Avoir un point de vue graphique sur les données de questionnaires : le logiciel libre *pointG*.

Stéphane Champely (1), Sylvain Férez (2), Brice Lefèvre (1) et Julie Thomas(2)
(1) *CRIS*, Université Lyon 1, France. (2) *Santésih*, Université Montpellier 1, France.
Correspondant : champely@univ-lyon1.fr

3 décembre 2011

Table des matières

«Cette procédure d'enquête est souvent coûteuse en temps et en moyens mobilisés mais le résultat est souvent décevant car ceux qui font des enquêtes ne disposent pas en général de méthodes pour les explorer en profondeur et se contentent donc de résultats superficiels.» (Philippe Cibois, 2009, Que sais-je ?)

1 Introduction

L'analyse statistique de questionnaires est une activité délicate pour le sociologue. La difficulté d'élaborer des stratégies efficaces d'étude statistique face à un jeu de données de grande taille, renforcée par une formation en statistique souvent sommaire et peu adaptée à ce contexte, à part le test du chi-carré d'indépendance, et la nécessité d'employer un logiciel complexe, payant et pas toujours documenté correctement pour ce type d'objectifs, expliquent pour partie certaines tendances de la sociologie française à parfois négliger les approches quantitatives.

Si le logiciel *TriDeux* de Philippe Cibois est une première solution à ces problèmes, nous en proposons une autre, plus visuelle, avec *pointG*, également libre de distribution, utilisable sur tout système d'exploitation et documentée sur un exemple « représentatif ». Ce logiciel, créé initialement pour les besoins d'enseignements de cours de statistique ou de méthodologie, vise à une approche simple mais cependant puissante du traitement de questionnaires, et assez spécifique de la sociologie française.

Cette simplicité découle *primo* de sa convivialité d'utilisation, par l'emploi de menus déroulants qui suivent la logique d'analyse d'un questionnaire : nettoyage de données, analyses univariées (de la signalétique en particulier), analyses bivariées puis modèles linéaires explicatifs et analyses factoriels exploratoires. *Secundo*, du choix de traduire d'abord les résultats sous forme graphique, ce qui permet une interprétation simplifiée et ouverte (un graphique peut toujours nous surprendre et générer des pistes inattendues). *Tertio*, de l'implémentation de deux concepts d'analyse : la « stratégie du saucisson » et « l'hétéro-statistique ». La stratégie du saucisson consiste à découper une « tranche de données » - un ensemble de variables faisant sens sociologiquement parlant - et de l'analyser soit de façon univariée simultanée, soit bivariée avec une variable externe soit bivariée avec elle-même. Le procédé permet à la fois de concevoir des stratégies d'analyse, de raisonner de façon plus multivariée grâce aux graphiques et de gagner du temps. L'hétéro-statistique est le fait de ne plus se soucier de varier le type d'analyse en fonction de la nature (numérique, ordonnée ou catégorielle) des variables concernées. Le logiciel détermine la méthode adaptée ou emploie des techniques sophistiquées qui tolèrent des types mixtes.

Malgré cette simplicité, le logiciel est conçu pour le traitement de questionnaires sociologiques. Il contient des procédures pour étudier les données manquantes, les échelles de type Likert, des graphes de relations, une cartographie en fonction des codes postaux et un lien internet direct avec quelques bases de données de l'INSEE. L'analyse de données « à la Française » est disponible : calcul du PEM (Cibois, 1993), analyses factorielles exploratoires avec projection

de variables supplémentaires.

PointG permet également de réaliser des analyses sophistiquées «sans le savoir» grâce à l'hétéro-statistique dans deux domaines. En ce qui concerne l'analyse factorielle, le choix de deux tranches - les variables actives (qui vont véritablement constituer l'analyse) et les variables passives (autrement appelées supplémentaires, qui vont l'enrichir) - permet de réaliser une analyse unique, dite de Hill et Smith (1976). De façon transparente pour l'utilisateur, cette analyse, selon la nature des variables en jeu, sera (1) une analyse en composantes principales (ACP, avec des variables toutes numériques), (2) une analyse des correspondances multiples (ACM, avec des variables toutes catégorielles) ou (3) une analyse véritablement mixte proche de l'analyse factorielle multiple (AFM, Escoffier & Pagès, 1994, voir aussi le package **R** *FactoMineR* de Husson et al, 2011). En ce qui concerne les modèles explicatifs, une fois choisies la variable à expliquer et la tranche de variables explicatives (avec éventuellement des interactions entre elles), le logiciel s'adapte à la variable à expliquer. Il utilise soit un modèle linéaire classique (variable numérique), un modèle logistique binomial (variable binaire), un modèle *proportional-odds* (variable ordonnée) ou un modèle logistique multinomial (variable purement catégorielle avec plus de deux catégories). Les tests de significativité sont identiques ainsi que l'interprétation graphique.

Les résultats issus de *pointG* sont essentiellement graphiques. C'est toujours la première analyse proposée. Les choix de présentation ont été optimisés vue la taille des jeux de données considérés (de 500 à 2000 sujets ?) selon les conseils de Cleveland (1993) et Tufte (2001). En ce qui concerne les résultats numériques, ils sont gardés à leur expression la plus simple et des méthodes de présentations originales sont proposées pour les analyses univariées et les tailles d'effet en relation bivariée.

Afin de montrer les aspects spécifiques de *pointG*, nous allons employer un jeu de données (inclus dans le logiciel) provenant d'une étude sur les pratiques sportives de personnes vivant avec le VIH. De cette lourde enquête, nous n'avons retenu que quelques variables provenant de la signalétique et d'un autre thème : le rapport au corps. Nous verrons sur la base d'une tranche de signalétique le traitement des données manquantes, l'analyse univariée et l'analyse bivariée. Ensuite, nous étudierons la variable (numérique) du poids de l'individu puis une autre (bianire) concernant le regard sur le corps afin de montrer l'unité de traitement des modélisations. Nous terminerons par diverses analyses factorielles concernant le rapport au corps. On trouvera en annexe les procédures d'installation du logiciel *pointG* et de son environnement.

2 Quelques considérations sur le jeu de données

L'enjeu de l'enquête est d'étudier l'accès aux activités physiques et/ou sportives (APS) des personnes vivant avec le VIH (PVVIH), en étudiant notamment l'effet du diagnostic ; l'effet de la socialisation au rôle/statut de PVVIH (comme « malade ») : des parcours de soins (en lien avec les services hospitaliers, les

médecins de ville, etc.), de la prise d'informations et de « l'éducation thérapeutique », de la fréquentation et de l'implication ou non dans les associations, etc. ; l'effet du rapport à son corps et au regard de l'autre (et inversement, l'effet du diagnostic sur cela).

Pour cela, hormis une cinquantaine d'entretiens non-directifs, nous disposons de 619 questionnaires. L'enjeu de l'échantillonnage a été de diversifier au maximum les profils de répondants (au regard de pratique des APS, mais aussi au regard de l'ancienneté du diagnostic et des modes de contamination, des modes de prise en charge du VIH, des situations sanitaires, professionnelles et sociales), pour tenter de montrer, par-delà les effets communs du diagnostic VIH sur la problématique de l'accès aux APS, l'hétérogénéité des ressources et des stratégies pour gérer cette problématique.

Le questionnaire comporte quatre parties : 1- Expérience en matière d'activités physiques et/ou de sport ; 2- Expérience du VIH, rapports au corps, traitements ; 3- Soin de soi et habitudes de vie ; 4- Informations générales (variables sociodémographiques, statut social, gestion globale de l'information sur le VIH auprès des proches, variables socio-politiques et religieuses).

3 Nettoyage et données manquantes

Une fois le logiciel *pointG* installé et lancé (voir annexe) ; l'écran apparait comme sur la Figure 1 avec, en particulier, un menu déroulant `Données` qui va être utile pour l'importation de données et surtout un menu `pointG`, ici déroulé afin d'avoir accès au traitement de données manquantes.

Le jeu de données VIH est disponible sous deux formes, soit par téléchargement sous forme de fichier Excel, soit intégré dans le logiciel *pointG*. Le menu déroulant `Données` du `R-Commander` offre plusieurs façons d'importer des données sous différents formats dont Excel. En l'espèce, nous utiliserons le mode intégré en allant dans le menu déroulant `Données`, puis `Données dans les packages`, ensuite `Lire des données depuis un package attaché`. Une fenêtre s'ouvre, double-cliquer à gauche sur `RcmdrPlugin.pointG` : des jeux de données apparaissent à droite. Double-cliquer sur `VIH` puis `OK`.

