

Méthodes quantitatives en sciences sociales avec R

Philippe Apparicio et Jérémy Gelb

2022-03-19

Table des matières

Liste des tableaux

Liste des figures

Préface

Ce livre vise à décrire une panoplie de méthodes quantitatives utilisées en sciences sociales avec le logiciel ouvert R. Il a d'ailleurs été écrit intégralement dans R avec rmarkdown¹. Le contenu est pensé pour être accessible à tous et toutes, même à ceux et celles n'ayant presque aucune base en statistique ou en programmation. Les personnes plus expérimentées y découvriront des sections sur des méthodes plus avancées comme les modèles à effets mixtes, les modèles multiniveaux, les modèles généralisés additifs ainsi que les méthodes factorielles et de classification. Ceux et celles souhaitant migrer progressivement d'un autre logiciel statistique vers R trouveront dans cet ouvrage les éléments pour une transition en douceur. La philosophie de ce livre est de donner toutes les clefs de compréhension et de mise en œuvre des méthodes abordées dans R. La présentation des méthodes est basée sur une approche compréhensive et intuitive plutôt que mathématique, sans pour autant que la rigueur statistique ne soit négligée. Servez-vous votre boisson chaude ou froide favorite et installez-vous dans votre meilleur fauteuil. Bonne lecture!

Un manuel sous la forme d'une ressource éducative libre

Pourquoi un manuel de statistique en sciences sociales sous licence libre? Les logiciels libres sont aujourd'hui très répandus. Comparativement aux logiciels propriétaires, l'accès au code source permet à quiconque de l'utiliser, de le modifier, de le dupliquer et de le partager. Le logiciel R, dans lequel sont mises en œuvre les méthodes quantitatives décrites dans ce livre, est d'ailleurs à la fois un langage de programmation et un logiciel libre (sous la licence publique générale GNU GPLv2²). Par analogie aux logiciels libres, il existe aussi des **ressources éducatives libres (REL)** « dont la licence accorde les permissions désignées par les 5R (**Retenir — Réutiliser — Réviser — Remixe — Redistribuer**) et donc permet nécessairement la modification» (*fabriqueREL*³). La licence de ce livre, CC BY-SA (figure ??), permet donc de :

- **Retenir**, c'est-à-dire télécharger et d'imprimer gratuitement le livre. Notez qu'il aurait été plutôt surprenant d'écrire un livre payant sur un logiciel libre et donc gratuit. Aussi, nous aurions été très embarrassés que des personnes étudiantes avec des ressources financières limitées doivent payer pour avoir accès au livre, sans pour autant savoir préalablement si le contenu est réellement adapté à leurs besoins.
- **Réutiliser**, c'est-à-dire utiliser la totalité ou une section du livre sans limitation et sans compensation financière. Cela permet ainsi à d'autres personnes enseignantes de l'utiliser dans le cadre d'activités pédagogiques.
- **Réviser**, c'est-à-dire modifier, adapter et traduire le contenu en fonction d'un besoin pédagogique précis puisqu'aucun manuel n'est parfait, tant s'en faut! Rappelons que le livre a d'ailleurs été écrit

¹<https://rmarkdown.rstudio.com/>

²https://fr.wikipedia.org/wiki/Licence_publique_g%C3%A9n%C3%A9rale_GNU

³<https://fabriquerel.org/rel/>

intégralement dans R avec rmarkdown⁴. Quiconque peut ainsi télécharger gratuitement le code source du livre sur github⁵ et le modifier à sa guise (voir l'encadré intitulé *Suggestions d'adaptation du manuel*).

- **Remixer**, c'est-à-dire « de combiner la ressource avec d'autres ressources dont la licence le permet aussi pour créer une nouvelle ressource intégrée » (*fabriqueREL*⁶).
- **Redistribuer**, c'est-à-dire distribuer en totalité ou partiellement le manuel ou une version révisée sur d'autres canaux que le site Web du livre (par exemple, sur le site Moodle de votre université ou en faire une version imprimée).



Illustration adaptée de *Les licences Creative Commons*, par la fabriqueREL sous licence CC BY.

FIG. 1 : Licence Creative Commons du livre

La licence de ce livre, CC BY-SA (figure ??), oblige donc de :

- Attribuer la paternité de l'auteur dans vos versions dérivées, ainsi qu'une mention concernant les grandes modifications apportées, en utilisant la formulation suivante : Apparicio, Philippe et Jérémie Gelb. 2022. *Méthodes quantitatives en sciences sociales avec R*. Institut national de la recherche scientifique. CC BY-SA (4.0).
- Utiliser la même licence ou une licence similaire à toutes versions dérivées.



Suggestions d'adaptation du manuel.

Notez que pour chaque méthode statistique abordée dans le livre sont disponibles une description détaillée et une mise en œuvre dans R. Par conséquent, plusieurs adaptations du manuel sont possibles :

- Conserver uniquement les chapitres sur les méthodes statistiques ciblées dans votre cours.
- En faire une version imprimée et la distribuer aux personnes étudiantes.
- Modifier la description d'une ou plusieurs méthodes en effectuant les mises à jour directement dans les chapitres.
- Insérer ses propres jeux de données dans les sections intitulées *Mise en œuvre dans R*.
- Modifier les tableaux et figures.
- Ajouter une série d'exercices.
- Rédiger un nouveau chapitre.
- Modifier des syntaxes R. Plusieurs *packages* R peuvent être utilisés pour mettre en œuvre telle ou telle méthode statistique. Ces derniers évoluent aussi très vite et de nouveaux *packages* sont proposés fréquemment! Par conséquent, il peut être judicieux de modifier une syntaxe R du livre en fonction de ses habitudes de programmation dans R (utilisation d'autres *packages* que ceux utilisés dans le manuel

⁴<https://rmarkdown.rstudio.com/>

⁵https://LAEQ.github.io/livre_statistique_Phil_Jere/

⁶<https://fabriquerel.org/rel/>

par exemple) ou de bien mettre à jour une syntaxe à la suite de la parution d'un nouveau *package* plus performant ou intéressant.

- Toute autre adaptation qui permet de répondre au mieux à un besoin pédagogique.

Un manuel conçu comme un projet collaboratif

Il existe actuellement de nombreux livres sous licence ouverte écrits avec rmarkdown⁷ avec le *package* bookdown (Xie 2016), répertoriés sur le site de <https://bookdown.org/>. Sans surprise, R étant un logiciel libre dédié aux méthodes statistiques et à la science des données, plusieurs abordent les méthodes quantitatives, notamment :

- Beyond Multiple Linear Regression : Applied Generalized Linear Models and Multilevel Models in R⁸ (Roback et Legler 2021), CC BY-NC-SA.
- Introduction to Econometrics with R⁹ (Handk et al. 2019), CC BY-NC-SA.
- Statistical Inference via Data Science : A ModernDive into R and the Tidyverse¹⁰ (Ismay et Kim 2019), CC BY-NC-SA.
- R Graphics Cookbook, 2nd edition¹¹ (Chang 2018), CC BY.

Par contre, la grande majorité de ces livres numériques rédigés avec bookdown sont en anglais. À notre connaissance, ce projet constitue le premier manuel numérique en français sur les méthodes quantitatives appliquées aux sciences sociales réalisé avec bookdown. La première version du livre étant lancée, il est grand temps de planifier les suivantes! Pour ce faire, nous considérons ce livre comme un **projet collaboratif visant à mobiliser la communauté universitaire francophone qui enseigne les statistiques en sciences sociales avec R**. Plusieurs raisons motivent cette vision collaborative :

- **Rien n'est parfait!** Cette première version comprend sûrement des coquilles et certaines sections mériteraient d'être améliorées. Les commentaires et suggestions visant à améliorer son contenu sont les bienvenus.
- **La table des matières doit être impérativement extensible!** De nombreuses méthodes statistiques très utilisées en sciences sociales ne sont pas abordées dans ce livre et mériteraient d'être ajoutées dans une version ultérieure : certaines extensions des régressions linéaires (régressions Rigge et Lasso, Tobit, quantile, etc.), les modèles d'équations simultanées, les analyses de données longitudinales (entre autres, modèles de survie, régression par panel), les modèles d'équations structurelles et bien d'autres! Par conséquent, si vous êtes intéressé(e)s, à ajouter un nouveau chapitre ou une partie du livre, nous vous invitons vivement à communiquer avec nous ou à diffuser sous une licence similaire votre version dérivée. L'objectif étant de continuer à faire tourner la roue du libre et, idéalement, que les futures versions soient corédigées par une communauté d'auteurs et d'autrices spécialistes en méthodes quantitatives.

Comment lire ce livre ?

Si vous googlez l'expression « comment lire un livre? », vous trouverez une multitude de conseils et astuces. Pour ce livre, nous vous conseillons de le lire de gauche à droite et page par page! Plus sérieusement, le livre comprend plusieurs types de blocs de texte qui, nous l'espérons, en facilitent la lecture.

⁷<https://rmarkdown.rstudio.com/>

⁸<https://bookdown.org/robact/bookdown-BeyondMLR/>

⁹<https://www.econometrics-with-r.org/>

¹⁰<https://moderndive.com/>

¹¹<https://r-graphics.org/>



Bloc packages. Habituellement localisé au début d'un chapitre, il comprend la liste des *packages R* utilisés pour un chapitre.



Bloc objectif. Il comprend une description des objectifs d'un chapitre ou d'une section.



Bloc notes. Il comprend une information secondaire sur une notion, un élément, une idée abordée dans une section.



Bloc pour aller plus loin. Il comprend des références ou des extensions d'une méthode statistique abordée dans une section.



Bloc astuce. Il décrit un élément qui vous facilitera la vie : une propriété statistique, un *package*, une fonction, une syntaxe R.



Bloc attention. Il comprend une notion ou un élément important à bien maîtriser.

Comment utiliser les données du livre pour reproduire les exemples ?

Ce livre propose des exemples détaillés et appliqués dans R pour chacune des méthodes abordées. Ces exemples se basent sur des jeux de données structurés et mis à disposition avec le livre. Ils sont disponibles sur le *repo github* dans le sous-dossier `data`, à l'adresse https://github.com/LAEQ/livre_statistique_Phil_Jere/tree/master/data.

Pour télécharger l'intégralité des données, vous pouvez utiliser le lien suivant : https://downgit.github.io/#/home?url=https://github.com/LAEQ/livre_statistique_Phil_Jere/tree/master/data. Cela est rendu possible grâce à l'outil DownGit¹².

Une autre option est de télécharger le *repo* complet du livre directement sur *github* (https://github.com/LAEQ/livre_statistique_Phil_Jere) en cliquant sur le bouton `Code`, puis le bouton `Download ZIP` (figure ??). Les données se trouvent alors dans le sous-dossier nommé `data`.

Structure du livre

Le livre est organisé autour de cinq grandes parties.

Partie 1. La découverte de R. Dans cette première partie, nous discutons brièvement de l'histoire et de la philosophie de R. Nous voyons ensuite comment installer R et RStudio. Les bases du langage R (particulièrement les principaux objets que sont le vecteur, la matrice, la liste et le *dataframe*) ainsi que la manipulation des données avec R sont aussi largement abordés dans le chapitre ??.

Partie 2. Analyses univariées et représentations graphiques. Cette seconde partie comprend deux chapitres. Dans le chapitre ??, nous décrivons dans un premier temps les différents types de données (primaires *versus* secondaires, transversales *versus* longitudinales, spatiales *versus* aspatiales, individuelles

¹²<https://downgit.github.io/#/home>

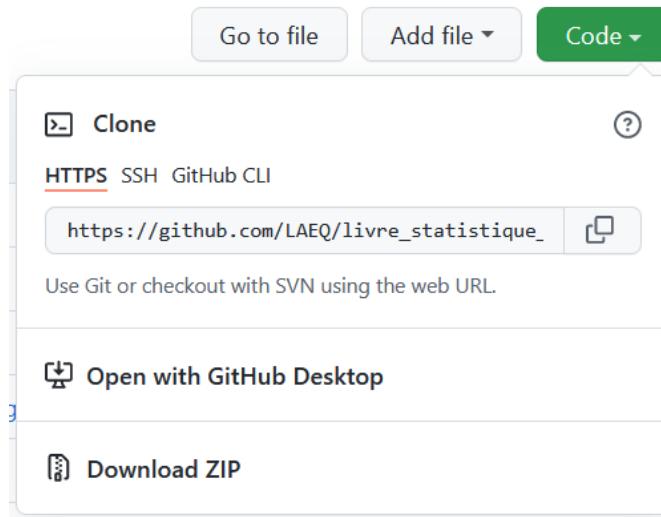


FIG. 2 : Téléchargement de l'intégralité du livre

versus agrégées), les différents types de variables quantitatives (discrètes et continues) et qualitatives (nominales et ordinaires) et les principales distributions de variables utilisées en sciences sociales (uniforme, Bernoulli, binomiale, géométrique, binomiale négative, poisson, poisson avec excès de zéros, gaussienne, gaussienne asymétrique, log-normale, Student, Cauchy, Chi-carré, exponentielle, gamma, bêta, Weibull et Pareto). Dans un second temps, nous abordons les statistiques descriptives pour des variables quantitatives (paramètres de tendance centrale, paramètres de position, paramètres de dispersion, paramètres de forme), puis qualitatives (fréquences absolues, relatives et cumulées).

Dans le chapitre ??, nous illustrons les incroyables capacités graphiques de R en mettant en œuvre les principaux graphiques (histogramme, graphique de densité, nuage de points, graphique en lignes, boîtes à moustache, graphique en violon, graphique en barre, graphique circulaire), quelques graphiques particuliers (graphique en radar, diagramme d'accord, nuage de mots, carte proportionnelle) et une initiation aux cartes choroplèthes.

Partie 3. Analyses bivariées. Cette troisième partie comprend trois chapitres dans lesquelles sont présentées les principales méthodes exploratoires et confirmatoires bivariées permettant d'évaluer la relation entre deux variables. Plus spécifiquement, nous présentons puis mettons en œuvre dans R les méthodes permettant d'explorer les relations entre deux variables quantitatives (covariance, corrélation et régression linéaire simple) dans le chapitre ??, deux variables qualitatives (tableau de contingence et test du khi-deux) dans le chapitre ?? et une variable quantitative avec une variable qualitative avec deux modalités (tests de Student, de Welch et de Wilcoxon) ou avec plus de deux modalités (ANOVA et test de Kruskal-Wallis) dans le chapitre ??.

Partie 4. Modèles de régression. Dans cette quatrième partie, sont présentées les principales méthodes de statistique inférentielle utilisées en sciences sociales : la régression linéaire multiple (chapitre ??), les régressions linéaires généralisées (chapitre ??), les régressions à effets mixtes (chapitre ??), les régressions à effets mixtes (chapitre ??), les régressions multiniveaux (chapitre ??) et les modèles généralisés additifs (chapitre ??).

Partie 5. Analyses exploratoires multivariées. Dans cette cinquième partie, sont abordées les méthodes de statistique exploratoire et descriptive permettant de décrire des tableaux de données comprenant plusieurs variables. Nous décrivons d'abord les méthodes de réduction de données : les méthodes factorielles dans le chapitre ?? (analyses de composantes principales, analyses factorielles de correspondances, analyses factorielles de correspondances multiples) et les méthodes de classification non supervisées dans le

chapitre ?? (classification ascendante hiérarchique, k-moyennes, k-médianes, k-médoïdes et leurs extensions en logique floue comme les c-moyennes et c-médianes).

Pourquoi faut-il programmer en sciences sociales ?

Vous contrasterez rapidement que R est un véritable langage de programmation. L'apprentissage de ce language de programmation est-il pour autant pertinent pour les étudiants et étudiantes en sciences sociales ? Il est vrai que la programmation n'est pas une compétence qui vient d'emblée à l'esprit lorsque l'on s'intéresse à la recherche aux sciences sociales. Pourtant, elle est de plus en plus importante, et ce, pour plusieurs raisons :

- Une part toujours plus grande des phénomènes sociaux se produisent ou peuvent s'observer au travers d'environnements numériques. Être capable d'exploiter efficacement ces outils permet d'extraire des données riches sur des phénomènes complexes, tel qu'en témoignent des études récentes sur la propagation de la désinformation sur les réseaux sociaux (Allcott et Gentzkow 2017), la migration des personnes (Spryatos et al. 2019), la propagation et les risques de contamination de la COVID-19 (Boulos et Geraghty 2020). Le plus souvent, les interfaces (API par exemple) permettant d'accéder à ces données nécessitent des habiletés en programmation.
- La quantité de données numériques ouvertes et accessibles en ligne croît chaque année sur des sujets très divers. La plupart des villes et des gouvernements ont maintenant leur portail de données ouvertes auxquelles s'ajoutent les données produites par des projets collaboratifs comme OpenStreetMap¹³ ou NoisePlanet¹⁴. Récupérer ces données et les structurer pour les utiliser à des fins de recherche nécessitent le plus souvent des compétences en programmation.
- Les méthodes quantitatives connaissent également un développement très important. Les logiciels propriétaires peinent à suivre la cadence de ce développement, contrairement aux logiciels à code source ouvert (comme R) qui permettent d'avoir accès aux dernières méthodes. Il est souvent long et coûteux de développer une interface graphique pour un logiciel, ce qui explique que la plupart de ces programmes en sont dépourvus et nécessitent alors de savoir programmer pour les utiliser.
- Savoir programmer donne une liberté considérable en recherche. Cette compétence permet notamment de ne plus être limité(e) aux fonctionnalités proposées par des logiciels spécifiques. Il devient possible d'innover tant en matière de structuration, d'exploration et d'analyse des données que de représentation des résultats en écrivant ses propres fonctions. Cette flexibilité contribue directement à la production d'une recherche de meilleure qualité et plus diversifiée.
- Programmer permet également d'automatiser des tâches qui autrement seraient extrêmement répétitives comme : déplacer et renommer une centaine de fichiers ; retirer les lignes inutiles dans un ensemble de fichiers et les compiler dans une seule base de données ; vérifier parmi des milliers d'adresses lesquelles sont valides ; récupérer chaque jour les messages postés sur un forum. Autant de tâches faciles à automatiser si l'on sait programmer.
- Dans un logiciel avec une interface graphique, il est compliqué de conserver un historique des opérations effectuées. Programmer permet au contraire de garder une trace de l'ensemble des actions effectuées au cours d'un projet de recherche. En effet, le code utilisé reste disponible et permet de reproduire (ou d'adapter) la méthode et les résultats obtenus, ce qui est essentiel dans le monde de la recherche. À cela s'ajoute le fait que chaque ligne de code que vous écrivez vient s'ajouter à un capital de code que vous possédez, car elles pourront être réutilisées dans d'autres projets !

