



Une chaire + R + Shiny : Vecteur d'innovation

Problématique

Savoir...

... Pour changer

Perspectives



typoChooseR

Comment utiliser Shiny pour la conduite du changement :
Un exemple avec des méthodologies de partitionnement

chaire.d-cc@univ-ubs.fr

<http://decisionnelclient.org>

<https://chairedecisionnelconnaissanceclient.shinyapps.io/typoChooseR/>





De quoi va-t-on parler ?

Problématique

Savoir...

... Pour changer

Perspectives

- ✓ Un problème méthodologique : partitionner à partir de variables qualitatives
- ✓ Une méthodologie d'aide à la décision
- ✓ Vraiment faire changer les pratiques : Shiny



Problématique

Contexte

Typologies

Le problème

Savoir...

... Pour changer

Perspectives

Problématique



Data-Miners en entreprise

Problématique

Contexte

Typologies

Le problème

Savoir...

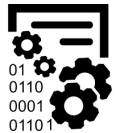
... Pour changer

Perspectives

Data-scientists = centre de service

- ✓ Charge Opérationnelle
- ✓ Pression des services demandeurs
- ✓ Logiciels propriétaires peu souples

⇒ Innover...





Data-Miners en entreprise

Problématique

Contexte

Typologies

Le problème

Savoir...

... Pour changer

Perspectives

Data-scientists = centre de service

- ✓ Charge Opérationnelle
- ✓ Pression des services demandeurs
- ✓ Logiciels propriétaires peu souples

⇒ Innover... mais sous contrainte





Un exemple de méthodologie perfectible

Problématique

Contexte

Typologies

Le problème

Savoir...

... Pour changer

Perspectives

- ✓ Population à segmenter : 10^6 clients
- ✓ Variables qualitatives
- ✓ Variables quantitatives

L'existant :

- ✓ discrétisation des variables quantitatives
- ✓ analyse factorielle des correspondances multiples
- ✓ calcul des distances à partir *d'un certain nombre* d'axes

Puis

Méthodes géométriques de partitionnement :

- ✓ K-means
- ✓ CAH
- ✓ mix des deux





Quelque chose qui cloche...

Problématique

Contexte

Typologies

Le problème

Savoir...

... Pour changer

Perspectives

- ✓ 20 variables qualitatives
- ✓ 2 à 4 modalités par variables
- ✓ Choisies pour ne pas être trop liées (V de Cramer)
- ✓ Méthodologie de typologie :
 - ✗ Deux axes de l'ACM retenus pour calcul des distances
 - ✗ K-means, avec K grand
 - ✗ CAH sur les K centres : choisir le nombre final de groupes



Quelque chose qui cloche...

Problématique

Contexte

Typologies

Le problème

Savoir...

... Pour changer

Perspectives

- ✓ 20 variables qualitatives
- ✓ 2 à 4 modalités par variables
- ✓ Choisies pour ne pas être trop liées (V de Cramer)
- ✓ Méthodologie de typologie :
 - ✗ Deux axes de l'ACM retenus pour calcul des distances
 - ✗ K-means, avec K grand
 - ✗ CAH sur les K centres : choisir le nombre final de groupes

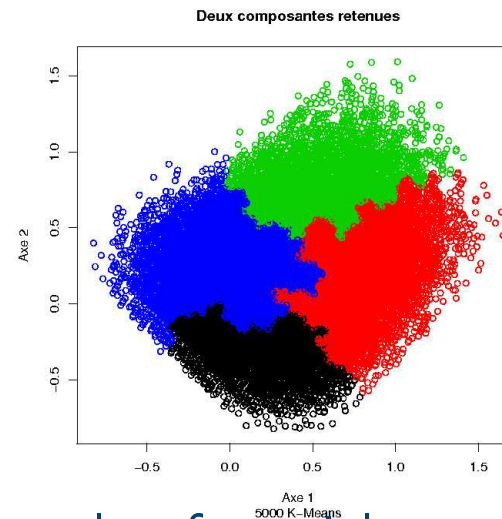


Figure 1: Le premier plan factoriel représente 96% de l'inertie



Diagnostic

[Problématique](#)

[Contexte](#)

[Typologies](#)

[Le problème](#)

[Savoir...](#)

[... Pour changer](#)

[Perspectives](#)



Un espace de dimension proche 40 rendu à 96% avec 2 axes de l'analyse des correspondances multiples ?