Au dessous du menu `Données` du *R-Commander*, on voit alors une icône rouge *Rcmdr* qui indique que le jeu de données actif pour les analyses statistiques est dorénavant VIH. Si vous souhaitez jeter un coup d'oeil aux données (toujours une bonne idée), cliquer sur le bouton `Visualiser` du `R-commander`.

La première option d'analyse `Que sais-je ?` est importante, elle permet de définir la nature présente des variables. Il est essentiel de la valider ou de procéder à des transformations/recodages [1] à cette étape puisque nombre d'opérations vont tenir compte de cette nature dans leur déroulement. En particulier, il est possible d'hésiter lorsque les données sont ordonnées. On a dans ce cas la possibilité de créer deux variables, l'une catégorielle ordonnée et l'autre numérique (de 1 à 6 par exemple). L'option donne également des premiers renseignements

1. Une prochaine note détaillera comment réaliser ces opérations à l'aide du R-Commander

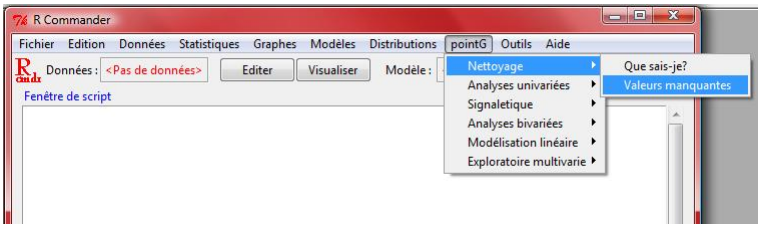

FIGURE 1 – L'interface du R-Commander avec le menu déroulant spécifique à *pointG* (ici choix de l'option `Nettoyage` et de la sous-option `Valeurs manquantes`.)

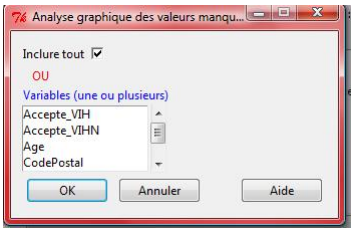

FIGURE 2 – Menu contextuel permettant la sélection d'une tranche (sous-option : `Valeurs manquantes`)

sur le nombre de valeurs manquantes et les valeurs minimum et maximum pour chaque variable.

Pour commencer l'analyse avec les données manquantes, aller dans le menu déroulant `pointG`, puis `Nettoyage` et enfin `Valeurs manquantes` (Figure 1). Un menu contextuel apparaît qui permet de choisir une tranche de données (Figure 2). Il y a alors dans ce cas deux possibilités, on peut ne rien choisir et cliquer sur le bouton `OK`, le jeu de données alors sélectionné est entièrement traité, ou plus stratégiquement et raisonnablement, ne retenir qu'une tranche de données. En l'espèce, cette tranche contient des variables de signalétique à savoir, l'âge, le sexe, le poids, la taille, le niveau de diplôme, le niveau de revenu et la profession des sondés.

On pourra constater dans le panneau de gauche [2] de la Figure 3 (qui provient de fonctions du package `VIM` de Temple et al., 2011) qu'il y a environ 5% de données manquantes dans les variables sélectionnées, ce qui n'est pas négligeable concernant la signalétique, mais surtout il y en a près de 40% en ce qui concerne la profession. C'est très important au regard d'une population de 20 ans et plus : nombre des personnes interrogées vivent dans des conditions précaires.

2. Nous ne nous soucierons pas dans cette introduction du panneau de droite le la figure.

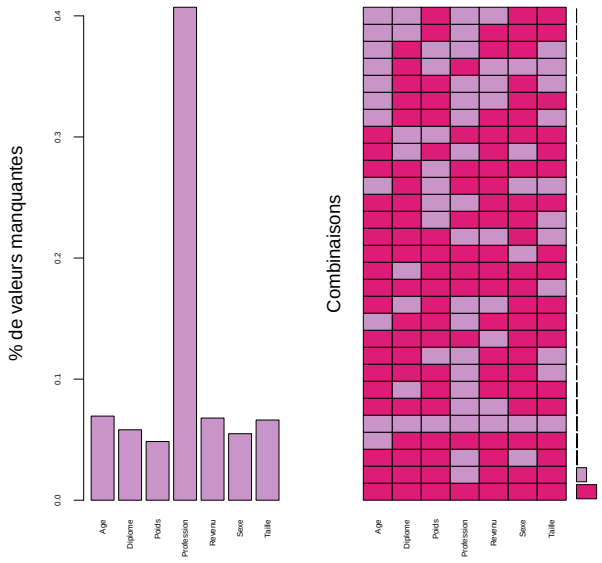

FIGURE 3 – Graphe de données manquantes

4 Analyses univariées

Nous allons alors passer à l'option `Analyses univariées`. La sous-option `Graphiques à plat`, appliquée à l'ensemble précédent de variables, renvoie la Figure 4. On peut voir que pour une variable numérique (l'âge) un histogramme est tracé. C'est le cas sauf si le nombre de valeurs prises est inférieur ou égal à 11 [3]. Un graphique en bâtons est alors préféré pour ce type de mesure qu'on appelle *discrète*. Pour les variables ordonnées, un graphe en barres est présenté. Pour les variables catégorielles, avec cinq catégories et moins c'est un camembert qui est produit, sinon c'est un graphique en points ordonnées.

Si les variables âge, taille et poids ne présentent pas de configurations particulières, on voit concernant le sexe, qu'il y a une petite surprise : deux catégories particulières (transexuel et intersexe) viennent troubler les repères habituels. Cette étape d'analyse graphique univariée est importante afin de prendre de nouvelles décisions de recodage : regrouper des catégories, corriger des données manifestement en erreur. En l'espèce, une nouvelle variable a été créée, `Sexe2`, qui présente comme manquantes les données des 13 individus concernés afin de simplifier les traitements par la suite.

On observe également, pour le graphique ordonné des professions, que ce qui se présente en premier est l'effectif le plus important (le nombre de cadres) et pas comme trop souvent le seul résultat de l'ordre alphabétique. Il y a bien

3. supposé être le cas extrême d'une échelle de Likert allant de 0 à 10

longtemps que les agriculteurs ne sont plus en premier que dans les questionnaires. On retrouvera ce parti-pris dans la présentation des statistiques. On constate également un « trou» en ce qui concerne les revenus. Ceci est lié à des problèmes d'échantillonnage, les cadres masculins homosexuels parisiens étant surreprésentés.

Une fois les opérations de transformations accomplies, il est possible de calculer des statistiques. Deux options s'offrent à nous. La première : `Bref.` donne simplement une valeur de localisation pour chaque variable : la catégorie la plus présente pour une variable catégorielle ou ordonnée et la moyenne pour une variable numérique. Cela permet de créer des tableaux résumés efficaces.

```
Age       Diplome              Profession        Revenu              Sexe       Taille
M:45.48   Secondaire (%): 32   Cadre... (%):34   800<R<1500 (%):29   H (%):67   M:172.1
```

Pour prolonger l'analyse et, en particulier, *se soucier de variabilité*, l'option `Statistiques à plat`, produit les statistiques classiques mais de façon peu classique. On constatera en effet des différences avec les résumés usuels de `R` : pour les variables numériques, l'écart-type est fourni, et pour les variables catégorielles, les catégories sont ordonnées de façon décroissante et des % sont présentés plutôt que des effectifs. En revanche, pour les variables ordinales, l'ordre est heureusement conservé.

```
 Age          Diplome          Profession     Revenu            Sexe    Taille
 Min.:20.00   Primaire  : 67   Cadre...:124   R<450      : 54   H:393   Min.:141.000
 Q1  :37.00   Secondaire:184   Employ  :114   450<R<800  :133   F:179   Q1  :166.000
 Q2  :45.00   Bac       : 99   Interm  : 64   800<R<1500 :168   T: 11   Q2  :171.000
 M   :45.48   Lic23     :115   Ouvri   : 34   1500<R<2000: 80   I:  2   M   :172.100
 Q3  :53.00   Master    :118   ArtCom  : 30   R>2000     :142   ?: 34   Q3  :178.000
 Max.:77.00   ?         : 36   Agric   :  1   ?          : 42           Max.:205.000
 S   :10.67                    ?       :252                             S   :  8.575
 ?   :43.00                                                             ?   : 41.000
```

Les variables de Likert sont généralement présentées en « batterie » dans les questionnaires. C'est aussi le cas dans l'enquête présentée avec les variables Corps_PasMoiN, Corps_MieuxN, Corps_TropGrosN, Corps_DeformeN, Corps_TropMaigreN et Corps_PasSeduisantN, qui décrivent sur une échelle de 1 à 5 la vision que les sondés ont de leur corps. Lorsque ces variables sont codées de façon numériques, on peut les analyser en une seule fois avec la sous-option spécifique `Graphiques pour échelles de Likert` (si elles sont en catégorielles, on peut employer la sous-option `Graphiques à plat` mais ils sont moins intéressants). On obtient la Figure 5. Il y a ici peu de différences sinon que la proposition Corps_PasSeduisantN présente une moyenne inférieure[4].