¹³<https://www.openstreetmap.org>

¹⁴https://noise-planet.org/map_noisecapture/index.html

Remerciements

De nombreuses personnes ont contribué à l'élaboration de ce manuel. Ce projet a bénéficié du soutien pédagogique et financier de la *fabriqueREL*¹⁵ (ressources éducatives libres). Les différentes rencontres avec le comité de suivi nous ont permis de comprendre l'univers des ressources éducatives libres (REL) et notamment leurs fameux 5R¹⁶ (Retenir — Réutiliser — Réviser — Remixer — Redistribuer), de mieux définir le besoin pédagogique visé par ce manuel, d'identifier des outils et des ressources pédagogiques pertinents pour son élaboration. Ainsi, nous remercions chaleureusement les membres de suivi de la *fabriqueREL* pour leur support inconditionnel :

- Myriam Beaudet, bibliothécaire à l'Université de Sherbrooke.
- Marianne Dubé, conseillère pédagogique à l'Université de Sherbrooke et coordonnatrice de la *fabriqueREL*.
- Myrian Grondin, bibliothécaire à l'Institut national de la recherche scientifique (INRS).
- Claude Potvin, conseiller en formation à l'Université Laval.
- Serge Allary, vice-recteur adjoint aux études de l'Université de Sherbrooke.

Nous tenons aussi à remercier sincèrement les étudiants et étudiantes du cours **Méthodes quantitatives appliquées aux études urbaines (EUR8219)** du programme de maîtrise en études urbaines de l'INRS. Leurs commentaires et suggestions nous ont permis d'améliorer grandement les versions préliminaires de ce manuel qui ont été utilisées dans le cadre de ce cours.

Nous remercions les membres du comité de révision pour leurs commentaires et suggestions très constructifs. Ce comité est composé de trois étudiantes et deux professeurs de l'INRS¹⁷ :

- Victoria Gay-Gauvin, étudiante à la maîtrise en études urbaines.
- Salomé Vallette, étudiante au doctorat en études urbaines.
- Diana Pena Ruiz, étudiante au doctorat en études des populations.
- Benoît Laplante¹⁸, professeur enseignant aux programmes de maîtrise et de doctorat en études des populations.
- Xavier Leloup¹⁹, professeur enseignant au programme de doctorat en études urbaines.

Finalement, nous remercions Denise Latreille, réviseure linguistique et chargée de cours à l'Université Sherbrooke, pour la révision du manuel.

Dédicace toute spéciale à Cargo et Ambrée

Fait cocasse, l'écriture de ce livre a démarré lorsque Philippe Apparicio était famille d'accueil d'un chiot de la Fondation Mira²⁰, un organisme à but non lucratif qui forme des chiens-guides et d'assistance pour accroître l'autonomie et l'inclusion sociale des personnes vivant avec un handicap visuel ou moteur, ainsi que des jeunes présentant un trouble du spectre de l'autisme (TSA). En fin de rédaction du livre, ce fût au tour de Jérémy Gelb d'être famille d'accueil d'un autre chiot Mira. Nous remercions chaleureusement la Fondation Mira²¹ pour nous avoir donné l'occasion de vivre cette expérience incroyable. Ce livre est donc dédié au beau Cargo et à la belle Ambrée qui nous ont tant supportés dans l'écriture du livre. Il n'y a rien de plus relaxant que d'écrire un livre de statistique avec un chiot qui dort à ses pieds!

¹⁵ <https://fabriquerel.org/>

¹⁶ <https://fabriquerel.org/rel/>

¹⁷ <https://inrs.ca/>

¹⁸ <https://inrs.ca/la-recherche/professeurs/benoit-laplante/>

¹⁹ <https://inrs.ca/la-recherche/professeurs/xavier-leloup/>

²⁰ <https://www.mira.ca/fr/>

²¹ <https://www.mira.ca/fr/>

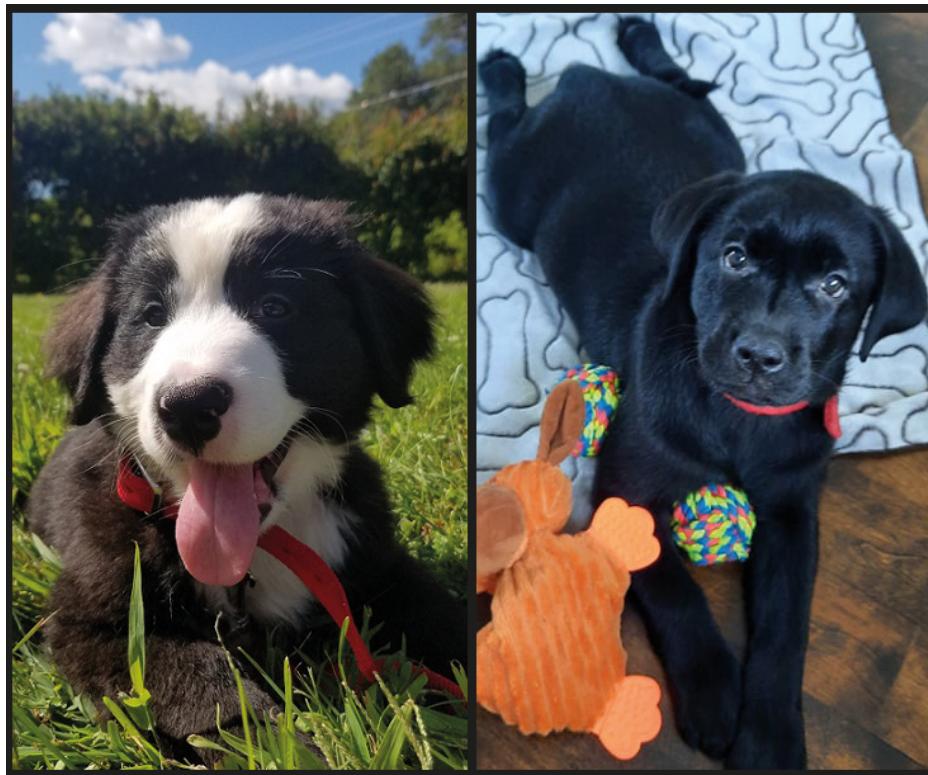


FIG. 3 : Cargo et Ambrée, chiots de la Fondation Mira

Première partie

Analyses exploratoires multivariées

Chapitre 1

Méthodes factorielles



Ce chapitre n'a pas encore fait l'objet d'une révision linguistique. Il comprend certainement plusieurs coquilles... Une version révisée sera mise à jour prochainement.

Dans le cadre de ce chapitre, nous présentons les trois méthodes factorielles les plus utilisées en sciences sociales : l'analyse en composantes principales (ACP, section ??), l'analyse factorielle des correspondances (AFC, section ??) et l'analyse factorielle des correspondances multiples (ACM, section ??). Ces méthodes qui permettent d'explorer et de synthétiser l'information de différents tableaux de données relèvent de la statistique exploratoire multidimensionnelle.



Dans ce chapitre, nous utilisons principalement les *packages* suivants :

- Pour créer des graphiques :
 - * `ggplot2`, le seul, l'unique!
 - * `ggpubr` pour combiner des graphiques.
- Pour les analyses factorielles :
 - * `FactoMineR` pour réaliser des ACP, AFC et ACM.
 - * `factoextra` pour réaliser des graphiques à partir des résultats d'une analyse factorielle.
 - * `explor` pour les résultats d'une ACP, d'une AFC ou d'une ACM avec une interface Web interactive.
- Autre package :
 - * `geocmeans` pour un jeu de données utilisé pour calculer une ACP.
 - * `ggplot2`, `ggpubr`, `stringr` et `corrplot` pour réaliser des graphiques personnalisés sur les résultats d'une analyse factorielle.
 - * `tmap` et `RColorBrewer` pour cartographier les coordonnées factorielles.
 - * `Hmisc` pour l'obtention d'une matrice de corrélation.



Réduction de données et identification de variables latentes

Les méthodes factorielles sont souvent dénommées des **méthodes de réduction de données**, en raison de leur objectif principal, à savoir résumer l'information d'un tableau en de nouvelles variables synthétiques (figure ??). Ainsi, elles permettent de réduire l'information d'un tableau volumineux — comprenant par exemple 1000 observations et 100 variables — en p nouvelles variables (par exemple cinq avec toujours 1000 observations) résumant X % de l'information contenue dans le tableau initial. Formulée plus mathématiquement, Lebart et al. (1995, pp. 13) signalent qu'avec les méthodes factorielles, « on cherche à réduire les dimensions du tableau de données en représentant les associations entre individus et entre variables dans des espaces de faibles dimensions ».

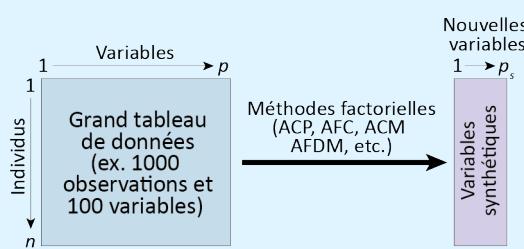


FIG. 1.1 : Principe de base des analyses factorielles

Ces nouvelles variables synthétiques peuvent être considérées comme des **variables latentes** puisqu'elles ne sont directement observées, mais plutôt produites par la méthode factorielle utilisée afin de résumer les relations/associations entre plusieurs variables mesurées initialement.

1.1 Aperçu des méthodes factorielles

1.1.1 Méthodes factorielles et types de données

En analyse factorielle, la nature même des données du tableau à traiter détermine la méthode à employer : l'analyse en composantes principales (ACP) est adaptée aux tableaux avec des variables continues (idéalement normalement distribuées), l'analyse factorielle des correspondances (AFC) s'applique à des tableaux de contingence tandis que l'analyse des correspondances multiples (ACM) permet de résumer des tableaux avec des données qualitatives (issues d'un sondage par exemple) (tableau ??). Sachez toutefois qu'il existe d'autres méthodes factorielles qui ne sont pas abordées dans ce chapitre, notamment : l'analyse factorielle de données mixtes (AFDM) permettant d'explorer des tableaux avec à la fois des variables continues et des variables qualitatives, l'analyse factorielle multiple hiérarchique (AFMH) permettant de traiter des tableaux avec une structure hiérarchique. Pour s'initier à ces deux autres méthodes factorielles plus récentes, consultez notamment l'excellent ouvrage de Jérôme Pagès (2013).

1.1.2 Bref historique des méthodes factorielles

Il existe une longue tradition de l'utilisation des méthodes factorielles dans le monde universitaire francophone puisque plusieurs d'entre elles ont été proposées par des statisticiens et des statisticiennes francophones à partir des années 1960. L'analyse en composantes principales (ACP) a été proposée dès les années 1930 par le statisticien américain Harold Hotelling (1933). En revanche, l'analyse des correspondances (AFC) et son extension (l'analyse des correspondances multiples, ACM) ont été proposées par le statisticien français Jean-Paul Benzécri (1973), tandis que l'analyse factorielle de données mixtes (AFDM) a été proposée par Brigitte Escofier et Jérôme Pagès (Escofier 1979; Pagès 2002).

Ainsi, plusieurs ouvrages de statistique sur les méthodes factorielles, désormais classiques, ont été publiés en français (Benzécri 1973; Escofier et Pagès 1998; Lebart, Morineau et Piron 1995; Pagès 2013).

TAB. 1.1 : Trois principales méthodes factorielles

Méthode factorielle	Abr.	Type de données	Type de distance
Analyse en composantes principales	ACP	Variables continues	Distance euclidienne
Analyse factorielle des correspondances	AFC	Tableau de contingence	Distance du khi-deux
Analyse factorielle des correspondances multiples	ACM	Variables qualitatives	Distance du khi-deux

Ils méritent grandement d'être consultés, notamment pour mieux comprendre les formulations mathématiques (matricielles et géométriques) de ces méthodes. À cela, s'ajoutent plusieurs ouvrages visant à « vulgariser ces méthodes » en sciences sociales ; c'est notamment le cas de l'excellent ouvrage de Léna Sanders (1989) en géographie.

1.2 Analyses en composantes principales (ACP)

D'emblée, notez qu'il existe deux types d'analyse en composantes principales (ACP) (*Principal Component Analysis, PCA* en anglais) :

- **l'ACP non normée** dans laquelle les variables quantitatives du tableau sont uniquement centrées (moyenne = 0).
- **l'ACP normée** dans laquelle les variables quantitatives du tableau sont préalablement centrées réduites (moyenne = 0 et variance = 1 ; section ??).

Puisque les variables d'un tableau sont souvent exprimées dans des unités de mesure différentes ou avec des ordres de grandeur différents (intervalles et écarts-types bien différents), l'utilisation de l'ACP normée est bien plus courante. Elle est d'ailleurs l'option par défaut dans les fonctions R permettant de calculer une ACP. Par conséquent, nous détaillons dans cette section uniquement l'ACP normée.

Autrement dit, le recours à une ACP non normée est plus rare et s'applique uniquement à la situation suivante : toutes les variables du tableau sont mesurées dans la même unité (par exemple, en pourcentage) ; il pourrait être ainsi judicieux de conserver leurs variances respectives.

1.2.1 Recherche d'une simplification

L'ACP permet d'explorer et de résumer un tableau constitué uniquement de variables quantitatives (figure ??), et ce, de trois façons : 1) en montrant les ressemblances entre les individus (observations), 2) en révélant les liaisons entre les variables quantitatives et 3) en résumant l'ensemble des variables du tableau par des variables synthétiques nommées composantes principales.

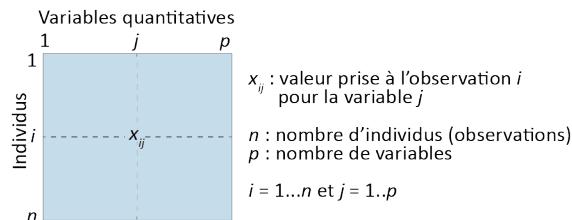


FIG. 1.2 : Tableau pour une ACP

Ressemblance entre les individus. Concrètement deux individus se ressemblent si leurs valeurs respectives pour les p variables du tableau sont similaires. Cette proximité/ressemblance est évaluée à partir de la distance euclidienne (eq. (??)). La notion de distance fait l'objet d'une section à part entière (section ??) que vous pouvez consulter dès à présent si elle ne vous est pas familière.

$$d^2(a, b) = \sum_{j=1}^p (x_{aj} - x_{bj})^2 \quad (1.1)$$

Prenons un exemple fictif avec trois individus (i , j et k) ayant des valeurs pour trois variables préalablement centrées réduites (V1 à V3) (tableau ??). La proximité entre les paires de points est évaluée comme suit :

$$d^2(i, j) = (-1,15 - 0,49)^2 + (-1,15 - 0,58)^2 + (0,83 + 1,11)^2 = 9,44 \quad d^2(i, k) = (-1,15 + 0,66)^2 + (-1,15 - 0,58)^2 + (0,83 - 0,28)^2 = 5,98 \quad d^2(j, k) = (0,49 + 0,66)^2 + (0,58 - 0,58)^2 + (-1,11 - 0,28)^2 = 1,97$$

Nous pouvons en conclure que i est plus proche de k que de j , mais aussi que la paire de points les plus proches est (i, k) . En d'autres termes, les deux observations i et k sont les plus similaires du jeu de données selon la distance euclidienne.

Liaisons entre les variables. Dans une ACP normée, les liaisons entre les variables deux à deux sont évaluées avec le coefficient de corrélation (section ??), soit la moyenne du produit des deux variables centrées réduites (eq. (??)). Notez que dans une ACP non normée, plus rarement utilisée, les liaisons sont alors évaluées avec la covariance puisque les variables sont uniquement centrées (eq. (??)).

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \sum_{i=1}^n \frac{Zx_i Zy_i}{n} \quad (1.2)$$

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} \quad (1.3)$$

Composantes principales. Au chapitre 4, nous avons abordé deux méthodes pour identifier des relations linéaires entre des variables continues normalement distribuées :

- la corrélation de Pearson (section ??), qu'il est possible d'illustrer graphiquement à partir d'un nuage de points ;
- la régression linéaire simple (section ??), permettant de résumer la relation linéaire entre deux variables avec une droite de régression de type $Y = a + bX$.

Brièvement, plus deux variables sont corrélées (positivement ou négativement), plus le nuage de points qu'elles forment est allongé et plus les points sont proches de la droite de régression (figure ??, partie a). À l'inverse, plus la liaison entre les deux variables normalement distribuées est faible, plus le nuage prend la forme d'un cercle et plus les points du nuage sont éloignés de la droite de régression (figure ??, partie b). Puisqu'en ACP normée, les variables sont centrées réduites, le centre de gravité du nuage de points est $(x=0, y=0)$ et il est toujours traversé par la droite de régression. Finalement, nous avons vu que la méthode des moindres carrés ordinaires (MCO) permet de déterminer cette droite en minimisant les distances entre les valeurs observées et celles projetées orthogonalement sur cette droite (valeurs prédictes). Dans le cas de deux variables uniquement, l'axe factoriel principal/la composante principale est donc la droite qui résume le mieux la liaison entre les deux variables (en rouge). L'axe 2 représente la seconde plus importante composante (axe, dimension) et il est orthogonal (perpendiculaire) au premier axe.

