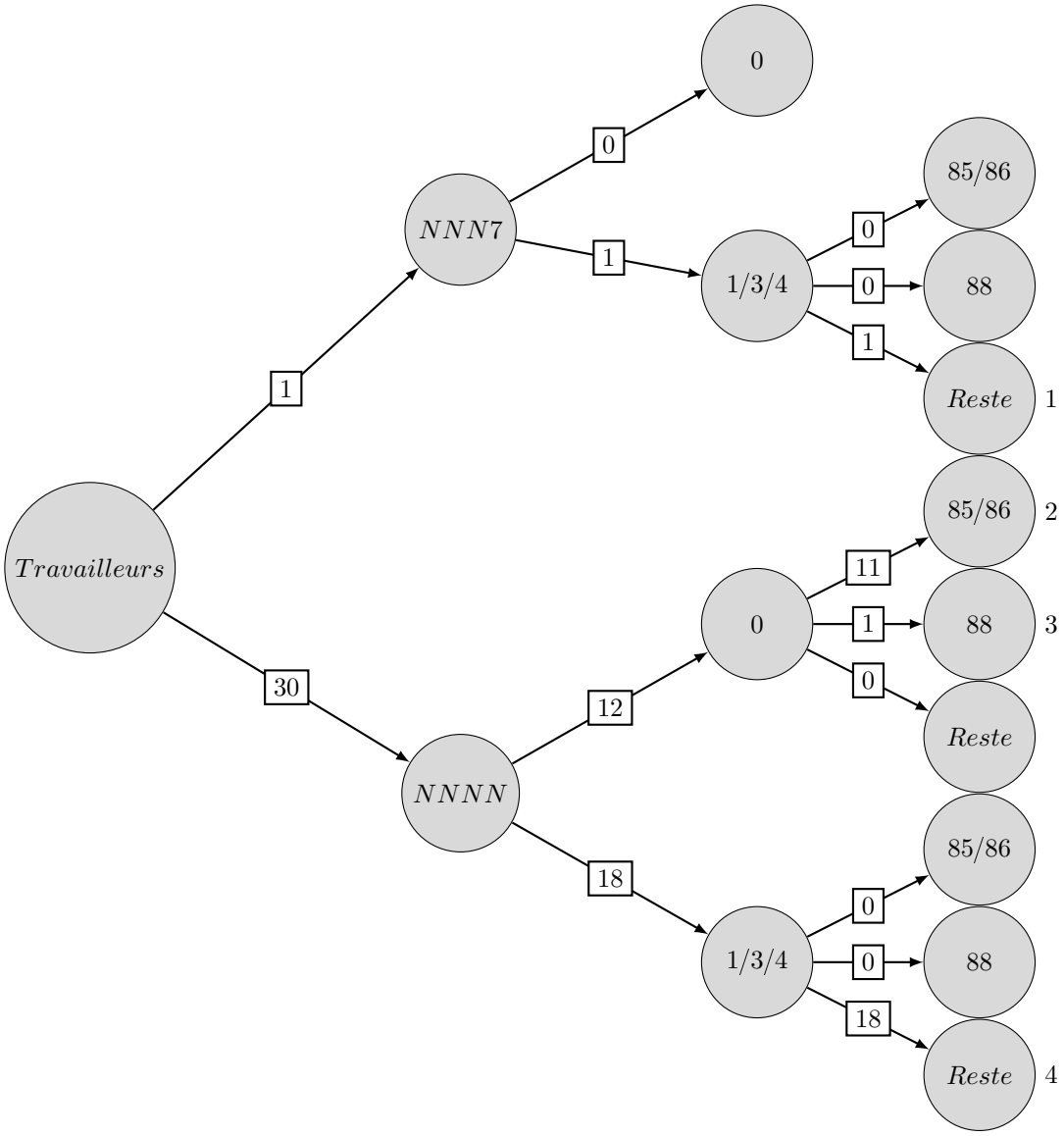


1 : Le sujet sur cette branche déclare être à la retraite et au foyer (N98N) avec un code de travail

correspondant bien à un retraité (77). Cependant il déclare également travailler (0). La CAH le classe en retraité malgré tout mais ce choix n'est pas judicieux.
2 : 3 sujets se déclarent à la retraite (NN8N), ne travaillant plus (Adm11 = 3) et ont un code travail qui correspond à des travailleurs (Adm10 = 32/36). Leur classification en Retraité n'est pas aberrante. Il est possible que ces sujets n'aient pas lu la liste des codes travail jusqu'à la fin (la où ceux concernant les retraités se trouvent) e qui expliquerait qu'ils aient indiqué par ce code qu'ils travaillent (32/36).
3 : 1 sujet déclare être retraité et pré retraité (NN87), travailler (Adm11 = 0) et a un code travaille de 77. Il y a une certaine forme de cohérence à l'ensemble des choix de ce sujet mais on ne peux pas le classer en tant que Retraité à cause de Adm11 = 0.