La dernière possibilité univariée avec le logiciel est de cartographier des données avec partir des codes postaux (ou de département). Pour l'instant, la seule option, basée sur le package `maps` (Becker et al., 2011) est de travailler au niveau national (celui de l'enquête). Il suffit, dans la sous-option

4. Nous y reviendrons en conclusion.

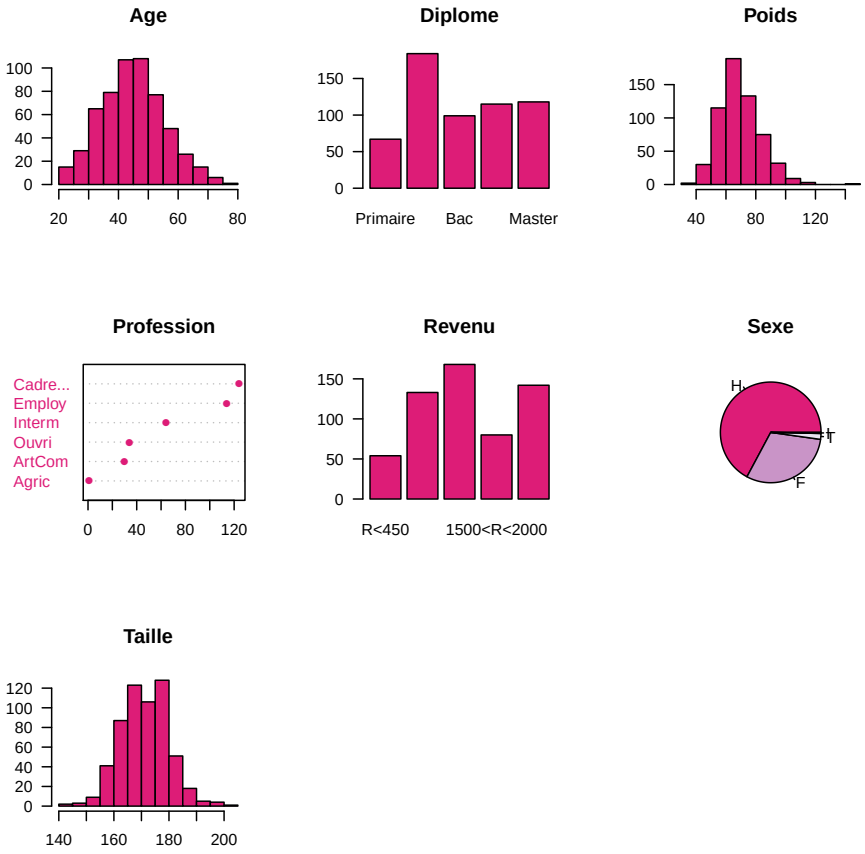

FIGURE 4 – Graphiques univariés d'une « tranche» de signalétique. Les variables numériques sont présentées en histogrammes ou graphes en bâtons (selon leur nombre de valeurs), les variables ordonnées en graphiques en barres et les variables catégorielles en camembert ou en graphiques en points ordonnées (selon leur nombre de catégories)

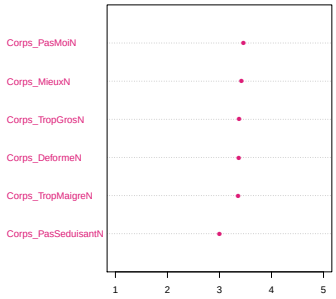

FIGURE 5 – Graphique des moyennes d'une batterie d'échelles de Likert, décrivant la vision de leur corps par les sondés.

`Cartographie chalandise par dép` de l'option `Signalétique`, de sélectionner la variable contenant les codes et de laisser l'option par défaut code de département plutôt que l'option code postal. On obtient la Figure 6 (les départements d'outremer pourtant importants dans ce sondage ne sont pas présentés [5]).

Dans l'option `Signalétique`, on trouvera également de quoi réaliser une pyramide des âges [6] (cf. Figure 7) et un menu contextuel (Figure 8) pour obtenir des données de cadrage. Il est souvent dit aux étudiants que les informations doivent être contextualisées, en alors cochant certaines options (*cf.* sous-option `Données de cadrage`), on est automatiquement dirigé vers des sites (souvent INSEE) donnant des informations sur la sociodémographie, la consommation des ménages, les pratiques sportives des Français.

5 Analyses bivariées

Puisque l'analyse d'une table de contingence (ou tableau à double entrée) est un des outils statistiques de base du sociologue, cette possibilité est aussi présentée dans l'option `Analyses bivariées` avec la sous-option `Table croisée`. Outre les résultats classiques (effectifs, profils et test du X^2), le PEM de Cibois (1993) est présenté ainsi qu'un carrousel de graphiques dont on trouvera un exemple de graphique d'association [7] en Figure 9 où diplôme et niveau de revenu sont croisés. On y retrouve sans surprise, mais avec un certain réconfort pour les étudiants, une relation très forte.

Si ce couplage simple a son intérêt, c'est de travailler sur une tranche de variables (stratégie du saucisson) qui est préféré dans *pointG*. Pour une tranche donnée, quelle que soit la nature des variables on peut étudier l'ensemble des

5. Seule la France métropolitaine et la Corse sont cartographiées. On limitera donc les codes de départements de 1 à 95.

6. programme basé sur la fonction `hisbackback` du package *Hmisc* (Harrel, 2011)

7. programme utilisant la fonction `strucplot` du package *vcd* de Meyer et al. (2006)

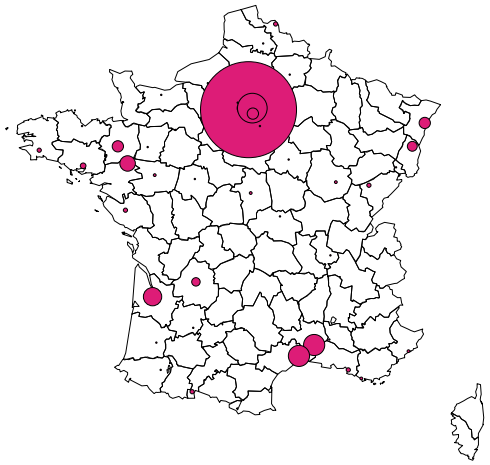

FIGURE 6 – Cartographie des effectifs de répondants

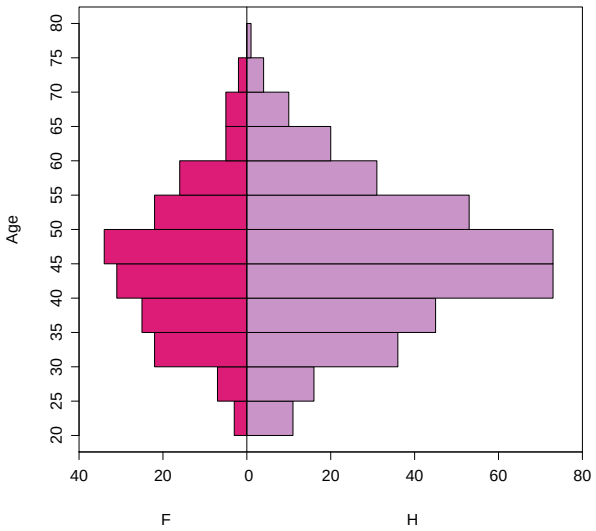

FIGURE 7 – Pyramide des âges des sondés de l'enquête VIH

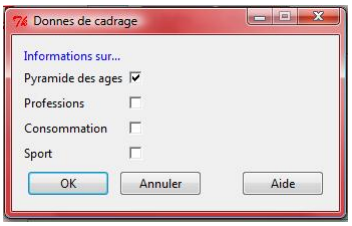

FIGURE 8 – Menu contextuel de la sous-option `Données de cadrage` permettant le lien avec des bases de données

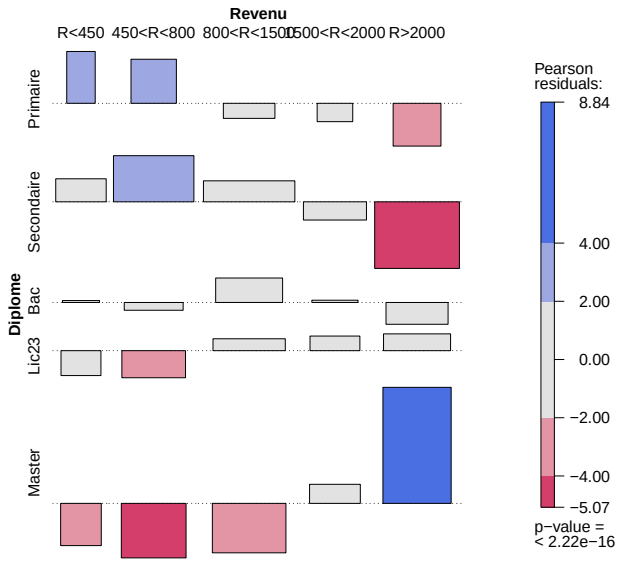

FIGURE 9 – Graphique d'association représentant une table de contingence par ses résidus au modèle d'indépendance, plus la couleur est vive, plus la case est en sur-effectif (en bleu) ou sous-effectif (en rouge) par rapport au modèle d'indépendance.