Imaginez maintenant trois variables pour lesquelles vous désirez identifier un axe, une droite qui résume le mieux les liaisons entre elles. Visuellement, vous passez d'un nuage de points en deux dimensions (2D) à trois dimensions (3D). Si les corrélations entre les trois variables sont très faibles, alors le nuage prendra la forme d'un ballon de football (soccer en Amérique du Nord). Par contre, plus ces liaisons seront fortes,

TAB. 1.2 : Données fictives

Individu	Variables centrées réduites		
	V1	V2	V3
i	-1,15	-1,15	0,83
j	0,49	0,58	-1,11
k	0,66	0,58	0,28

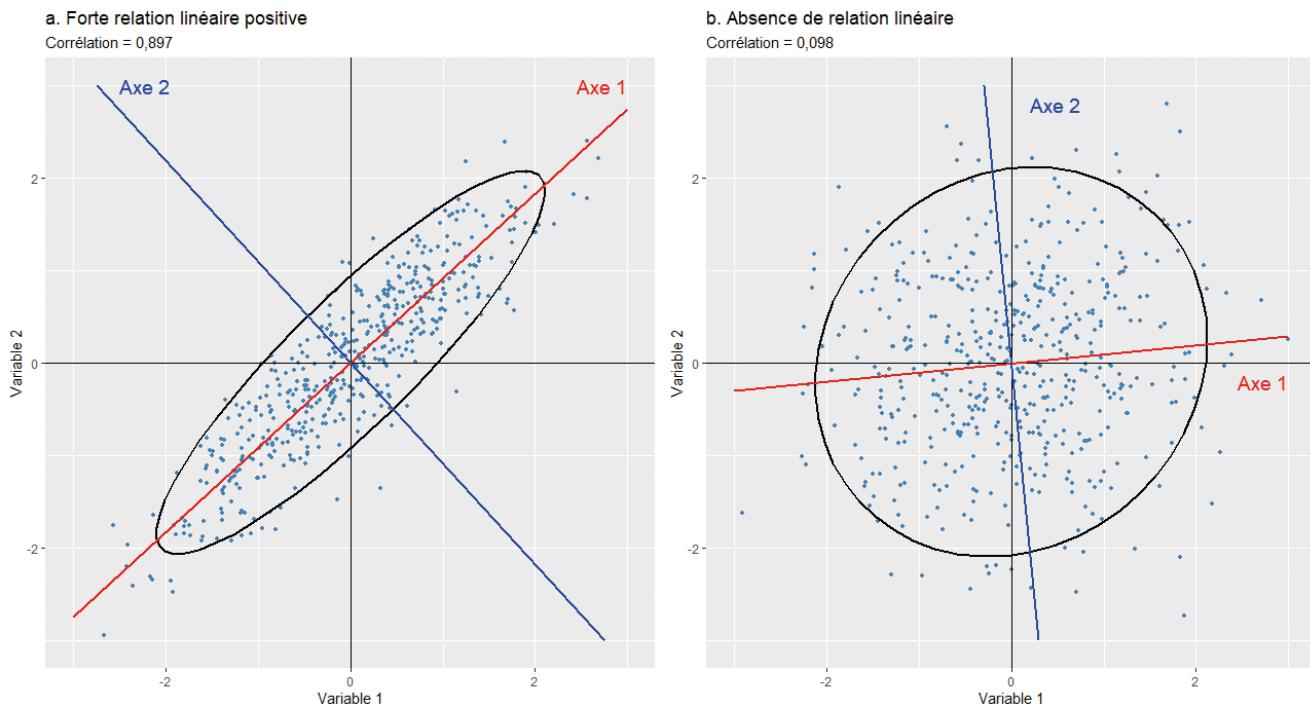


FIG. 1.3 : Corrélation, allongement du nuage de points et axe factoriel

plus la forme sera allongée telle celle d'un ballon de rugby (ou football américain) et plus les points seront proches de l'axe traversant le ballon.

Ajouter une autre variable revient alors à ajouter une quatrième dimension qu'il est impossible de visualiser, même pour les plus fervents adeptes de science-fiction. Pourtant le problème reste le même, identifier dans un plan en p dimensions (variables), les axes factoriels, les composantes principales qui concourent le plus à résumer liaisons entre les variables continues préalablement centrées réduites, et ce, en utilisation la méthode des moindres carrés ordinaires.



Les termes **composantes principales** et **axes factoriels** sont des synonymes employés pour référer aux nouvelles variables synthétiques produites par l'ACP et résumant l'information du tableau initial.

1.2.2 Aides à l'interprétation

Pour illustrer les aides à l'interprétation de l'ACP, nous utilisons un jeu de données spatiales tiré d'un article sur l'agglomération lyonnaise en France (Gelb et Apparicio 2021). Ce jeu de données comprend dix variables, dont quatre environnementales (EN) et six socioéconomiques (SE), pour les îlots regroupés pour l'information statistique (IRIS) de l'agglomération lyonnaise (tableau ?? et figure ??). Sur ces dix variables, nous calculons une **ACP normée**.

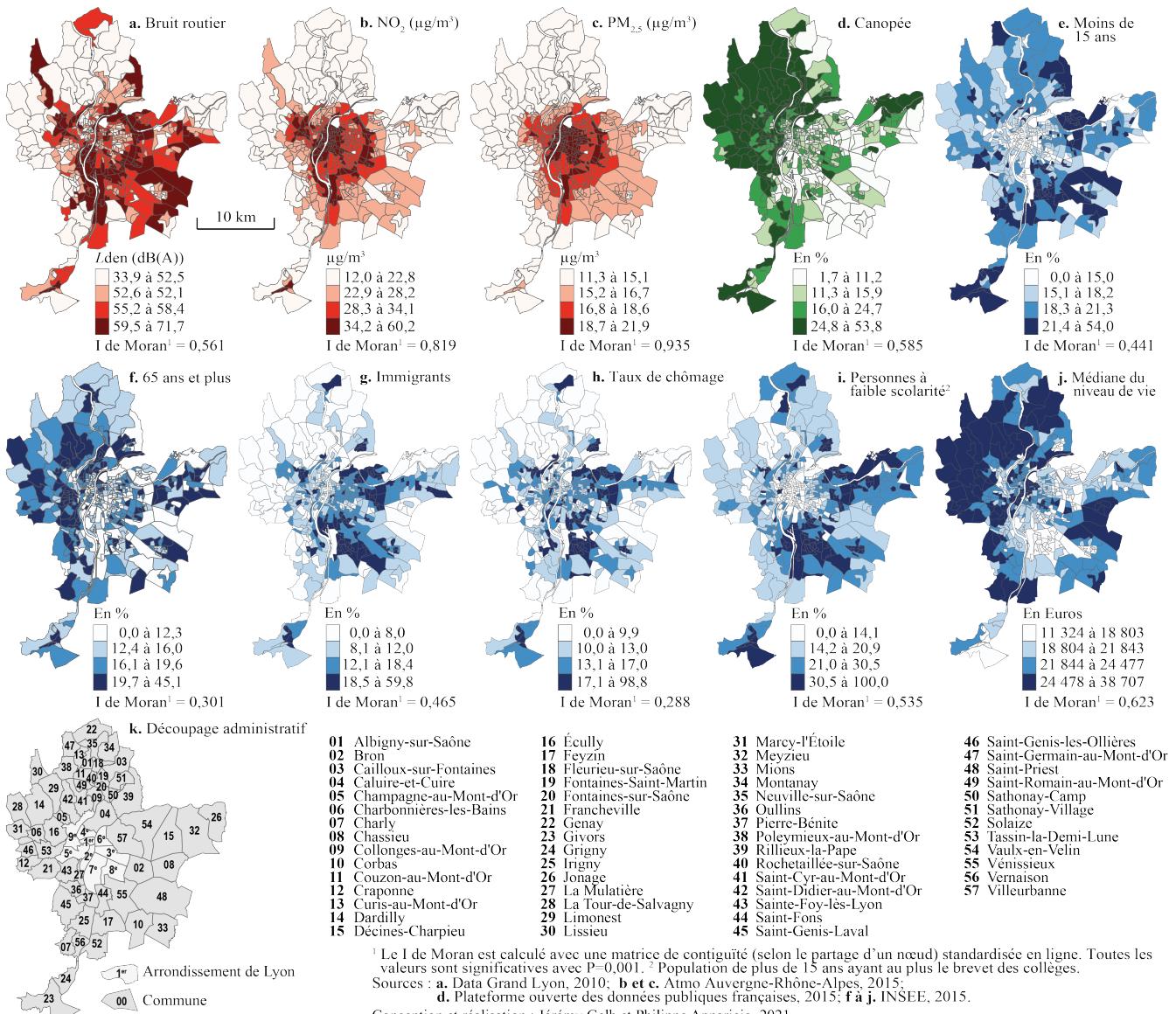


Trois étapes pour bien analyser une ACP et comprendre la signification des axes factoriels :

1. Interprétation des résultats des valeurs propres pour identifier le nombre d'axes (de composantes principales) à retenir. L'enjeu est de garder un nombre d'axes limité qui résume le mieux le tableau initial (réduction des données).
2. Analyse des résultats pour les variables (coordonnées factorielles, cosinus carrés et contributions sur les axes retenus).

Tab. 1.3 : Statistiques descriptives pour le jeu de données utilisé pour l'ACP

Nom	Intitulé	Type	Moy.	E.-T.	Min.	Max.
Lden	Bruit routier (Lden dB(A))	EN	55,6	4,9	33,9	71,7
NO2	Dioxyde d'azote ($\mu\text{g}/\text{m}^3$)	EN	28,7	7,9	12,0	60,2
PM25	Particules fines ($\text{PM}_{2,5}$)	EN	16,8	2,1	11,3	21,9
VegHautPrt	Canopée (%)	EN	18,7	10,1	1,7	53,8
Pct_14	Moins de 15 ans (%)	SE	18,5	5,7	0,0	54,0
Pct_65	65 ans et plus (%)	SE	16,2	5,9	0,0	45,1
Pct_Img	Immigrants (%)	SE	14,5	9,1	0,0	59,8
TxChom1564	Taux de chômage	SE	14,8	8,1	0,0	98,8
Pct_brevet	Personnes à faible scolarité (%)	SE	23,5	12,6	0,0	100,0
NivVieMed	Médiane du niveau de vie (Euros)	SE	21 804,5	4 922,5	11 324,0	38 707,0

**Fig. 1.4 : Cartographie des dix variables utilisées pour l'ACP**

3. Analyse des résultats pour les individus (coordonnées factorielles, cosinus carrés et contributions sur les axes retenus).

Les deux dernières étapes permettent de comprendre la signification des axes retenus et de les qualifier. Cette étape d'interprétation est essentielle en sciences sociales. En effet, nous avons vu dans l'introduction du chapitre que les méthodes factorielles permettent de résumer l'information d'un tableau en quelques nouvelles variables synthétiques, souvent considérées comme des variables latentes dans le jeu de données. Il convient alors de bien comprendre ces variables synthétiques (latentes) si nous souhaitons les utiliser dans une autre analyse subséquente (par exemple, les introduire dans une régression).

1.2.2.1 Résultats de l'ACP pour les valeurs propres

À titre de rappel, une ACP normée est réalisée sur des variables préalablement centrées réduites (équation (??)), ce qui signifie que pour chaque variable :

- Nous soustrayons à chaque valeur la moyenne de la variable correspondante (centrage); la moyenne est donc égale à 0.
- Nous divisons cette différence par l'écart-type de la variable correspondante (réduction); la variance est égale à 1.

$$z = \frac{x_i - \mu}{\sigma} \quad (1.4)$$

Par conséquent, la variance totale (ou inertie totale) d'un tableau sur lequel est calculée une ACP normée est égale au nombre de variables qu'il comprend. Puisque nous l'appliquons ici à dix variables, la variance totale du tableau à réduire – c'est-à-dire à résumer en K nouvelles variables synthétiques, composantes principales, axes factoriels – est donc égale à 10. Trois mesures reportées au tableau ?? permettent d'analyser les valeurs propres :

- VP_k , la valeur propre (*eigenvalue* en anglais) de l'axe k c'est-à-dire la quantité de variance du tableau initial résumé par l'axe.
- VP_k/P avec P étant le nombre de variables que comprend le tableau initial. Cette mesure représente ainsi le pourcentage de la variance totale du tableau résumé par l'axe k , autrement dit, la quantité d'informations du tableau initial résumée par l'axe, la composante principale k . Cela nous permet ainsi d'évaluer le pouvoir explicatif de l'axe.
- Le pourcentage cumulé pour les axes.

Avant d'analyser en détail le tableau ??, notez que la somme des valeurs propres de toutes les composantes de l'ACP est toujours égale au nombre de variables du tableau initial. Aussi, la quantité de variance

TAB. 1.4 : Résultats de l'ACP pour les valeurs propres

Composante	Valeur propre	Pourcentage	Pourc. cumulé
1	3,543	35,425	35,425
2	2,760	27,596	63,021
3	1,042	10,422	73,443
4	0,751	7,511	80,954
5	0,606	6,059	87,013
6	0,388	3,880	90,893
7	0,379	3,788	94,681
8	0,244	2,441	97,122
9	0,217	2,167	99,289
10	0,071	0,711	100,000

expliquée (les valeurs propres) décroît de la composante 1 à la composante K .

Combien d'axes d'une ACP faut-il retenir? Pour ce faire, deux approches sont possibles :

- **Approche statistique** (avec le critère de Kaiser (1960)). Nous retenons uniquement les composantes qui présentent une valeur propre supérieure à 1. Rappelez-vous qu'en ACP normée, les variables sont préalablement centrées réduites et donc que leur variance respective est égale à 1. Par conséquent, une composante ayant une valeur propre inférieure à 1 a un pouvoir explicatif inférieur à celui d'une variable. À la lecture du tableau, nous retenons les trois premières composantes si nous appliquons ce critère.
- **Approche empirique** basée sur la lecture des pourcentages et des pourcentages cumulés. Nous pourrons retenir uniquement les deux premières composantes. En effet, ces deux premiers facteurs résument près des deux tiers de la variance totale du tableau (63,02 %). Cela démontre bien que l'ACP, comme les autres méthodes factorielles, est bien une méthode de réduction de données puisque nous résumons dix variables avec deux nouvelles variables synthétiques (axes, composantes principales). Pour faciliter le choix du nombre d'axes, il est fortement conseillé de construire des histogrammes à partir des valeurs propres, des pourcentages et des pourcentages cumulés (figure ??). Or, à la lecture de ces graphiques, nous constatons que la variance expliquée chute drastiquement après les deux premières composantes. Par conséquent, nous pouvons retenir uniquement les deux premiers axes.



Lecture du diagramme des valeurs propres

Plus les variables incluses dans l'ACP sont corrélées entre elles, plus l'ACP sera intéressante : plus les valeurs propres des premiers axes sont fortes et plus il y a des sauts importants dans le diagramme des valeurs propres. À l'inverse, lorsque les variables incluses dans l'ACP sont peu corrélées entre elles, il n'y aura pas de sauts importants dans l'histogramme, autrement dit, les valeurs propres sont uniformément décroissantes.

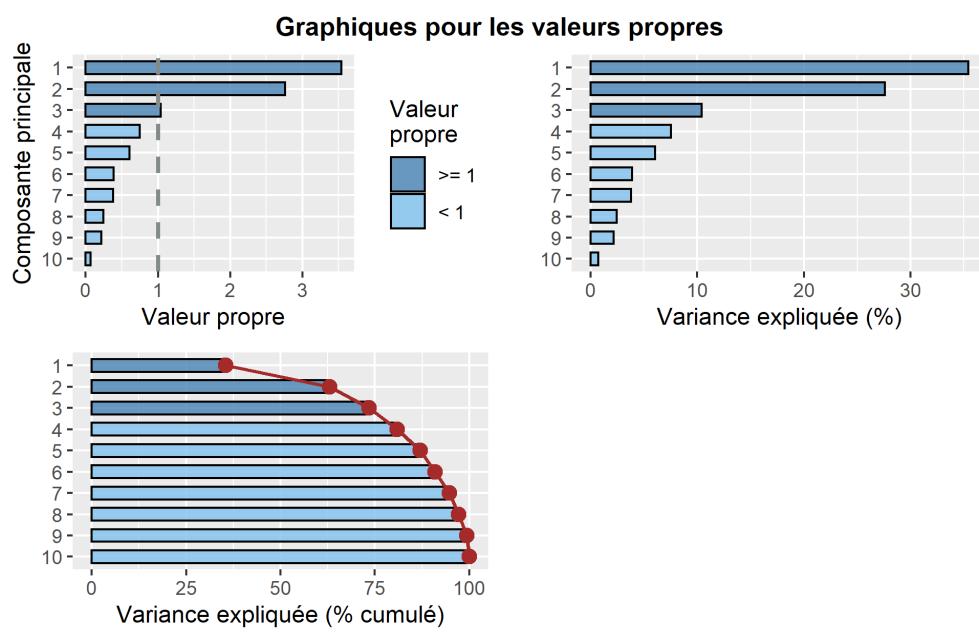


FIG. 1.5 : Graphiques personnalisés pour les valeurs propres pour l'ACP

1.2.2.2 Résultats de l'ACP pour les variables

Pour qualifier les axes, quatre mesures sont disponibles pour les variables :

- **Les coordonnées factorielles des variables** sont simplement les coefficients de corrélation de Pearson des variables sur l'axe k et varient ainsi de -1 à 1 (relire au besoin la section ??). Pour qualifier un axe, il convient alors de repérer les variables les plus corrélées positivement et négativement sur l'axe, autrement dit, de repérer les variables situées aux extrémités l'axe.
- **Les cosinus carrés des variables** (Cos^2) (appelées aussi les qualités de représentation des variables sur un axe) permettent de repérer le ou les axes qui concourent le plus à donner un sens à la variable. Elles sont en fait les coordonnées des variables mises au carré. La somme des cosinus carrés d'une variable sur tous les axes de l'ACP est donc égale à 1 (sommation en ligne). **La qualité de représentation d'une variable sur les n premiers axes** est simplement la somme des cosinus carrés d'une variable sur les axes retenus. Par exemple, pour la variable `Lden`, la qualité de représentation de la variable sur le premier axe est égale : $0,42^2 = 0,17$. Pour cette même variable, la qualité de la `Lden` sur les trois premiers axes est égale à : $0,17 + 0,32 + 0,26 = 0,75$.
- **Les contributions des variables** permettent de repérer celles qui participent le plus à la formation d'un axe. Elles s'obtiennent en divisant les cosinus carrés par la valeur propre de l'axe multiplié par 100. La somme des contributions des variables pour un axe donné est donc égale à 100 (sommation en colonne). Par exemple, pour la variable `Lden`, la contribution sur le premier axe est égale : $0,174 / 3,543 \times 100 = 4,920$.

Les résultats de l'ACP pour les variables sont présentés au tableau ??.

