Projet Data Mining

Axel-Cleris Gailloty

Table of Contents

# Introduction

Aux Etats-Unis l’Enquête sur les bibliothèques publiques, en anglais Public Library Survey (PLS) est menée chaque année par l’Institut des services de muséologie et de bibliothèque dans le but d’examiner comment les bibliothèques publiques changent pour répondre aux besoins des usagers.

Les enquêtes réalisées sont destinées à l’endroit des politiques publiques afin de les aider à prendre des décisions. Les enquêtes permettent de collecter une importante quantité de données. Toutefois les analyses qui sont faites de ces données depuis la mise en place de l’enquête ne sont que descriptives et graphiques. Les analyses ne s’intéressent qu’à un phénomène à la fois et ne prennent pas compte les interactions qui peuvent exister entre les caractéristiques des bibliothèques.

C’est pourquoi dans le cadre de ce projet de Data Mining, je me propose de faire une étude approfondie du sujet en allant au-delà de simples statistiques descriptives pour réaliser des analyses factorielles des données.

Le travail est divisé en trois parties principales : dans la première partie nous analysons les données sous l’angle des statistiques descriptives. Nous présentons notamment les indicateurs de tendance centrale et les corrélations entre les variables. Ensuite nous analysons les données grâce aux méthodes d’analyses factorielles. A partir des méthodes d’analyses factorielles nous faisons dans la troisième partie une classification ascendante hiérarchique des données.

# Les données

Les données sur lesquelles j’ai décidé de travailler sont de 2014. Elles portent sur 9255 bibliothèques publiques pour 74 variables.

Dans les faits il y a deux jeux de données fournis par l’Institute of Museum and Library. Le premier correspond aux données brutes de l’enquête et des estimations faites sur les caractéristiques des bibliothèques. Le second jeu de données contient des données sur les Etats américains et fournissent notamment des données démographiques et économiques des Etats dans lesquels se trouvent les bibliothèques. Les deux jeux de données sont reliés par une clé étrangère qui permet de les fusionner si besoin.

Vu que mon intérêt porte principalement sur les bibliothèques en question, j’ai décidé de me focaliser sur le premier jeu de donnée. J’ai fusionné les deux jeux de données pour ne recueillir que quelques informations démographiques et économiques des Etats pour aider à situer le contexte géographique et économique des bibliothèques.

Pour le projet, je me suis limité à 22 variables. La liste des variables que j’ai sélectionnées est en annexe de ce projet.

## Actions réalisées pour nettoyer les données

Bien que le nettoyage des données ait pris une grande partie du temps consacré au projet, voilà sommairement les actions que j’ai réalisées pour permettre l’analyse des données.

Remplacement des données numériques négatives par des valeurs manquantes : les valeurs manquantes contenues dans le jeu de données sont codées par des nombres négatifs comme -3 et -1. Je les ai remplacées par des NA que R comprend.

Normalisation des noms des colonnes : certaines colonnes contenaient des caractères spéciaux dans leurs noms.

Suppression des données manquantes : Bien qu’il existe des méthodes d’imputation pour remplacer les données, j’ai jugé nécessaire de ne pas les utiliser car les valeurs sont souvent des caractéristiques qu’on ne peut pas prédire (l’Etat dans lequel se trouve la bibliothèque, sa structure administrative, le nombre de collection audio …)

# Statistiques descriptives

Les statistiques descriptives nous permettent de comprendre la distribution de nos données. Nous allons présenter les indicateurs de moments et de tendances centrales tels le minimum, la moyenne, l’écart-type, la médiane et le maximum.  
Pour les variables qualitatives, nous affichons le nombre de modalités et la modalité la plus fréquente pour la colonne (mode).

Statistiques descriptives des colonnes numériques

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min | Moyenne | Ecart\_type | Max | Médiane |
| print\_collection | 0 | 9.1e+04 | 414998 | 2.2e+07 | 2.9e+04 |
| audio\_collection | 0 | 5.4e+03 | 22733 | 7.5e+05 | 1.1e+03 |
| county\_population | 262 | 3.0e+05 | 782641 | 1.0e+07 | 6.3e+04 |
| downloadable\_audio | 0 | 6.1e+03 | 12665 | 2.6e+05 | 2.5e+03 |
| physical\_video | 0 | 7.3e+03 | 27651 | 8.3e+05 | 2.0e+03 |
| hours\_open | 104 | 4.3e+03 | 9389 | 2.2e+05 | 2.4e+03 |
| print\_subscriptions | 0 | 1.6e+02 | 1136 | 7.5e+04 | 5.0e+01 |
| registered\_users | 0 | 2.1e+04 | 86631 | 2.9e+06 | 4.2e+03 |
| interlibrary\_loans\_provided | 0 | 7.5e+03 | 26598 | 6.0e+05 | 3.8e+02 |
| interlibrary\_loans\_received | 0 | 7.5e+03 | 27798 | 9.6e+05 | 5.1e+02 |
| childrens\_programs | 0 | 3.1e+02 | 1003 | 2.9e+04 | 1.0e+02 |
| young\_adult\_programs | 0 | 5.4e+01 | 280 | 1.1e+04 | 8.0e+00 |
| internet\_computer\_use | 0 | 3.9e+04 | 184719 | 6.5e+06 | 7.1e+03 |
| ST\_state\_population | 582658 | 1.0e+07 | 8969511 | 3.8e+07 | 6.6e+06 |
| wireless\_internet\_sessions | 0 | 2.7e+04 | 272179 | 1.4e+07 | 1.3e+03 |
| branch\_libraries | 0 | 9.7e-01 | 4 | 9.1e+01 | 0.0e+00 |
| local\_government\_operating\_revenue | 0 | 1.2e+06 | 5507524 | 1.4e+08 | 1.7e+05 |
| local\_cooperative\_agreements | 0 | 9.5e+00 | 23 | 6.4e+02 | 3.0e+00 |
| digital\_collection | 0 | 2.0e+04 | 39157 | 1.0e+06 | 6.1e+03 |
| mls\_librarians | 0 | 3.8e+00 | 16 | 4.9e+02 | 5.0e-01 |

Ce qu’on peut constater de l’observation de ce tableau est que l’amplitude des colonnes numériques varie grandement. Une chose est importante à noter aussi : l’écart-type de chacune des colonnes est très élevée, cela indique une grande disparité autour de la moyenne des colonnes. A priori nous ne pouvons pas dire si ces colonnes sont normalement distribuées.

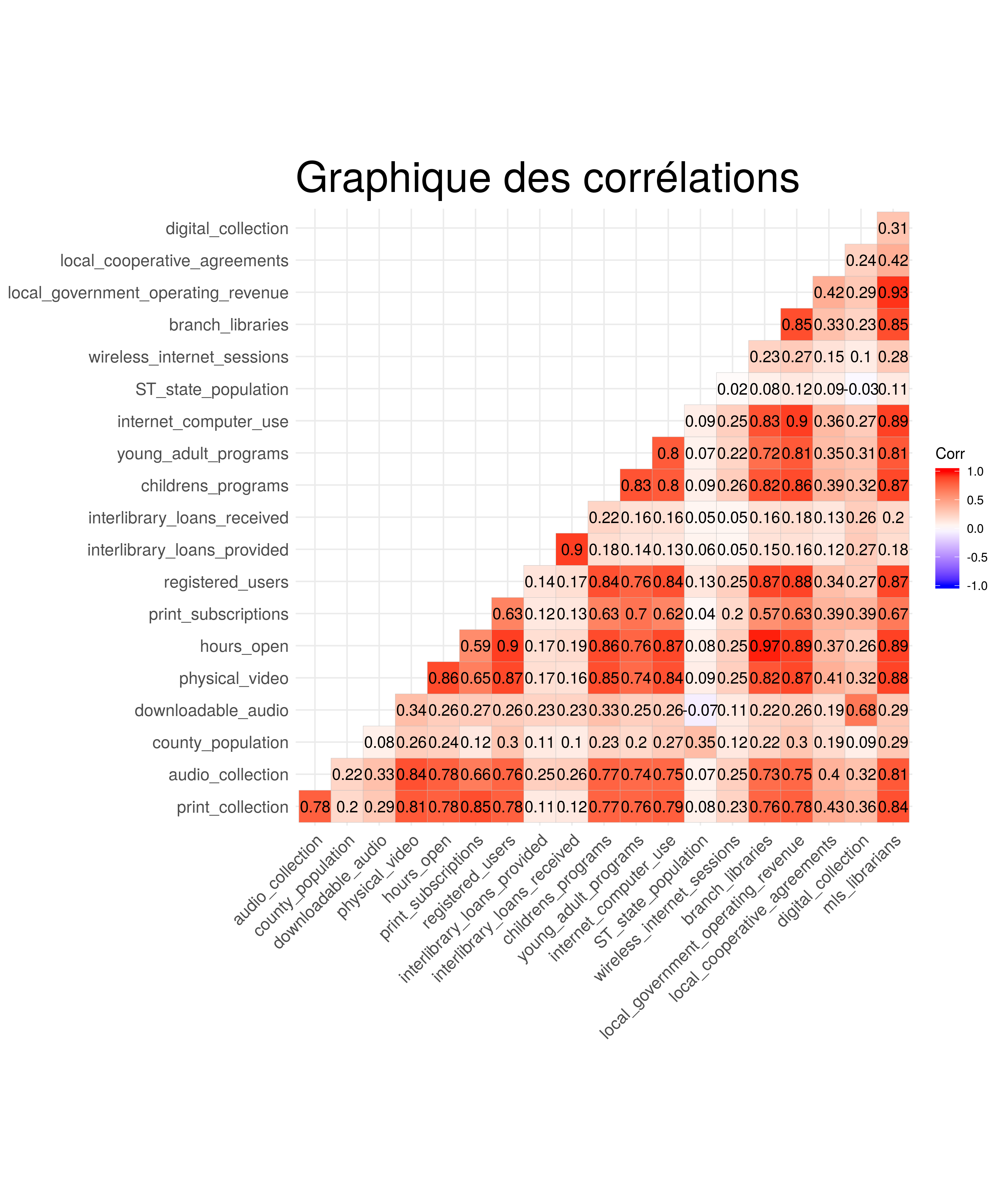
Il existe aussi dans le jeu de données des variables catégorielles. Nous allons afficher les statistiques associées.

Liste des variables qualitatives

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Modalities | MostFrequent | NbFrequent | Missing |
| library\_name | 5583 | ANCHOR POINT PUBLIC LIBRARY | 1 | 0 |
| administrative\_structure | 3 | SO | 4419 | 0 |
| interlibrary\_relationship | 3 | ME | 4020 | 0 |
| ST\_region\_code | 8 | st\_2 | 1306 | 0 |
| legal\_basis | 9 | CI | 2583 | 0 |

# Analyse des corrélations

L’analyse des corrélations nous permet de déterminer le niveau de dépendance qui existe entre les variables numériques du jeu. Pour chaque paire de variable nous calculons le coefficient de corrélation qui résume leur dépendance. Les corrélations entre les variables s’afficheront dans une matrice des corrélations. Pour faciliter la lecture de la matrice des corrélations, nous pouvons la représenter graphiquement.



Le nombre d’individus étant supérieur à 30, les coefficients de corrélation que nous obtenons sur ce graphique sont significatifs.  
L’analyse des corrélations entre les variables quantitatives nous montre que globalement la distribution des corrélations est assez variée. D’une part nous avons des corrélations assez élevées (plus de 0.75) et d’autres qui sont faibles. Toutefois nous observons que toutes les corrélations sont positives, autrement dit toutes les variables sont **positivement** liées dans une relation linéaire. A partir d’ici nous pouvons prédire que cela pourra éventuellement avoir un impact sur les analyses factorielles que nous effectuerons.