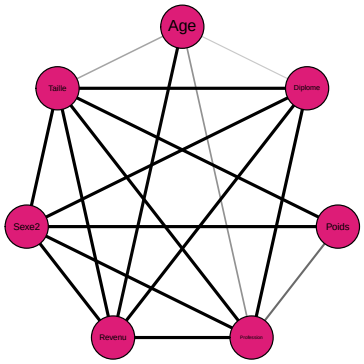

FIGURE 10 – Graphe de relation des variables de signalétique. Les relations statistiquement significatives au seuil de 5% sont indiquées par une ligne reliant les deux variables.

relations deux à deux grâce au *graphe des relations*. Pour chacune des relations le test d'hypothèses adéquat est pratiqué (χ^2, F de Fisher ou r de Pearson) et la présence d'un trait plus ou moins épais indique que les relations sont plus ou moins significatives au seuil conventionnel de 5%. La sous-option `Graphique des relations`[8] donne la Figure 10, et signale ici un bon nombre de relations significatives ; par exemple, si le poids n'est relié qu'à la taille, au sexe et, dans une moindre mesure, à la profession, les relations entre les autres variables sont plus nombreuses.

À partir de là, il convient d'interpréter les relations significatives et pour cela deux façons de procéder sont possibles : soit on utilise une tranche et on la croise avec elle-même, soit on croise cette tranche avec une variable externe servant de pivot. En ce qui concerne la première façon, nous allons choisir les variables Taille, Poids, Sexe2 et Revenu, c'est-à-dire deux variables numériques (notées N) et deux variables catégorielles (notées C, dont une ordonnée). La sous-option `Graphiques croisés complet` génère la Figure 11 et reprend simplement le graphique généralisé par paires d'Emerson et Green (2011). Les quatre variables sont croisées deux fois, une fois en abscisses et une fois en ordonnées. Cela permet de présenter deux façons de visualiser la relation, une dans le triangle supérieur, l'autre dans le triangle inférieur, différentes selon la nature des deux variables. Pour le croisement C*C, nous avons deux graphiques en mosaïque exprimant les distributions conditionnelles (par exemple le revenu par rapport au sexe et le sexe par rapport au revenu) ; pour le croisement N*N, deux nuages de points indiquant de même les distributions conditionnelles (par exemple Poids par rapport à Taille et Taille par rapport à Poids). Le croisement N*C donne lieu à une collection de boîtes de dispersion (ou boîtes à moustaches) dans le triangle inférieur ou de barres codes dans le triangle supérieur.

Il existe des relations fortes entre le sexe et les autres variables que nous

8. Le dessin du graphe a été produit par le package *qgraph* de Epskamp et al., 2011

approfondirons plus avant infra, ainsi qu'une relation positive et forte entre la taille et le poids. Plus intéressantes sont les relations entre le revenu et les autres variables (en colonne 2 ou ligne 2 sur la Figure 11). La taille augmente (et dans une moindre mesure le poids) avec les revenus. Mais surtout les hommes et les femmes n'ont pas les mêmes revenus, et les relations « morphologiques » précédentes mériteraient d'être corrigées par cet élément dans une modélisation.

L'autre façon de procéder s'obtient en employant une variable pivot, que l'on va relier à une tranche par la sous-option `Graphiques croisés en saucisson`. Le sexe est généralement un bon candidat pour cette opération. Nous obtenons la Figure 12 avec une tranche constituée de la taille, du poids et du revenu et du diplôme. Les « inégalités » morphologiques et sociales entre les sexes sont claires.

À ces relations graphiques peut être ajoutée une analyse statistique à partir de la sous-option `Statistiques croisées en saucisson`. On obtient la sortie ci-dessous :

```
$names
[1] "Diplome" "Poids"   "Revenu"  "Taille"
$RC
[1] 0.04 0.05 0.09 0.38
$ES
[1] 0.20 0.22 0.31 0.78
$size
[1] "M"  "M"  "M"  "XL"
$p.value
[1] 0 0 0 0
```

On regarde ici essentiellement les appréciations qualitatives qui montrent que la relation Sexe/Taille est classée XL alors que les trois autres sont seulement classées M (toutes sont statistiquement significatives).

Quatre types d'information sont donnés détaillés dans le tableau 1 : $R2$ le carré d'une corrélation entre les deux variables, ES une taille d'effet standardisée, accompagnée d'une appréciation qualitative de ces tailles d'effet et p la probabilité critique correspondante. Selon la nature des deux variables, ces calculs changent. Le $R2$ est le carré du coefficient de corrélation entre les deux variables. Les tailles d'effet ES sont celles employées par Cohen (1988). Cohen a proposé trois appréciations qualitatives de ces quantités (classant un effet en petit, moyen ou grand) pour chacune de ces situations, que nous avons étendues à cinq niveaux donnant des tailles imagées : XS, S, M, L et XL (voir Tableau 1).

6 Modèle linéaire explicatif

Une fois les relations bivariées explorées, il est possible de procéder à une modélisation pour certaines variables centrales dans nos hypothèses. Il s'agit

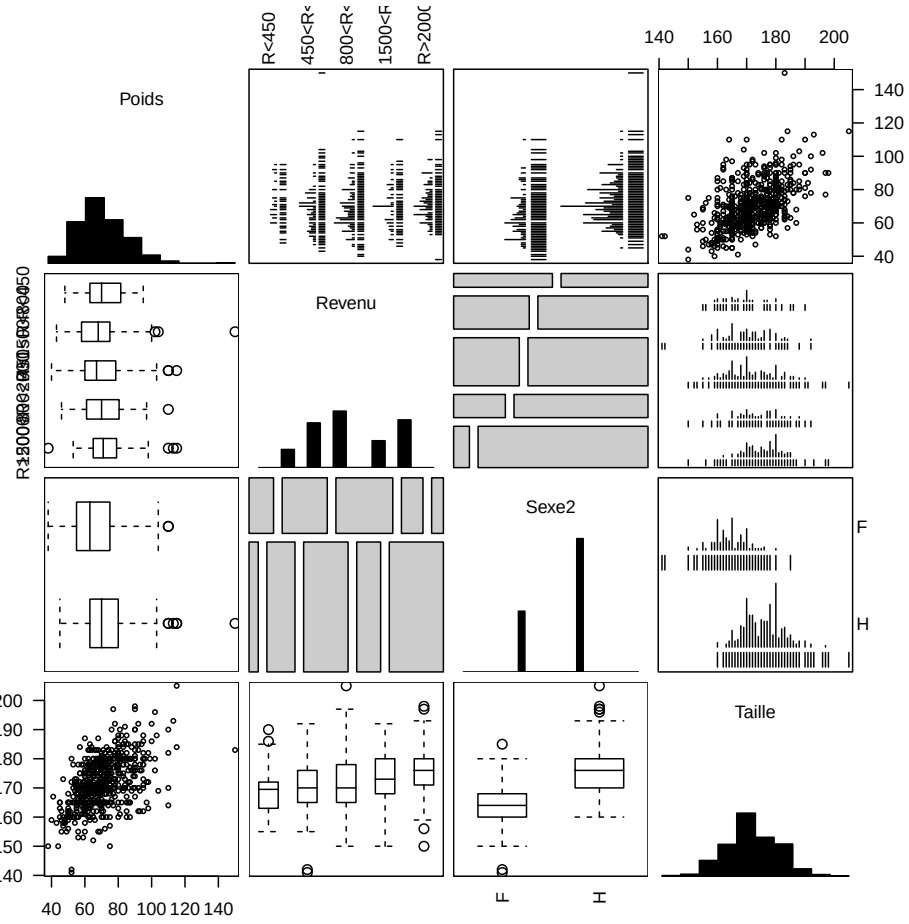

FIGURE 11 – Graphique en paires généralisé pour exprimer les relations bivariées (ici entre variables de signalétique)

