

Représenter des analyses géométriques de données

Des résultats aux graphiques avec R/RStudio et ggplot2

Anton Perdoncin

ERC Lubartworld, EHESS, Cens

04 juin 2021

- 1 Introduction
- 2 Une famille de méthodes graphiques
- 3 Quelques règles pratiques
- 4 Réalisation de l'ACM et premières visualisations
- 5 Mise en forme des résultats statistiques
- 6 Graphiques sur-mesure pas-à-pas

Section 1

Introduction

Objectifs

- Montrer **comment réaliser**, avec R et RStudio, des **graphiques** et des **tableaux statistiques sur-mesure**, permettant de présenter et visualiser les résultats d'une analyse géométrique des données.
- Donner un exemple d'**espace de travail quantitatif reproductible**.
- Plaider pour l'usage des solutions de **contrôle de version git** (intégré à RStudio).

Avant de commencer

- Rendez-vous sur le **dépôt GitHub** dédié à cette présentation :
https://github.com/APerdoncin/visual_agd
- Les usagers de GitHub peuvent cloner le dépôt sur leur ordinateur (git clone ...)
- Les autres peuvent simplement télécharger le contenu du dépôt et le copier dans un dossier dédié sur leur ordinateur (Code > Download ZIP)
- **Deux scripts à exécuter :**
 - 00-setup : packages (à installer éventuellement) et options
 - 01-data : téléchargement, “dézipage” et importation des données de l'EEC 2018

Un espace de travail reproductible

- Des données brutes aux résultats et à la rédaction **sans intervention parallèle ou externe sur les données**
- Une **organisation logique** qui distingue les **types de fichiers** et les **types d'opérations réalisées**
- La garantie de **retrouver le même résultat** ... à condition d'**appliquer les mêmes procédures**.

Git : un ami qui vous veut du bien

- En finir avec les V1, V2, V142, VDEF ... VDEFDEF ... VDERDESDERS
- Sourcer et retrouver les modifications réalisées sur les fichiers
- Travailler de façon collaborative en minimisant les risques de conflits
- Un petit coût d'entrée... très nettement diminué par l'intégration de Git à RStudio

Pourquoi le tutoriel sur Quanti ?

- Nouvelles fonctionnalités et nouveaux packages disponibles :
 - `explor` (Julien Barnier) ;
 - `FactoMineR` ;
 - `factoextra` ;
 - `GDAtools` (Nicolas Robette) : `ggcloud_variables`, `ggcloud_indiv`, `ggadd_ellipses`, `ggadd_interaction`, `ggadd_supvar`.
- Mais... comment faire pour pouvoir *tout* paramétrer : modalités à représenter, couleurs, symboles, etc.
- `ggplot2` : fonctionnalités graphiques surpuissantes pour des graphiques infiniment paramétrables
- mettre au propre des routines pouvant être adaptées à une diversité de données d'enquête et aux objectifs d'administration de la preuve statistique.

Les données

- Un objet “bac à sable” : le travail intérimaire... sans aucune ambition sociologique (du moins pas aujourd'hui !)
- De “vraies” données d'enquête : fichier détail de l'EEC 2018 (<http://insee.fr/fr/statistiques/4191029#consulter>)
- On travaille directement sur les données recodées, “prêtes-à-jouer” (mais le script de recodage est disponible sur GitHub) : 4632 individus et 13 variables.

Section 2

Une famille de méthodes graphiques

Résumer l'information le mieux possible

Analyses géométriques des données incluent notamment :

- Analyse en composantes principales (ACP)
- Analyse des correspondances multiples (ACM)
- Classifications (ascendantes hiérarchies notamment, CAH)
- Analyse factorielle de tableaux multiples (AFM), etc.