Diagnostic

[Problématique](#)

[Contexte](#)

[Typologies](#)

[Le problème](#)

[Savoir...](#)

[... Pour changer](#)

[Perspectives](#)



Un espace de dimension proche 40 rendu à 96% avec 2 axes de l'analyse des correspondances multiples ?

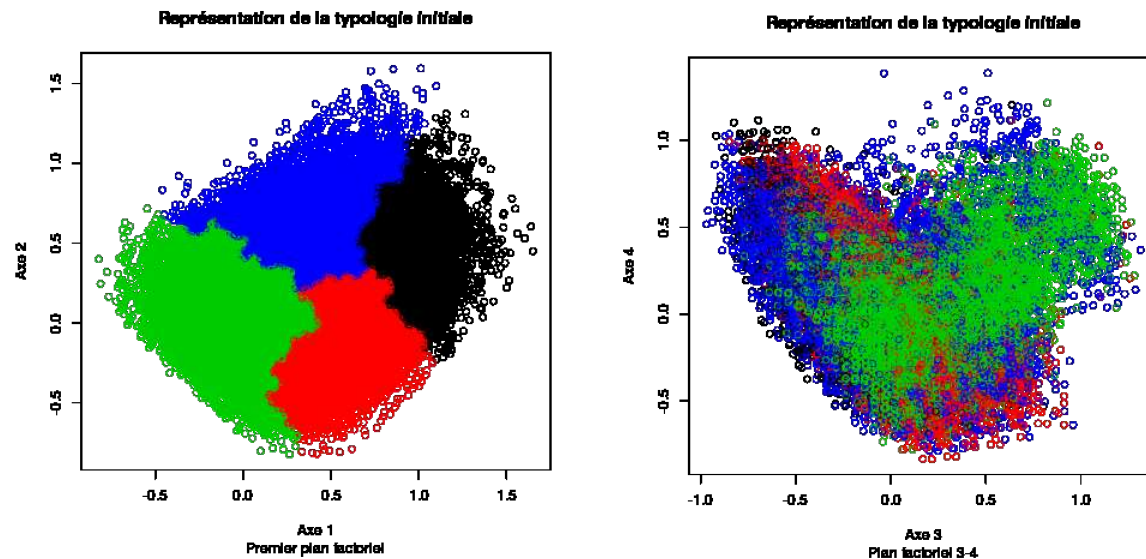


Figure 2: Les valeurs propres des deux premiers axes représentent 96% du total des valeurs propres... Après transformation !!



Problématique

Savoir...

ACM et distances
Les deux fonctions
de l'ACM
D'autres méthodes
de partitionnement
Livrables

... Pour changer

Perspectives

Savoir n'est pas changer



Tout est dans la distance

[Problématique](#)

[Savoir...](#)

ACM et distances

Les deux fonctions
de l'ACM

D'autres méthodes
de partitionnement

Livrables

[... Pour changer](#)

[Perspectives](#)

La distance du Chi2 permet de calculer des distances entre individus à partir de variables qualitatives

$$d_{\chi^2}^2(i, i') = \frac{n}{Q} \sum_{q=1}^Q \sum_{l=1}^{m_q} \delta_{ii'}^{ql} \frac{1}{n_l^q}$$

Mais peu utilisée par les data-miners en entreprise. La pratique :

- ✓ Faire une Analyse des correspondances multiples...
- ✓ Pour récupérer *un certains nombre* d'axes quantitatifs...
- ✓ Calculer une distance euclidienne classique.

Si on garde tous les axes : distance du Chi2.

Mais...



Les deux fonctions de l'ACM

[Problématique](#)

[Savoir...](#)

[ACM et distances](#)

**Les deux fonctions
de l'ACM**

[D'autres méthodes
de partitionnement](#)

[Livrables](#)

[... Pour changer](#)

[Perspectives](#)

Les praticiens ne conservent pas tous les axes.

Il y a souvent confusion entre :

- ✓ Fonction de représentation :
 - ✗ On ne voit qu'en 2D → "meilleurs plans"
- ✓ Fonction de réduction de la dimensionnalité
 - ✗ "Filtrer" une information non pertinente → "conjurier" l'influence de variables trop liées, éviter du bruit

Certaines transformations de valeurs propres utilisées font des hypothèses implicites très fortes, et amènent une surestimation grave de "l'information" expliquée par les premiers axes.