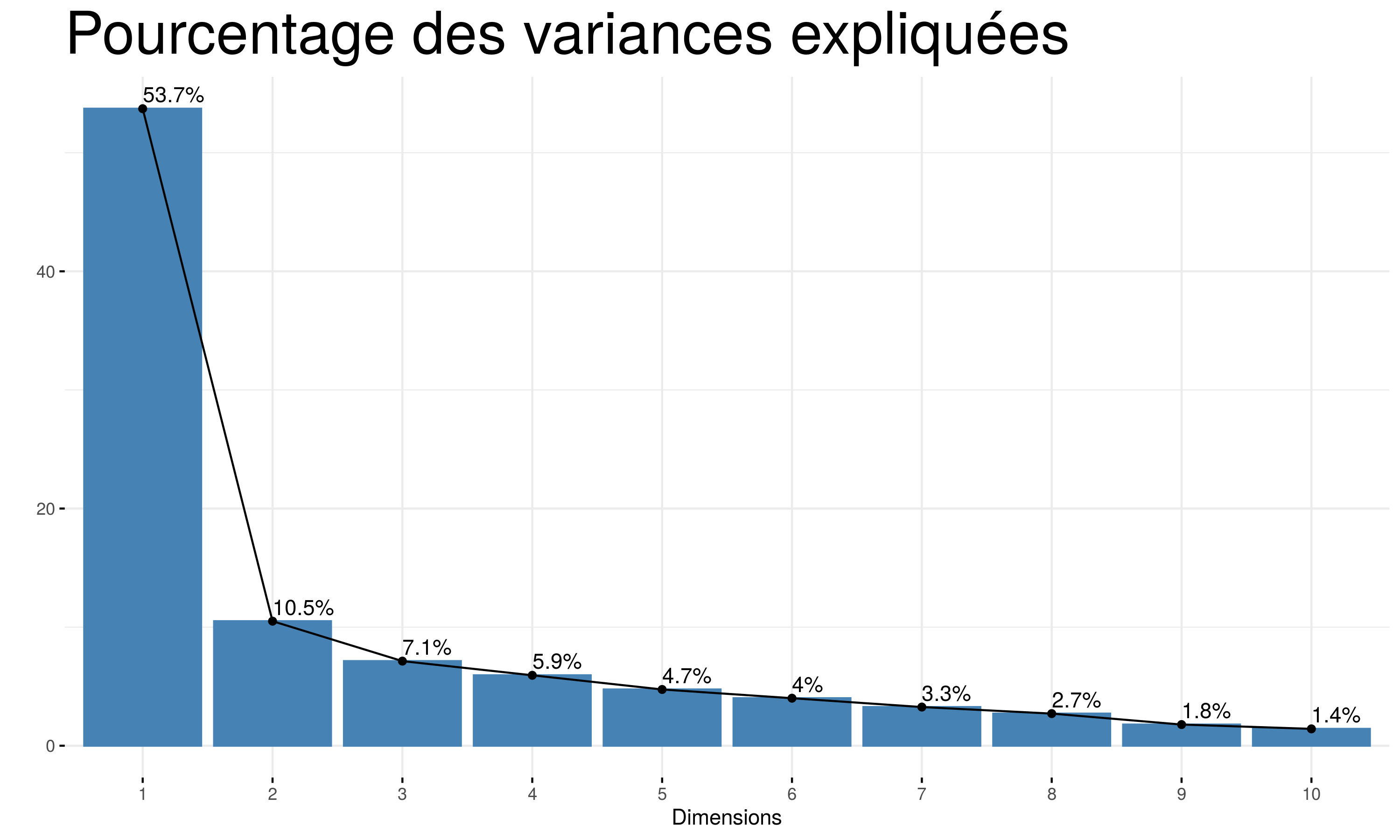
# Analyse de données

## L’analyse en composantes principales

L’analyse en composantes principales (ACP) est une méthode factorielle de réduction de dimension des données quantitatives complexes. L’ACP se base sur la matrice des variances-covariances. Elle cherche à résumer la dispersion des données. Le jeu de données sur lequel ce projet est réalisé est composé à majorité des données quantitatives, comme nous venons de voir dans les statistiques descriptives et les corrélations. Donc le choix d’une ACP pour analyser les données est justifié. L’analyse en composante principale que nous effectuons est centrée et réduite. Le but de centrer et réduire les données est de donner le même poids à chaque variable présente dans le jeu de données.

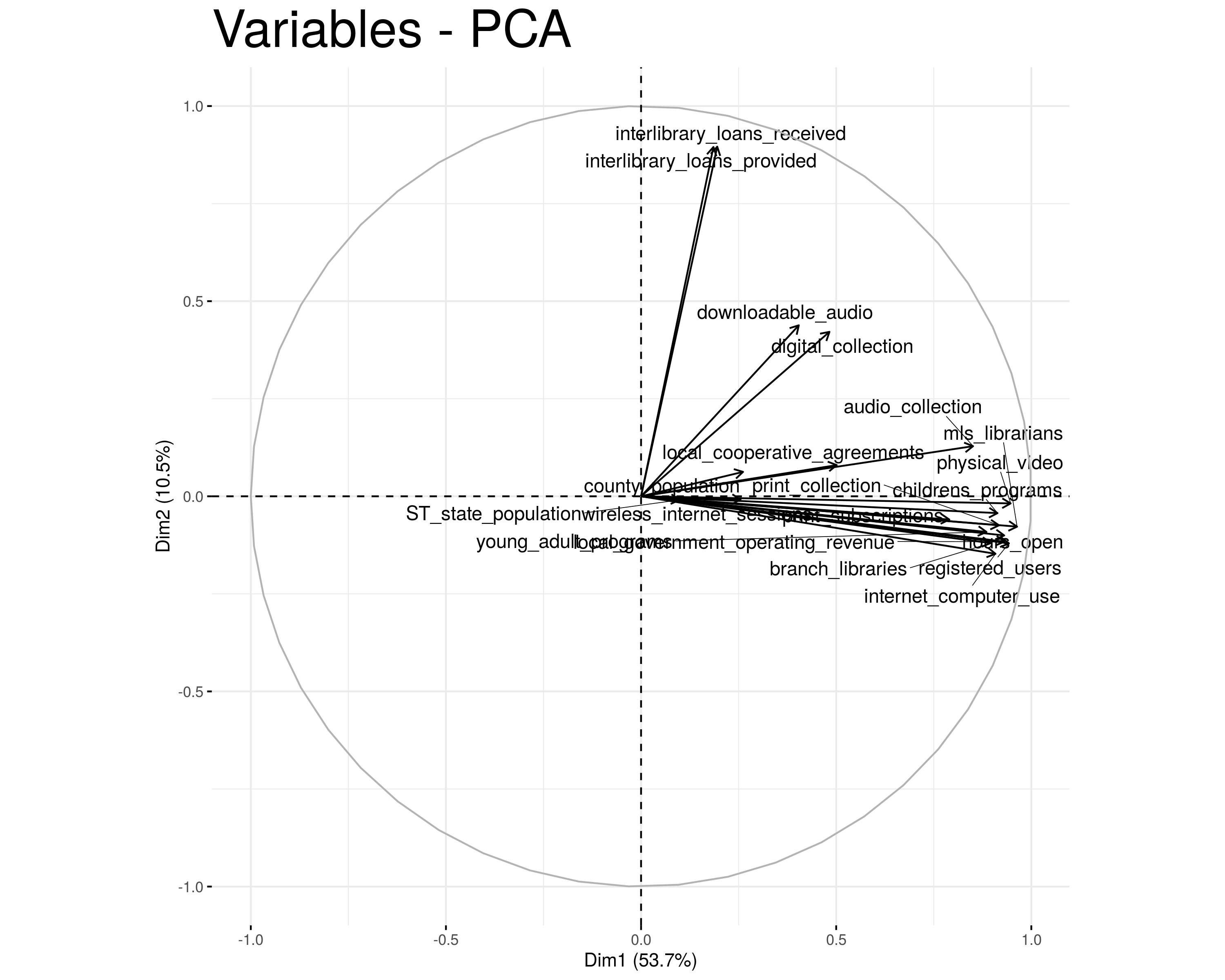
Le jeu de données initial contient 9225 observations, mais pour des raisons d’efficience (notamment pour la classification ascendante hiérarchique) nous n’allons utiliser qu’un échantillon tiré aléatoirement de 3000 individus du jeu de données. Ainsi nous travaillons sur un jeu de données de 3000 individus et 22 variables.

Nous allons réaliser une première ACP sur le jeu de données, en prenant toutes les colonnes numériques comme des colonnes actives (colonnes qui entrent dans le calcul) et les colonnes *catégoriques structure administratives*, *les relations avec les autres bibliothèques*, *le code de la région* et *le statut juridique des bibliothèques*. Voici les résultats de cette première ACP :



Ce graphique présente le pourcentage des variances expliquées par les axes factoriels. Le premier axe explique à lui seul 50% des variances, autrement dit, si on ne retient que cet axe on peut expliquer 50% des différences qui existent entre les bibliothèques. Or nous observons que le deuxième axe n’explique que 10%, on peut se douter qu’il y a un effet de taille dans l’analyse en composante principale.

Regardons le cercle des corrélations de l’ACP. Ce cercle montre la force de l’association des variables à deux axes factoriels (ici les axes 1 et 2).



On observe que toutes les variables sont positivement corrélées à l’axe 1 et très peu sont négativement corrélées à l’axe 2. Il s’agit d’un effet de taille.

## Observation d’un effet de taille

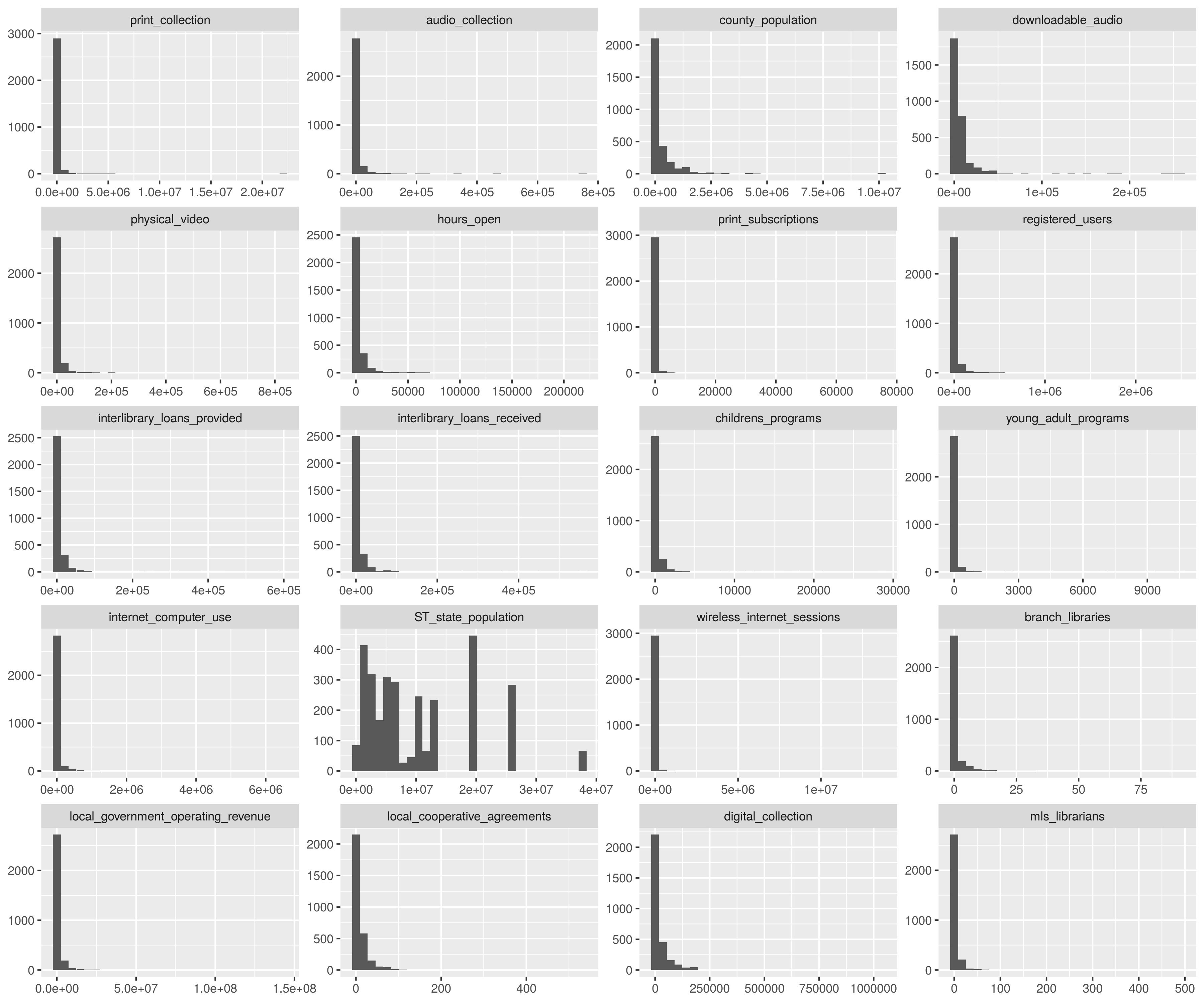
On observe un effet de taille dans une ACP sur les variables lorsque tous les points sont regroupés du même côté d’un point factoriel. La première analyse que nous venons de faire montre clairement un effet de taille car tous les points sont positivement corrélés à l’axe 1 du plan factoriel.