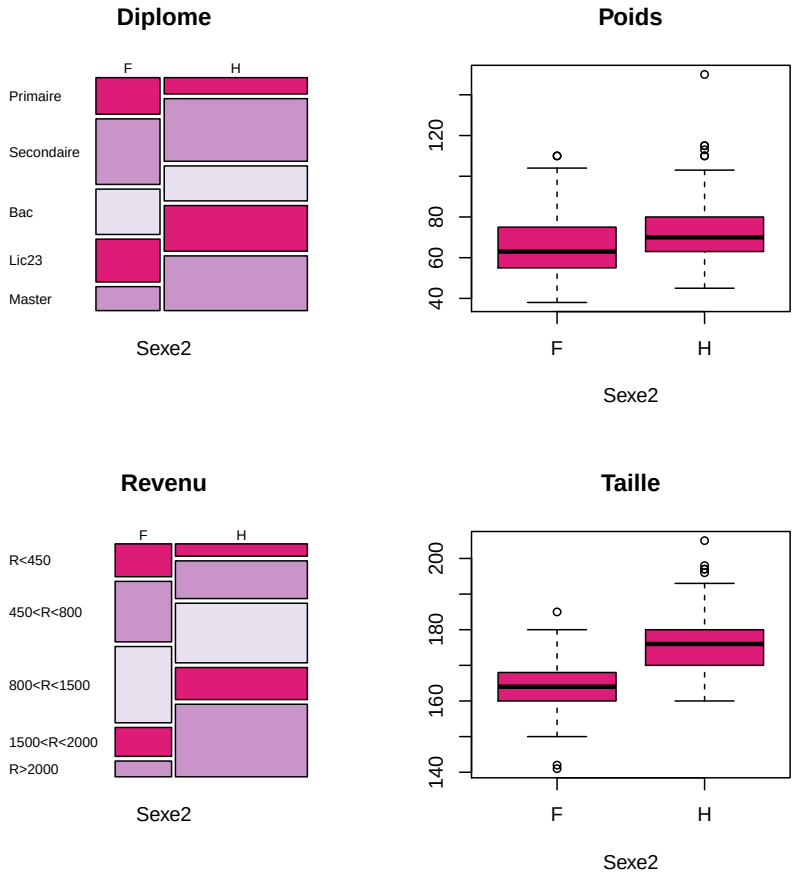

FIGURE 12 – Série de graphiques croisant le sexe (x) avec une tranche de variables (y : Diplôme, Poids, Revenu, Taille). Selon le type des variables une représentation graphique adaptée est automatiquement générée.

TABLE 1 – Résumés utilisés pour les trois types de relations possibles (N*N, N*C et C*C)

Relation	Statistique	R2	ES	Niveaux (Cohen, 1988)	Niveaux (*pointG*)
N*N	r (Pearson)	r^2	r	0.1 (petit), 0.3 (moyen), 0.5 (grand)	[0,0.05] (XS) [0.05,0.20] (S) [0.20,0.40] (M) [0.40,0.75] (L) [0.75,1] (XL)
N*C	RC, rapport de corrélation	RC	$f = \sqrt{\frac{RC}{1-RC}}$	0.1 (petit), 0.25 (moyen), 0.40 (grand)	[0,0.05] (XS) [0.05,0.175] (S) [0.175,0.325] (M) [0.325,0.70] (L) [0.70,...] (XL)
C*C	X^2 (Pearson)	λ_1 première valeur propre d'une AFC	$w = \sqrt{\frac{X^2}{n}}$	0.1 (petit), 0.3 (moyen), 0.5 (grand)	[0,0.05] (XS) [0.05,0.20] (S) [0.20,0.40] (M) [0.40,0.75] (L) [0.75,...] (XL)

d'expliquer cette variable par une fonction linéaire d'autres variables. Ceci permet en particulier de corriger certains effets. Nous commencerons par une étude peu intéressante sur le plan biologique ou sociologique car bien connue, mais intéressante sur le plan statistique car elle permet de montrer le déroulement des opérations et de saisir plus facilement comment en interpréter le résultat.

Nous avons vu que le poids des sondés est relié à leur taille mais aussi à leur sexe, les hommes étant plus lourds que les femmes. Que se passe-t-il lorsque les deux effets sont pris en compte simultanément ? Pour ce faire, on utilise l'option `Modèle linéaire explicatif` et la sous-option `Ajustement`. Il faut à ce niveau créer une formule pour ce modèle. À gauche de la formule, on met la variable à expliquer, ici le Poids et à droite de la formule, les variables explicatives, ici Sexe2 et Taille. Notons qu'à aucun moment nous ne nous soucions de la nature de ces trois variables. Une table de significativité (dite d'analyse de variance) est générée, laquelle indique l'apport de chaque variable dans le modèle, une fois toutes les autres prises en compte. En cas de variables inutiles, il faut alors recommencer la modélisation en les éliminant. Ici, les deux variables sont intéressantes, la taille plus que le sexe.

```
Anova Table (Type II tests)
Response: Poids
          Sum Sq  Df F value    Pr(>F)
Sexe2       2002   1  14.726 0.0001389 ***
Taille     26444   1 194.530 < 2.2e-16 ***
```

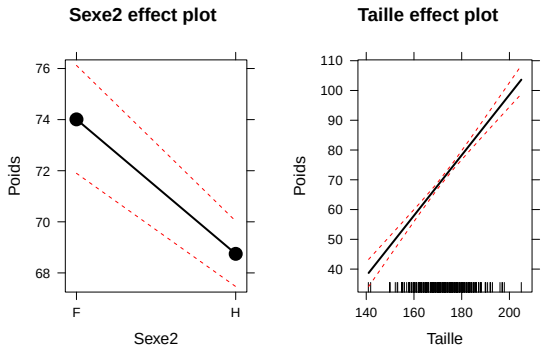

FIGURE 13 – Graphique des effets de la modélisation linéaire (gaussienne) du poids par le sexe et la taille.

```
Residuals  73813 543
```

Pour interpréter les effets significatifs, on emploie la sous-option `Graphique des effets` qui produit la Figure 13. Ce graphique, proposé par Fox (2003), donne pour chaque variable son effet, *une fois toutes les autres fixées à leur moyenne*, autrement dit toutes choses égales par ailleurs. Ainsi, on constate que quelqu'un qui a une taille de 1m60 a en moyenne un poids de 55 kgs alors qu'une personne avec une taille de 1m80 approche plutôt des 80 kgs. Plus intéressant, pour une taille donnée, on s'aperçoit que les femmes sont plus lourdes que les hommes d'environ 5 kilos. La relation est inversée par rapport au lien direct que l'on avait vu précédemment. Si ce phénomène d'inversion qui s'appelle « effet Simpson » peut paraître extrême, les méthodes linéaires servent fondamentalement à corriger les effets des variables entre elles sur la variable à expliquer.

Lorsque la variable à expliquer n'est plus de nature numérique, la modélisation change mais de façon complètement transparente pour l'utilisateur. Le Tableau 2 indique les techniques statistiques employées unifiées dans le cadre du modèle linéaire généralisé (McCullagh et Nelder, 1989).

TABLE 2 – Modèle linéaire utilisé selon la nature de la variable dépendante (N, B, O ou C)

Variable dépendante	Modèle	McCullagh et Nelder	Fonctions R
Numérique	Linéaire classique (gaussien)	Chapitre 3	lm
Binaire	Logistique binomial	Chapitre 4	glm
Catégorielle ordonnée	Proportional-odds	Chapitre 6, p. 159	polr (package *MASS*)
Catégorielle	Logit multinomial	Chapitre 6, p.151	multinom (package *nnet*)

Pour exemple, nous allons nous intéresser à la variable binaire RegardCorps (« La séropositivité a-t-elle transformée le regard que vous avez sur votre corps :

(1) NON et (2) OUI »). Nous allons essayer de la modéliser sur la base des variables Sexe2, Age, Taille, Poids et Diplôme. Le modèle s'ajuste sans problème et la table correspondante de significativité (dite dans ce cas d'analyse de déviance) donne trois variables significatives.

```
Analysis of Deviance Table (Type II tests)
Response: Regards_Corps
        LR Chisq Df Pr(>Chisq)
Sexe2     4.2691  1   0.038812 *
Age       2.4714  1   0.115932
Poids     4.5000  1   0.033894 *
Taille    0.8855  1   0.346704
Diplome  15.6861  4   0.003471 **
```

Nous simplifions le modèle en recommençant avec ces variables significatives : Sexe2, Poids et Diplôme (dans ce cas, le sexe et le poids sont marginalement significatifs). Le graphique des effets (Figure 14) donne la probabilité de répondre OUI[9] selon les trois explicatives. Cette probabilité est plus élevée chez les femmes, elle est reliée négativement au poids mais surtout il y a un fort effet diplôme (et donc revenu) : les personnes ayant un niveau d'étude primaire ont une probabilité prédite de 40% à répondre OUI alors que pour les autres elle s'élève à 60-70%[10].