D'autres méthodes de partitionnement

[Problématique](#)

[Savoir...](#)

ACM et distances
Les deux fonctions
de l'ACM

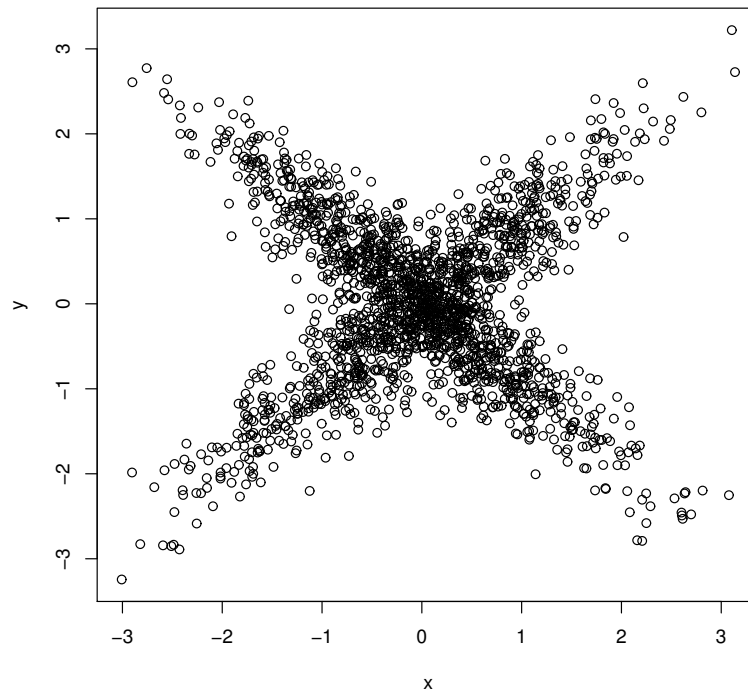
**D'autres méthodes
de partitionnement**

[Livrables](#)

[... Pour changer](#)

[Perspectives](#)

D'autres modèles possibles sont aussi à envisager, par exemple les modèles de mélange.





D'autres méthodes de partitionnement

[Problématique](#)

[Savoir...](#)

ACM et distances
Les deux fonctions
de l'ACM

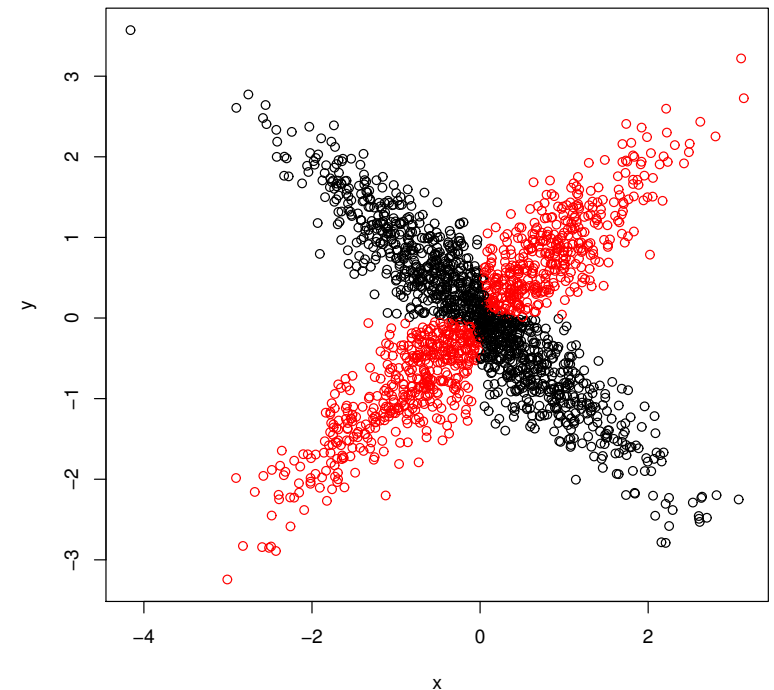
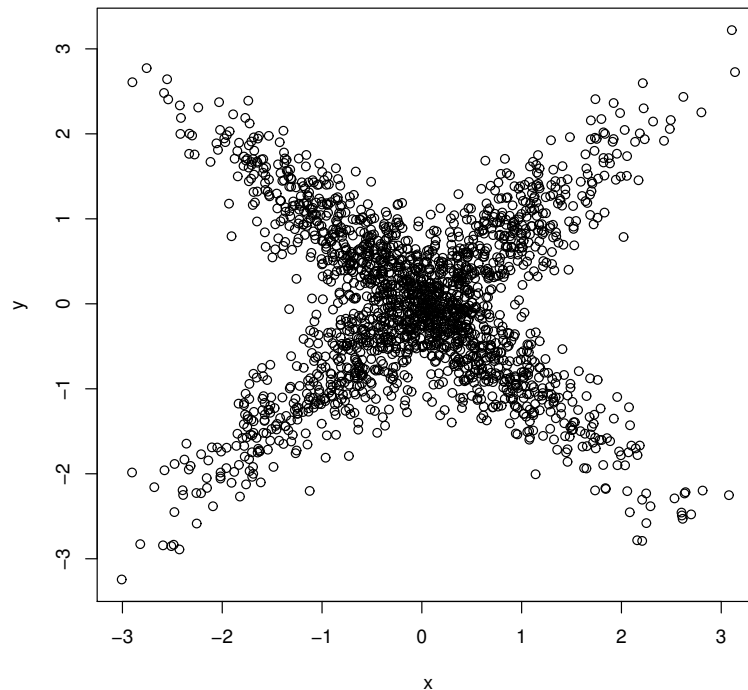
**D'autres méthodes
de partitionnement**