Il existe plusieurs façons de corriger un effet de taille. La méthode la plus utilisée pour corriger l’effet de taille est de transformer les variables quantitatives en valeurs catégoriques et utiliser une autre méthode d’analyse factorielle, en l’occurence l’analyse des correspondances multiples (ACM).

Nous allons commencer dans un premier temps, à nous inspirer des statistiques descriptives des variables quantitatives et de la distribution des colonnes pour transformer les données numériques en données qualitatives.

## Correction de l’effet de taille

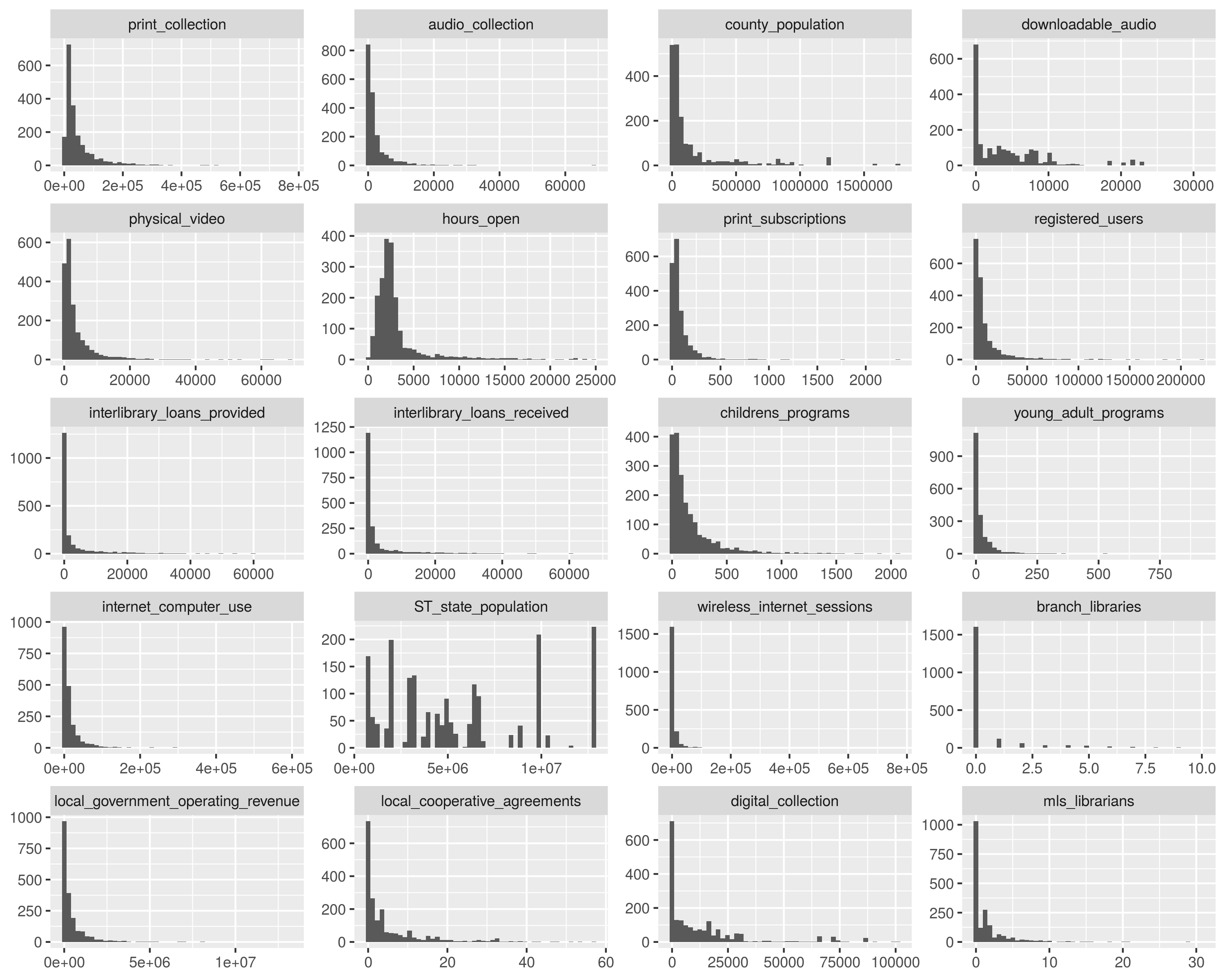
Il convient dans un premier d’observer la distribution de chaque variable pour se faire une idée du nombre de modalités qu’on peut créer pour chaque variable.



Au regard de ce graphique nous observons une polarisation extrême de la distribution de colonnes numériques. Notre but dans la correction de l’effet de taille est de transformer les colonnes numériques en des catégories, autrement dit, nous cherchons à discrétiser ces colonnes. **Or en ACM la part de l’inertie totale due à une modalité est d’autant plus grande que la modalité est rare**, autrement dit les modalités peu fréquentes ont plus de poids que les modalités fréquentes, donc il est important que nous nous assurions que les intervalles que nous allons créer aient des fréquences à peu près équivalentes. Cela implique que d’enlever les bibliothèques qui prennent des valeurs extrêmes afin de ne pas biaiser les indicateurs de moments (quartiles) sur lesquels nous nous basons pour diviser les colonnes numériques et que ces bibliothèques constituent à elles seules des catégories du fait des valeurs qu’elles prennent.

Pour enlever les bibliothèques ayant les valeurs extrêmes, nous allons utiliser la méthode de 3 sigmas, bien que les colonnes ne soient pas distribuées selon une loi normale. Cette méthode permet d’enlever les observations qui ont des valeurs supérieures à 3 fois l’écart-type de la colonne.

Après avoir appliqué la méthode des 3 sigmas, voilà la nouvelle distribution des colonnes.



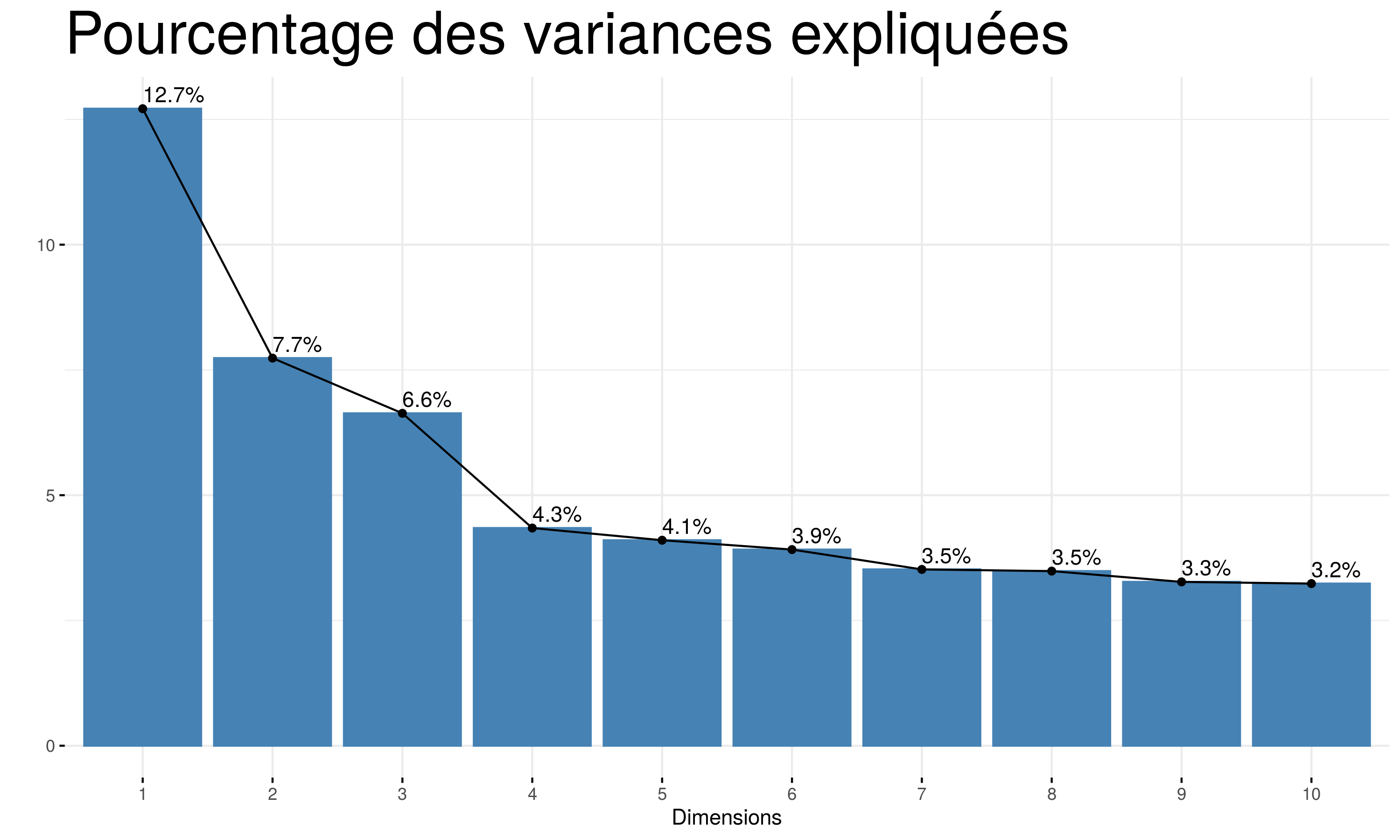
Le fait d’enlever les observations extrêmes a permis de réduire l’amplitude des colonnes, toutefois la distribution n’est pas normale.

Certaines colonnes numériques peuvent être facilement converties en colonnes catégoriques binaires. La variable branch\_libraries qui représente le nombre de branches qu’une bibliothèque possède peut être transformée pour prendre la valeur 1 si elle possède des branches et 0 sinon.

Nous ferons ainsi pour les colonnes similaires. Nous transformons les variables interlibrary\_loans\_provided et interlibrary\_loans\_received en une seule variable qui indique si oui ou non la bibliothèque prête plus qu’elle n’en emprunte. Et pour les autres variables, nous allons les discrétiser selon leurs indicateurs de moments, principalement la médiane et les autres quartiles.

# L’analyse des correspondances multiples

L’analyse des correspondances multiples est une technique descriptive visant à résumer l’information contenu dans un grand nombre de variables afin de faciliter l’interprétation des corrélations existantes entre ces différentes variables. On cherche à savoir quelles sont les modalités corrélées entre elles.