TRAVAILLEURS

	NNN7	NNNN	0	1	3	4	NNNNNN	32/36/47/48/ 51/54/55/56	85/86	88
NNN7	1	∅								
NNNN	∅	30								
0	0	12	12	∅	∅	∅				
1	0	1	∅	1	∅	∅				
3	0	1	∅	∅	1	∅				
4	1	16	∅	∅	∅	17				
NNNNNN	1	30	12	1	1	171	31			
32/36/47/48/ 51/54/55/56	1	18	0	1	1	17	19	19	∅	∅
85/86	0	11	11	0	0	0	11	∅	11	∅
88	0	1	1	0	0	0	1	∅	∅	1



1 : 1 sujet se déclare pré retraité (NNN7), travaille (Adl11 = 0) et a un code travail correspondant à quelqu'un ayant encore une activité professionnelle. On peut accepter la classification en travailleur de ce sujet par la CAH.
2 : 11 sujets ne cochent aucune case pour Adm12a (soit ils ont oublié cette question soit aucun des choix ne les satisfont ce qui correspondrait à des personnes qui travaillent puisque ce choix n'est pas proposé dans Adm12a).En outre ils déclarent travailler (Adm11 = 0),mais leur code travail est incohérent :85/86 correpond à des inactifs.
3 : 1 sujet travaille et a un code travail qui vaut 88. Ce code n'est pas présent dans la liste, il s'agit probablement d'une faute de frappe.

4 : Enfin, 18 sujets déclarent ne plus travailler (Adm11 = 1/3/4)mais ont un code travail indiquant qu'ils ont encore une activité professionnelle. Ces sujets ne peuvent pas être classés comme des travailleurs.

A présent que les deux classifications sont établies et clairement expliquées, il y a lieu de se demmander si utiliser l'une ou l'autre dans un futur modèle changera les résultats.
Autrement dit construisons un test permettant de vérifier si les deux distributions de la variable activpro sont similaires.

Test du Khi deux d'ajustement Il s'agit de vérifier si les écarts d'effectifs entre activproManuelle et activproCAH sont "trop" importants pour chacune des classes.
2tant donné que la CAH classe TOUS les individus, la classe NSP de activproCAH est vide. Or le test du chi deux d'ajustement n'est appicable que si toutes les classes contiennent au moins 5 sujets.
Puisque nous nous inreressont particulièrement aux TRavailleurs, retraités et Chômeurs, on rassemble les inactifs, les NSP et les NA dans une classe "AUTRE". Ona alors :

	T	R	C	Autre
activproManuelle	5666	3209	883	299
activproCAH	5460	3410	842	445

Notations :
k : nombre de modalités.
ϕ_i : probabilité qu'un individus soit dans la classe i de la distribution observée.
ϕ_{hi} : probabilité qu'un individu soit dans la classe i de la distribution théorique.
O_i : nombre d'individus observés dans la modalité i.
A_i : nombre d'individus attendus dans la modalité i.
On a : $\forall i \in [\![1, k]\!], A_i = n\phi_{hi}$

Hypothéses :

$$\begin{aligned} H_0 &: \forall i \in [\![1, k]\!], \phi_i = \phi_{hi} \\ H_1 &: \exists i \in [\![1, k]\!] tq \phi_i \neq \phi_{hi} \end{aligned} \tag{3}$$

Statistique de test :

$$Q = \sum_{i=1}^{k} \frac{(O_i - A_i)^2}{A_i} \tag{4}$$

D'aprés le Théorème de pearson :

$$\begin{aligned} &Si\ \forall i \in [\![1, k]\!],\ A_i \geqslant 5 : \\ &Q \underset{H_0}{\hookrightarrow} \chi^2(k-1),\ n \to +\infty \end{aligned} \tag{5}$$

Application numérique :
$Q_{obs} = 69.52$ Au niveau de confiance 95 % le χ^2 à 3 degrés de liberté vaux 12.84.
On rejette donc H_0
Puisqu'on ne peut pas conclure que les deux variable on une distribution similaire nous feront

tourner les modèles de régression avec les deux variables et nous interpréterons les résultats en conséquence.

3.2 Variable d'intérêt : Self Rated Health (SRH)

La santé perçue est décrit dans la litterature comme un bon indicateur de la mortalité. Il s'agit d'une mesure subjective puisque le sujet s'auto-évalue : il est son propre témoin.
Dans le questionnaire d'inclusion il est demandé au sujets d'indiquer par une note comprise entre 0 et 10 leur état de santé tel qu'il le ressente : 0 pour mauvais et 10 pour excellent.

	0	1	2	3	4	5	6	7	8	9	10	NA
Effectfs	11	166	30	85	194	901	828	2175	3309	1698	734	26
Effectifs cumulés	11	177	207	292	486	1387	2215	4390	7699	9397	10131	10157

On remarque que globalement les sujets de la cohorte EPP3 s'estiment en bonne santé.
Du fait du grand nombre de classe de cette variable, il peut être interessante de la binariser. On choisi de couper la variable à la médiane : ainsi les niveaux de santé perçue entre 0 et 7 inclus sont considérés comme mauvais et de 8 à 10 comme bon :

0	1	NA
4390	5741	26

3.3 Variables d'ajustement et facteurs de confusion

3.4 Gestion des données manquantes

4 Statistiques

4.1 Analyses univariées

4.1.1 Variables qualitatives

4.1.2 Variables quantitatives

4.2 Analyses multivariées

4.2.1 Régression logistique

4.2.2 Régression linéaire

4.2.3 Régression polytomique

4.3 Analyse en sous-groupes

4.3.1 Dépression

4.3.2 Pathologies lourdes

4.3.3 Sexe

4.3.4 Âge

5 Perspectives

...

5.1 Le monde professionnel

5.2 Nouveaux outils

5.3 Les données : théorie vs pratique

Conclusion

6 Annexes

7 Bibliographie