[Livrables](#)

[... Pour changer](#)

[Perspectives](#)

D'autres modèles possibles sont aussi à envisager, par exemple les modèles de mélange.





Livrables

Problématique

Savoir...

ACM et distances
Les deux fonctions
de l'ACM
D'autres méthodes
de partitionnement

Livrables

... Pour changer

Perspectives

- | | |
|-----------------------------|--|
| ✓ Rapport de Recherche | ✓ "Mmm, oui, intéressant" |
| ✓ Guide de Bonnes Pratiques | ✓ "Ah, c'est comme ça qu'on va faire" |
| ✓ Démonstrateur | ✓ " Mais est-ce qu'on pourrait aussi faire varier ceci et voir cela" |



Problématique

Savoir...

... Pour changer

Import

Options

Variables

algorithmes

Perspectives

Jouer avec le réel pour Changer



Premier onglet : import de fichier

[Problématique](#)

[Savoir...](#)

[... Pour changer](#)

Import

[Options](#)

[Variables](#)

[algorithmes](#)

[Perspectives](#)

Shiny application window: /data/A1_CreditAgricole/TypoACMPropre/typoChooseR - Shiny

URL: http://127.0.0.1:5495 | Open in Browser | Publish

Navigation tabs: typoChooseR | Choix des variables actives | Choix de la méthode de classification | Exploration de la Typo Choisie

InputType tooltip: Seuls les fichiers csv avec un séparateur de champ point-virgule sont pour le moment pris en charge

Text: Veuillez choisir un fichier pour votre jeu de données qualitatives ?

Buttons: Browse... | No file chosen

Slider: Pourcentage échantillon pour représentation de l'ACM (0.1 to 100, current value 12.1)

Text: Veuillez charger un fichier



Premier onglet : panneaux conditionnels

[Problématique](#)

[Savoir...](#)

[... Pour changer](#)


[Import](#)

[Options](#)

[Variables](#)

[algorithmes](#)

[Perspectives](#)

typoChooseR

Choix des variables actives

Choix de la méthode de classification

Exploration de la Typo Choisie

Veuillez choisir un fichier pour votre jeu de données qualitatives ?

Browse...

dataCf

Upload complete

Pourcentage échantillon pour représentation de l'ACM

0.1

25

100

0.1

20.1

40.1

60.1

80.1

100

Variables à prendre en compte pour le partitionnement

☒ nb_op

☒ DET_prev

☒ det_auto

☒ det_mrh

☒ det_sante

☒ det_pj

☒ det_cconso

Ce jeu de données contient 168816 données manquantes, veuillez choisir la façon de les prendre en compte

Choisir la méthode de prise en compte des NA

<- Panneau Conditionnel, n'apparaît que si NA

Veuillez choisir le mode de gestion des données manquantes

Choisir le mode de prise en compte des données manquantes.



Premier onglet : explorer les variables



typoChooseR

Choix des variables actives

Choix de la méthode de classification

Exploration de la Typo Choisie

[Problématique](#)

[Savoir...](#)

[... Pour changer](#)

[Import](#)

[Options](#)

[Variables](#)

[algorithmes](#)

[Perspectives](#)

Veillez choisir un fichier pour votre jeu de données qualitatives ?