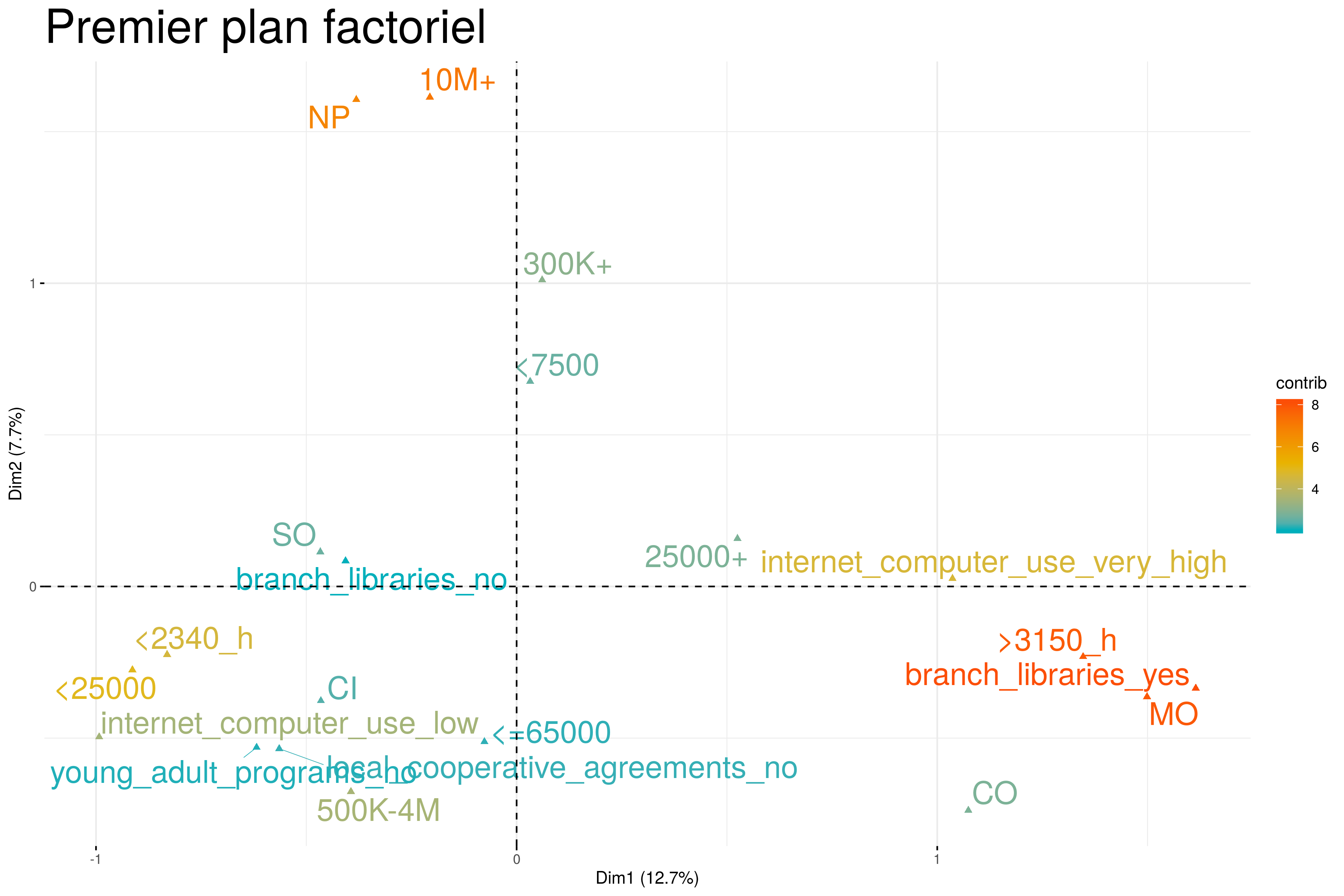
En ACM, les pourcentages expliqués par les dimensions sont en général plus faibles que pour une ACP. Cela est due principalement au fait que toutes les modalités des colonnes sont transformées elles-mêmes en colonnes. Ce qui fait exploser le nombre des colonnes.

Nous allons dans le cadre de ce projet retenir 4 dimensions pour expliquer la variance entre les bibliothèques. Ces 4 dimensions expliquent en total 31.3% de la variance totale. Le choix de 4 dimensions nous permet également de représenter 2 plans factoriels sur lesquels nous observons la forme que prend le nuage des points des individus.

Etant donné le grand nombre des variables créées par l’ACM, nous n’allons représenter pour chaque plan factoriel que les 20 premières contributions pour ne pas rendre illisible le graphique.

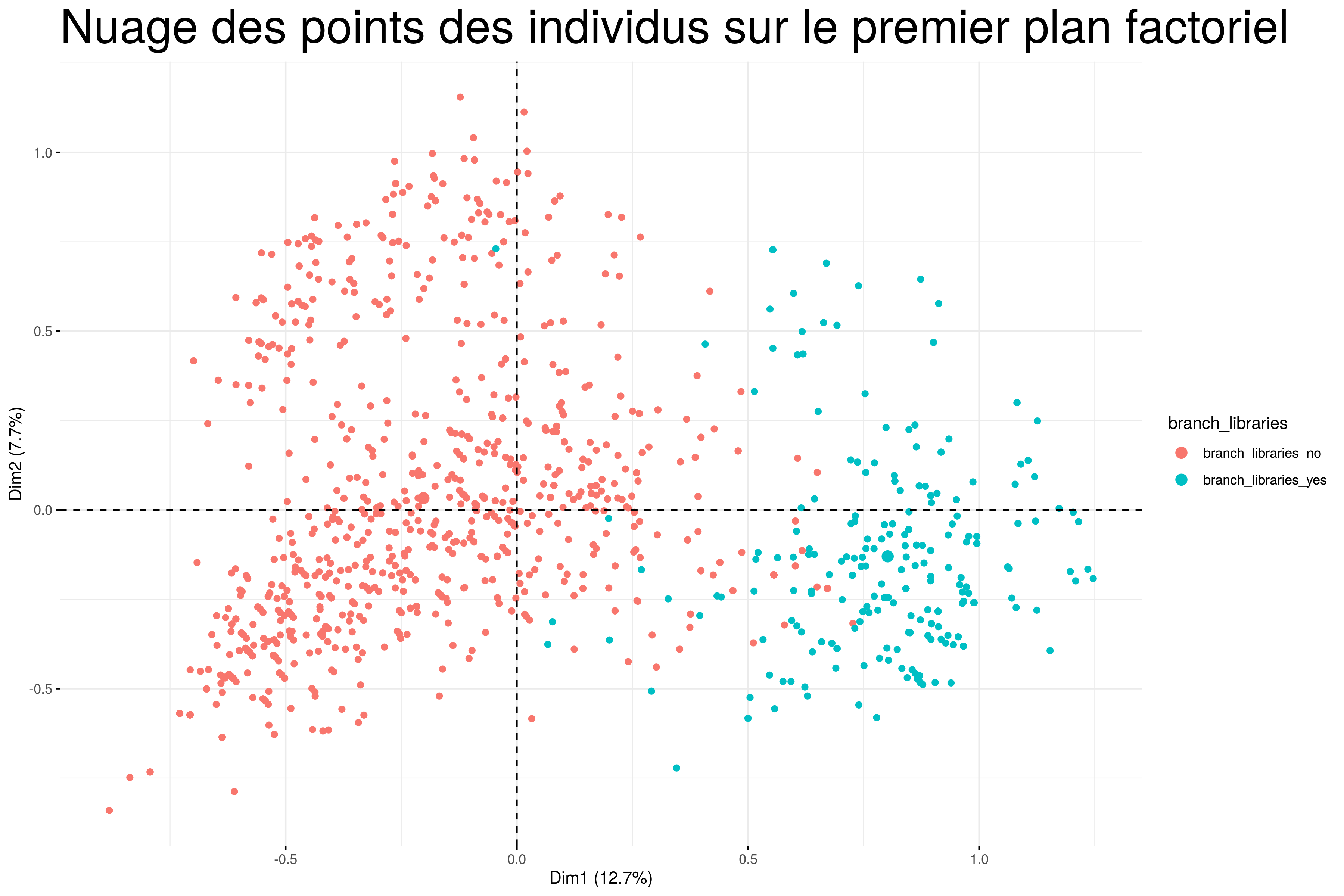
## Analyse graphique

### Premier plan factoriel



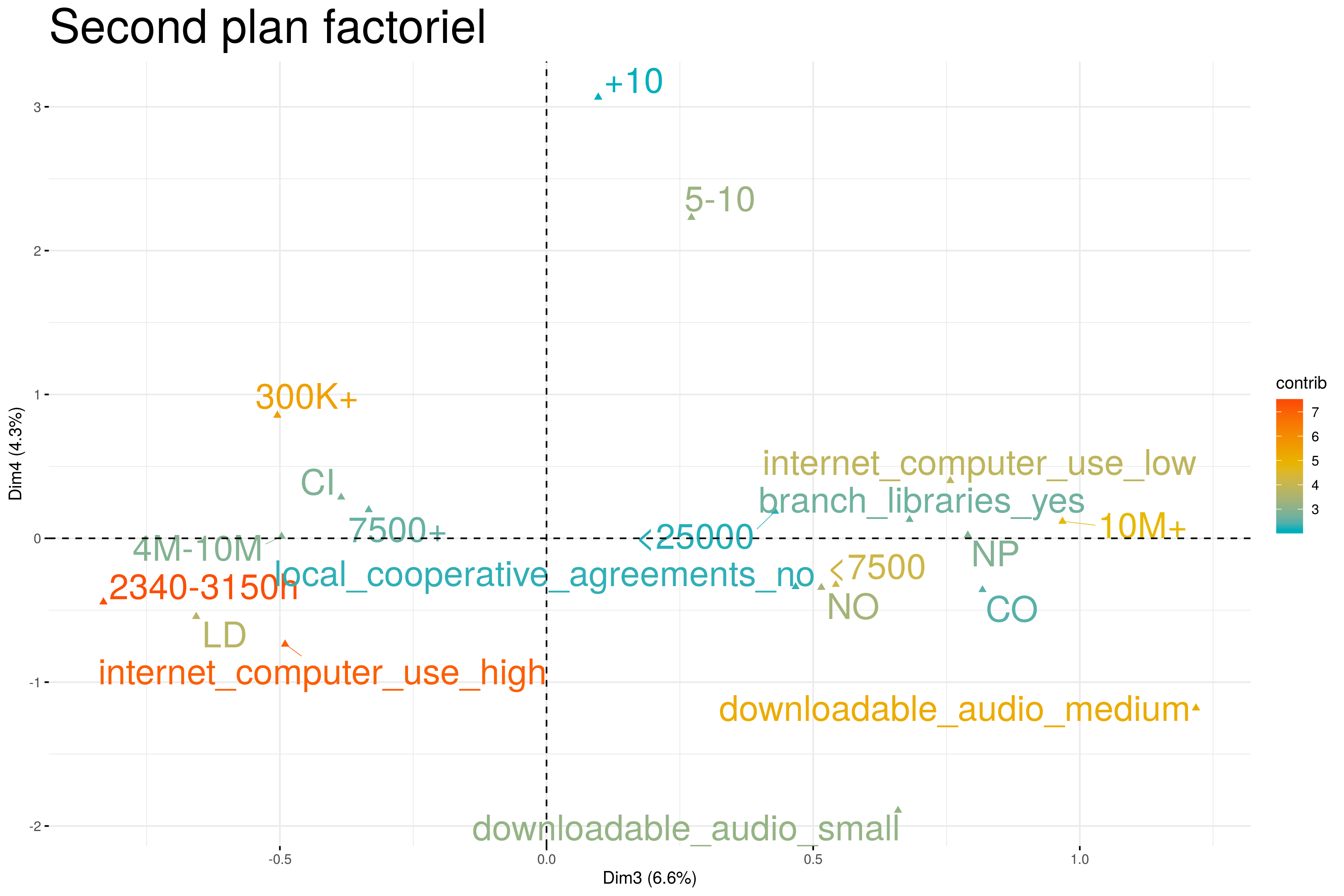
Le premier plan factoriel fait sortir une opposition assez nette entre les bibliothèques deux catégories de bibliothèques : d’un côté, on observe des bibliothèques possédant des branches (branch\_libraries\_yes), qui ont un catalogue de plus de 25000 oeuvres imprimées et une utilisation très importante d’Internet situées à droite du cadran et de l’autre côté des bibliothèques qui n’ont pas de branches, qui ouvrent moins 2340 heures (hours\_open <2340\_h) et qui sont des bibliothèques à guichet unique (administrative\_structure = Single Outlet (S0)).

