# Projet mmrClustVar

## Clustering de variables et implémentation

## R6


**Auteurs**

Marin NAGY
Mazilda ZEHRAOUI
Rina RAZAFIMAHEFA

**Encadrement**

Ricco RAKOTOMALALA


Master 2 SISE |

2025-2026

# Contexte et Objectifs du Projet

## Objectif Pédagogique

Maîtriser la programmation orientée objet sous R via le système **R6**.

- Encapsulation des données et méthodes.
- Architecture logicielle modulaire.
- Création d'un package structuré et documenté.

## Objectif Méthodologique

Implémenter des algorithmes de **classification de variables**.

- Regrouper les variables homogènes/redondantes.
- Réduction de dimension et création de facteurs latents.
- Traitement des données mixtes (numériques et catégorielles).

# Validation du Cahier des Charges

## Exigences du Sujet

- **Implémentation POO :** Utilisation stricte de R6.
- **Diversité des méthodes :** Au moins deux algorithmes distincts.
- **Réallocation :** Une méthode permettant la réaffectation dynamique (type K-means).
- **Gestion Mixte :** Support des variables quantitatives et qualitatives.
- **Interface :** Application Shiny complète pour l'utilisateur final.

## Réalisations

Le package **mmrClustVar** valide l'intégralité des points avec 4 méthodes implémentées :

1. **K-means** (Numérique)
2. **K-modes** (Catégoriel)
3. **K-prototypes** (Mixte)
4. **K-medoids** (Mixte & Dissimilarités)

# Architecture Orientée Objet (R6)

Le package suit une architecture hiérarchique stricte pour garantir la maintenabilité et l'extensibilité.

**Classe Mère (ClusterBase)**

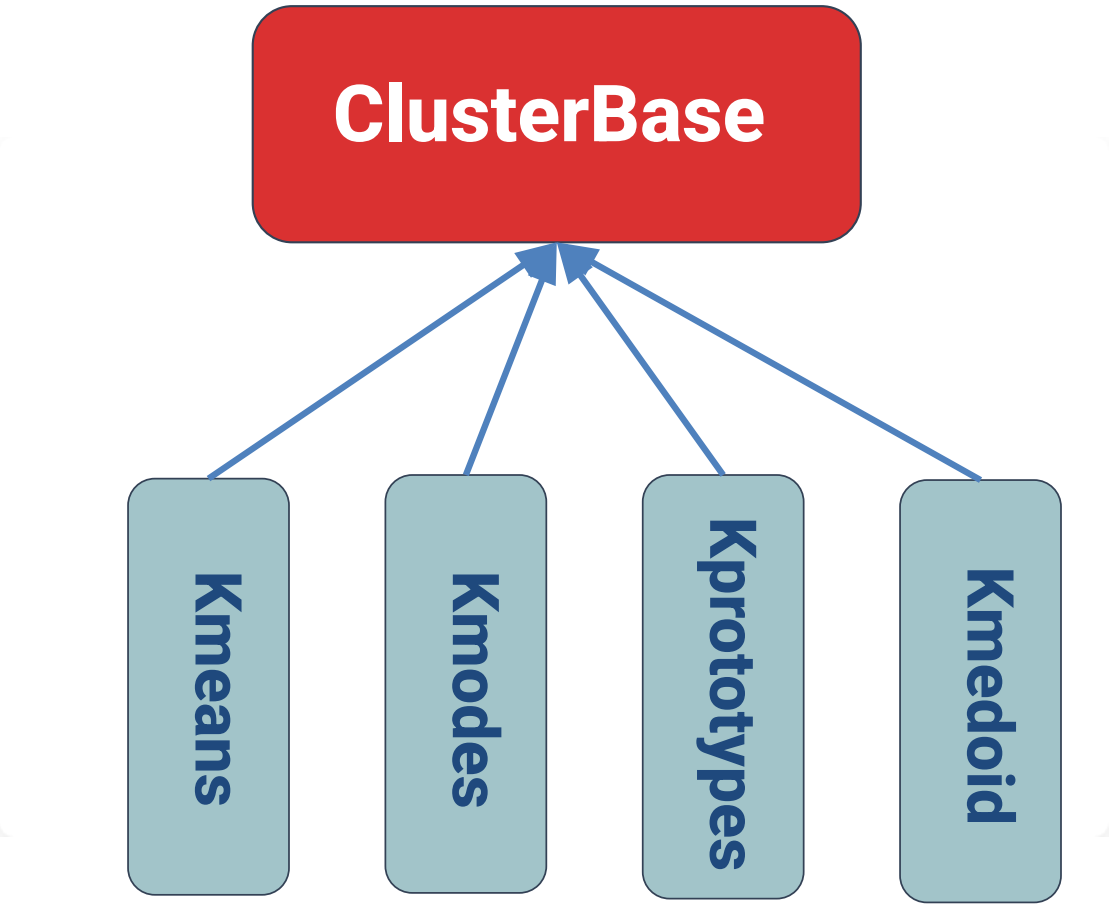
Gère le pipeline commun : fit(), summary(), plot() et la préparation des données.