Browse...

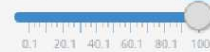
dataC4

Upload complete

Pourcentage échantillon pour représentation de l'ACM

0.1

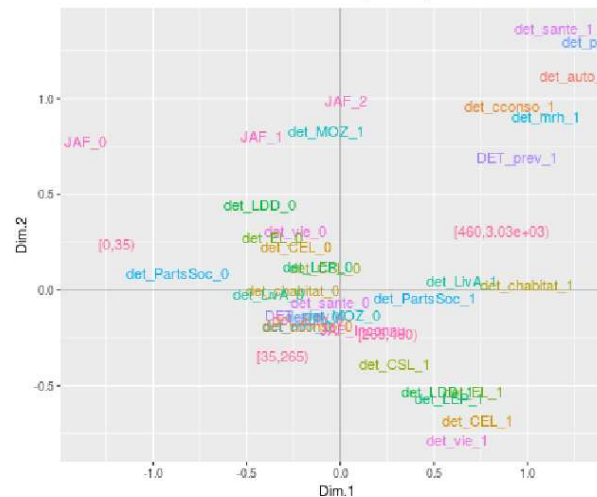
100



Variables à prendre en compte pour le partitionnement

- ☒ nb_op
- ☒ DET_prev
- ☒ det_auto
- ☒ det_mrh
- ☒ det_sante
- ☒ det_pj
- ☒ det_cconso
- ☒ det_chabitat
- ☒ det_vie
- ☒ det_EL
- ☒ det_CEL
- ☒ det_CSL
- ☒ det_LivA
- ☒ det_LDD
- ☒ det_LEP
- ☒ det_MOZ
- ☒ det_PartsSoc
- ☒ JAF

Modalités des variables retenues sur le premier plan factoriel

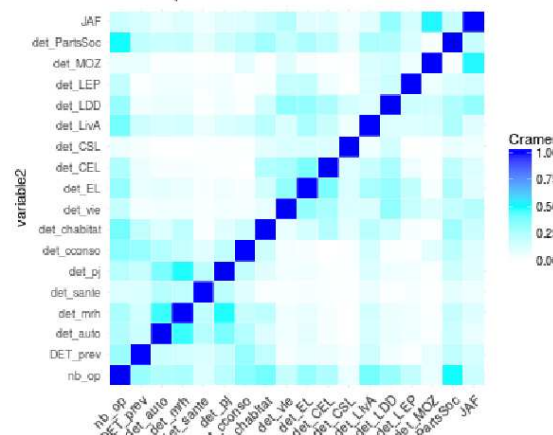


Variable

- det_auto
- det_cconso
- det_CEL
- det_chabitat
- det_CSL
- det_EL
- det_LDD
- det_LEP
- det_LivA
- det_MOZ
- det_mrh
- det_PartsSoc
- det_pj
- DET_prev
- det_sante
- det_vie
- JAF
- nb_op

Représentation :
Premier plan ACM

Liens entre les variables retenues mesuré par le V de Cramer



Estimation du lien
entre variables : V de
Cramer



Deuxième onglet : explorer les méthodes



typoChooseR

Choix des variables actives

Choix de la méthode de classification

Exploration de la Typo Choisie

TypoChooseR

Pourcentage
échantillon pour
exploration



Méthode de
Partitionnement :

K-means seulement

Kmeans+CAH

Modèle de
mélange

K-means seulement

Explorer nombre
de groupes

Pour un nombre de
composantes donné

☒ Explorer

Nombre de
composantes à retenir

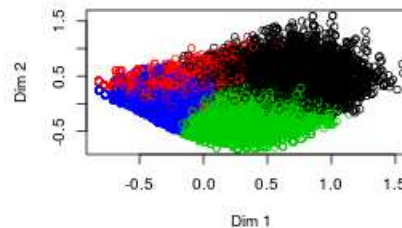


Plan Factoriel
supplémentaire à
représenter

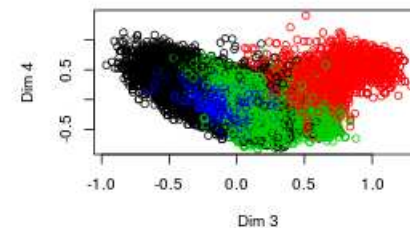


On échantillonne 44127 lignes, soit 21.1 pourcent
du jeu de données qui a 18 variables

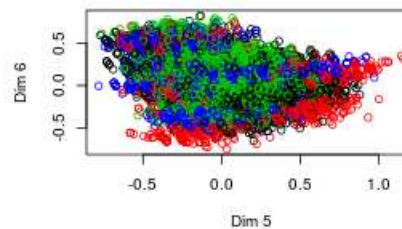
Classes sur le premier plan fact.