Analyse de la première composante principale (valeur propre de 3,54, 35,43 %)

- À la lecture des contributions, il est clair que quatre variables contribuent grandement à la formation de l'axe 1 : `NivVieMed` (22,06 %), `Pct_Img` (21,56 %), `TxChom1564` (16,89 %) et `Pct_brevet` (14,94 %). Il convient alors d'analyser en détail leurs coordonnées factorielles et leurs cosinus carrés.
- À la lecture des coordonnées factorielles, nous constatons que trois variables socioéconomiques sont fortement corrélées positivement avec l'axe 1, soit le *pourcentage d'immigrants* (0,87), le *taux de chômage* (0,77) le *pourcentage de personnes avec une faible scolarité* (0,73). À l'autre extrémité, la *médiane du niveau de vie* (en Euros) est négativement corrélée avec l'axe 1. Comment interpréter ce résultat ? Premièrement, cela signifie que plus la valeur de l'axe 1 est positive et élevée, plus celles des trois variables (`Pct_Img`, `TxChom1564` et `Pct_brevet`) sont aussi élevées (corrélations positives) et plus la valeur de `NivVieMed` est faible (corrélation négative). Inversement, plus la valeur de l'axe 1 est négative et faible, les valeurs de `Pct_Img`, `TxChom1564` et `Pct_brevet` sont faibles et plus celle de `NivVieMed` est

Tab. 1.5 : Résultats de l'ACP pour les variables

Variable	Coordonnées			Cosinus carrés				Contributions		
	1	2	3	1	2	3	Qualité	1	2	3
<code>Lden</code>	0,42	0,57	0,51	0,17	0,32	0,26	0,75	4,92	11,64	24,80
<code>NO2</code>	0,15	0,93	0,19	0,02	0,86	0,04	0,92	0,66	31,07	3,54
<code>PM25</code>	0,19	0,92	0,03	0,04	0,84	0,00	0,87	1,01	30,36	0,12
<code>VegHautPrt</code>	-0,40	-0,42	0,40	0,16	0,18	0,16	0,50	4,63	6,35	15,46
<code>Pct0_14</code>	0,55	-0,53	0,08	0,30	0,28	0,01	0,59	8,61	10,28	0,55
<code>Pct_65</code>	-0,41	-0,27	0,72	0,17	0,07	0,51	0,75	4,73	2,66	49,26
<code>Pct_Img</code>	0,87	-0,09	0,11	0,76	0,01	0,01	0,78	21,56	0,29	1,08
<code>TxChom1564</code>	0,77	-0,09	-0,07	0,60	0,01	0,00	0,61	16,89	0,27	0,45
<code>Pct_brevet</code>	0,73	-0,43	0,22	0,53	0,19	0,05	0,77	14,94	6,81	4,61
<code>NivVieMed</code>	-0,88	0,09	0,04	0,78	0,01	0,00	0,79	22,06	0,28	0,14

forte. Deuxièmement, cela signifie que les trois variables (`Pct_Img`, `TxChom1564` et `Pct_brevet`) sont fortement corrélées positivement entre elles puisqu'elles se situent sur la même extrémité de l'axe et qu'elles sont toutes trois négativement corrélées avec la variable `NivVieMed`. Cela peut être rapidement confirmé avec la matrice de corrélation entre les dix variables (tableau ??).

- À la lecture des cosinus carrés de l'axe 1, nous constatons que plus des trois quarts de la dispersion/de l'information des variables `NivVieMed` (0,78) et `Pct_Img` (0,76) est concentrée sur l'axe 1.

Analyse de la deuxième composante principale (valeur propre de 2,76, 27,60 %)

- À la lecture des contributions, trois variables environnementales contribuent à la formation de l'axe 1 : principalement, celles sur la pollution de l'air (`NO2` = 31,07 % et `PM25` = 30,36 %) et secondairement, sur le bruit routier (`Lden` = 11,64 %).
- À la lecture des coordonnées factorielles, ces trois variables sont fortement corrélées positivement avec l'axe 2 : `NO2` (0,93), `PM25` (0,92) et `Lden` (0,57). À l'autre extrémité de l'axe, la variable `Pct0_14` est négativement, mais pas fortement corrélée négativement (-0,53). La lecture de la matrice de corrélation au tableau ?? confirme que ces trois variables environnementales sont fortement corrélées positivement entre elles (par exemple, un coefficient de corrélation de Pearson de 0,90 entre `NO2` et `PM25`).
- À la lecture des cosinus carrés de l'axe 2, nous constatons que près de 90 % de la dispersion/de l'information des variables `NO2` (0,86) et `PM25` (0,84) est concentré sur l'axe 1.

Analyse de la troisième composante principale (valeur propre de 1,042, 10,42 %)

- Le pourcentage de personnes âgées (`Pct_65`) contribue principalement à la formation de l'axe avec lequel est corrélée positivement (contribution de 49,26 % et coordonnée factorielle de 0,72). S'en suit, la variable `Lden` qui joue un rôle beaucoup moins important (contribution de 24,80 % et coordonnée factorielle de 0,51).



Lien entre la valeur propre d'un axe et le nombre de variables contribuant à sa formation

Vous avez compris que plus la valeur propre d'un axe est forte, plus il y a potentiellement de variables qui concourent à sa formation. Cela explique que pour la troisième composante qui a une faible valeur propre (1,042), seule une variable contribue significativement à sa formation.

Analyse de la qualité de représentation des variables sur les premiers axes de l'ACP

À titre de rappel, la qualité est simplement la somme des cosinus carrés d'une variable sur les axes retenus. Si nous retenons trois axes, les six variables qui sont le mieux résumées – et qui ont donc le plus d'influence sur les résultats de l'ACP – sont : `NO2` (0,92), `PM25` (0,87), `NivVieMed` (0,79), `Pct_Img` (0,78), `Pct_brevet`

TAB. 1.6 : Matrice de corrélation de Pearson entre les variables utilisées pour l'ACP

Variable	A	B	C	D	E	F	G	H	I	J
A. Lden		0,62	0,49	-0,23	0,04	-0,09	0,28	0,19	0,14	-0,26
B. NO2	0,62		0,90	-0,28	-0,34	-0,21	0,07	0,04	-0,25	-0,04
C. PM25	0,49	0,90		-0,39	-0,34	-0,26	0,12	0,07	-0,25	-0,10
D. VegHautPrt	-0,23	-0,28	-0,39		0,04	0,32	-0,22	-0,18	-0,14	0,32
E. Pct0_14	0,04	-0,34	-0,34	0,04		-0,12	0,46	0,36	0,54	-0,45
F. Pct_65	-0,09	-0,21	-0,26	0,32	-0,12		-0,24	-0,30	0,00	0,32
G. Pct_Img	0,28	0,07	0,12	-0,22	0,46	-0,24		0,66	0,64	-0,73
H. TxChom1564	0,19	0,04	0,07	-0,18	0,36	-0,30	0,66		0,47	-0,62
I. Pct_brevet	0,14	-0,25	-0,25	-0,14	0,54	0,00	0,64	0,47		-0,67
J. NivVieMed	-0,26	-0,04	-0,10	0,32	-0,45	0,32	-0,73	-0,62	-0,67	

(0,77) et Lden (0,75).

Qualification, dénomination d'axes factoriels

L'analyse des coordonnées, contributions et cosinus carrés doit vous permettre de formuler un intitulé pour chacun des axes retenus. Nous pouvons ainsi proposer les intitulés suivants :

- *Niveau de défavorisation socioéconomique* (axe 1). Plus la valeur de l'axe est élevée, plus le niveau de défavorisation de l'entité spatiale (IRIS) est élevé.
- *Qualité environnementale* (axe 2). Plus la valeur de l'axe est forte, plus les niveaux de pollution atmosphérique (dioxyde d'azote et particules fines) et de bruit (Lden) sont élevés.

Le recours à des graphiques pour analyser les résultats de l'ACP pour des variables

Plus le nombre de variables utilisées pour calculer l'ACP est important, plus l'analyse des coordonnées factorielles, des cosinus carrés et des contributions reportés dans un tableau devient fastidieuse. Puisque l'ACP a été calculée sur uniquement dix variables, l'analyse des valeurs du tableau ?? a donc été assez facile et rapide. Imaginez maintenant que nous réalisons une ACP sur une centaine de variables, la taille du tableau des résultats pour les variables sera considérable... Par conséquent, il est recommandé de construire plusieurs graphiques qui facilitent l'analyse des résultats pour les variables.

Par exemple, à la figure ??, nous avons construit des graphiques avec les coordonnées factorielles sur les trois premiers axes de l'ACP. En un coup d'œil, il est facile de repérer les variables plus corrélées positivement ou négativement avec chacun d'entre eux. Aussi, il est fréquent de construire un nuage de points avec les coordonnées des variables sur les deux premiers axes factoriels, soit un graphique communément appelé **nuage de points des variables sur le premier plan factoriel** sur lequel est représenté le cercle des corrélations (figure ??). Bien entendu, cet exercice peut être fait avec d'autres axes factoriels (les axes 3 et 4 par exemple).

1.2.2.3 Résultats de l'ACP pour les individus

Comme pour les variables, nous retrouvons les mêmes mesures pour les individus : les coordonnées factorielles, les cosinus carrés et les contributions. Les coordonnées factorielles des individus sont les projections orthogonales des observations sur l'axe. Puisqu'en ACP normée, les variables utilisées pour l'ACP sont centrées réduites, la moyenne des coordonnées factorielles des individus pour un axe est toujours égale à zéro. En revanche, contrairement aux coordonnées factorielles pour les variables, les coordonnées pour les individus ne varient pas de -1 à 1! Les cosinus carrés quantifient à quel point chaque axe représente chaque individu. Enfin, les contributions quantifient la contribution de chaque individu à la formation d'un axe.

Si le jeu de données comprend peu d'observations, il est toujours possible de créer un **nuage de points des individus sur le premier plan factoriel** sur lequel vous pouvez ajouter les étiquettes permettant d'identifier les observations (figure ??). Ce graphique est rapidement illisible lorsque le nombre d'observations est important. Il peut rester utile si certaines des observations du jeu de données doivent faire l'objet d'une analyse spécifique.

Lorsque les observations sont des unités spatiales, il est très intéressant de cartographier les coordonnées factorielles des individus (figure ??). À la lecture de la carte choroplète de gauche (axe 1), nous pouvons constater que le niveau de défavorisation socioéconomique est élevé dans l'est (IRIS en vert), et inversement, très faible à l'ouest de l'agglomération (IRIS en rouge). À la lecture de la carte de droite (axe 2), sans surprise, la partie centrale de l'agglomération est caractérisée par des niveaux de pollution atmosphérique et de bruit routier bien plus élevés qu'en périphérie.

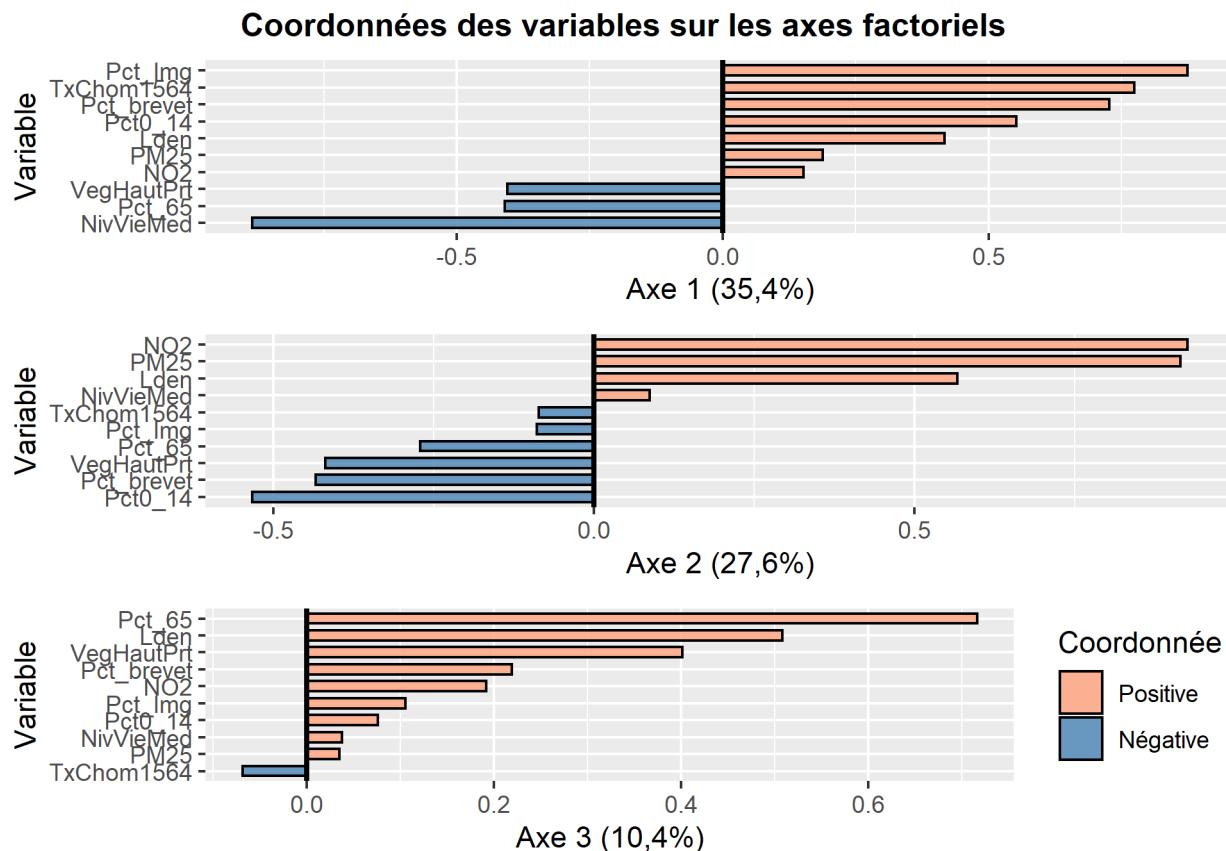


FIG. 1.6 : Coordonnées factorielles des variables

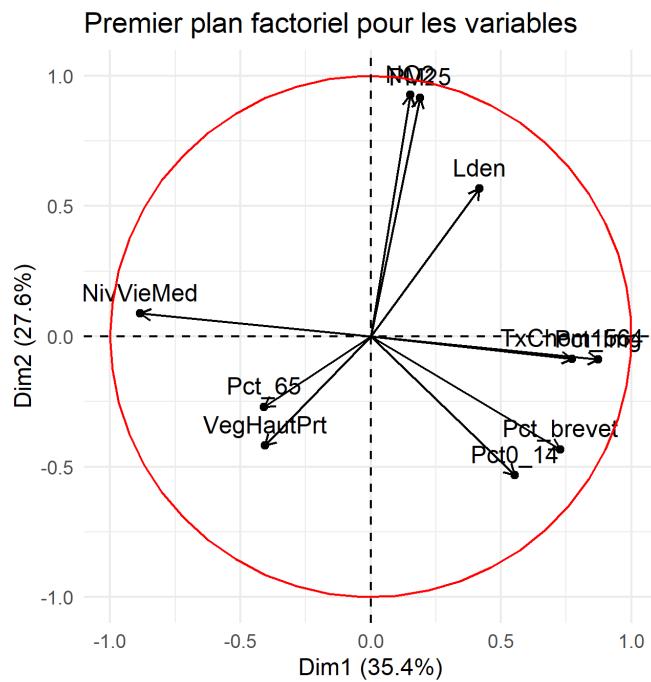


FIG. 1.7 : Premier plan factoriel de l'ACP pour les variables

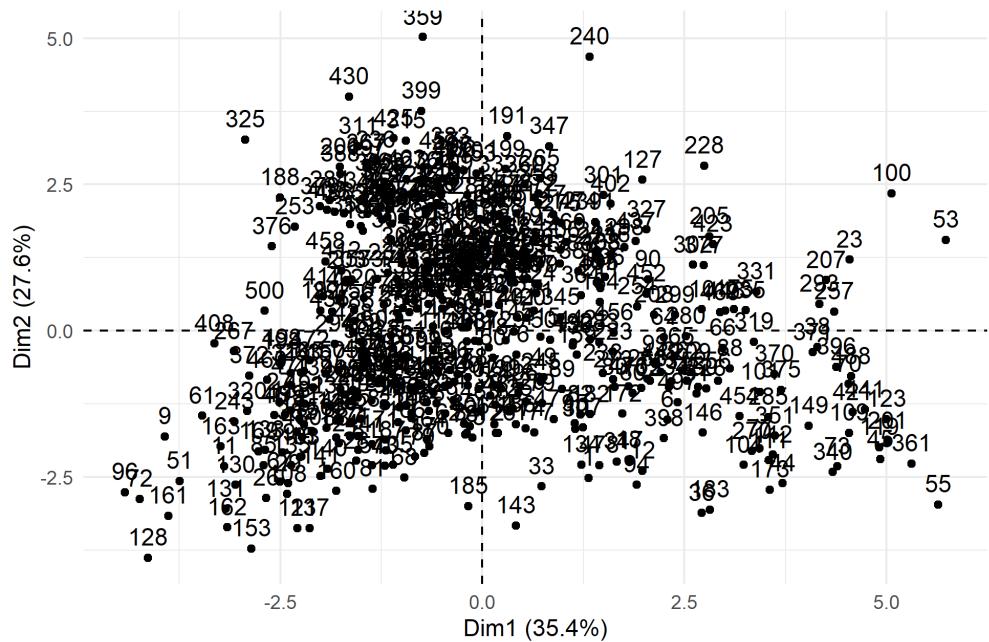
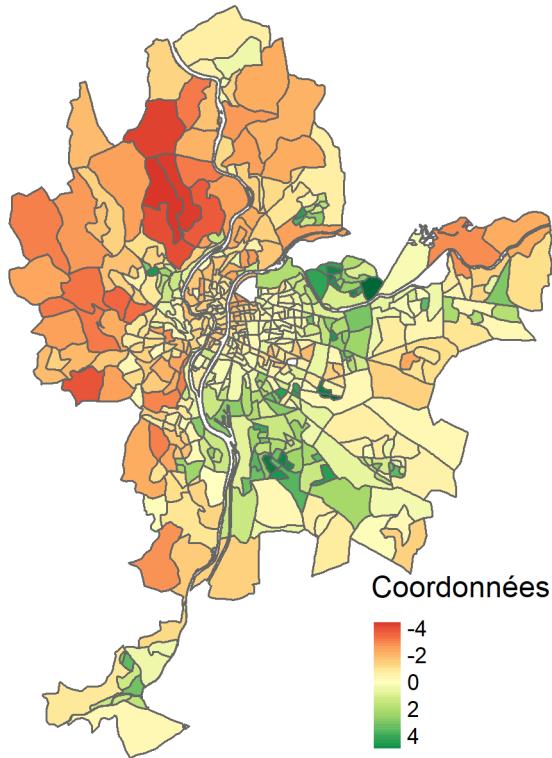


FIG. 1.8 : Premier plan factoriel pour les individus

Axe 1 : Défavorisation socioéco.