Un principe et résultat commun : fournir la **meilleure description possible des corrélations** dans un jeu de données. (Cibois, 2000 ; Le Roux et Rouanet, 2014 ; Volle, 1997)

Résumer l'information le mieux possible

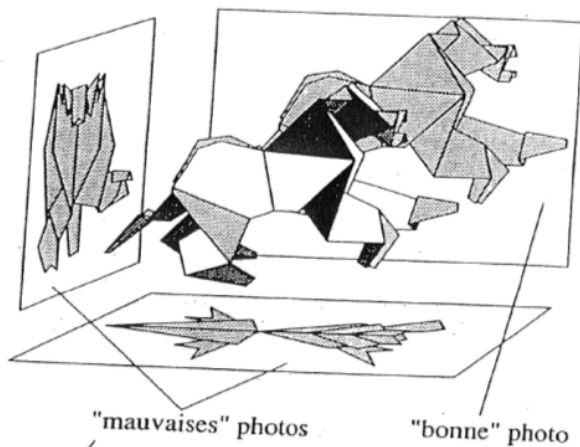


FIGURE 1 – La forme cheval

Le parti-pris graphique

- Un usage répandu dans les enseignements et les publications : ne présenter que le fameux **graphique en croix**
 - un ou plusieurs plans factoriels et les nuages des modalités actives et/ou supplémentaires
 - parfois (rarement) les nuages d'individus encore plus rarement habillé ou structuré
- Les **résultats statistiques** (fréquences, contributions, coordonnées, cosinus carrés) sont rarement présentés.

Le parti-pris graphique

- Usage cohérent avec l'**intérêt intrinsèque des méthodes de réduction de dimensionnalité** : représenter sur un plan une information très multidimensionnelle. . .
- A **deux condition** :
 - que le graphique soit démonstratif ;
 - que l'on prenne garde aux erreurs d'interprétations graphiques (Cibois, 1997).

Section 3

Quelques règles pratiques

Un graphique démonstratif ?

Schématiquement : un graphique utile à l'administration de la preuve doit être lisible, structuré, *self-explaning* et esthétique.

- ❶ Sélectionner quelles modalités représenter : quel que soit le critère retenu, tout n'est pas bon à représenter !
- ❷ Les noms des variables et les libellés des modalités doivent être présentés en “langage naturel” et non en code hiéroglyphique.
- ❸ Les libellés des modalités ne doivent pas se chevaucher.
- ❹ Les labels des axes doivent être clairement présentés, et indiquer le pourcentage d'inertie conservé par chaque axe.

Un graphique démonstratif ?

- 5 La légende (si nécessaire) doit être positionnée de façon à ne pas empiéter sur le graphique.
- 6 Distinguer clairement les types de modalités (actives ou illustratives).
- 7 Distinguer clairement les variables ou groupes de variables.
- 8 Ne pas oublier l'esthétisme : gammes de couleur permettant de distinguer ce qui doit l'être ; pouvoir aisément passer de la couleur au noir et blanc (ou nuances de gris).

Un graphique démonstratif ?

C'est tout ? Non...

Des graphiques démonstratifs ne suffisent pas à asseoir statistiquement l'argumentation : il faut aussi présenter lisiblement les résultats statistiques de l'analyse géométrique.

Quels résultats statistiques présenter ?

- Dépend du type d'analyse
- ACP : inerties et corrélations des variables aux axes ;
- AFM : inerties, fréquences, coordonnées, v-test ;
- ACM : en plus des inerties :
 - fréquences : repérer les modalités à petits effectifs ;
 - contributions, coordonnées, cosinus carrés, v-test pour chacun des axes interprétés ;
 - éventuellement sommer les contributions par variable, ou par groupe de variables ;
 - pour chaque modalité illustrative : effectif brut, fréquence, puis coordonnées et cosinus carrés sur chacun des axes interprétés.

Quels résultats statistiques présenter ?

- Là encore, la mise en forme des tableaux doit être faite avec attention :
 - sur LibreOffice Calc (ou son avatar non libre) ;
 - dans LaTeX (package `xtable` ou macro `CalctoLatex` / `Excel2Latex`) ;
 - RMarkdown : `kable` et `kableExtra` ou `flextable`

Une bonne nouvelle : structurer le tableau des résultats et rassembler les informations utiles à la réalisation des graphiques vont de pair...