Représentons également le nuage de points des individus sur le premier plan factoriel en coloriant les individus (bibliothèques) selon qu’ils possèdent ou non une branche.



Nous comprenons mieux à l’aide de ce nuage des points des individus colorié pourquoi la variable branch\_libraries est d’une grande contribution. On observe qu’elle discrimine nettement les bibliothèques. A gauche se trouvent des bibliothèques qui n’ont pas de branches et à droite celles qui en ont.

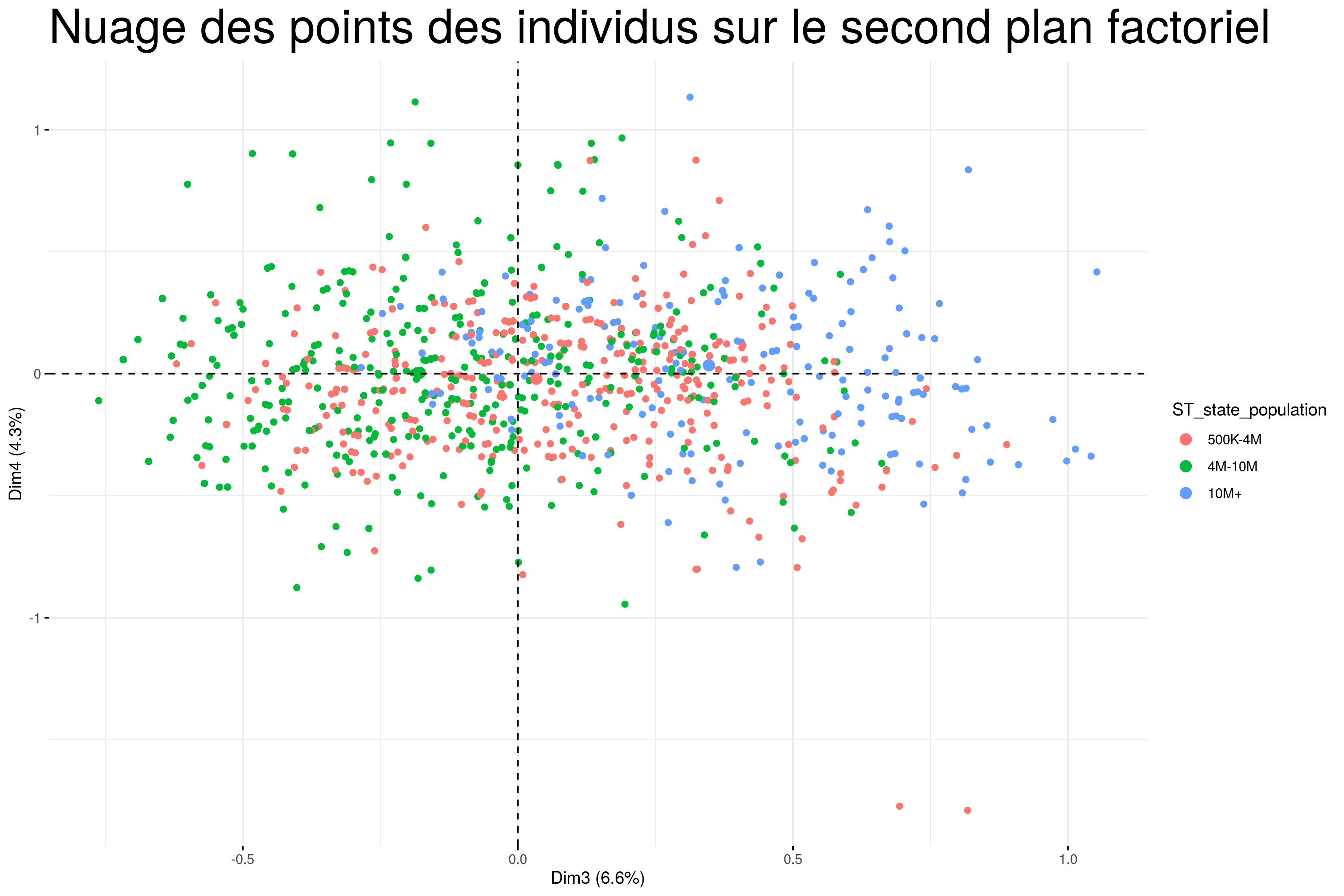
### Second plan factoriel



Le second plan factoriel met en opposition la taille des Etats dans lesquels se trouvent les bibliothèques, le nombre de libraires diplômés, le téléchargement des fichiers audio, la structure administrative des bibliothèques et leurs statuts juridiques.  
A droite du plan se trouvent des bibliothèques qui se situent dans les Etats de plus de plus de 10 million d’habitats (10M+). Ce sont des bibliothèques gérées par les comtés (CO) et qui ont un statut d’association à but non lucratif (NP). A gauche se trouvent des bibliothèques plutôt modernes, qui utilisent Internet, qui sont gérées par la municipalité (CI) ou par les quartiers (LD).

Nous détaillerons davantage les caractéristiques des bibliothèques lorsqu’elles seront classifiées hiérarchiquement en fonction de leurs similitudes.

Représentons le nuage des points bibliothèques sur le deuxième axe factoriel.



Ce second nuage colorié par la taille des Etats montre une saturation. Cela signifie que le pouvoir explicatif de ce plan est plus faible que le premier. En effet le 3e axe résume 6.6% de l’inertie du nuage et le 4e résume 4.3%. Les résultats sont donc à relativiser. Il y a beaucoup de chevauchement et on ne voit pas de tendance dominante.

## Résultats tabulaires

Le package que nous utilisons pour l’analyse des données faire ressortir dans de nombreux tableaux les résultats intermédiaires à l’interprétation des graphiques que nous venons de commenter. Les tableaux des résultats contiennent par exemple la qualité de représentation des modalités et des individus. La plupart des informations contenues dans ces tableaux aident à affiner l’interprétation mais elles peuvent également être redondantes. Nous allons afficher uniquement le qui représente la qualité de représentation d’une modalité sur un axe factoriel. Le reste des résultats est placé dans les annexes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 5 |
| MA | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 |
| MO | 0.65 | 0.04 | 0.08 | 0.00 | 0.00 |
| SO | 0.71 | 0.04 | 0.08 | 0.00 | 0.00 |
| <25000 | 0.48 | 0.04 | 0.11 | 0.02 | 0.03 |
| 25000+ | 0.48 | 0.04 | 0.11 | 0.02 | 0.03 |
| audio\_collection\_no | 0.01 | 0.02 | 0.01 | 0.00 | 0.00 |
| audio\_collection\_yes | 0.01 | 0.02 | 0.01 | 0.00 | 0.00 |
| <=65000 | 0.01 | 0.31 | 0.00 | 0.16 | 0.00 |
| 65K-300K | 0.00 | 0.03 | 0.03 | 0.00 | 0.01 |
| 300K+ | 0.00 | 0.26 | 0.06 | 0.18 | 0.03 |
| HQ | 0.01 | 0.00 | 0.01 | 0.01 | 0.10 |
| ME | 0.08 | 0.04 | 0.14 | 0.07 | 0.09 |
| NO | 0.07 | 0.04 | 0.13 | 0.05 | 0.05 |
| downloadable\_audio\_small | 0.00 | 0.00 | 0.01 | 0.10 | 0.01 |
| downloadable\_audio\_medium | 0.00 | 0.08 | 0.10 | 0.10 | 0.05 |
| downloadable\_audio\_high | 0.00 | 0.05 | 0.11 | 0.20 | 0.06 |
| physical\_video\_no | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 |
| physical\_video\_yes | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 |
| <2340\_h | 0.49 | 0.04 | 0.10 | 0.03 | 0.04 |
| 2340-3150h | 0.01 | 0.12 | 0.30 | 0.08 | 0.07 |
| >3150\_h | 0.71 | 0.02 | 0.04 | 0.01 | 0.00 |
| young\_adult\_programs\_no | 0.10 | 0.08 | 0.07 | 0.02 | 0.01 |
| young\_adult\_programs\_yes | 0.10 | 0.08 | 0.07 | 0.02 | 0.01 |
| internet\_computer\_use\_low | 0.23 | 0.06 | 0.14 | 0.04 | 0.01 |
| internet\_computer\_use\_medium | 0.11 | 0.00 | 0.01 | 0.00 | 0.03 |
| internet\_computer\_use\_high | 0.00 | 0.04 | 0.12 | 0.26 | 0.05 |
| internet\_computer\_use\_very\_high | 0.44 | 0.00 | 0.00 | 0.10 | 0.00 |
| 500K-4M | 0.10 | 0.29 | 0.01 | 0.00 | 0.22 |
| 4M-10M | 0.15 | 0.00 | 0.18 | 0.00 | 0.31 |
| 10M+ | 0.01 | 0.58 | 0.21 | 0.00 | 0.02 |
| CC | 0.00 | 0.01 | 0.00 | 0.00 | 0.07 |
| CI | 0.18 | 0.12 | 0.12 | 0.07 | 0.04 |
| CO | 0.15 | 0.07 | 0.08 | 0.02 | 0.02 |
| LD | 0.08 | 0.00 | 0.10 | 0.07 | 0.36 |
| MJ | 0.04 | 0.01 | 0.03 | 0.00 | 0.05 |
| NL | 0.00 | 0.00 | 0.01 | 0.06 | 0.01 |
| NP | 0.03 | 0.52 | 0.12 | 0.00 | 0.01 |
| OT | 0.00 | 0.06 | 0.03 | 0.03 | 0.00 |
| SD | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 |
| wireless\_internet\_sessions\_no | 0.00 | 0.00 | 0.03 | 0.00 | 0.03 |
| wireless\_internet\_sessions\_yes | 0.00 | 0.00 | 0.03 | 0.00 | 0.03 |
| branch\_libraries\_no | 0.66 | 0.03 | 0.12 | 0.00 | 0.00 |
| branch\_libraries\_yes | 0.66 | 0.03 | 0.12 | 0.00 | 0.00 |
| local\_cooperative\_agreements\_no | 0.10 | 0.09 | 0.07 | 0.04 | 0.01 |
| local\_cooperative\_agreements\_yes | 0.10 | 0.09 | 0.07 | 0.04 | 0.01 |
| <7500 | 0.00 | 0.28 | 0.18 | 0.06 | 0.04 |
| 7500+ | 0.00 | 0.28 | 0.18 | 0.06 | 0.04 |
| <5 | 0.12 | 0.00 | 0.00 | 0.19 | 0.01 |
| 5-10 | 0.08 | 0.00 | 0.00 | 0.12 | 0.01 |
| +10 | 0.04 | 0.00 | 0.00 | 0.08 | 0.00 |

Le montre la qualité de représentation des modalités sur les axes factoriels où ils sont projettés. Le cosinus carré prend une valeur entre 0 et 1. Nous voyons grâce à ce tableau que plusieurs des modalités ont de très faibles , seules quelques unes sont bien représentées. Sur l’axe 1, les modalités MO, SO, <2500, 2500+, <2340h, >3150h, internet\_use, branch\_libraries (yes et no) ont des cosinus carré supérieurs à 0.4, cela montre qu’ils sont assez bien représentés et cela signifie que ces modalités caractérisent bien l’axe. Ce qu’on observe également c’est que la valeur des diminue à mesure qu’on passe à des axes qui ont de faibles contributions. Sur l’axe 2, le cosinus carré le plus élevé est de 0.52 et est prise par la modalité NP. Les autres modalités ayant des qualités de représentation assez élevée sont <7500 et 7500+ qui représente le niveau de collection digitale des bibliothèques.