Axe 2 : Qualité environnementale

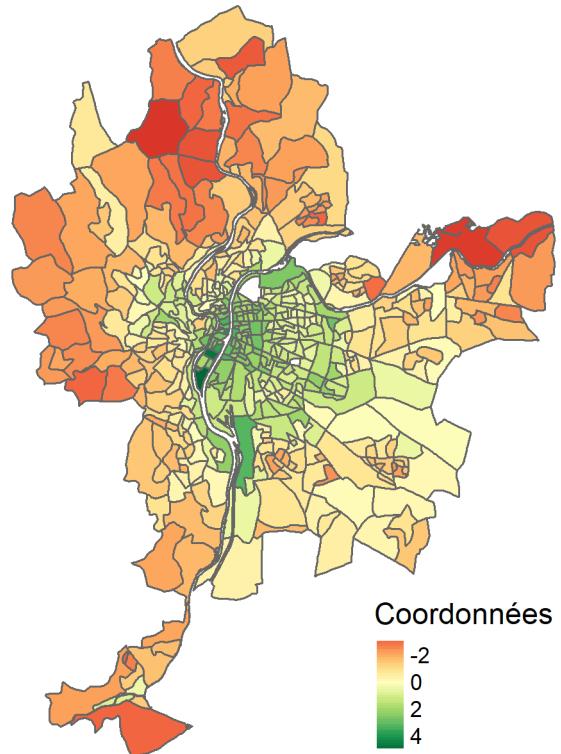


FIG. 1.9 : Cartographie des coordonnées factorielles des individus



Nous n'avons pas abordé plusieurs autres éléments intéressants de l'ACP.

Ajout de variables ou d'individus supplémentaires.

Premièrement, il est possible d'ajouter des variables continues ou des individus supplémentaires qui n'ont pas été pris en compte dans le calcul de l'ACP (figure ??). Concernant les variables continues supplémentaires, il s'agit simplement de calculer leurs corrélations avec les axes retenus de l'ACP. Concernant les individus, il s'agit de les projeter sur les axes factoriels. Pour plus d'informations sur le sujet, consultez les excellents ouvrages de Ludovic Lebart, Alain Morineau et Marie Piron (1995, pp. 42-45) ou encore Jérôme Pagès (2013, pp. 22-24).

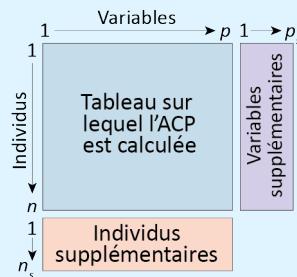


FIG. 1.10 : Variables et individus supplémentaires pour l'ACP

Pondération des individus et des variables.

Deuxièmement, il est possible de pondérer à la fois les individus et plus rarement les variables lors du calcul du l'ACP.

Analyse en composantes principales non paramétrique

Troisièmement, il est possible de calculer une ACP sur des variables préalablement transformées en rangs (section ??). Cela peut être justifié lorsque les variables sont très anormalement distribuées en raison de valeurs extrêmes. Les coordonnées factorielles pour les variables sont alors le coefficient de Spearman (section ??) et non de Pearson. Aussi, les variables sont centrées non pas sur leurs moyennes respectives, mais sur leurs médianes. Pour plus d'informations sur cette approche, consultez de nouveau Lebart et al. (1995, pp. 51-52).

Analyse en composantes principales robuste

D'autres méthodes plus avancées qu'une ACP non paramétrique peuvent être utilisées afin d'obtenir des composantes principales qui ne sont pas influencées par des valeurs extrêmes : les ACP robustes (Rivest et Plante 1988 ; Hubert, Rousseeuw et Vanden Branden 2005) qui peuvent être mises en œuvre, entre autres, avec le package *roscpca*.

1.2.3 Mise en œuvre dans R

Plusieurs *packages* permettent de calculer une ACP dans R, notamment *psych* (fonction *principal*), *ade4* (fonction *dudi.pca*) et *FactoMineR* (fonction *pca*). Ce dernier est certainement le plus abouti. De plus, il permet également de calculer une analyse des correspondances (AFC), une analyse des correspondances multiples (ACM) et une analyse factorielle de données mixtes (AFDM). Nous utilisons donc *FactoMineR* pour mettre en œuvre les trois types de méthodes factorielles abordées dans ce chapitre (ACP, AFC et ACM). Pour l'ACP, nous exploitons un jeu de données issu du *package geocmeans* qu'il faut préalablement charger à l'aide des lignes de code suivantes.

```
library(geocmeans)
data(LyonIris)
Data <- LyonIris@data[c("CODE_IRIS","Lden","N02","PM25","VegHautPrt",
                      "Pct0_14","Pct_65","Pct_Img",
                      "TxChom1564","Pct_brevet","NivVieMed")]
```

1.2.3.1 Calcul et exploration d'une ACP avec FactoMineR

Pour calculer l'ACP, il suffit d'utiliser la fonction `PCA` de `FactoMineR`, puis la fonction `summary`(`MonACP`) qui renvoie les résultats de l'ACP pour :

- Les valeurs propres (section `Eigenvalues`) pour les composantes principales (`Dim.1` à `Dim.n`) avec leur variance expliquée brute (`Variance`), en pourcentage (% of var.) et en pourcentage cumulé (Cumulative % of var.).
- Les dix premières observations (section `Individuals`) avec les coordonnées factorielles (`Dim.1` à `Dim.n`), les contributions (`ctr`) et les cosinus carrés (`cos2`). Pour accéder aux résultats pour toutes les observations, utilisez les fonctions `res.acp$ind` ou encore `res.acp$ind$coord` (uniquement les coordonnées factorielles), `res.acpindcontrib` (uniquement les contributions) et `res.acpindcos2` (uniquement les cosinus carrés).
- Les variables (section `Variables`) avec les coordonnées factorielles (`Dim.1` à `Dim.n`), les contributions (`ctr`) et les cosinus carrés (`cos2`).

```
library(FactoMineR)
# Version classique avec FactoMineR
# Construction d'une ACP sur les colonnes 2 à 11 du datafram Data
res.acp <- PCA(Data[,2:11], scale.unit=TRUE, graph=F)
# Affichage des résultats de la fonction PCA
print(res.acp)

## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 506 individuals, described by 10 variables
## *The results are available in the following objects:
##
##      name          description
## 1  "$eig"        "eigenvalues"
## 2  "$var"        "results for the variables"
## 3  "$var$coord"  "coord. for the variables"
## 4  "$var$cor"    "correlations variables - dimensions"
## 5  "$var$cos2"   "cos2 for the variables"
## 6  "$var$contrib" "contributions of the variables"
## 7  "$ind"        "results for the individuals"
## 8  "$ind$coord"  "coord. for the individuals"
## 9  "$ind$cos2"   "cos2 for the individuals"
## 10 "$ind$contrib" "contributions of the individuals"
## 11 "$call"       "summary statistics"
## 12 "$call$centre" "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"   "weights for the individuals"
## 15 "$call$col.w"   "weights for the variables"
```

```
# Résumé des résultats (valeurs propres, individus, variables)
summary(res.acp)
```

```
##
## Call:
## PCA(X = Data[, 2:11], scale.unit = TRUE, graph = F)
##
##
## Eigenvalues
##                               Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance                 3.543   2.760   1.042   0.751   0.606   0.388   0.379
## % of var.                35.425  27.596  10.422   7.511   6.059   3.880   3.788
## Cumulative % of var.    35.425  63.021  73.443  80.954  87.013  90.893  94.681
##                               Dim.8   Dim.9   Dim.10
## Variance                  0.244   0.217   0.071
## % of var.                 2.441   2.167   0.711
## Cumulative % of var.    97.122  99.289 100.000
##
## Individuals (the 10 first)
##          Dist   Dim.1   ctr   cos2   Dim.2   ctr   cos2   Dim.3
## 1 | 3.054 | 1.315  0.096  0.185 | -2.515  0.453  0.678 | 0.221
## 2 | 1.882 | 0.193  0.002  0.011 | -1.744  0.218  0.859 | 0.082
## 3 | 2.820 | 2.338  0.305  0.687 | -0.860  0.053  0.093 | -0.765
## 4 | 2.816 | -0.740  0.031  0.069 |  2.265  0.367  0.647 | 1.293
## 5 | 3.210 | -2.208  0.272  0.473 | -1.597  0.183  0.248 | 1.471
## 6 | 3.016 | 2.287  0.292  0.575 | -1.515  0.164  0.252 | 0.390
## 7 | 3.022 | -1.540  0.132  0.260 | -1.803  0.233  0.356 | 0.465
## 8 | 3.122 | -1.536  0.132  0.242 | -2.038  0.298  0.426 | -0.120
## 9 | 4.743 | -3.930  0.862  0.687 | -1.806  0.234  0.145 | 0.993
## 10 | 3.055 | 2.713  0.411  0.789 |  0.368  0.010  0.014 | -0.391
##          ctr   cos2
## 1 | 0.009  0.005 |
## 2 | 0.001  0.002 |
## 3 | 0.111  0.074 |
## 4 | 0.317  0.211 |
## 5 | 0.411  0.210 |
## 6 | 0.029  0.017 |
## 7 | 0.041  0.024 |
## 8 | 0.003  0.001 |
## 9 | 0.187  0.044 |
## 10 | 0.029  0.016 |
##
## Variables
##          Dim.1   ctr   cos2   Dim.2   ctr   cos2   Dim.3   ctr   cos2
## Lden | 0.417  4.920  0.174 | 0.567 11.640  0.321 | 0.508 24.799  0.258
## NO2  | 0.153  0.657  0.023 | 0.926 31.068  0.857 | 0.192 3.540  0.037
## PM25 | 0.189  1.007  0.036 | 0.915 30.355  0.838 | 0.035 0.117  0.001
## VegHautPrt | -0.405  4.630  0.164 | -0.419  6.353  0.175 | 0.401 15.459  0.161
## Pct0_14  | 0.552  8.605  0.305 | -0.533 10.281  0.284 | 0.076  0.553  0.006
```

```

## Pct_65      | -0.409  4.730  0.168 | -0.271  2.658  0.073 |  0.716 49.258  0.513
## Pct_Img     |  0.874 21.559  0.764 | -0.089  0.288  0.008 |  0.106  1.077  0.011
## TxChom1564 |  0.774 16.893  0.598 | -0.086  0.267  0.007 | -0.068  0.450  0.005
## Pct_brevet  |  0.727 14.936  0.529 | -0.434  6.813  0.188 |  0.219  4.612  0.048
## NivVieMed   | -0.884 22.062  0.782 |  0.088  0.278  0.008 |  0.038  0.136  0.001
##
## Lden        |
## NO2         |
## PM25        |
## VegHautPrt  |
## Pct0_14     |
## Pct_65      |
## Pct_Img     |
## TxChom1564 |
## Pct_brevet  |
## NivVieMed   |

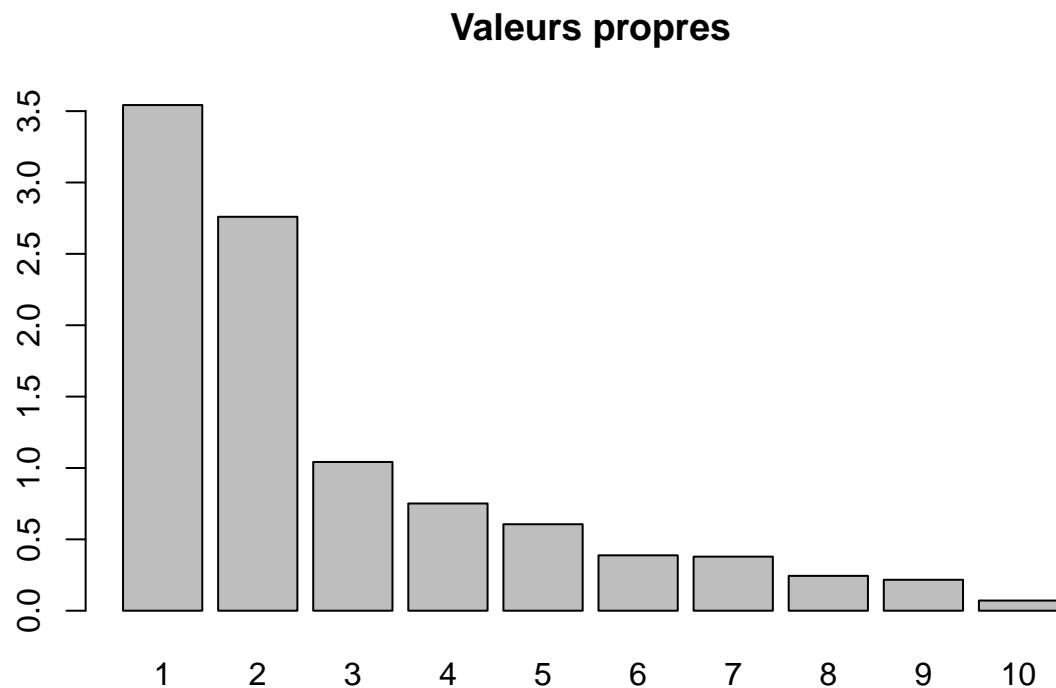
```

Avec les fonctions de base `barplot` et `plot`, il est possible de construire rapidement des graphiques pour explorer les résultats de l'ACP pour les valeurs propres, les variables et les individus.

```

# Graphiques pour les valeurs propres
barplot(res.acp$eig[,1], main="Valeurs propres", names.arg=1:nrow(res.acp$eig))

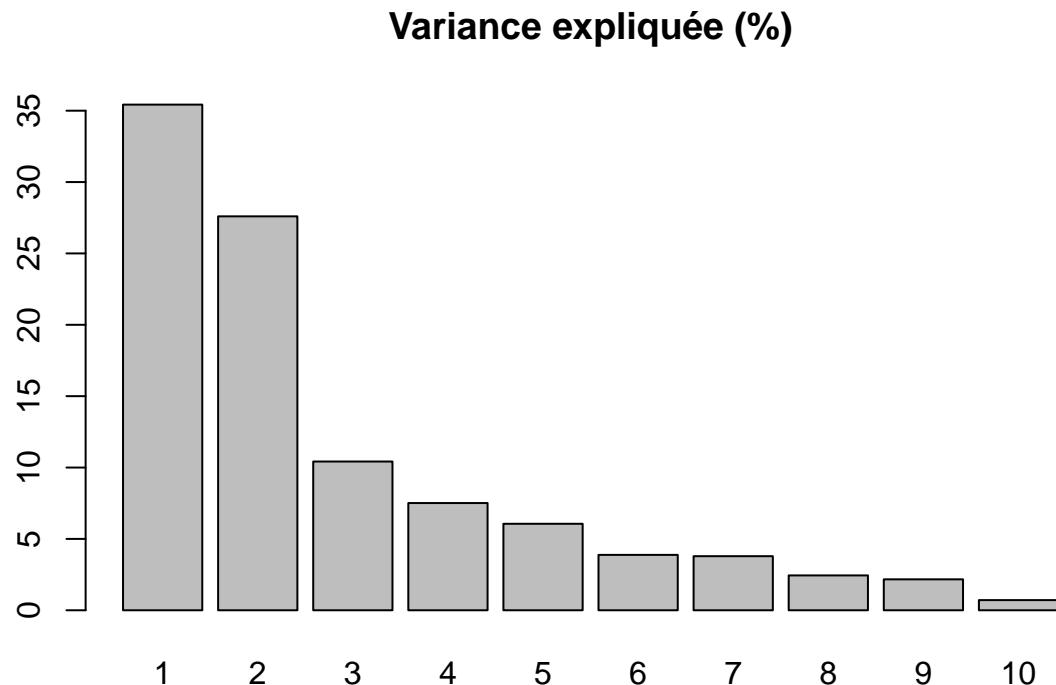
```