7 Analyse factorielle exploratoire

La dernière méthode que nous allons employer pour analyser une tranche de variables est l'analyse factorielle. Elle consiste à représenter les relations entre un ensemble de variables de façon graphique en passant par la création intermédiaire de variables de synthèse. Ces dernières permettent également de situer graphiquement d'autres variables dites supplémentaires qui n'ont pas participé à l'analyse mais qu'il est intéressant de visualiser simultanément (Cibois, 1990).

De la même façon que le modèle linéaire explicatif a été unifié par la théorie du modèle linéaire généralisé, des analyses apparemment distinctes (ACP, ACM) ont été mises en perspective par la théorie du schéma de dualité (Escoufier, 1987) programmée dans le package `ade4` de `R` (Chessel et al, 2004) dont nous avons utilisé et combiné trois fonctions pour produire notre programme. L'analyse de Hill et Smith (1976) peut être vue dans cette même perspective et permet de mêler variables catégorielles et numériques au sein d'une même analyse. Celle-ci recherche des variables de synthèse les plus reliées possibles aux variables de départ, au sens du rapport de corrélation pour les variables catégorielles et du coefficient de détermination pour les variables numériques (qui sont en fait le même calcul géométrique R2). Les représentations graphiques des résultats sont

9. C'est toujours la probabilité de la deuxième catégorie (dans l'ordre alphabétique) qui est modélisé, soit ici OUI

10. Sur le graphe des effets, des pointillés en rouge donnent des intervalles de confiance de ces prédictions, importants pour éviter une sur-interprétation des différences observées.

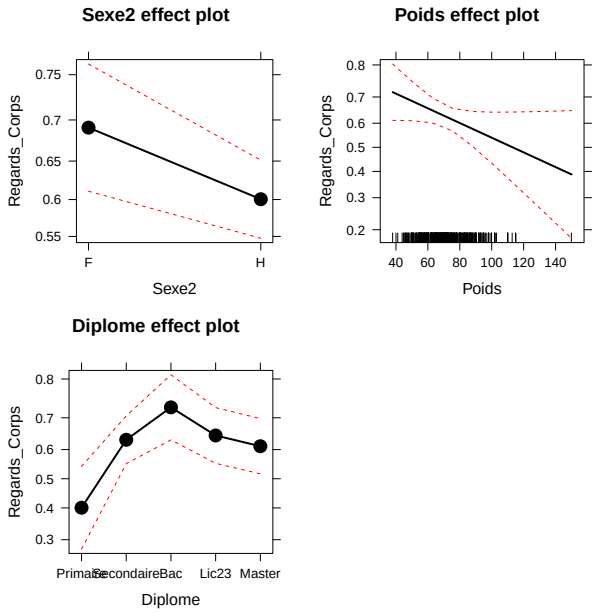

FIGURE 14 – Graphique des effets de la modélisation linéaire (logistique binomiale) du regard sur le corps (Oui vs. Non) par le sexe, le poids et le diplôme.

séparées avec un classique cercle de corrélation pour les variables numériques et des ellipses de gravité pour les modalités des variables catégorielles.

Le premier exemple traite du rapport au cours des sondés et use des six échelles de Likert suivantes : CorpsPasMoiN, CorpsTropMaigreN, CorpsTropGrosN, CorsPasSéduisantN, CorpsDéforméN et CorspMieuxN. Nous les sélectionnons comme variables actives par le truchement de l'option `Exploratoire multivariée` dont la seule sous-option est pour l'heure `Factoriel multivarié`. Elle produit le classique éboulis des valeurs propres (Figure 15) qui montre qu'un axe prédomine mais que le second dépasse 1, donc présente un intérêt. Le graphique produit est un cercle de corrélation (Figure 16), ici l'analyse n'est en fait qu'une classique analyse en composantes principales (ACP). Ce cercle montre deux dimensions sous-jacentes et relativement indépendantes puisque leurs directions sont perpendiculaires. D'une part une corrélation positive entre corps trop maigre et mieux qu'avant ; d'autre part des corrélations positives entre des visions négatives et le surpoids.

Le deuxième exemple concerne d'autres variables mesurant le rapport au corps : ImpactVIH, PenséeVIH, EvolLoisir, EvolProf, RegardsCorps (toutes catégorielles). En répétant la même opération, l'analyse se transforme automatiquement en une analyse des correspondances multiples (ACM). L'éboulis des valeurs propres (non présenté ici) montre qu'un axe domine mais la représentation est en deux dimensions (seule option pour l'instant). La structuration est

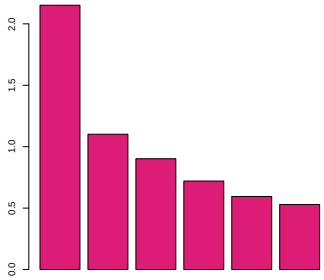

FIGURE 15 – Éboulis des valeurs propres d'une analyse factorielle exploratoire de variables numériques (donc une ACP)

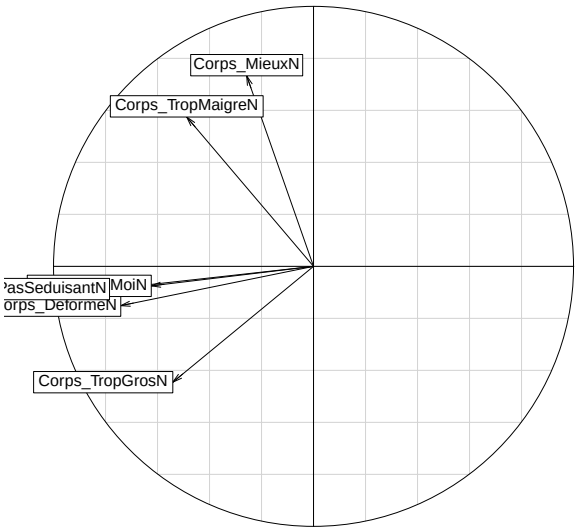

FIGURE 16 – Représentation graphique des variables (exemple 1) sur le cercle de corrélation de l'analyse

remarquable (Figure 17) entre l'impact et les pensées au VIH, les évolutions de pratiques de loisirs et professionnelles et le regard porté sur le corps.

Enfin, il est possible de combiner les deux types de variables dans une même analyse dans un troisième exemple. Nous éliminerons trois variables de Likert manifestement redondantes (Corps déformé et Corps pas moi et Corps Mieux) et évacuerons le regard sur le corps et l'impact sur l'existence, soit au final six variables actives. On constate (Figure 18 et 19) que le premier axe reflète l'impact sur l'existence et le second la vision du corps, qui sont donc non corrélées.

En variables supplémentaires, nous avons employé des variables de signalétiques : Age, Sexe, Taille, Poids, Revenu et Diplôme. On peut les interpréter à partir des Figures 20 et 21. Le message du cercle de corrélation est simple, il n'y a pas de relations entre les structures observées des variables actives et la taille, le poids, l'âge. Pour ce qui est des variables catégorielles, le sexe n'est pas relié à ces structures. En revanche le diplôme et, dans une moindre mesure, le revenu le sont, mais de façon différente. Le diplôme est plutôt corrélé aux structures des variables de Likert (deuxième facteur) et montre que la vision négative du corps est reliée à de faibles diplômes, alors que le faible revenu est relié aux personnes ayant cessé toute activité de loisirs ou professionnelle et qui pensent sans arrêt à leur séropositivité.

8 Conclusion

La programmation du package *pointG* est au stade expérimental et en cours d'amélioration. Les graphiques et statistiques ne sont pas encore tous optimisés. Ainsi, dans les graphiques pour échelles de Likert, l'ajout d'intervalles de confiance pour les moyennes serait bienvenu. Toutefois, il ne s'agit pas de raffiner les aspects esthétiques car ce logiciel n'est pas conçu pour communiquer les résultatsobtenus mais bien pour obtenir des résultats par exploration ouverte. Pour produire d'élégants graphiques, il existe de nombreuses possibilités dans le logiciel `R`, par exemple : `lattice` (Deepayan, 2008), `ggplot2` (Wickham, 2009) ou `grid` (Murrell, 2002).