Section 4

Réalisation de l'ACM et premières visualisations

Sélection des variables

“All in all, doing a data analysis, in good mathematics, is simply searching eigenvectors (*valeurs propres*) ; all the science (or the art) of it is just to find the right matrix to diagonalize.” (J.-P. Benzécri)

Sélection des variables

```
d_acm <- d %>%  
  select("age", "genre", "diplome", "nationalite", "gsp",  
         "sociopro", "menage",  
         "position", "choix", "horaires", "heures_plus",  
         "public_prive", "secteur") %>%  
  mutate_all(factor) %>%  
  drop_na()
```

Réalisation de l'ACM

```
res_acm <- MCA(d_acm, quali.sup = 5:7)
```


Premières visualisations : graphiques par défaut

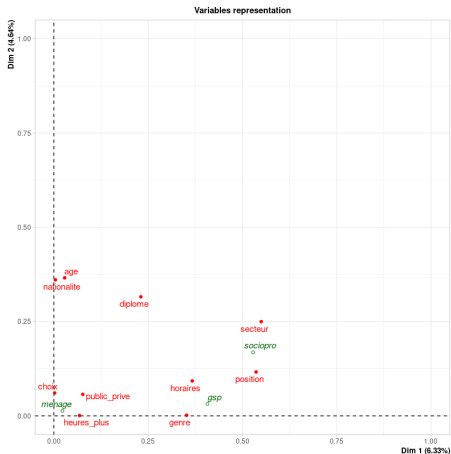


FIGURE 2 – Nuage des variables

Premières visualisations : graphiques par défaut

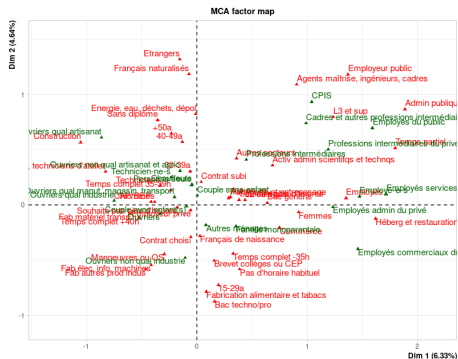


FIGURE 3 – Nuage des modalités

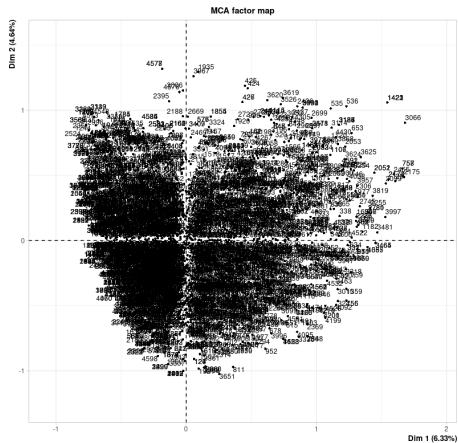


FIGURE 4 – Nuage des individus

Alternatives

- FactoMineR : possibilité de paramétrer ces graphiques... dans certaines limites ; difficulté principale : représenter ensemble modalités actives et supplémentaires.
- factoextra : fonctions utiles pour des représentations rapides.
- explor : paramétrage interactif des graphiques.

Section 5

Mise en forme des résultats statistiques

Remarques

- Objectif : construire un gros tableau qui comporte toutes les informations pertinentes pour les modalités actives et supplémentaires.
- Modus operandi : on part de l'objet liste `res_acm` qui stocke les résultats de l'analyse, on y prend les infos dont on a besoin, et on les manipule pour obtenir les “sous-tableaux” qui sont in fine assemblés.
- Go to 03-visu-agd.R !