**Classes Filles**

Implémentations spécifiques : *KMeans, KModes, KPrototypes, KMedoids*.

**Façade (Interface)**

Point d'entrée unique pour l'utilisateur, sélectionnant automatiquement la méthode adaptée.



# Méthodes de Clustering Implémentées

## 1. K-means

**Données :** Numériques
**Centre :** Composante principale (ACP locale)
**Critère :** Maximiser la corrélation au carré.

## 2. K-modes

**Données :** Catégorielles
**Centre :** Profil modal (Mode par individu)
**Critère :** Minimiser la distance de Hamming.

## 3. K-prototypes

**Données :** Mixtes
**Centre :** Combinaison (ACP + Mode)
**Paramètre :** Pondération λ pour l'équilibre.

## 4. K-medoids

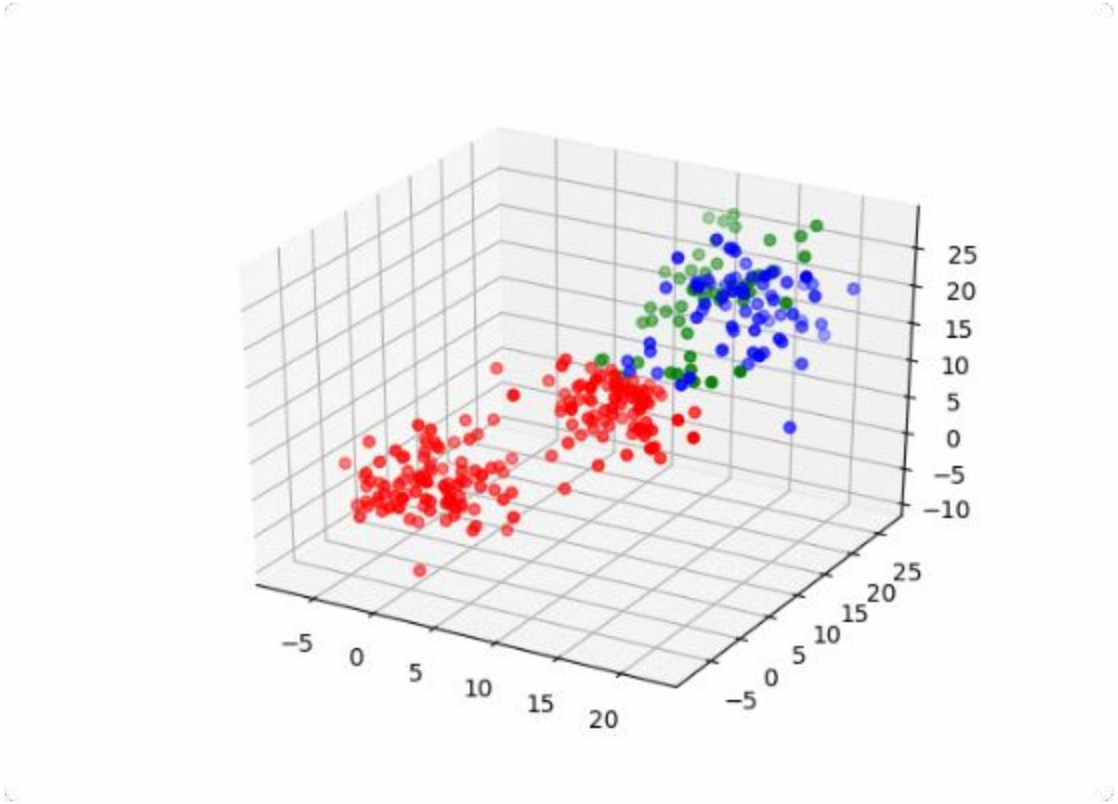
**Données :** Mixtes (via dissimilarités)
**Centre :** Variable réelle du cluster (Médoïde)
**Avantage :** Interprétabilité directe.

# Structure Algorithmique & Spécificités

Tous les algorithmes suivent la structure itérative de type EM (Expectation-Maximization)


| Algorithme | Étape 2 : Critère d'Affectation | Étape 3 : Mise à jour du Centre |
|---|---|---|
| **K-means (Var)** | Distance corrélative | 1ère composante principale (ACP locale) |
| **K-modes** | Distance de Hamming (Différences) | Mode (Catégorie la plus fréquente) |
| **K-prototypes** | Distance mixte et lambda | Combinaison ACP + Mode |
| **K-medoids** | Matrice de dissimilarité (Gower) | Médoïde (Variable réelle la plus centrale) |

# Indicateurs et Diagnostics




## Outils d'interprétation

Pour valider la qualité de la partition, nous fournissons plusieurs métriques :

- **Inertie Expliquée :** Ratio Inertie Inter / Inertie Totale.
- **Courbe "Coude" :** Aide à la décision du nombre de clusters K.
- **Dendrogramme :** Pour visualiser la hiérarchie (méthodes mixtes).
- **Factor Maps :** Projection des variables sur les plans factoriels (pour K-means).