Il est en revanche prévu à brève échéance d'introduire une gestion préalable de la nature des variables et de recodage/transformation [11], des options de cartographies plus riches et par communes, un traitement des réponses multiples, des méthodes pour étudier la satisfaction et les attentes d'usagers ou de clients, méthodes classiques dans les approches socio-marketing et enfin la possibilité de sélectionner une sous-population aisément [12]. A moyenne échéance, l'introduction de pondérations semble nécessaire pour disposer d'un outil complet d'analyse de données de sondage ou nécessitant une pondération.

Bien d'autres options seraient envisageables mais le parti-pris du package

11. Largement présentes dans le menu déroulant `Données`, option `Gérer les variables dans le jeu de donnnées actif` du *R-Commander*.

12. C'est en fait possible avec une manipulation assez simple du `R-Commander` dans le menu `Données`, `Jeu de données actif` et `Sous-ensemble` mais qui s'avère assez fastidieux à la longue...

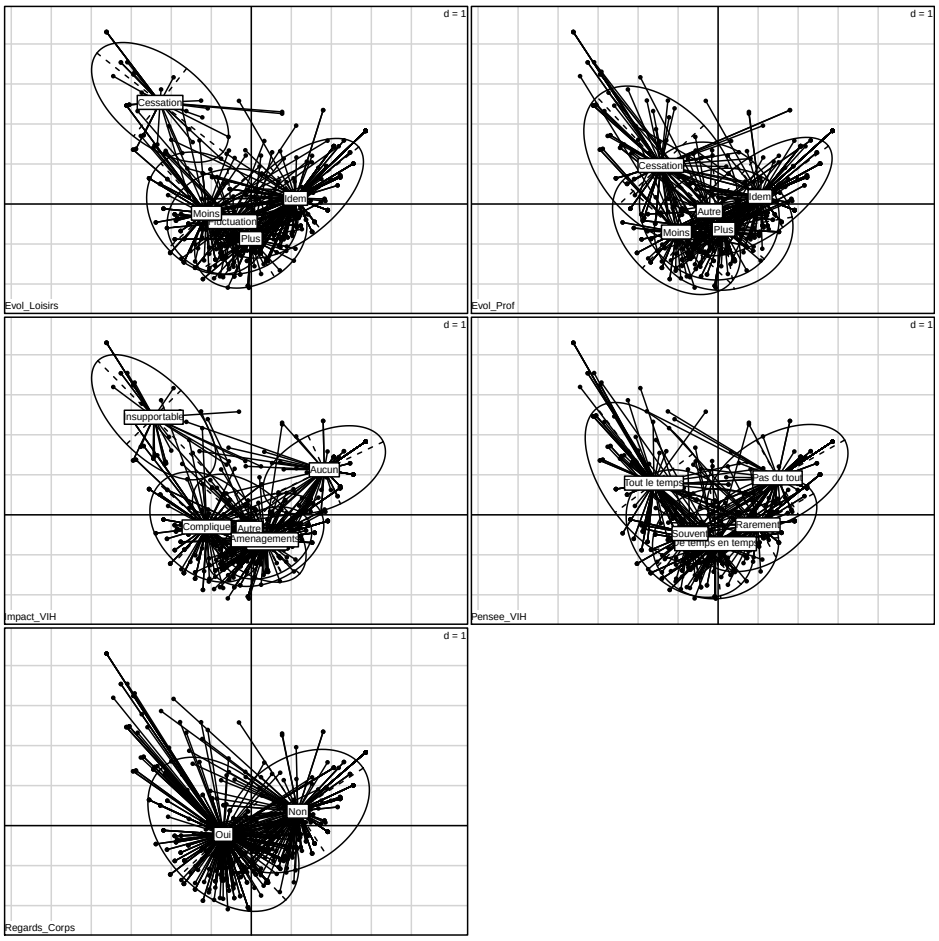

FIGURE 17 – Représentation des variables (exemple 2) pour une analyse exploratoire composée de variables catégorielles (donc une ACM)

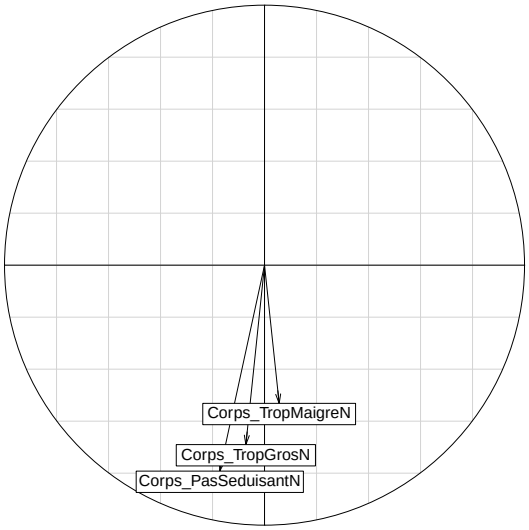

FIGURE 18 – Représentation graphique des variables (exemple 3) actives numériques dans une analyse factorielle exploratoire mixte (analyse de type Hill & Smith)

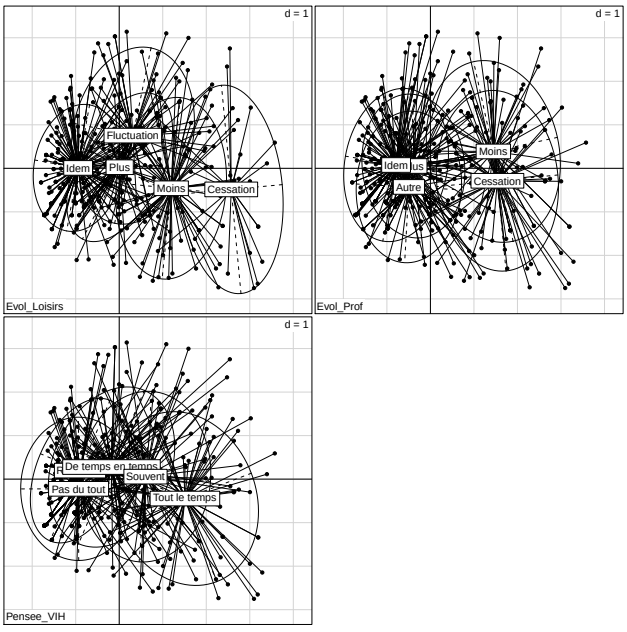

FIGURE 19 – Variables actives catégorielles (exemple 3)

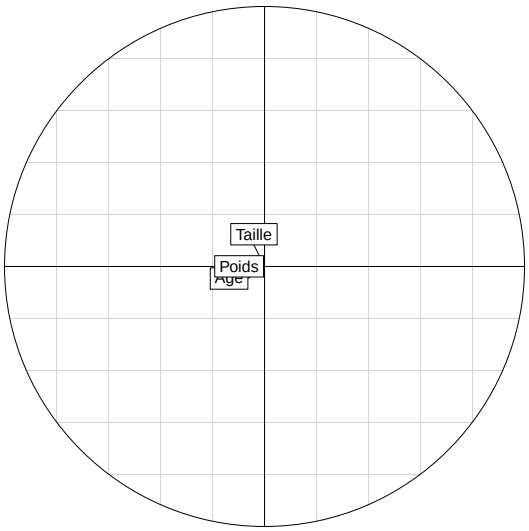

FIGURE 20 – Variables supplémentaires numériques (exemple 3)

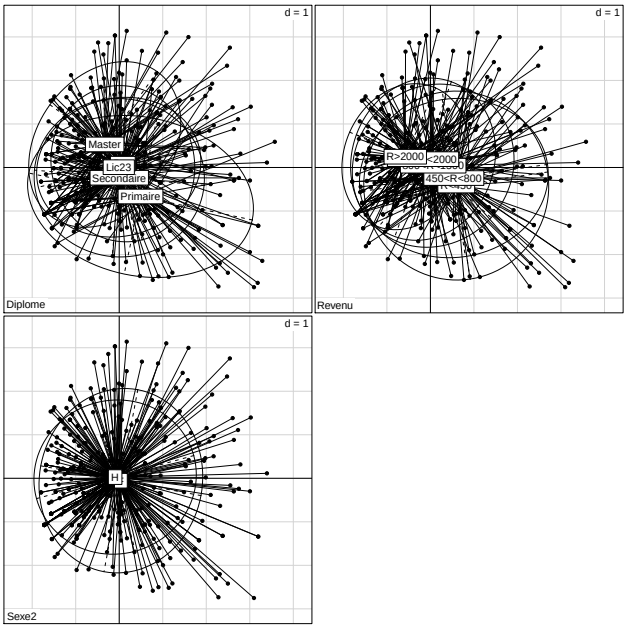

FIGURE 21 – Variables supplémentaires catégorielles

pointG est de garder les choses simples et de ne pas multiplier les options. Il s'agit donc plus d'être à présent à l'écoute des utilisateurs pour améliorer la convivialité du logiciel, sa documentation et la mise à disposition d'autres jeux de données utiles en particulier pour des enseignants.