Section 6

Graphiques sur-mesure pas-à-pas

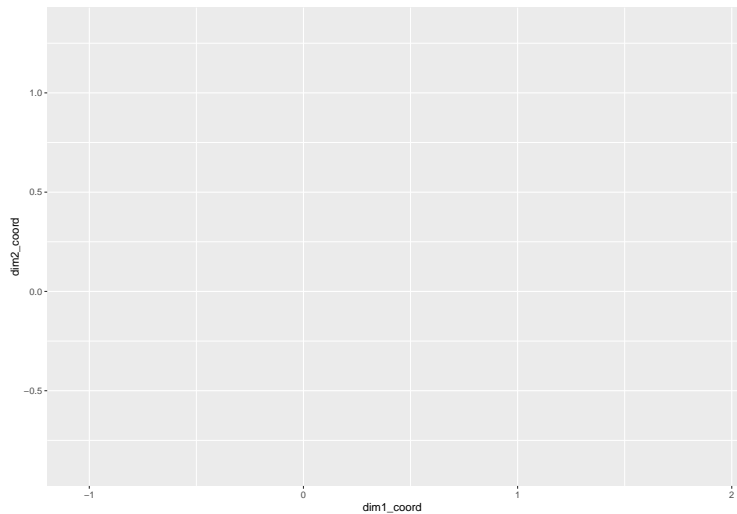
Nuage des modalités actives : sélection des modalités

```
resultats_actives %>%  
  filter(dim1_contrib > seuil |  
         dim2_contrib > seuil) %>%
```


Nuage des modalités actives : initialisation

```
resultats_actives %>%  
  filter(dim1_contrib > seuil |  
         dim2_contrib > seuil) %>%  
  
  ggplot(aes(x = dim1_coord, y = dim2_coord,  
             label = modalites,  
             shape = variables))
```

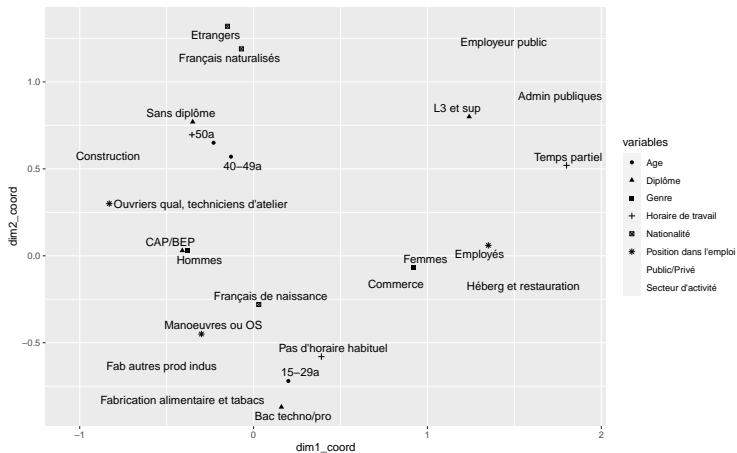
Nuage des modalités actives : initialisation



Nuage des modalités actives : points et labels

```
geom_point() +  
coord_fixed() +  
geom_text_repel(segment.alpha = 0.5)
```

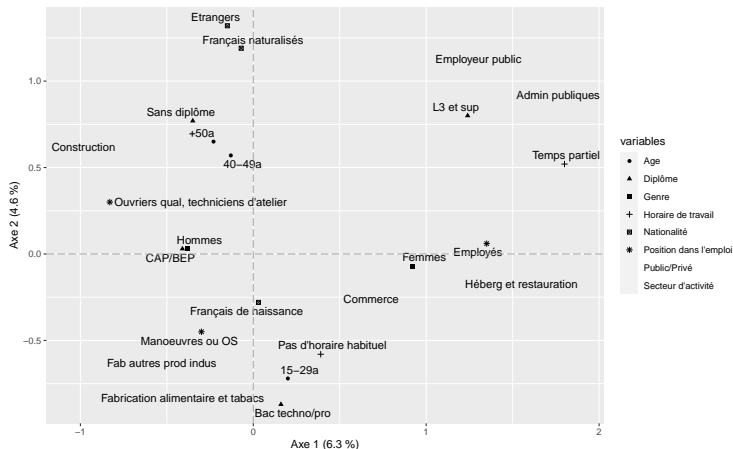
Nuage des modalités actives : points et labels