### Variables quantitatives supplémentaires

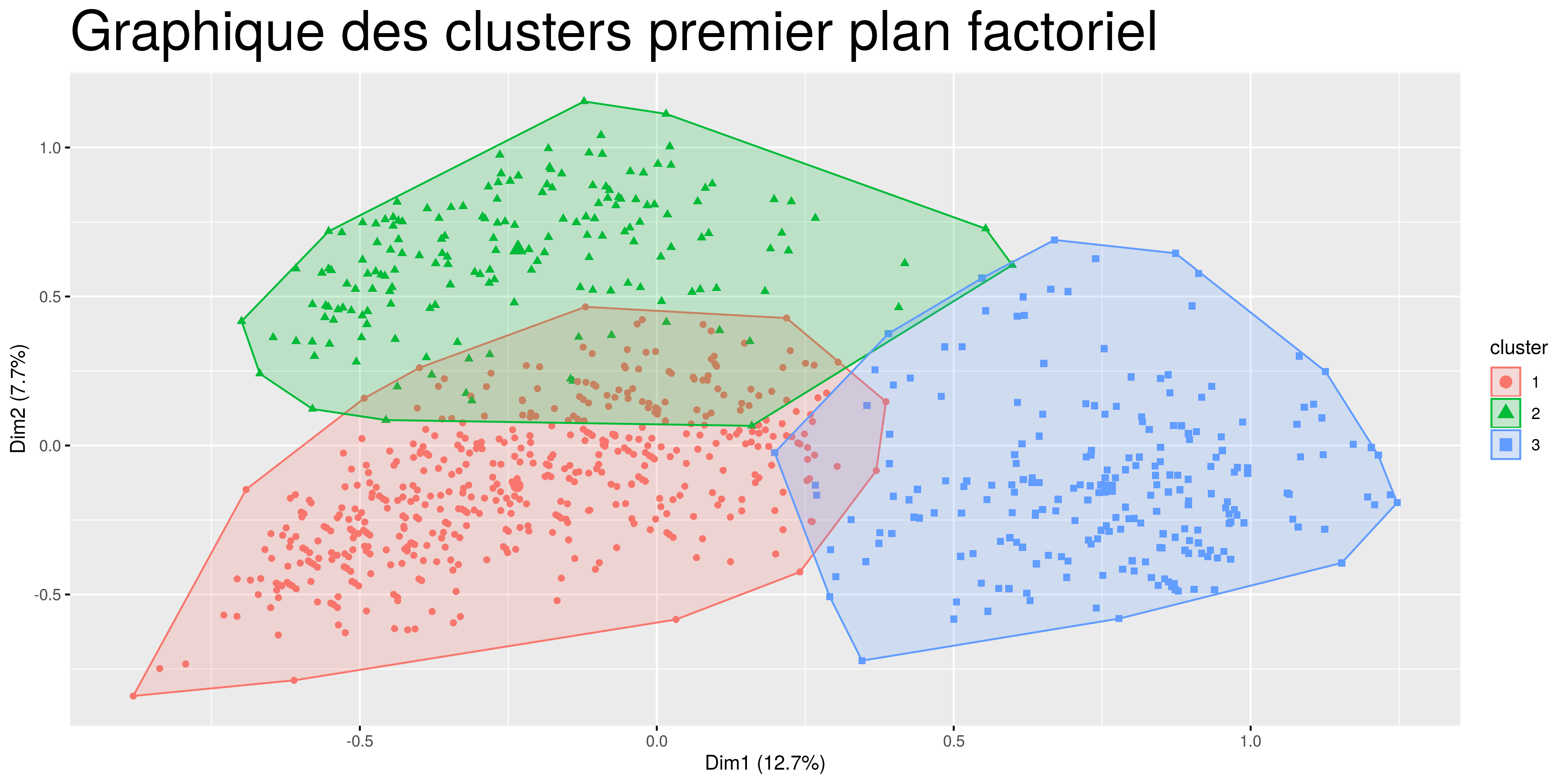
Nous avons spécifié dans l’ACM des variables quantitatives supplémentaires (ou encore illustratives) pour permettre d’affiner l’interprétation. Une variable illustrative n’intervient pas dans le calcul mais aide à placer dans les résultats de l’ACM dans un contexte.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 5 |
| print\_subscriptions | 0.53 | 0.04 | -0.06 | 0.23 | -0.01 |
| registered\_users | 0.62 | 0.02 | 0.08 | 0.31 | -0.05 |
| childrens\_programs | 0.60 | 0.11 | -0.04 | 0.22 | -0.02 |
| local\_government\_operating\_revenue | 0.61 | -0.01 | -0.05 | 0.36 | 0.07 |

On observe que l’effet de taille constatée dans l’ACP persiste avec les variables quantitatives illustratves. Toutes les variables sont positivement corrélées avec le premier axe factoriel.

# Classification ascendante hiérarchique (CAH)

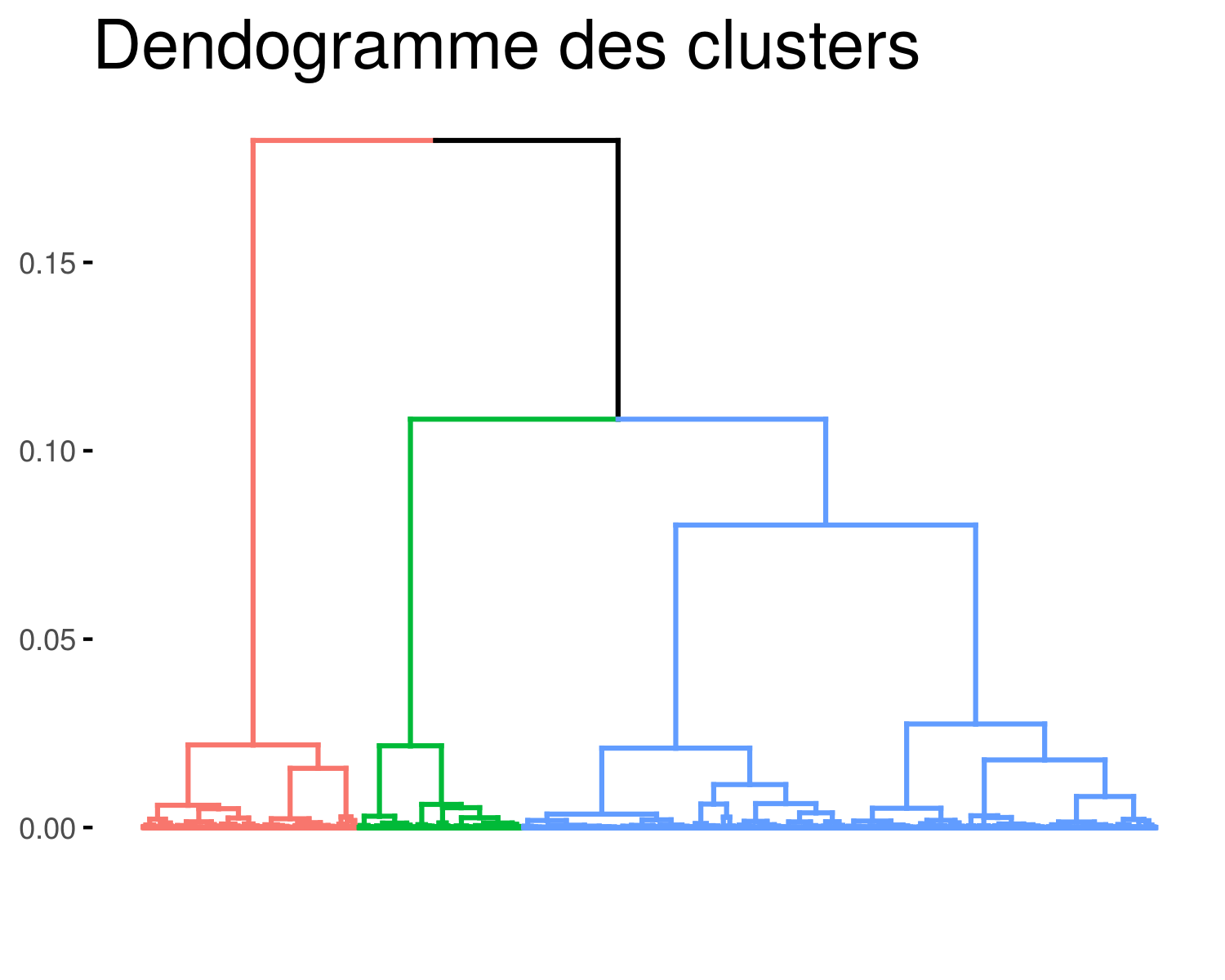
La classification ascendante hiérarchique (CAH) est une technique statistique visant à partitionner une population en différents sous-groupes, appelé aussi classes ou clusters. La CAH cherche à ce que les individus au sein d’une même classe soient les plus proches possibles (homogénéité intra-classe) tandis que les classes soient les plus dissemblables possibles. [[1]](#footnote-46)



Le nombre des clusters est déterminé automatiquement lors de la classification, car c’est ce qui sépare le mieux les bibliothèques. Toutefois on observe qu’il y a quelques chevauchements.



A mesure que le pouvoir explicatif des axes décroit, la classification des individus se chevauche.



Affichons le nombre d’individus dans chaque cluster.

|  |  |
| --- | --- |
| Cluster | Individus |
| 1 | 791 |
| 2 | 231 |
| 3 | 315 |

## Caractéristiques des clusters

## Cluster 1

Le tableau des résultats décrivant les caractéristiques des clusters contient 5 colonnes et autant de lignes que les variables et modalités qui influencent ce cluster. Voici une description de ce que représente chaque colonne :

* *Cla/mod* indique quelle part (pourcentage) de tous les individus présentant cette modalité se retrouve dans cette classe (ce cluster, cette catégorie). Autrement dit, c’est la fréquence du cluster dans la modalité.
* *Mod/cla* indique quelle part (pourcentage) de tous les individus du cluster présentent cette modalité. C’est la fréquence de la modalité dans le cluster. [^3]
* *Global* : indique le nombre total d’occurrence de la modalité.
* *p.value* : indique la significativité de la modalité dans la construction de la classe.
* *v.test* : indique la statistique du test. C’est un moyen alternatif pour lire la significativité de la modalité. Une v.test supérieure en valeur absolue à 1.96 indique la modalité est significative à un seuil d’erreur inférieur à 5%.