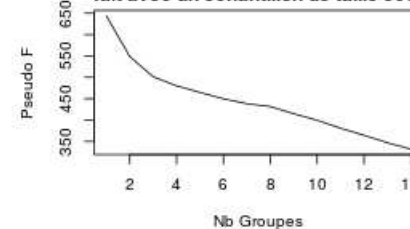
Classes sur le deuxième plan fact.



Classes sur le plan fact. 3



Pas de critère scalable
pour le moment
fait avec un échantillon de taille 5000





Deuxième Onglet : explorer les méthodes



typoChooseR

Choix des variables actives

Choix de la méthode de classification

Exploration de la Typo Choisie

TypoChooseR

Pourcentage
échantillon pour
exploration



Méthode de
Partitionnement :

K-means seulement

Kmeans+CAH

Modèle de
mélange

K-means seulement

Explorer nombre
de groupes

Pour un nombre de
composantes donné

☒ Explorer

Nombre de
composantes à retenir

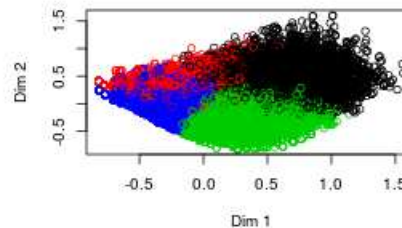


Plan Factoriel
supplémentaire à
représenter

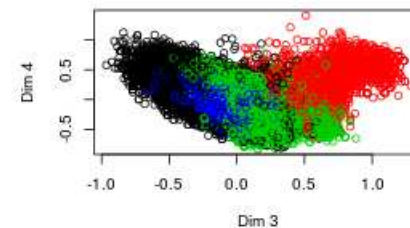


On échantillonne 44127 lignes, soit 21.1 pourcent
du jeu de données qui a 18 variables

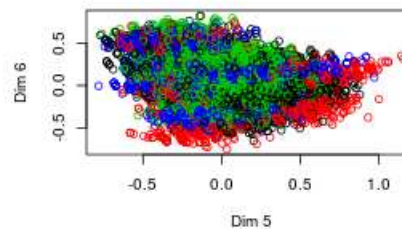
Classes sur le premier plan fact.



Classes sur le deuxième plan fact.



Classes sur le plan fact. 3



Pas de critère scalable
pour le moment
fait avec un échantillon de taille 5000



Mais comment choisir le nombre de
composantes qu'on retient ?

Pourquoi ne pas représenter leur similarité
(mesurée par l'indice de Rand) et voir les zones
de stabilité ?



Deuxième Onglet : explorer les méthodes

typoChooseR

Choix des variables actives

Choix de la méthode de classification

Exploration de la Typo Choisie

Problématique

Savoir...

... Pour changer

Import

Options

Variables

algorithmes

Perspectives

TypoChooseR

Pourcentage échantillon pour exploration
0.1 21.1 100
0.1 10.1 20.1 30.1 40.1 50.1 60.1 70.1 80.1 90.1 100

Méthode de Partitionnement :
K-means seulement

Nombre de groupes
2 4 10
2 3 4 5 6 7 8 9 10

Explorer nombre de groupes
Pour un nombre de composantes donné
☒ Explorer

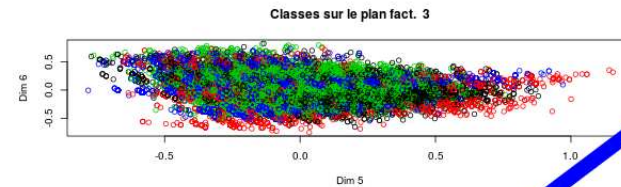
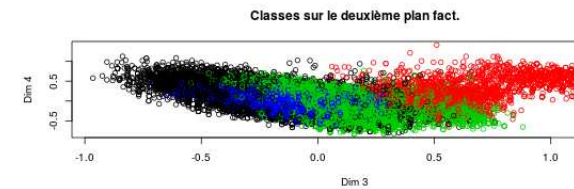
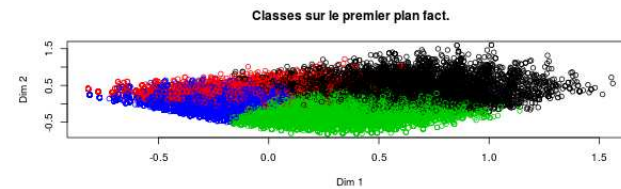
Nombre de composantes à retenir
2 4 22
2 4 6 8 10 12 14 16 18 20 22