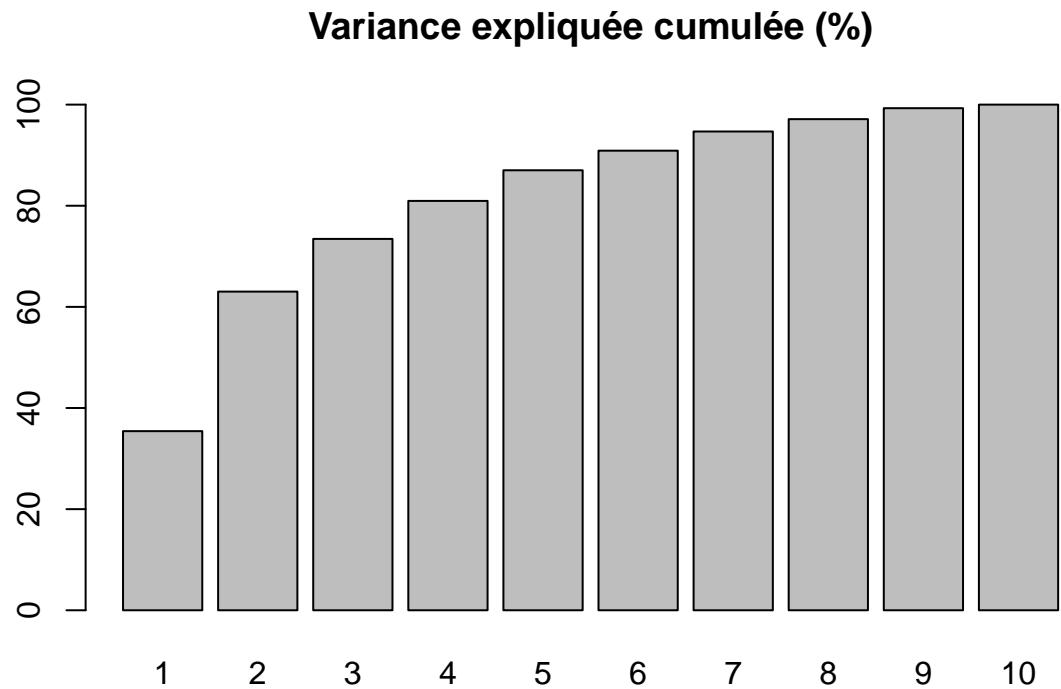
```

barplot(res.acp$eig[,2], main="Variance expliquée (%)", names.arg=1:nrow(res.acp$eig))

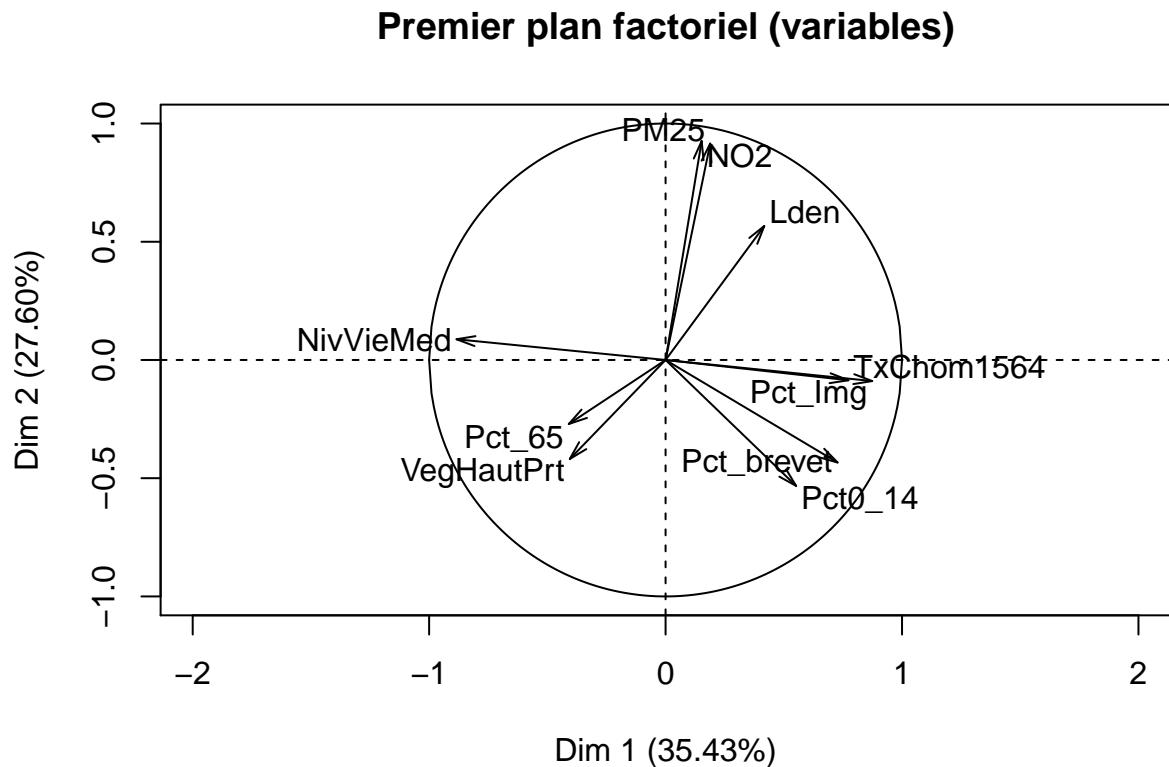
```



```
barplot(res.acp$eig[,3], main="Variance expliquée cumulée (%)",
        names.arg=1:nrow(res.acp$eig))
```

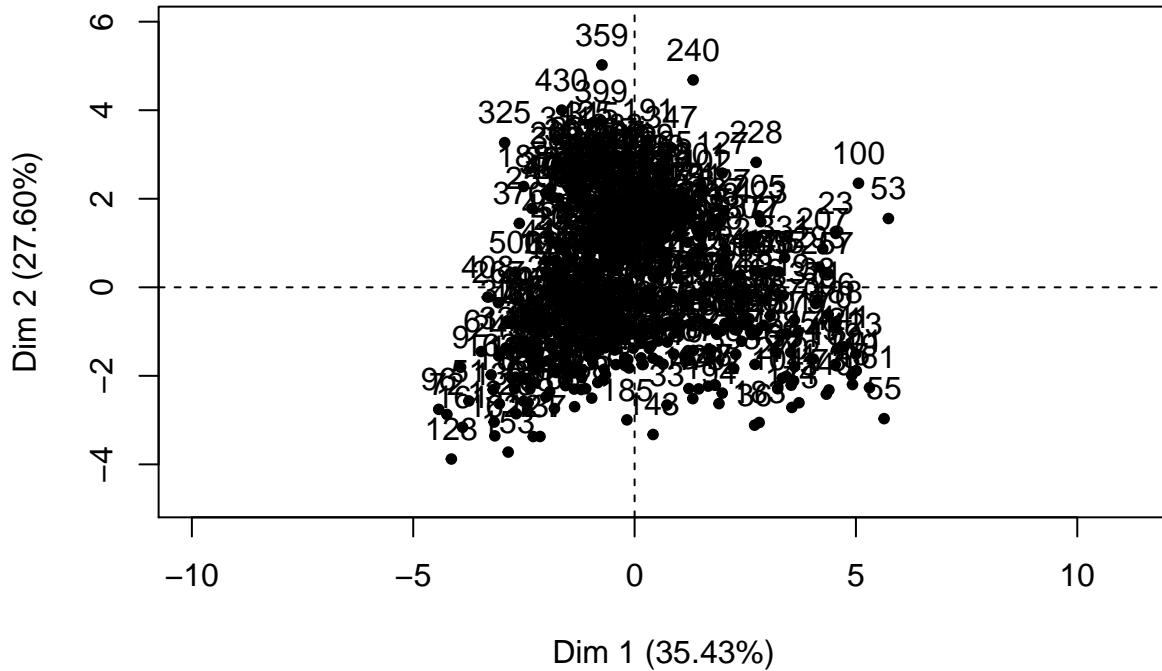


```
# Nuage du points du premier plan factoriel pour les variables et les individus
plot(res.acp, graph.type = "classic", choix="var", axes = 1:2,
     title = "Premier plan factoriel (variables)")
```



```
plot(res.acp, graph.type = "classic", choix="ind", axes = 1:2,
     title = "Premier plan factoriel (individus)")
```

Premier plan factoriel (individus)



Nous avons vu dans un encadré ci-dessus qu'il est possible d'ajouter des variables et des individus supplémentaires dans une ACP, ce que permet la fonction `PCA` de `FactoMineR` avec les paramètres `ind.sup` et `quanti.sup`. Aussi, pour ajouter des pondérations aux individus ou aux variables, utiliser les paramètres `row.w` et `col.w`. Pour plus d'informations sur ces paramètres, consulter l'aide de la fonction en tapant `?PCA` dans la console de Rstudio.

1.2.3.2 Exploration graphique des résultats de l'ACP avec `factoextra`

Visuellement, vous avez pu constater que les graphiques ci-dessus (pour les valeurs propres et pour le premier plan factoriel pour les variables et les individus) réalisés avec les fonctions de base `barplot` et `plot` sont peu attrayants. Avec le package `factoextra`, quelques lignes de code suffisent pour construire des graphiques bien plus esthétiques.

Premièrement, la syntaxe ci-dessous renvoie deux graphiques pour analyser les résultats des valeurs propres (figure ??).

```
library(factoextra)
library(ggplot2)
library(ggpubr)

# Graphiques pour les variables propres avec factoextra
G1 <- fviz_screenplot(res.acp, choice ="eigenvalue", addlabels = TRUE,
                      x="Composantes",
                      y="Valeur propre",
                      title="")
```

```
G2 <- fviz_screenplot(res.acp, choice = "variance", addlabels = TRUE,
                      x="Composantes",
                      y="Pourcentage de la variance expliquée",
                      title="")
ggarrange(G1, G2)
```

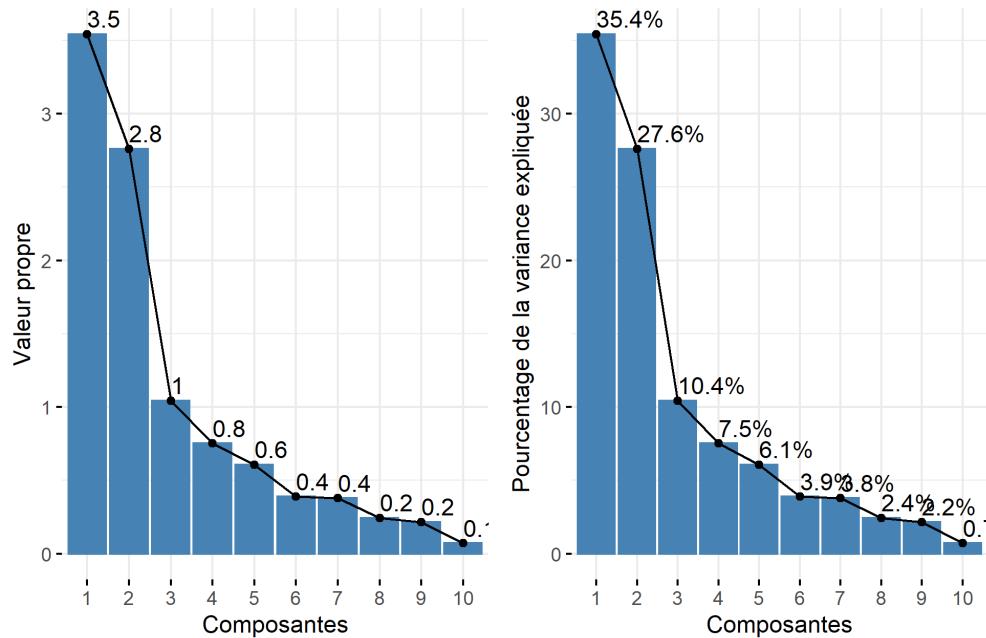


FIG. 1.11 : Graphiques pour les valeurs propres de l'ACP avec factoextra

Deuxièmement, la syntaxe ci-dessous renvoie trois graphiques pour analyser les contributions de chaque variable aux deux premiers axes de l'ACP (figures ?? et ??) et la qualité de représentation des variables sur les trois premiers axes (figure ??), c'est-à-dire la somme des cosinus carrés sur les trois axes retenus.

```
# Contributions des variables aux deux premières composantes avec factoextra
fviz_contrib(res.acp, choice = "var", axes = 1, top = 10,
             title = "Contributions des variables à la première composante")
fviz_contrib(res.acp, choice = "var", axes = 2, top = 10,
             title = "Contributions des variables à la deuxième composante")
fviz_cos2(res.acp, choice = "var", axes = 1:3)+
  labs(x="", y="Somme des cosinus carrés sur les 3 axes retenus",
       title ="Qualité de représentation des variables sur les axes retenus de l'ACP")
```

Troisièmement, le code ci-dessous renvoie un nuage de points pour le premier plan factoriel de l'ACP (axes 1 et 2) pour les variables (figure ??) et les individus (figure ??).

```
# Premier plan factoriel pour les variables avec factoextra
fviz_pca_var(res.acp, col.var="contrib",
              title = "Premier plan factoriel pour les variables")+
  scale_color_gradient2(low="#313695", mid="#ffffbf", high="#a50026",
                        midpoint=mean(res.acp$var$contrib[,1]))
# Premier plan factoriel pour les individus avec factoextra
```

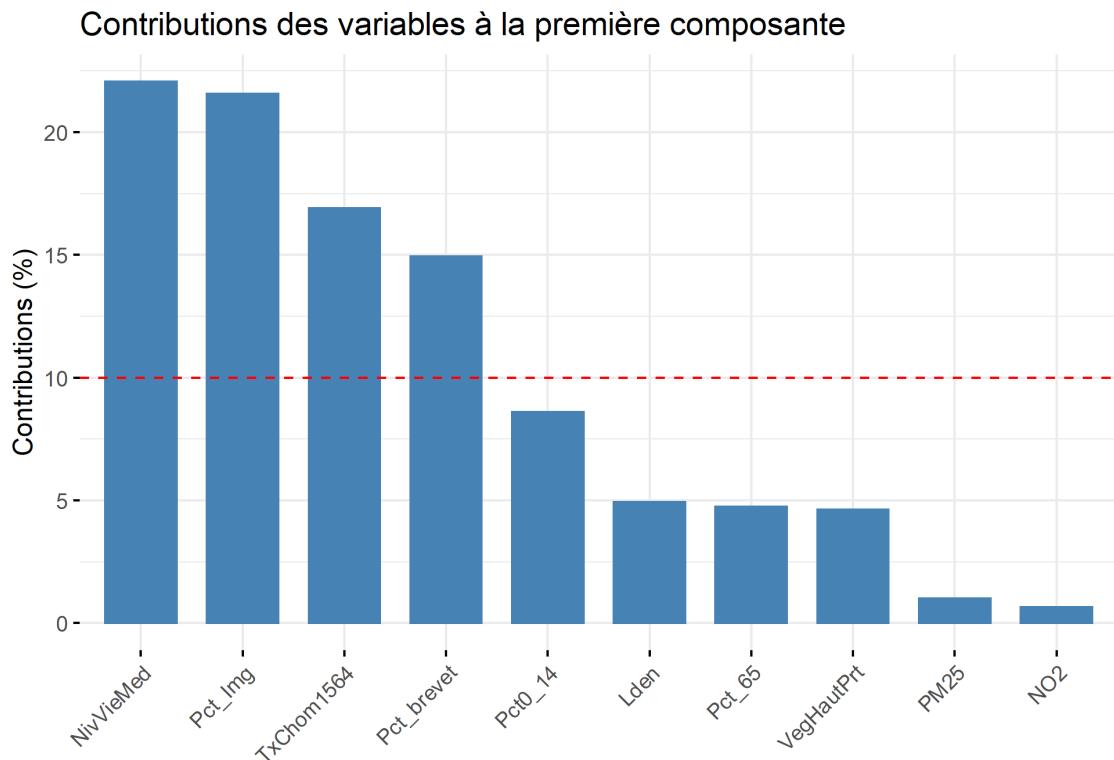


FIG. 1.12 : Contributions des variables à la première composante avec factoextra

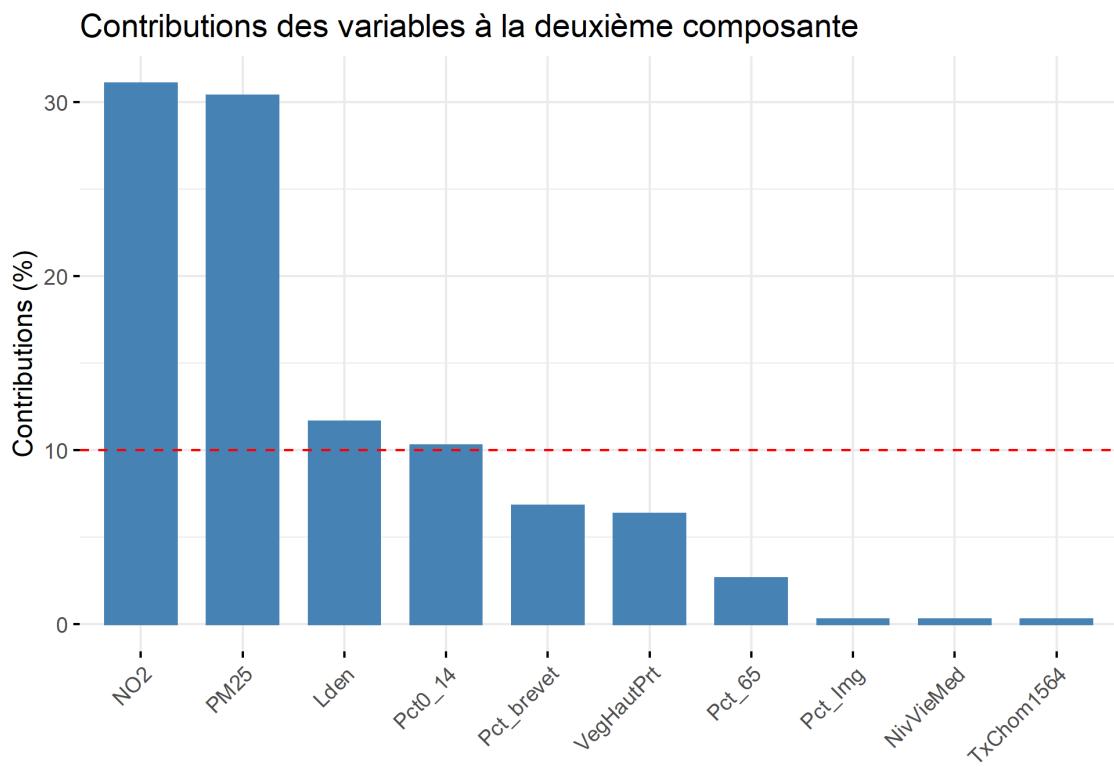


FIG. 1.13 : Contributions des variables à la deuxième composante avec factoextra

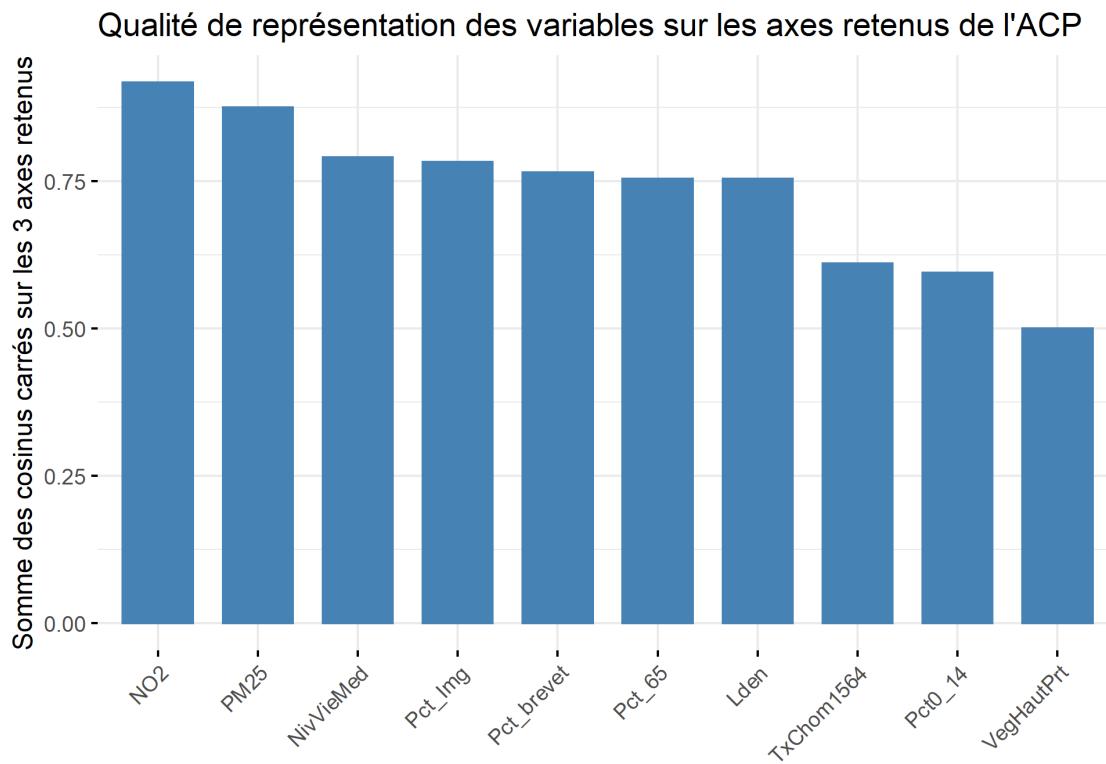


FIG. 1.14 : Qualité des variables sur les trois premières composantes avec factoextra

```
fviz_pca_ind(res.acp, label="none")
fviz_pca_ind(res.acp, col.ind="cos2") +
  scale_color_gradient2(low="blue", mid="white", high="red", midpoint=0.50)
```

1.2.3.3 Personnalisation des graphiques avec les résultats de l'ACP

Avec un peu plus de code et l'utilisation d'autres *packages* (`ggplot2`, `ggpubr`, `stringr`, `corrplot`), vous pouvez aussi construire des graphiques personnalisés.