Nuage des modalités actives : axes, ordonnées et abscisses

```
geom_hline(yintercept = 0, colour = "darkgrey",  
           linetype="longdash") +  
geom_vline(xintercept = 0, colour = "darkgrey",  
           linetype="longdash") +  
  
xlab(paste0("Axe 1 (", round(variances[1, 3], 1), " %)")) +  
ylab(paste0("Axe 2 (", round(variances[2, 3], 1), " %)"))
```

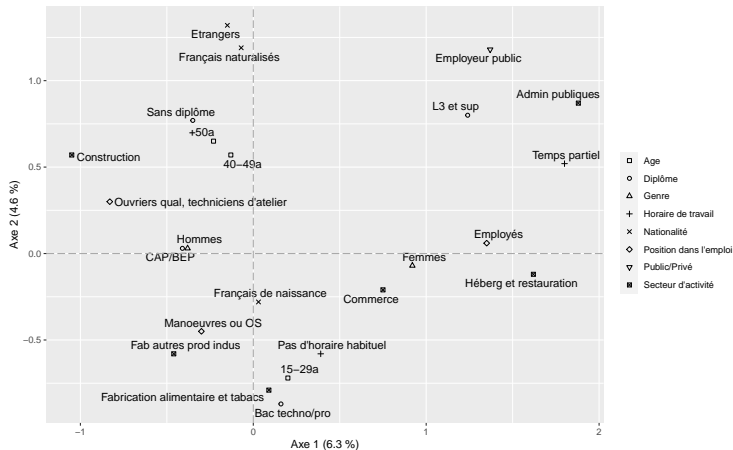
Nuage des modalités actives : axes, ordonnées et abscisses



Nuage des modalités actives : paramétrage des points et de la légende

```
scale_shape_manual(name = "", values = 0:20) +  
guides(shape=guide_legend(title = ""))
```

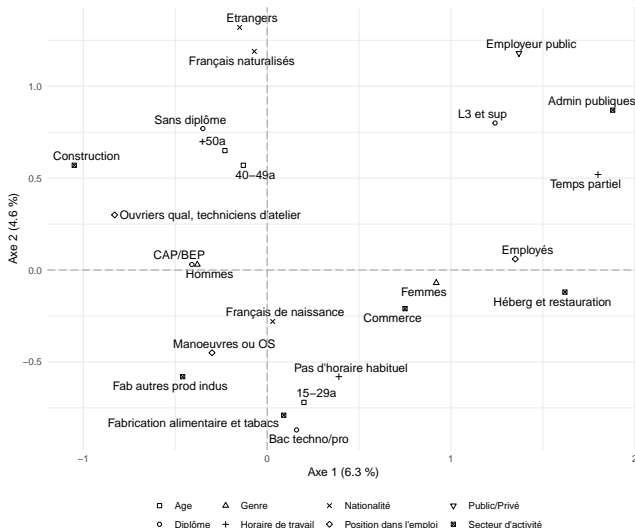
Nuage des modalités actives : paramétrage des points et de la légende



Nuage des modalités actives : cosmétique générale

```
theme_minimal() +  
theme(legend.position="bottom")
```

Nuage des modalités actives : résultat final



Autres graphiques

Go to 03-visu-agd.R

This is the end !

Merci de votre attention !

Mail : anton.perdoncin@ehess.fr

Twitter : @AntonPerdoncin

GitHub : <https://github.com/APerdoncin>

Bibliographie I

Cibois Philippe, 1997, « Les pièges de l'analyse des correspondances », *Histoire & Mesure*, 12(3), p. 299-320.

Cibois Philippe, 2000, *L'analyse factorielle. Analyse en composantes principales et analyse des correspondances*, Paris, PUF.

Le Roux Brigitte, **Rouanet** Henry, 2014, *Analyse géométrique des données multidimensionnelles*, Paris, Dunod.

Volle Michel, 1997, *Analyse des données*, Paris, Economica.