Plan Factoriel supplémentaire à représenter
3 11
3 4 5 6 7 8 9 10 11

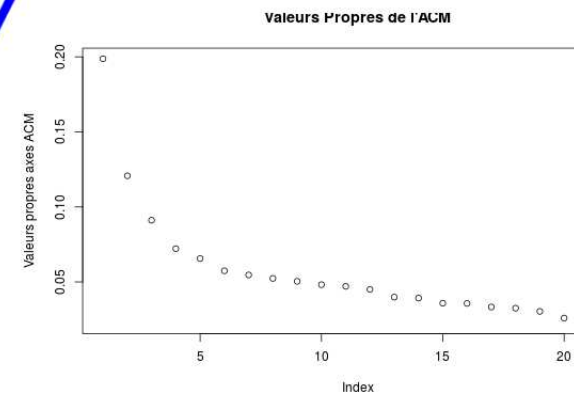
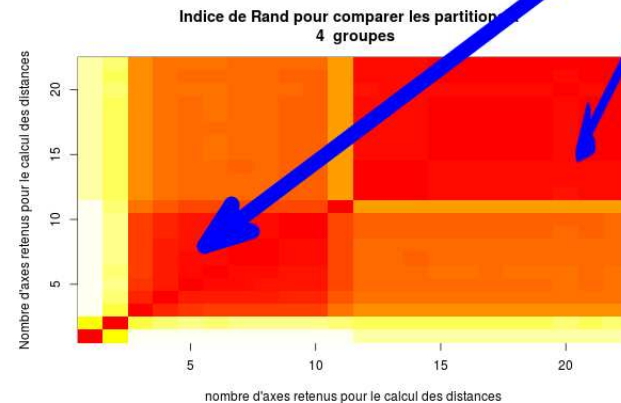
Explorer le nombre de composantes
Comparer les partitions

Attention ! Comparaison = temps de calcul qui augmente avec la taille de l'échantillon

On échantillonne 44127 lignes, soit 21.1 pourcent du jeu de données qui a 18 variables



Pas de critère scalable
Deux Zones de stabilité :
Choisir une partition dans chaque zone et l'explorer plus en détail





Troisième onglet : explorer les groupes

[Problématique](#)

[Savoir...](#)

[... Pour changer](#)