Premièrement, la syntaxe ci-dessous permet de réaliser trois graphiques pour analyser les valeurs propres (figure ??). Notez que, d'un coup d'œil, nous pouvons identifier les composantes principales avec une valeur propre égale ou supérieure à 1.

```
library(ggplot2)
library(ggpubr)
library(stringr)
library(corrplot)

# Calcul de l'ACP
res.acp <- PCA(Data[,2:11], ncp=5, scale.unit=TRUE, graph=F)
print(res.acp)

# Construction d'un dataframe pour les valeurs propres
dfACPvp <- data.frame(res.acp$eig)
names(dfACPvp) <- c("VP", "VP_pct", "VP_cumupct")
```

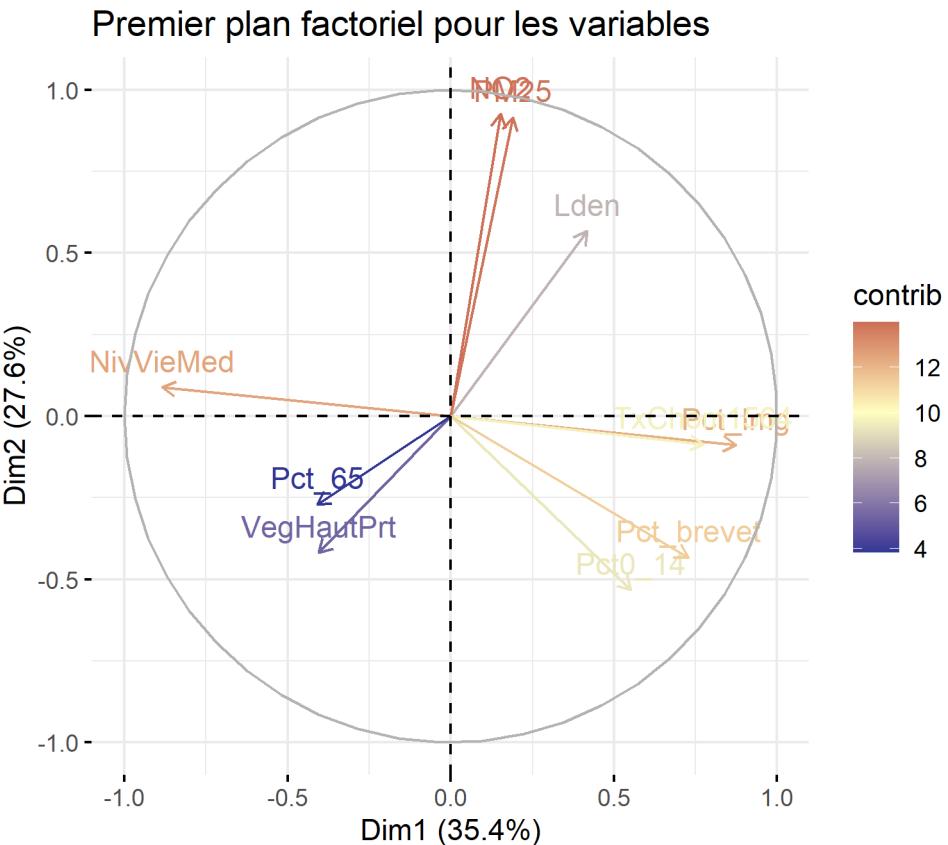


FIG. 1.15 : Premier plan factoriel de l'ACP pour les variables avec factoextra

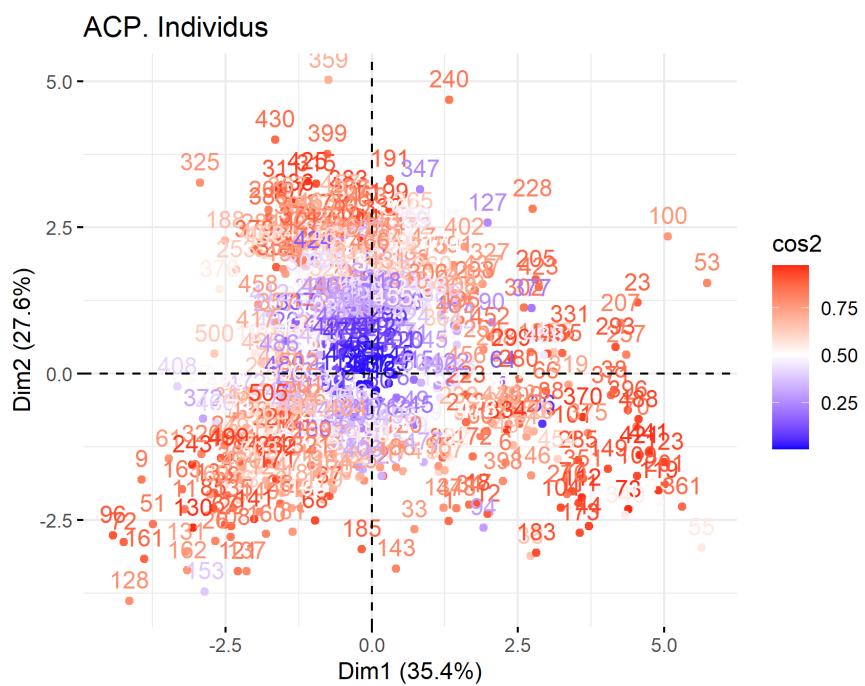


FIG. 1.16 : Premier plan factoriel de l'ACP pour les individus avec factoextra

```

dfACPvp$Composante <- factor(1:nrow(dfACPvp), levels=rev(1:nrow(dfACPvp)))
couleursAxes <- c("steelblue","skyblue2")
vpsup1 <- round(sum(subset(dfACPvp, VP >= 1)$VP),2)
vpsup1cumul <- round(sum(subset(dfACPvp, VP >= 1)$VP_pct),2)

plotVP1 <- ggplot(dfACPvp,aes(x=VP, y=Composante,fill=VP<1))+  

  geom_bar(stat="identity", width = .6, alpha=.8, color="black")+
  geom_vline(xintercept=1, linetype="dashed", color = "azure4", size=1)+  

  scale_fill_manual(name="Valeur\npropre",values=couleursAxes,labels = c(">= 1","< 1"))+
  labs(x="Valeur propre", y="Composante principale")
plotVP2 <- ggplot(dfACPvp, aes(x=VP_pct, y=Composante,fill=VP<1))+  

  geom_bar(stat="identity", width = .6, alpha=.8, color="black")+
  scale_fill_manual(name="Valeur\npropre",values=couleursAxes,labels = c(">= 1","< 1"))+
  theme(legend.position="none")+
  labs(x="Pourcentage de la variance expliquée", y="")
plotVP3 <- ggplot(dfACPvp, aes(x=VP_cumupct, y=Composante,fill=VP<1, group=1))+  

  geom_bar(stat="identity", width = .6, alpha=.8, color="black")+
  scale_fill_manual(name="Valeur\npropre",values=couleursAxes,labels = c(">= 1","< 1"))+
  geom_line(colour="brown", linetype="solid", size=.8) +
  geom_point(size=3, shape=21, color="brown", fill="brown")+
  theme(legend.position="none")+
  labs(x="Pourcentage cumulé de la variance expliquée", y="")

text1 <- paste0("Somme des valeurs propres supérieures à 1 : ",  

  vpsup1,  

  ".\nPourcentage cumulé des valeurs propres supérieures à 1 : ",  

  vpsup1cumul, "%.")

annotate_figure(ggarrange(plotVP1, plotVP2, plotVP3, ncol=2),  

  text_grob("Analyse des valeurs propres",  

    color = "black", face = "bold", size = 12),  

  bottom = text_grob(text1,  

    color = "black", hjust = 1, x = 1, size = 10))

```

Deuxièmement, la syntaxe ci-dessous permet de :

- Construire un *dataframe* avec les résultats des variables.
- Construire des histogrammes avec les coordonnées des variables sur les axes factoriels (figure ??). Notez que les coordonnées négatives sont indiquées avec des barres bleues et celles négatives avec des barres de couleur saumon.
- Un graphique avec les contributions des variables sur les axes retenus (figure ??).
- Un graphique avec les cosinus carrés des variables sur les axes retenus (figure ??).
- Un histogramme avec la qualité des variables sur les axes retenus (figure ??), soit la sommation de leurs cosinus carrés sur les axes retenus.

```

# Analyse des résultats de L'ACP pour les variables
library(corrplot)
library(stringr)
library(ggplot2)
library(ggpubr)

# Indiquer le nombre d'axes à conserver suite à l'analyse des valeurs propres
nComp <- 3
# Variance expliquée par les axes retenus

```

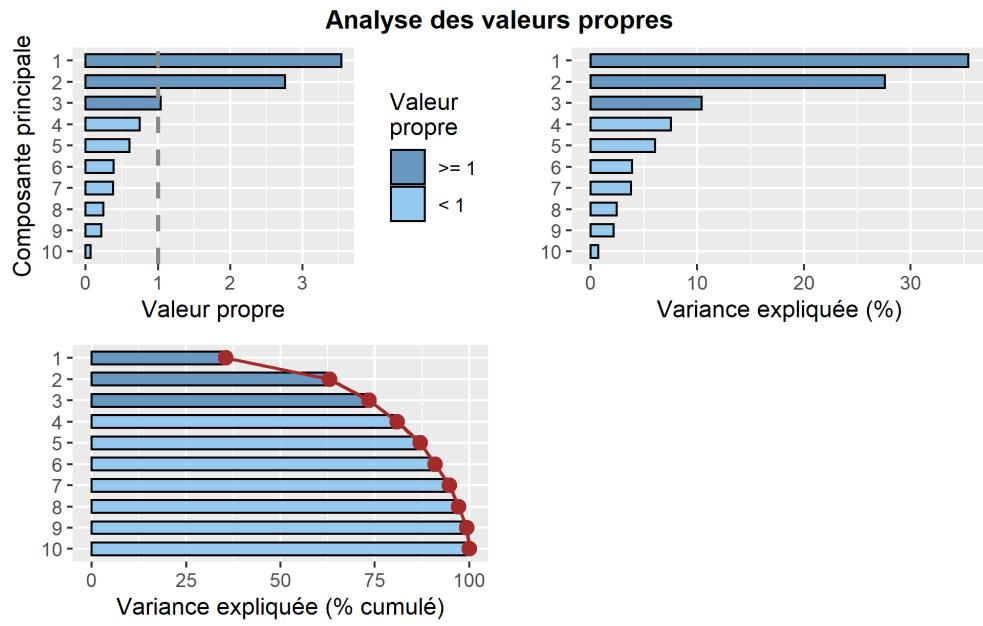


FIG. 1.17 : Graphiques personnalisés pour les valeurs propres

```
vppct <- round(dfACPvp[1:nComp,"VP_pct"],1)
# Dataframe des résultats pour les variables
CoordsVar <- res.acp$var$coord[, 1:nComp]
Cos2Var   <- res.acp$var$cos2[, 1:nComp]
CtrVar    <- res.acp$var$contrib[, 1:nComp]
dfACPVars <- data.frame(Variable = row.names(res.acp$var$coord[, 1:nComp]),
                         Coord = CoordsVar,
                         Cos2 = Cos2Var,
                         Qualite = rowSums(Cos2Var),
                         Ctr = CtrVar)
row.names(dfACPVars) <- NULL
names(dfACPVars) <- str_replace(names(dfACPVars), ".Dim.", "Comp")
dfACPVars

# Histogrammes pour les coordonnées
couleursCoords <- c("lightsalmon","steelblue")
plotCoordF1 <- ggplot(dfACPVars,
                       aes(y = reorder(Variable, CoordComp1),
                           x = CoordComp1, fill=CoordComp1<0))+ 
  geom_bar(stat="identity", width = .6, alpha=.8, color="black")+
  geom_vline(xintercept=0, color = "black", size=1)+ 
  scale_fill_manual(name="Coordonnée",values=couleursCoords,
                    labels = c("Positive","Négative"))+
  labs(x=paste0("Axe 1 (", vppct[1],"%)"), y="Variable")+
  theme(legend.position="none")
plotCoordF2 <- ggplot(dfACPVars,
                       aes(y = reorder(Variable, CoordComp2),
                           x = CoordComp2, fill=CoordComp2<0))+ 
  geom_bar(stat="identity", width = .6, alpha=.8, color="black")+
  geom_vline(xintercept=0, color = "black", size=1)+
```

```

scale_fill_manual(name="Coordonnée", values=couleursCoords,
                  labels = c("Positive","Négative"))+
  labs(x=paste0("Axe 2 (", vppct[2],"%)", y="Variable")+
    theme(legend.position="none")
plotCoordF3 <- ggplot(dfACPVars,
                      aes(y = reorder(Variable, CoordComp3),
                          x = CoordComp3, fill=CoordComp3<0))+ 
  geom_bar(stat="identity", width = .6, alpha=.8, color="black")+
  geom_vline(xintercept=0, color = "black", size=1)+ 
  scale_fill_manual(name="Coordonnée", values=couleursCoords,
                    labels = c("Positive","Négative"))+
  labs(x=paste0("Axe 3 (", vppct[3],"%)", y="Variable"))

annotate_figure(ggarrange(plotCoordF1, plotCoordF2, plotCoordF3, nrow=nComp),
               text_grob("Coordonnées des variables sur les axes factoriels",
                         color = "black", face = "bold", size = 12))

# Contributions des variables à la formation des axes
couleurs <- colorRampPalette(c("#ffffd4","#993404"))
corrplot(CtrVar, is.corr=FALSE, method ="square", col = couleurs(20),
         addCoef.col = 1, cl.pos = FALSE)

# La qualité des variables sur les composantes retenues : cosinus carrés
corrplot(Cos2Var, is.corr=FALSE, method ="square", col = couleurs(20),
         addCoef.col = 1, cl.pos = FALSE)

ggplot(dfACPVars)+ 
  geom_bar(aes(y=reorder(Variable, Qualite), x=Qualite),
           stat="identity", width = .6, alpha=.8, fill="steelblue")+
  labs(x="", y="Somme des cosinus carrés sur les axes retenus",
       title ="Qualité de représentation des variables sur les axes retenus de l'ACP",
       subtitle = paste0("Variance expliquée par les ", nComp,
                        " composantes : ", sum(vppct), "%"))

```

Troisièmement, lorsque les observations sont des unités spatiales, il convient de cartographier les coordonnées factorielles des individus. Dans le jeu de données utilisé, les observations sont des polygones délimitant les îlots regroupés pour l'information statistique (IRIS) pour l'agglomération de Lyon (France). Nous utilisons les *packages* tmap et RColorBrewer pour réaliser des cartes choroplèthes avec les coordonnées deux premières composantes (figure ??).

```

library("tmap")
library("RColorBrewer")
# Analyse des résultats de l'ACP pour les individus
# Dataframe des résultats pour les individus
CoordsInd <- res.acp$ind$coord[, 1:nComp]
Cos2Ind   <- res.acp$ind$cos2[, 1:nComp]
CtrInd    <- res.acp$ind$contrib[, 1:nComp]
dfACPIInd <- data.frame(Coord = CoordsInd, Cos2 = Cos2Ind, Ctr = CtrInd)
names(dfACPIInd) <- str_replace(names(dfACPIInd), ".Dim.", "Comp")
# Fusion du tableau original avec les résultats de l'ACP pour les individus
CartoACP <- cbind(LyonIris, dfACPIInd)
# Cartographie des coordonnées factorielles pour les individus pour les
# deux premières composantes
Carte1 <- tm_shape(CartoACP) +