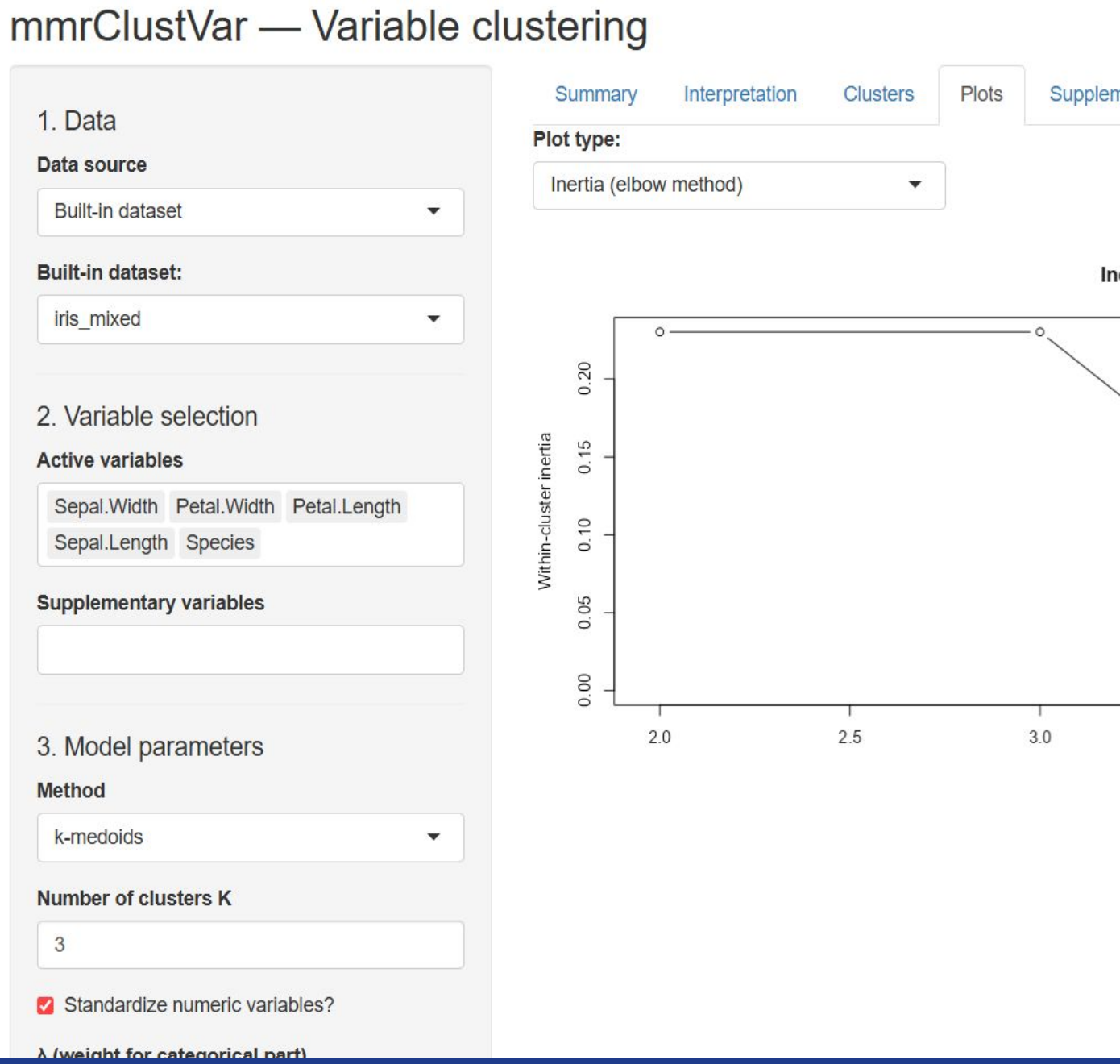
*"Ces indicateurs permettent d'identifier les variables les plus représentatives et de détecter les clusters instables."*

# Application Shiny

L'interface utilisateur a été conçue pour être intuitive et complète, permettant l'analyse sans code.

## Fonctionnalités Clés

- **Import Flexible :** CSV, Excel et datasets internes.
- **Configuration :** Choix interactif des variables actives et supplémentaires.
- **Paramétrage :** Sélection de la méthode, du nombre de clusters **K**, et du **lambda**.
- **Reporting :** Export des résultats et génération de rapports automatisés.



# Limites et Perspectives

## Paramètre Lambda

Dans *K-prototypes*, le choix du poids lambda entre la partie numérique et catégorielle est manuel. Une optimisation automatique serait une amélioration majeure.

## Métriques de Stabilité

Ajout d'indicateurs de stabilité comme **le *Bootstrap*** ou **le coefficient de Silhouette** adapté aux variables pour valider la robustesse des groupes.

## Approche Hiérarchique

Intégration d'une méthode HCPC (Hierarchical Clustering on Principal Components) spécifique aux variables pour compléter le partitionnement.

# Conclusion

Le projet **mmrClustVar** fournit une solution robuste et modulaire pour l'analyse multidimensionnelle. Il combine la rigueur de la programmation R6 avec une interface utilisateur accessible.

## Merci de votre attention

Avez-vous des questions ?