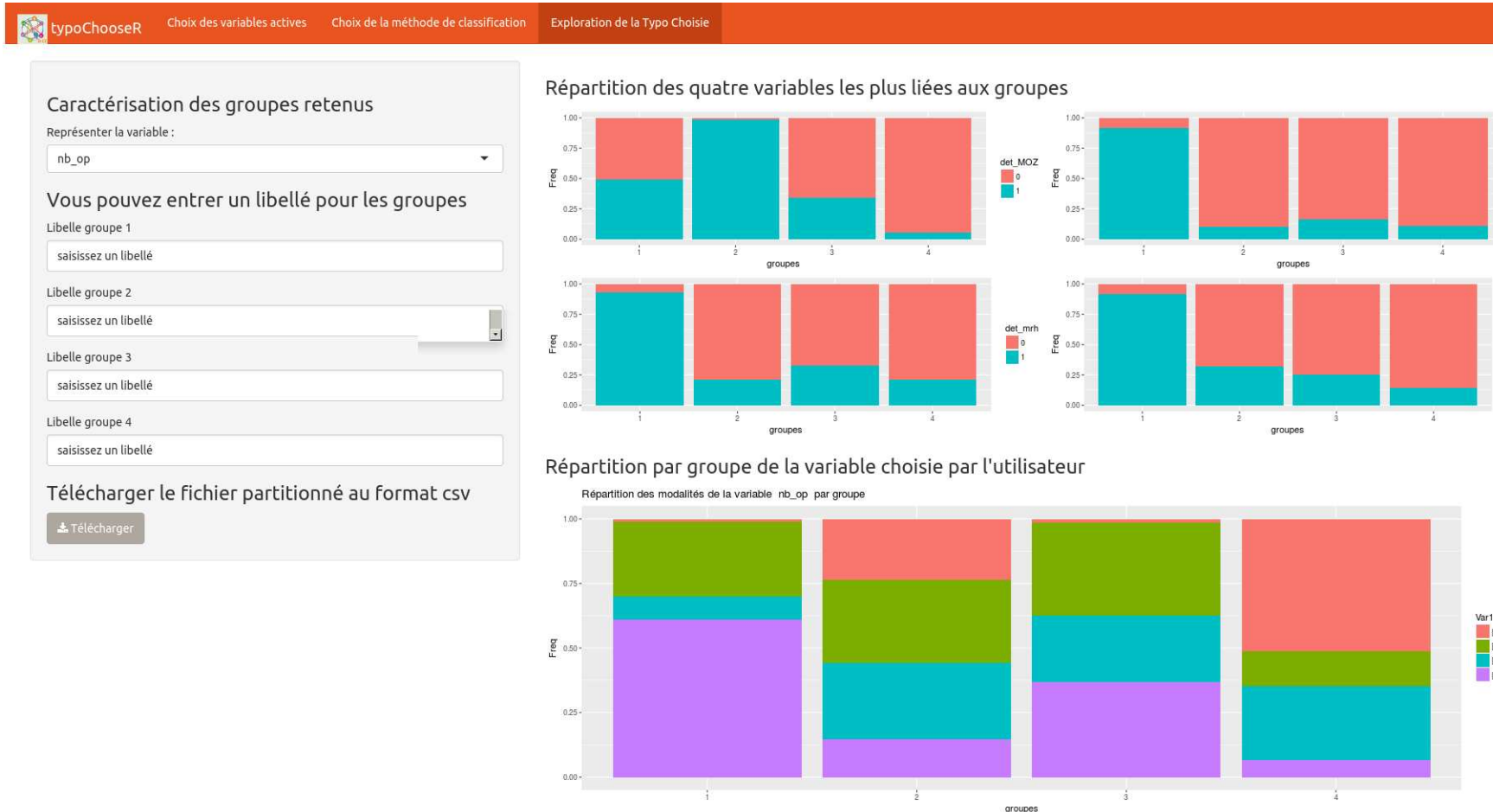
[Import](#)

[Options](#)

[Variables](#)

[algorithmes](#)

[Perspectives](#)



Au final : nommer les groupes s'ils font sens, ou itérer jusqu'à ce qu'ils fassent sens.



Problématique

Savoir...

... Pour changer

Perspectives

Du problème au
changement

Production

Remerciements

Conclusions et Perspectives



Récapitulatif : du problème au changement de pratiques

[Problématique](#)

[Savoir...](#)

[... Pour changer](#)

[Perspectives](#)

[Du problème au
changement](#)

[Production](#)

[Remerciements](#)

- ✓ Il est difficile de remettre en cause une pratique en place
- ✓ Les arguments théoriques ne suffisent pas
- ✓ Une chaire partenariale permet une intervention extérieure différente de la prestation
- ✓ R/Shiny permet le développement agile d'interfaces à même de convaincre.





Mise en production : deux options

[Problématique](#)

[Savoir...](#)

[... Pour changer](#)

[Perspectives](#)

[Du problème au
changement](#)

[Production](#)

[Remerciements](#)

- ✓ Optimisation du code R et déploiement : Shiny comme outil de production.
 - ✗ Pontage aux bases de données SI entreprise
 - ✗ Parallélisation des algorithmes qui peuvent l'être
 - ✗ Tests/Recette, généralisation des formats de données...
- ✓ Identification de l'algorithme souhaité et utilisation des outils "maison : Shiny comme outil de démonstration"

Dans tous les cas : conduite du changement en mode agile



Remerciements

[Problématique](#)

[Savoir...](#)

[... Pour changer](#)

[Perspectives](#)

[Du problème au
changement](#)

[Production](#)

[Remerciements](#)



Packages utilisés :

- ✓ clusterCrit, **FactoMineR**, fields, **ggplot2**, gridExtra, mclust, textbfMixAll, vcd
- ✓ shiny, shinyBS, shinyccsloaders, shinythemes