```

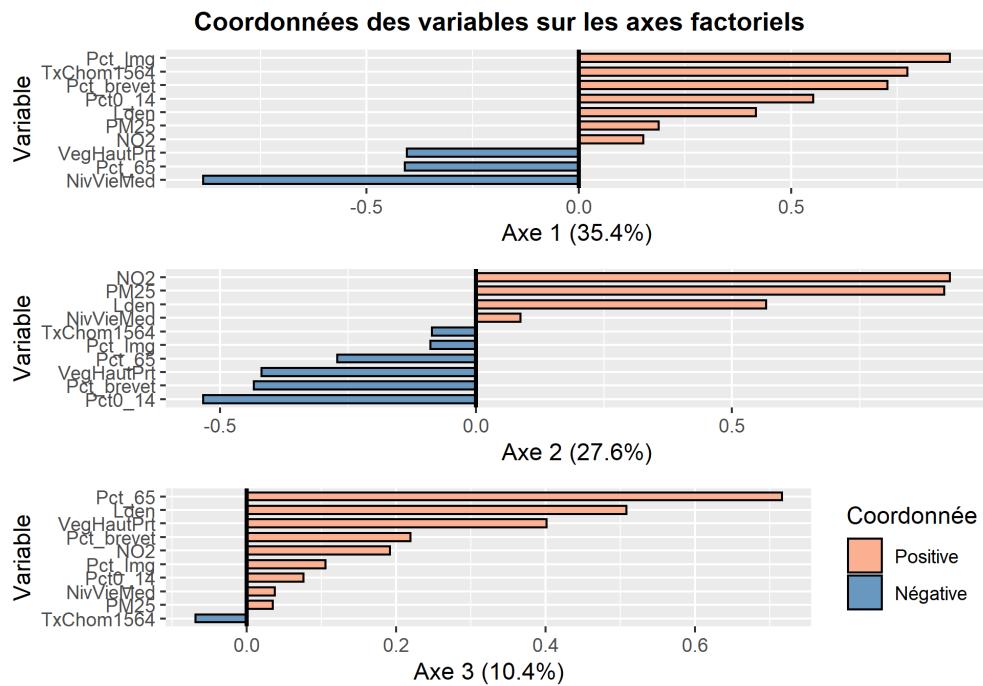


FIG. 1.18 : Histogrammes personnalisés avec les coordonnées factorielles pour les variables

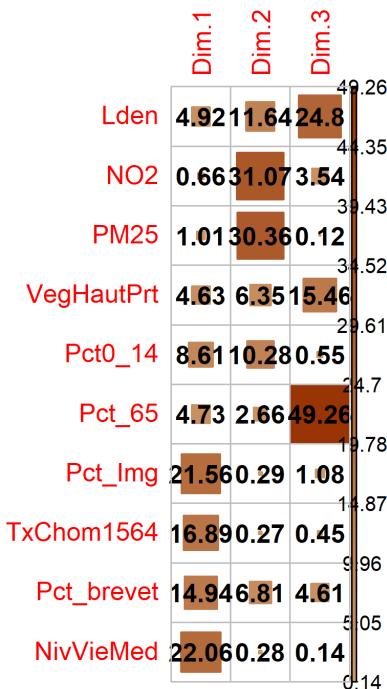


FIG. 1.19 : Graphiques personnalisés avec les contributions des variables

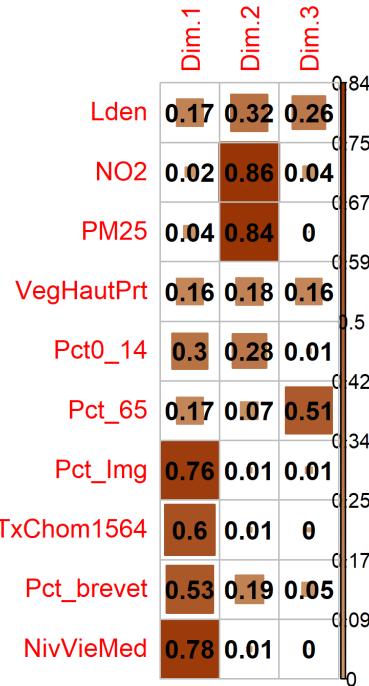


FIG. 1.20 : Graphiques personnalisés avec les cosinus carrés des variables

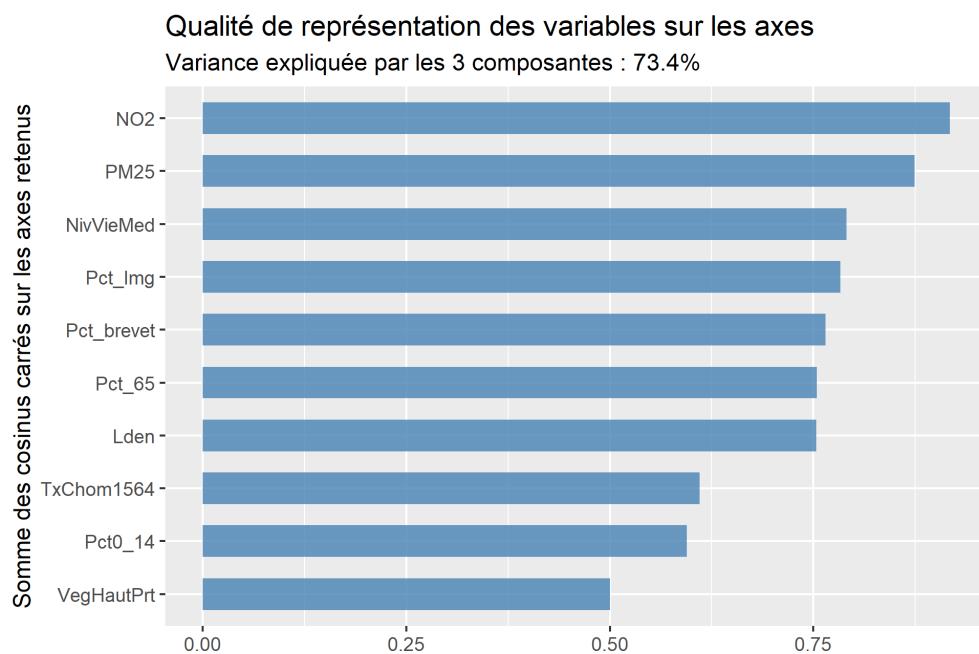


FIG. 1.21 : Graphique personnalisé avec la qualité des variables sur les axes retenus de l'ACP

```

tm_polygons(col = "CoordComp1", style = "cont",
            midpoint = 0, title = 'Coordonnées')+
  tm_layout(main.title = paste0("Axe 1 (", vppct[1],"%)"),
            attr.outside = TRUE, frame = FALSE, main.title.size = 1)
Carte2 <- tm_shape(CartoACP) +
  tm_polygons(col = "CoordComp2", style = "cont",
              midpoint = 0, title = 'Coordonnées')+
  tm_layout(main.title = paste0("Axe 2 (", vppct[2],"%)"),
            attr.outside = TRUE, frame = FALSE, main.title.size = 1)
tmap_arrange(Carte1, Carte2)

```

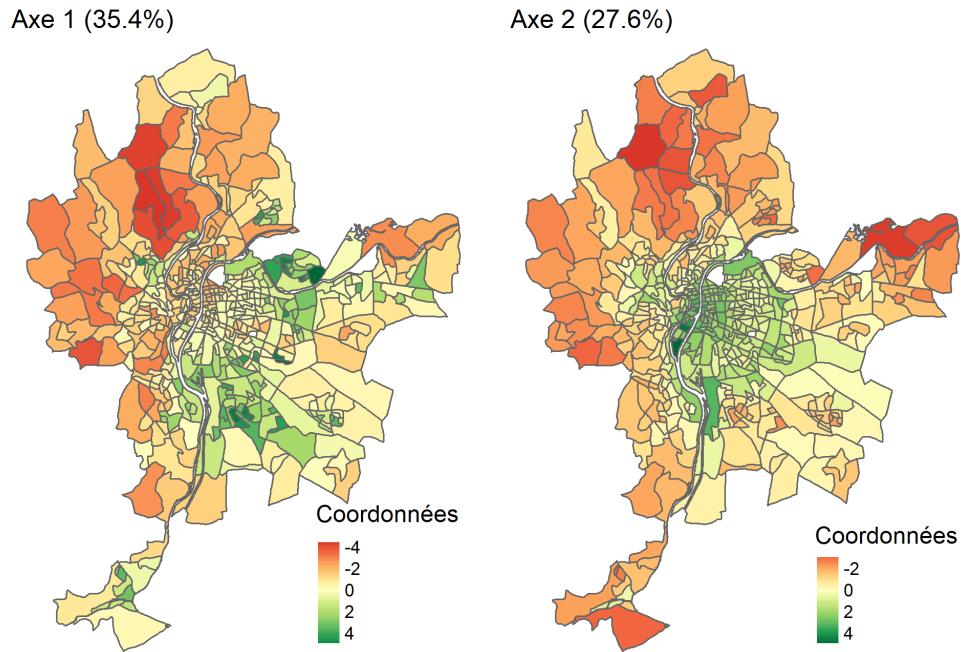


FIG. 1.22 : Cartographie des coordonnées factorielles des individus



Exploration interactive des résultats d'une ACP avec le package *explor*.

Vous avez compris qu'il ne suffit pas de calculer une ACP, il faut retenir les n premiers axes de l'ACP qui nous semblent les plus pertinents, puis les interpréter à la lecture des coordonnées factorielles, les cosinus carrés et les contributions des variables et des individus sur les axes. Il faut donc bien explorer les résultats à l'aide de tableaux et de graphiques! Cela explique que nous avons mobilisé de nombreux graphiques dans les deux sections précédentes (?? et ??). L'exploration des données d'une ACP peut aussi être réalisée avec des graphiques interactifs. Or, un superbe package dénommé *explor* (<https://juba.github.io/explor/>), reposant sur Shiny (<https://shiny.rstudio.com/>), permet d'explorer de manière interactive les résultats de plusieurs méthodes factorielles calculés avec FactorMinerR. Pour cela, il vous suffit de lancer les deux lignes de code suivantes :

```

library(explor)
explor(res.acp)

```

Finalement, *explor* permet également d'explorer les résultats d'une analyse des correspondances (AFC) et d'une analyse des correspondances multiples (ACM).

1.3 Analyse factorielle des correspondances (AFC)



Pour bien comprendre l'AFC, il est essentiel de bien maîtriser les notions de tableau de contingence (marges du tableau, fréquences observées et théoriques, pourcentages en ligne et en colonne, contributions au khi-deux) et de distance du khi-deux. Si ce n'est pas le cas, il est conseillé de (re)lire le chapitre ??.

Dans le chapitre ??, nous avons vu comment construire un tableau de contingence (figure ??) à partir deux variables qualitatives comprenant plusieurs modalités, puis comment vérifier s'il y a dépendance entre les deux variables qualitatives avec le test du khi-deux. Or, s'il y a bien dépendance, il est peut-être judicieux de résumer l'information que contient le tableau de contingence en quelques nouvelles variables synthétiques, objectif auquel répond l'analyse factorielle des correspondances (AFC).

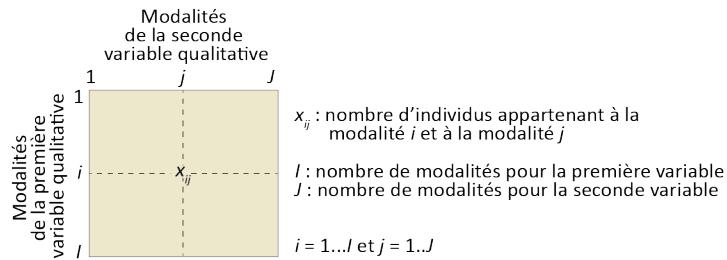


FIG. 1.23 : Tableau de contingence pour une AFC

À titre de rappel (section ??), l'AFC a été développée par le statisticien français Jean-Paul Benzécri (1973). Cela explique qu'elle est souvent enseignée et utilisée en sciences sociales dans les universités francophones, mais plus rarement dans les universités anglophones. Pourtant, les applications de l'AFC sont nombreuses dans différentes disciplines des sciences sociales comme illustrées avec les exemples suivants :

- En géographie, les modalités de la première variable du tableau de contingence sont souvent des entités spatiales (par exemple, régions, municipalités, quartiers, etc.) croisées avec les modalités d'une autre variable qualitative (catégories socioprofessionnelles, modes de transport, tranches de revenu des ménages ou des individus, etc.).
- En économie régionale, nous pourrions vouloir explorer un tableau de contingence croisant des entités spatiales (par exemple, MRC au Québec, départements en France) et les effectifs d'emplois pour différents secteurs d'activité.
- En sciences politiques, le recours à l'AFC peut être intéressant pour explorer les résultats d'une élection. Les deux variables qualitatives pourraient être les *circonscriptions électoralles* et les *partis politiques*. Le croisement des lignes et des colonnes du tableau de contingence représenterait le nombre de votes obtenus par un parti politique j dans la circonscription électorale i . Appliquer une AFC sur un tel tableau de contingence permettrait de révéler les ressemblances entre les différents partis politiques et celles entre les circonscriptions électorales.



Application d'une ACP sur un tableau de contingence transformé en un tableau avec les pourcentages en ligne : un bien mauvais calcul...

Il pourrait être tentant de transformer le tableau de contingence initial (tableau ??) en un tableau avec les pourcentages en lignes (tableau ??) afin de lui appliquer une analyse en composantes principales. Une telle démarche a deux inconvénients majeurs : chacune des modalités de la première variable qualitative (I) aurait alors le même poids ; chacune des modalités de la deuxième variable (J) aurait aussi le même poids. Or, à la lecture des marges en ligne et en colonne au tableau ??, il est clair que la modalité j_1 et i_1 comprennent plus bien d'individus que les autres modalités respectives.

Si nous reprenons le dernier exemple applicatif, cela signifierait que le même poids sera accordé à chaque parti puisque les variables sont centrées réduites en ACP (moyenne = 0 et variance = 1). Autrement dit, les grands partis traditionnels seraient ainsi sur le pied d'égalité que les autres partis. Aussi, chaque circonscription électorale aurait le même poids bien que certaines comprennent bien plus d'électeurs et d'électrices que d'autres.

TAB. 1.7 : Exemple de tableau de contingence pour l'ACF

	j1	j2	j3	j4	j5	Marge (ligne)
i1	357 060	22 010	276 625	65 000	29 415	750 110
i2	427 530	26 400	295 860	69 410	30 645	849 845
i3	147 500	6 545	34 545	4 415	1 040	194 045
i4	128 520	6 405	42 925	6 565	2 670	187 085
Marge (colonne)	1 060 610	61 360	649 955	145 390	63 770	1 981 085

\begin{table}[H]

\caption{Exemple d'un tableau de contingence transformé (% en ligne) pour l'ACP}

	V1	V2	V3	V4	V5
i1	47,6	2,9	36,9	8,7	3,9
i2	50,3	3,1	34,8	8,2	3,6
i3	76,0	3,4	17,8	2,3	0,5
i4	68,7	3,4	22,9	3,5	1,4

\end{table}

1.3.1 Recherche d'une simplification basée sur la distance du khi-deux

Sur le plan mathématique et des objectifs visés, l'ACF est similaire à l'ACP puisqu'elle permet d'explorer un tableau de trois façons : 1) en montrant les ressemblances entre un ensemble d'individus (I), 2) en révélant les liaisons entre les variables (J) et 3) en résumant le tout avec des variables synthétiques. Toutefois avec l'ACF, les ensembles I et J sont les modalités de deux variables qualitatives (dont le croisement forme un tableau de contingence) et elle est basée sur la distance du khi-deux (et non sur la distance euclidienne comme en ACP).

Ainsi, avec la distance du khi-deux, la proximité (ressemblance) entre deux lignes (i et l) et deux colonnes (j et k) est mesurée comme suit :

$$d_{\chi^2_{il}} = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{lj}}{f_{l.}} \right)^2 \quad (1.5)$$

$$d_{\chi^2_{jk}} = \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ik}}{f_{.k}} \right)^2 \quad (1.6)$$

Prenons un exemple fictif pour calculer ces deux distances. Le tableau ?? comprend trois modalités en ligne (I) et trois autres en colonnes (J). Le total des effectifs de ce tableau de contingence est égal à 1 665.

À partir des données brutes, il est facile de construire deux tableaux : le profil des lignes et le profil des colonnes (tableau ??), c'est-à-dire les proportions en ligne et en colonne.

En divisant les valeurs du tableau ?? par le grand total (1 665), nous obtenons tous les termes utilisés dans les équations (??) et (??) au tableau ?? :

- Les fréquences relatives dénommées f_{ij} .
- La marge $f_{i.}$ est égale à la somme des fréquences relatives en ligne.

TAB. 1.8 : Données brutes du tableau de contingence

	j1	j2	j3	Total (ligne)
i1	360	65	275	700
i2	420	70	290	780
i3	145	5	35	185
Total (colonne)	925	140	600	1 665

TAB. 1.9 : Profils des lignes et des colonnes

	j1	j2	j3	Total
Profil des lignes				
i1	0,514	0,093	0,393	1
i2	0,538	0,090	0,372	1
i3	0,784	0,027	0,189	1
Profil des colonnes				
i1	0,389	0,464	0,458	
i2	0,454	0,500	0,483	
i3	0,157	0,036	0,058	
Total	1,000	1,000	1,000	

- La marge $f_{.j}$ est égale à la somme des fréquences relatives en colonne.
- La somme de toutes les fréquences relatives est donc égale à 1, soit $\sum f_{i.}$ ou $\sum f_{.j}.$

Il est possible de calculer les distances entre les différentes modalités de I en appliquant l'équation (??); par exemple, la distance entre les observations i1 et i2 est égale à :

$$d_{(i1,i2)} = \frac{1}{0,556}(0,216 - 0,252)^2 + \frac{1}{0,084}(0,039 - 0,042)^2 + \frac{1}{0,360}(0,165 - 0,174)^2 = 0,003$$

Avec l'équation (??), la distance entre les modalités j1 et j2 de J est égale à :

$$d_{(j1,j2)} = \frac{1}{0,420}(0,216 - 0,039)^2 + \frac{1}{0,468}(0,252 - 0,042)^2 + \frac{1}{0,111}(0,087 - 0,003)^2 = 0,233$$

À la lecture du tableau ??, les modalités les plus sont semblables sont i1 et i2 (0,003) pour I et j1 et j3 (0,058) pour J .

Finalement, l'approche pour déterminer les axes factoriels de l'AFC est similaire à celle de l'AFC : les axes factoriels sont les droites orthogonales qui minimisent les distances aux points du profil des lignes, excepté que la métrique pour mesurer ces distances est celle du khi-deux (et non celle la distance euclidienne comme ACP). Pour plus détail sur le calcul de ces axes (notamment les formulations matricielles), consultez notamment Benzécri (1973), Escofier et Pagès (1998) et Lebart et al. (1995).

TAB. 1.10 : Données relatives du tableau de contingence (fij)

	j1	j2	j3	Total (fi.)
i1	0,216	0,039	0,165	0,420
i2	0,252	0,042	0,174	0,468
i3	0,087	0,003	0,021	0,111
Total (fij)	0,556	0,084	0,360	1,000