Or ce tableau peut être très long car il contient la liste de toutes les modalités qui contribuent à définir le cluster et comme nous avons vu dans le graphique des clusters, il y a des chevauchements donc il serait pertinent de ne relever que les modalités les plus exclusives à ce cluster. Nous déterminons ces modalités à partir de la colonne *Cla/mod* qui indique la part des individus qui possèdent la modalité étudiée se trouvant dans le cluster. La logique est que si un fort pourcentage des individus possèdent la modalité c’est que cette modalité leur est propre.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cla.Mod | Mod.Cla | Global | p.value | v.test |
| audio\_collection=audio\_collection\_no | 100 | 1.01 | 0.60 | 0.01 | 2.4 |
| legal\_basis=SD | 100 | 0.88 | 0.52 | 0.03 | 2.2 |
| ST\_region\_code=st\_4 | 93 | 28.07 | 17.80 | 0.00 | 12.9 |
| legal\_basis=CI | 93 | 70.54 | 44.88 | 0.00 | 24.1 |
| ST\_region\_code=st\_1 | 87 | 13.78 | 9.42 | 0.00 | 7.0 |
| ST\_state\_population=500K-4M | 83 | 54.49 | 39.04 | 0.00 | 14.3 |
| ST\_region\_code=st\_8 | 82 | 7.59 | 5.46 | 0.00 | 4.3 |
| ST\_region\_code=st\_6 | 81 | 4.30 | 3.14 | 0.00 | 3.0 |
| hours\_open=<2340\_h | 78 | 54.74 | 41.51 | 0.00 | 12.0 |
| internet\_computer\_use=internet\_computer\_use\_low | 77 | 24.78 | 19.07 | 0.00 | 6.6 |
| local\_cooperative\_agreements=local\_cooperative\_agreements\_no | 77 | 31.61 | 24.38 | 0.00 | 7.6 |
| print\_collection=<25000 | 76 | 47.16 | 36.50 | 0.00 | 9.9 |
| administrative\_structure=SO | 76 | 98.48 | 76.44 | 0.00 | 24.1 |
| hours\_open=2340-3150h | 76 | 38.69 | 30.22 | 0.00 | 8.3 |
| branch\_libraries=branch\_libraries\_no | 74 | 99.62 | 79.88 | 0.00 | 23.3 |
| internet\_computer\_use=internet\_computer\_use\_medium | 73 | 23.39 | 19.00 | 0.00 | 5.0 |
| young\_adult\_programs=young\_adult\_programs\_no | 73 | 26.17 | 21.32 | 0.00 | 5.3 |
| digital\_collection=7500+ | 72 | 75.60 | 61.93 | 0.00 | 12.4 |
| ST\_region\_code=st\_7 | 72 | 9.86 | 8.08 | 0.00 | 2.9 |
| ST\_region\_code=st\_3 | 69 | 19.47 | 16.75 | 0.00 | 3.2 |

Le cluster 1 est constitué de bibliothèques ayant dans leur catalogue moins de 25000 œuvres imprimées. Ce sont des bibliothèques situées dans des petites comtés (moins de 65000 habitants) et qui ouvrent en moyenne 6 heures et demi par jour (week-end y compris). Ces bibliothèques n’ont pas dans leur catalogue des vidéos ou des fichiers audio. Elles ne sont pas spécifiques à une région en particulier car elles sont étendue dans 4 régions différentes.font partie des Etats de taille moyenne (500.000 à 4 million d’habitants. Elles ont une faible utilisation d’Internet et sont à grande partie des structures administratives à guichet unique (administrative\_structure=SO).

## Cluster 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cla.Mod | Mod.Cla | Global | p.value | v.test |
| legal\_basis=NP | 86 | 83 | 16.7 | 0.00 | 26.6 |
| ST\_state\_population=10M+ | 86 | 90 | 18.2 | 0.00 | 28.3 |
| legal\_basis=OT | 84 | 12 | 2.4 | 0.00 | 8.4 |
| ST\_region\_code=st\_2 | 77 | 90 | 20.0 | 0.00 | 26.6 |
| downloadable\_audio=downloadable\_audio\_medium | 60 | 23 | 6.5 | 0.00 | 9.3 |
| digital\_collection=<7500 | 38 | 84 | 38.1 | 0.00 | 16.0 |
| county\_population=300K+ | 37 | 43 | 20.0 | 0.00 | 8.8 |
| county\_population=65K-300K | 25 | 38 | 25.8 | 0.00 | 4.4 |
| internet\_computer\_use=internet\_computer\_use\_medium | 23 | 26 | 19.0 | 0.01 | 2.7 |
| hours\_open=2340-3150h | 23 | 40 | 30.2 | 0.00 | 3.4 |
| print\_collection=<25000 | 22 | 47 | 36.5 | 0.00 | 3.7 |
| administrative\_structure=SO | 22 | 97 | 76.4 | 0.00 | 9.3 |
| hours\_open=<2340\_h | 22 | 52 | 41.5 | 0.00 | 3.6 |
| branch\_libraries=branch\_libraries\_no | 21 | 97 | 79.9 | 0.00 | 8.2 |
| internet\_computer\_use=internet\_computer\_use\_high | 20 | 39 | 32.8 | 0.05 | 2.0 |

Les bibliothèques qui se trouvent dans le cluster 2 ouvrent en moyenne entre 6.5 heures et 9h par jour (week-end y compris). Elles ont une utilisation très importante d’Internet. Elle sont localisées dans les comtés de plus de 300 mille habitants et dans les grands Etats américains (+ 10 million d’habitants). Le nombre de leur catalogue d’oeuvres imprimées dépasse les 25000. Elles ont dans leurs catalogues des fichiers vidéo et audio. Ce sont des bibliothèques qui font partie d’une fédération des bibliothèques (interlibrary\_relationship = ME). Légalement ces bibliothèques ont le statut d’associations à but non lucratif (legal\_basis = NP)

## Cluster 3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cla.Mod | Mod.Cla | Global | p.value | v.test |
| administrative\_structure=MA | 100 | 4.8 | 1.12 | 0.00 | 6.3 |
| mls\_librarians=+10 | 100 | 3.5 | 0.82 | 0.00 | 5.3 |
| branch\_libraries=branch\_libraries\_yes | 96 | 82.2 | 20.12 | 0.00 | 30.5 |
| administrative\_structure=MO | 94 | 89.2 | 22.44 | 0.00 | 31.9 |
| mls\_librarians=5-10 | 87 | 8.6 | 2.32 | 0.00 | 7.5 |
| hours\_open=>3150\_h | 81 | 97.8 | 28.27 | 0.00 | 31.8 |
| legal\_basis=CO | 71 | 33.6 | 11.22 | 0.00 | 13.1 |
| legal\_basis=MJ | 61 | 11.1 | 4.26 | 0.00 | 6.2 |
| ST\_region\_code=st\_5 | 60 | 49.2 | 19.30 | 0.00 | 14.4 |
| internet\_computer\_use=internet\_computer\_use\_very\_high | 57 | 69.8 | 29.09 | 0.00 | 17.6 |
| interlibrary\_relationship=HQ | 48 | 3.2 | 1.57 | 0.02 | 2.4 |
| legal\_basis=LD | 43 | 34.0 | 18.55 | 0.00 | 7.6 |
| interlibrary\_relationship=NO | 40 | 54.6 | 32.01 | 0.00 | 9.6 |
| print\_collection=25000+ | 36 | 98.1 | 63.50 | 0.00 | 16.7 |
| ST\_state\_population=4M-10M | 36 | 64.8 | 42.78 | 0.00 | 9.0 |

Ces bibliothèques se situent à la fois dans des Etats de taille moyennes (4 à 10 millions d’habitants) et dans des grands Etats (plus de 10 million). Elles sont à la fois des bibliothèques de quartiers. Ce qui est significativement spécifique aux bibliothèques de ce cluster c’est le fait d’avoir des branches externes et que ce sont à majorité des bibliothèques ayant plusieurs guichets. Ces bibliothèques ouvrent plus de 9h par jours (y compris les week-end) et n’ont pas de relation avec les autres bibliothèques. Elles ont un grand nombre de libraires ayant des diplômes supérieures (mls\_librarians >10). Les bibliothèques de ce cluster ne sont pas caractérisées par un seul statut juridique, car elles peuvent être multi-juridictionnelles (legal\_basis = MJ), gérée par le comté (legal\_basis = CO) ou sont encore des bibliothèques de quartier (legal\_basis = LD).

## Individus parangons et individus spécifiques

Le commentaire des clusters ressort les caractéristiques des bibliothèques. Or nous observons qu’il y a des chevauchements entre les classes. En effet, plusieurs des modalités sont partagées par les 3 classes. Nous serons donc intéressés de savoir les caractéristiques de “l’individu moyen” de chaque classe. Autrement dit l’individu parangon de chaque classe.

Un individu parangon est un individu dont les coordonnées sont les plus proches du centre de gravité du groupe. Le profil de cet individu caractérise alors le groupe auquel il appartient. [[2]](#footnote-55)

Nous allons afficher pour chaque classe les individus parangons.

Les individus parangons du cluster 1 :

|  |  |
| --- | --- |
| Distance | Individus |
| 0.12 | HUDSON PUBLIC LIBRARY.2 |
| 0.13 | CLIFTON FORGE PUBLIC LIBRARY |
| 0.17 | CHELSEA PUBLIC LIBRARY |
| 0.17 | WABASHA PUBLIC LIBRARY |
| 0.18 | JASPER PUBLIC LIBRARY |

Les parangons du cluster 2 :

|  |  |
| --- | --- |
| Distance | Individus |
| 0.13 | SAYRE PUBLIC LIBRARY.1 |
| 0.14 | PRIESTLEY FORSYTH MEM LIBRARY |
| 0.15 | GERMAN MASONTOWN PUB LIBRARY |
| 0.17 | SOUTH BUTLER COMMUNITY LIBRARY |
| 0.17 | DR WILLIAM B KONKLE MEM LIB |

Les parangons du cluster 3

|  |  |
| --- | --- |
| Distance | Individus |
| 0.15 | PORTSMOUTH PUBLIC LIBRARY.2 |
| 0.19 | DICKINSON COUNTY LIBRARY |
| 0.23 | TRANSYLVANIA COUNTY LIBRARY |
| 0.25 | OTSEGO COUNTY LIBRARY |
| 0.25 | SUMMIT COUNTY LIBRARY |

Trouvons les caractéristiques des clusters dans le jeu de données.

|  |  |  |  |
| --- | --- | --- | --- |
|  | HUDSON PUBLIC LIBRARY.2 | SAYRE PUBLIC LIBRARY.1 | PORTSMOUTH PUBLIC LIBRARY.2 |
| administrative\_structure | SO | SO | MO |
| print\_collection | 25000+ | 25000+ | 25000+ |
| audio\_collection | yes | yes | yes |
| county\_population | 65K-300K | <=65000 | 65K-300K |
| interlibrary\_relationship | ME | ME | NO |
| downloadable\_audio | high | high | high |
| physical\_video | yes | yes | yes |
| hours\_open | <2340\_h | <2340\_h | >3150\_h |
| print\_subscriptions | 48 | 30 | 332 |
| registered\_users | 3667 | 8199 | 63417 |
| childrens\_programs | 28 | 120 | 461 |
| young\_adult\_programs | no | yes | yes |
| internet\_computer\_use | high | medium | very\_high |
| ST\_state\_population | 4M-10M | 10M+ | 4M-10M |
| ST\_region\_code | st\_3 | st\_2 | st\_5 |
| legal\_basis | CI | NP | CI |
| wireless\_internet\_sessions | no | yes | yes |
| branch\_libraries | no | no | yes |
| local\_government\_operating\_revenue | 73083 | 30000 | 2289159 |
| local\_cooperative\_agreements | no | yes | yes |
| digital\_collection | 7500+ | <7500 | <7500 |
| mls\_librarians | <5 | <5 | <5 |

# Conclusion

Cette analyse sur les bibliothèques nous a permis de ressortir que les caractéristiques non directement observables par une analyse statistique descriptive simple peuvent ressortir si nous appliquons une analyse factorielle qui prend simultanément en compte plusieurs attributs des bibliothèques. Les axes factoriels que nous trouvés montrent une

Il ressort tout de même une faiblesse dans cette analyse à cause de l’extrême distrbution des colonnes du départ. Cette distribution des colonnes et le fait que toutes les colonnes sont corrélées positivement a conduit a un effet de taille où le premier axe résume plus de 50% de la variance totale. L’analyse des correspondances multiples que nous avons réalisée ensuite montre aussi des faiblesse car j’ai réalisé après avoir refait plusieurs fois la même analyse sur des échantillons différents et sur des segmentations différentes qu’il existe une très grande différence dans les résultats. Cela montre que la variance des colonnes n’est pas stable et qu’il serait mieux de prendre l’exhaustivité des données dans l’analyse.