Lorsque l'utilisateur se sera acclimaté, probablement rapidement, à l'univers de *pointG*, il va pouvoir explorer de nouvelles possibilités. *pointG* est en effet inclus dans une interface conviviale plus générale - le *R-commander* de Fox et al. (2011) -, elle-même incluse et extensible par le logiciel `R` (R Development Core Team, 2011) libre de distribution. À partir du *R-commander*, il pourra facilement réaliser des zooms sur des variables grâce à d'autres procédures ou reprendre les modèles générés pour les étudier plus complètement : intervalles de confiance des paramètres, nouveaux tests, étude de résidus, *etc.* L'immersion dans l'environnement plus général de `R` (Barnier, 2009) sera facilitée pour les utilisateurs les plus attirés par les approches quantitatives.

Fondamentalement, ce qui rend possible le logiciel *pointG* est ce remarquable environnement de logiciel collaboratif. Sont à disposition des méthodes fort bien programmées (par exemple, Venables et Ripley, 2002) que l'on peut assembler selon des besoins spécifiques. Ici, il s'agit de l'étude de questionnaires pour des sociologues mais nous avons construit une version bien différente (et non publique) pour des cours de marketing en management du sport.

Références

[1] Barnier, J. (2009) R pour les sociologues et assimilés. http ://cran.r-project.org/.

[2] Chessel, D., Dufour, A.B. et Thioulouse, J. (2004) The ade4 package-I- One-table methods. R News, 4, 5-10.

[3] Cibois P. (1990) L'analyse de données en sociologie. Paris : PUF.

[4] Cibois, P. (1993) Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence. Bulletin de Méthodologie Sociologique, 40, 43-63.

[5] Cleveland, W. (1993) Visualizing data. Hobart Press.

[6] Cohen, J. (1988) Statistical power analysis for the behavioral sciences (2ème édition). Hillsdale,NJ : Lawrence Erlbaum.

[7] McCullagh P. et Nelder, J. A. (1989) Generalized Linear Models. London : Chapman and Hall.

[8] Deepayan, S. (2008) Lattice : Multivariate Data Visualization with R. Springer, New York.

[9] Emerson, J.W. et Green, W.A. (2011) gpairs : The Generalized Pairs Plot. R package version 1.0. http ://CRAN.R-project.org/package=gpairs

[10] Epskamp, S., Cramer, A.O.J., Waldorp, L.J., Schmittmann, V.D. et Borsboom, D. (2011) qgraph : Network representations of relationships in data. R package version 0.5.3. http ://CRAN.R-project.org/package=qgraph

[11] Escoffier, B. et Pages, J. (1994) Multiple Factor Analysis (AFMULT package). Computational Statistics and Data Analysis, 18, 121-140.

[12] Escoufier, Y. (1987) The duality diagram : a means of better practical applications In Development in numerical ecology, Legendre, P. & Legendre, L. (Eds.) NATO advanced Institute, Serie G. Springer Verlag, Berlin, 139-156.

[13] Fox, J. (2003). Effect Displays in R for Generalised Linear Models. Journal of Statistical Software, 8(15), 1-27. URL : http ://www.jstatsoft.org/v08/i15/.

[14] Fox, J. (2011) <jfox@mcmaster.ca>, with contributions from Liviu Andronic, Michael Ash, Theophilius Boye, Stefano Calza, Andy Chang, Philippe Grosjean, Richard Heiberger, G. Jay Kerns, Renaud Lancelot, Matthieu Lesnoff, Uwe Ligges, Samir Messad, Martin Maechler, Robert Muenchen, Duncan Murdoch, Erich Neuwirth, Dan Putler, Brian Ripley, Miroslav Ristic and Peter Wolf. Rcmdr : R Commander. R package version 1.7-0. http ://CRAN.R-project.org/package=Rcmdr

[15] Harrell Jr, F.E. (2001) <f.harrell@vanderbilt.edu> and with contributions from many other users. (2011). Hmisc : Harrell Miscellaneous. R package version 3.9-0. http ://CRAN.R-project.org/package=Hmisc

[16] Hill, M. O., et A. J. E. Smith. (1976) Principal component analysis of taxonomic data with multi-state discrete characters. Taxon, 25, 249-255.

[17] Husson, F., Josse, J., Le, S. et Mazet, J. (2011). FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R. R package version 1.16. http ://CRAN.R-project.org/package=FactoMineR

[18] Meyer, D., Zeileis, D. et Hornik, K. (2006) The Strucplot Framework : Visualizing Multi-Way Contingency Tables with vcd. Journal of Statistical Software, 17(3), 1-48. URL http ://www.jstatsoft.org/v17/i03/

[19] Murrel, P. (2002)l. The grid graphics package. R News, 2, 14-19.

[20] R Development Core Team (2011). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http ://www.R-project.org/.

[21] Temple, M., Alfons, A. et Kowarik, A. (2011). VIM : Visualization and Imputation of Missing Values. R package version 2.0.4. http ://CRAN.R-project.org/package=VIM.

[22] Tufte, E.R. (2001) The visual display of quantitative information. Graphics Press.

[23] Venables, W.N. et Ripley, B.D. (2002) Modern Applied Statistics with S (4ème édition). Springer : New York.

[24] Wickham, H. (2009) ggplot2 : elegant graphics for data analysis. Springer : New York.

[25] Becker, R.A., Wilks, A.R., Brownrigg, R. et Minka, T.P. (2011) Original S code by Richard A. Becker and Allan R. Wilks. R version by Ray Brown-

rigg. Enhancements by Thomas P Minka <surname@stat.cmu.edu>. <surname@stat.cmu.edu> (2011). maps : Draw Geographical Maps. R package version 2.2-2. http ://CRAN.R-project.org/package=maps

A Installations informatiques

Afin d'installer les éléments suivants, il faut disposer des droits d'administrateur sur l'ordinateur. Sous windows, cliquer sur le programme à lancer avec le bouton droit de la souris, choisir `Exécuter en tant qu'administrateur`.

A.1 Installer `R`

1. Aller sur le site : `http://cran.univ-lyon1.fr/`
2. Cliquer sur `Download R for windows` (ou un autre système d'exploitation de meilleure qualité)
3. Cliquer sur `base`
4. Cliquer sur `Download R 2.14.0 for windows`[13]
5. Cliquer sur `Enregistrer` et déposer le fichier exécutable d'installation `R-2.14.0-win.exe` sur le bureau.
6. Double-cliquer sur ce fichier `R-2.14.0-win.exe` et se laisser guider.

A.2 Installer `Rcmdr`

1. Il faut tout d'abord lancer le logiciel **R** afin d'installer les extensions appelées *packages*. Pour lancer **R**, il suffit de cliquer sur l'icône bleue **R** du bureau ou alors par le biais du menu `Démarrer`. Une fenêtre `RGui` apparaît alors.
2. Aller dans le menu déroulant `Package` et choisir l'option `Installer le(s) package(s)`. Une liste de sites de téléchargement apparaît, prendre un site proche de chez vous (ex : Lyon 1). Une longue liste de package s'ouvre, choisir `Rcmdr`.
3. Pour lancer le package, aller dans le menu déroulant `Package` et choisir `Charger le package`. Choisir `Rcmdr` dans la liste des packages disponibles (puisqu'il vient d'être installé). Attention, lors de la première installation (uniquement) il demandera l'autorisation de charger des extensions, il faut répondre `OK` à ces demandes. L'opération prend un petit moment et la fenêtre interactive du R-commander s'ouvre.

A.3 Installer `pointG`

Les procédures d'installation sont les mêmes pour tous les packages. Voir donc le point précédent. Attention, le nom du package à télécharger est `RcmdrPlugin.pointG` !

A.4 Lancer `pointG`

Pour les utilisations usuelles de *pointG* (une fois qu'il est installé par les trois phases précédentes), il suffit de lancer **R** (icône bleu sur votre bureau) puis

13. Les versions changent mais la procédure d'installation reste la même.

d'aller dans le menu déroulant `Package`, le sous-menu `Charger le package` et de choisir `RcmdrPlugin.pointG` dans la liste.

Une fenêtre *R-Commander* classique s'ouvrira d'abord, puis se fermera. Une seconde fenêtre *R-Commander* apparaitra, agrémentée du menu déroulant *pointG*. A vous de jouer !