# Bibliographie

Alboukadel Kassambara and Fabian Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. <http://www.sthda.com/english/rpkgs/factoextra>

*Datafile Documentation* <https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey>

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2019). rmarkdown: Dynamic Documents for R. R package version 1.16. URL <https://rmarkdown.rstudio.com>.

Yihui Xie and J.J. Allaire and Garrett Grolemund (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC. ISBN 9781138359338. URL <https://bookdown.org/yihui/rmarkdown>.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

# Annexes

*Annexe 1 : Coordonnées des modalités des colonnes*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 5 |
| MA | 1.82 | -0.50 | 0.62 | 0.33 | -0.09 |
| MO | 1.50 | -0.36 | 0.52 | 0.04 | 0.09 |
| SO | -0.47 | 0.11 | -0.16 | -0.02 | -0.02 |
| <25000 | -0.91 | -0.27 | 0.43 | 0.19 | 0.23 |
| 25000+ | 0.53 | 0.16 | -0.25 | -0.11 | -0.13 |
| audio\_collection\_no | -1.47 | -1.75 | 1.30 | 0.65 | -0.68 |
| audio\_collection\_yes | 0.01 | 0.01 | -0.01 | 0.00 | 0.00 |
| <=65000 | -0.08 | -0.51 | 0.05 | -0.37 | 0.04 |
| 65K-300K | 0.11 | 0.29 | 0.29 | 0.11 | 0.17 |
| 300K+ | 0.06 | 1.01 | -0.50 | 0.86 | -0.32 |
| HQ | 0.68 | -0.28 | 0.94 | -0.74 | -2.47 |
| ME | -0.20 | 0.14 | -0.27 | 0.18 | 0.22 |
| NO | 0.39 | -0.27 | 0.52 | -0.34 | -0.33 |
| downloadable\_audio\_small | -0.07 | -0.17 | 0.66 | -1.89 | 0.57 |
| downloadable\_audio\_medium | -0.06 | 1.06 | 1.22 | -1.18 | 0.82 |
| downloadable\_audio\_high | 0.01 | -0.07 | -0.11 | 0.14 | -0.08 |
| physical\_video\_no | -0.41 | 0.13 | 1.39 | 1.24 | 1.26 |
| physical\_video\_yes | 0.00 | 0.00 | -0.01 | -0.01 | -0.01 |
| <2340\_h | -0.83 | -0.22 | 0.38 | 0.20 | 0.23 |
| 2340-3150h | -0.12 | 0.52 | -0.83 | -0.44 | -0.40 |
| >3150\_h | 1.35 | -0.23 | 0.33 | 0.17 | 0.09 |
| young\_adult\_programs\_no | -0.62 | -0.53 | 0.50 | 0.25 | 0.14 |
| young\_adult\_programs\_yes | 0.17 | 0.14 | -0.13 | -0.07 | -0.04 |
| internet\_computer\_use\_low | -0.99 | -0.50 | 0.76 | 0.40 | 0.17 |
| internet\_computer\_use\_medium | -0.69 | -0.02 | 0.17 | 0.10 | 0.36 |
| internet\_computer\_use\_high | 0.06 | 0.28 | -0.49 | -0.74 | -0.31 |
| internet\_computer\_use\_very\_high | 1.04 | 0.03 | -0.06 | 0.50 | 0.00 |
| 500K-4M | -0.39 | -0.68 | 0.09 | -0.07 | -0.58 |
| 4M-10M | 0.45 | -0.07 | -0.50 | 0.01 | 0.65 |
| 10M+ | -0.21 | 1.61 | 0.97 | 0.12 | -0.26 |
| CC | 0.54 | -0.61 | 0.57 | -0.60 | -2.23 |
| CI | -0.47 | -0.38 | -0.39 | 0.29 | -0.22 |
| CO | 1.07 | -0.74 | 0.82 | -0.36 | -0.35 |
| LD | 0.61 | -0.09 | -0.66 | -0.54 | 1.26 |
| MJ | 0.98 | -0.57 | 0.78 | -0.03 | -1.02 |
| NL | -0.94 | -0.10 | 2.11 | -6.13 | 1.83 |
| NP | -0.38 | 1.61 | 0.79 | 0.02 | -0.27 |
| OT | -0.14 | 1.56 | 1.02 | 1.12 | 0.10 |
| SD | -0.96 | -0.60 | 0.47 | 0.12 | 2.89 |
| wireless\_internet\_sessions\_no | -0.17 | 0.02 | 0.47 | 0.07 | -0.43 |
| wireless\_internet\_sessions\_yes | 0.02 | 0.00 | -0.06 | -0.01 | 0.06 |
| branch\_libraries\_no | -0.41 | 0.08 | -0.17 | -0.03 | -0.01 |
| branch\_libraries\_yes | 1.61 | -0.34 | 0.68 | 0.13 | 0.02 |
| local\_cooperative\_agreements\_no | -0.56 | -0.53 | 0.47 | -0.34 | 0.20 |
| local\_cooperative\_agreements\_yes | 0.18 | 0.17 | -0.15 | 0.11 | -0.07 |
| <7500 | 0.03 | 0.68 | 0.54 | -0.32 | 0.24 |
| 7500+ | -0.02 | -0.42 | -0.33 | 0.20 | -0.15 |
| <5 | -0.06 | -0.01 | -0.01 | -0.08 | -0.02 |
| 5-10 | 1.80 | 0.39 | 0.27 | 2.23 | 0.58 |
| +10 | 2.17 | -0.28 | 0.10 | 3.07 | 0.44 |

*Annexe 2 : Les v-tests des modalités*

Certaines des modalités ne sont significatives sur certains axes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Dim 1 | Dim 2 | Dim 3 | Dim 4 | Dim 5 |
| MA | 7.10 | -1.94 | 2.43 | 1.29 | -0.35 |
| MO | 29.46 | -7.15 | 10.21 | 0.72 | 1.75 |
| SO | -30.73 | 7.51 | -10.64 | -1.03 | -1.63 |
| <25000 | -25.31 | -7.62 | 11.88 | 5.17 | 6.43 |
| 25000+ | 25.31 | 7.62 | -11.88 | -5.17 | -6.43 |
| audio\_collection\_no | -4.16 | -4.96 | 3.69 | 1.84 | -1.94 |
| audio\_collection\_yes | 4.16 | 4.96 | -3.69 | -1.84 | 1.94 |
| <=65000 | -3.04 | -20.34 | 1.95 | -14.76 | 1.39 |
| 65K-300K | 2.45 | 6.24 | 6.23 | 2.48 | 3.75 |
| 300K+ | 1.11 | 18.50 | -9.24 | 15.66 | -5.83 |
| HQ | 3.12 | -1.30 | 4.33 | -3.41 | -11.39 |
| ME | -10.36 | 7.16 | -13.91 | 9.36 | 11.26 |
| NO | 9.66 | -6.90 | 12.93 | -8.56 | -8.36 |
| downloadable\_audio\_small | -0.42 | -1.04 | 4.12 | -11.82 | 3.54 |
| downloadable\_audio\_medium | -0.62 | 10.24 | 11.74 | -11.37 | 7.87 |
| downloadable\_audio\_high | 0.76 | -8.08 | -12.30 | 16.38 | -8.69 |
| physical\_video\_no | -1.47 | 0.46 | 5.04 | 4.49 | 4.57 |
| physical\_video\_yes | 1.47 | -0.46 | -5.04 | -4.49 | -4.57 |
| <2340\_h | -25.59 | -6.91 | 11.62 | 6.26 | 7.19 |
| 2340-3150h | -2.85 | 12.63 | -19.98 | -10.64 | -9.65 |
| >3150\_h | 30.91 | -5.31 | 7.67 | 4.00 | 1.97 |
| young\_adult\_programs\_no | -11.76 | -10.10 | 9.46 | 4.72 | 2.73 |
| young\_adult\_programs\_yes | 11.76 | 10.10 | -9.46 | -4.72 | -2.73 |
| internet\_computer\_use\_low | -17.61 | -8.80 | 13.43 | 7.11 | 3.05 |
| internet\_computer\_use\_medium | -12.23 | -0.31 | 3.05 | 1.81 | 6.36 |
| internet\_computer\_use\_high | 1.48 | 7.03 | -12.53 | -18.79 | -7.88 |
| internet\_computer\_use\_very\_high | 24.26 | 0.61 | -1.30 | 11.72 | 0.02 |
| 500K-4M | -11.52 | -19.79 | 2.75 | -2.07 | -17.09 |
| 4M-10M | 14.13 | -2.16 | -15.70 | 0.45 | 20.39 |
| 10M+ | -3.55 | 27.80 | 16.67 | 2.04 | -4.53 |
| CC | 2.29 | -2.62 | 2.42 | -2.57 | -9.53 |
| CI | -15.35 | -12.44 | -12.70 | 9.43 | -7.13 |
| CO | 13.95 | -9.59 | 10.62 | -4.66 | -4.56 |
| LD | 10.60 | -1.63 | -11.46 | -9.47 | 22.00 |
| MJ | 7.53 | -4.36 | 6.04 | -0.20 | -7.88 |
| NL | -1.33 | -0.14 | 2.98 | -8.68 | 2.59 |
| NP | -6.23 | 26.26 | 12.92 | 0.37 | -4.35 |
| OT | -0.78 | 8.91 | 5.86 | 6.38 | 0.60 |
| SD | -2.55 | -1.59 | 1.25 | 0.31 | 7.66 |
| wireless\_internet\_sessions\_no | -2.22 | 0.31 | 6.32 | 0.89 | -5.81 |
| wireless\_internet\_sessions\_yes | 2.22 | -0.31 | -6.32 | -0.89 | 5.81 |
| branch\_libraries\_no | -29.62 | 6.15 | -12.49 | -2.38 | -0.42 |
| branch\_libraries\_yes | 29.62 | -6.15 | 12.49 | 2.38 | 0.42 |
| local\_cooperative\_agreements\_no | -11.71 | -11.10 | 9.69 | -7.01 | 4.25 |
| local\_cooperative\_agreements\_yes | 11.71 | 11.10 | -9.69 | 7.01 | -4.25 |
| <7500 | 0.92 | 19.38 | 15.55 | -9.24 | 7.00 |
| 7500+ | -0.92 | -19.38 | -15.55 | 9.24 | -7.00 |
| <5 | -12.47 | -1.43 | -1.49 | -16.13 | -3.55 |
| 5-10 | 10.11 | 2.22 | 1.53 | 12.56 | 3.24 |
| +10 | 7.22 | -0.95 | 0.32 | 10.21 | 1.46 |

*Annexe 3 : Statistiques descriptives des colonnes discrétisées.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Modalities | MostFrequent | NbFrequent | Missing |
| administrative\_structure | 3 | SO | 1022 | 0 |
| print\_collection | 2 | 25000+ | 849 | 0 |
| audio\_collection | 2 | yes | 1329 | 0 |
| county\_population | 3 | <=65000 | 724 | 0 |
| interlibrary\_relationship | 3 | ME | 888 | 0 |
| downloadable\_audio | 3 | high | 1212 | 0 |
| physical\_video | 2 | yes | 1324 | 0 |
| hours\_open | 3 | <2340\_h | 555 | 0 |
| young\_adult\_programs | 2 | yes | 1052 | 0 |
| internet\_computer\_use | 4 | high | 439 | 0 |
| ST\_state\_population | 3 | 4M-10M | 572 | 0 |
| ST\_region\_code | 8 | st\_2 | 268 | 0 |
| legal\_basis | 9 | CI | 600 | 0 |
| wireless\_internet\_sessions | 2 | yes | 1179 | 0 |
| branch\_libraries | 2 | no | 1068 | 0 |
| local\_cooperative\_agreements | 2 | yes | 1011 | 0 |
| digital\_collection | 2 | 7500+ | 828 | 0 |
| mls\_librarians | 3 | <5 | 1295 | 0 |

1. Classification ascendante hiérarchique (CAH), analyse-R [↑](#footnote-ref-46)
2. Eléments de classification, François Chesneau [↑](#footnote-ref-55)