# Régressions SOM-Clusterwise et SOM-PLS-Clusterwise

**Encadrants :**
Ndeye Niang
Sylvie Thiria
**Stagiaire :**
Rémi Bernhard

CNRS/Locean

Mars-Juin 2018

# Introduction
## Objectif

- **Objectif :**
  Développement d'un nouvel algorithme de régression clusterwise qui hérite des propriétés des cartes topologiques de Kohonen

- **Enjeux :**
  1. Performances en généralisation
  2. Robustesse
  3. Convergence

# Plan
## Démarche

# Notations

**n** : nombre d'individus

**p** : nombre de variables explicatives

**X** : matrice des variables explicatives ($\mathbf{n} \times \mathbf{p}$)

**y** : vecteur variable réponse ($\mathbf{n} \times \mathbf{1}$)

**G** : nombre de groupes

# Régression PLS(*Partial Least Squares*)

**Objectif de la régression PLS** : Pouvoir faire de la régression linéaire quand il y a un grand nombre de variables explicatives ou qu'elles sont très corrélées entre elles

**Principe** : On cherche une matrice $T$ $(n \times q)$
avec :

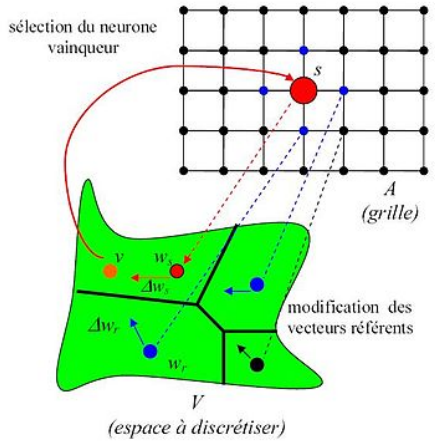
- $q < p$
- Chaque colonne de T est une combinaison linéaire des colonnes de X
- Les colonnes de T sont orthogonales entre elles
- On tient compte de la relation entre $X$ et $y$

# Cartes topologiques de Kohonen

Classification non-supervisée

$G$ groupes : $A$ fonction d'affectation

$G$ référents : $\Omega = (\omega_1, \omega_2, ..., \omega_G)$




Fonction de coût :

$$J(A, \Omega) = \sum_{i=1}^{n} \sum_{c=1}^{G} K_t(\delta(A(X_i), c)) \|X_i - \omega_c\|_2$$

$X_i$ : individu $i$

$\delta(A(X_i), c)$ : distance sur la carte entre les neurones $A(X_i)$ et $c$

$K_t$ : fonction de voisinage gaussienne

# Régression clusterwise (Motivation et principe)

**Motivation** : On pense que les individus sont classés en $G$ groupes et que la relation entre $y$ et $X$ est différente selon le groupe

**Objectif** : Classer les individus en $G$ groupes tout en ayant le meilleur modèle linéaire au sein de chacun des groupes.

**Principe :** Classification et régression réalisées en simultané

# Régression clusterwise (Illustration)

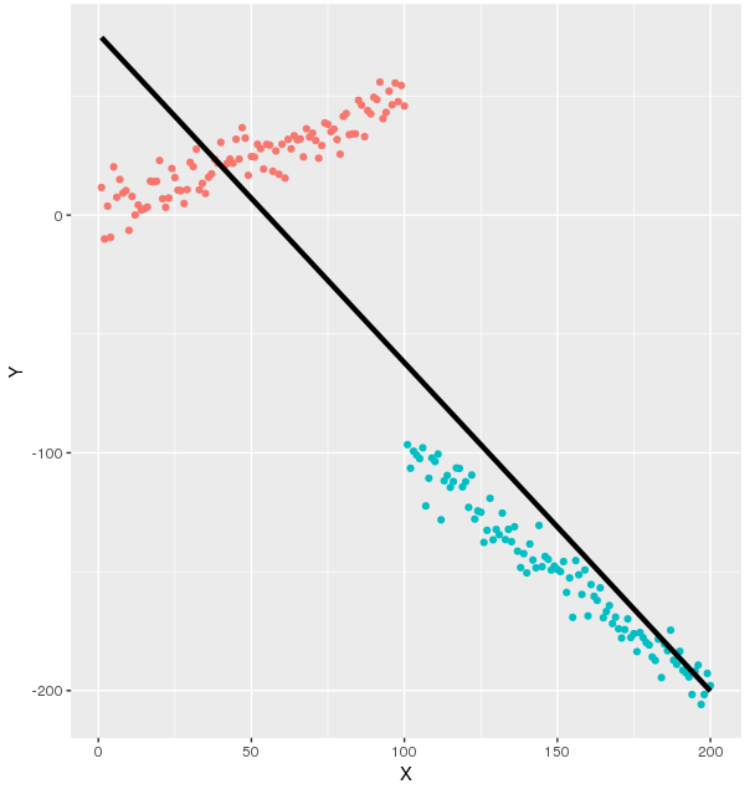



FIGURE 1 – Régression linéaire
$R^2 = 0.04$

FIGURE 2 – Régression clusterwise
$R^2 = 0.94 - 0.96$

# Régression clusterwise (Méthodes existantes)

**Méthodes existantes** :

1. méthode géométrique (minimisation d'une fonction de coût)
   *Défaut : faible robustesse ($p > n$ dans un groupe)*

2. méthode probabiliste (maximisation de la vraisemblance)
   *Défaut : choix d'une loi de probabilité pour $y$ obligatoire*

# Rappels

## Régression clusterwise (algorithme kreg)

**Fonction de coût** :

$$J(A, \beta) = \sum_{i=1}^{n} (y_i - X_i^T \beta^{A(X_i)})^2$$

avec :
$\beta^c$ le vecteur de coefficients pour le groupe $c$
$A$ une fonction qui affecte à chaque individu un numéro de groupe

**Principe** (batch) :
Affectation à la droite de régression la plus proche
Apprentissage d'un modèle linéaire au sein de chaque groupe

**Inconvénients :**

- Sensibilité aux groupes où $n < p$
- Pas de notion d'ordre

# Algorithme SOM-Clusterwise

## Motivation et principe

**Motivations** :

- Tirer parti des avantages des cartes topologiques de Kohonen
- Pouvoir traiter le cas où $n < p$ dans un groupe
- Ne pas avoir à choisir de loi de densité pour $y$

**Principe** :
Adapter la fonction de coût des cartes de Kohonen à un problème de régression linéaire

# Algorithme SOM-Clusterwise

## Présentation de l'algorithme

*Kohonen*

$$J(A,\Omega) = \sum_{i=1}^{n}\sum_{c=1}^{G} K_t(\delta(A(X_i),c))\|X_i - \omega_c\|_2$$

*SOM – Clusterwise*

$$J(A,\beta) = \sum_{i=1}^{n}\sum_{c=1}^{G} K_t(\delta(A(X_i),c))(y_i - X_i^T\beta^c)^2$$

avec :

- les référents qui sont notés $\beta = (\beta^1, \beta^2, ..., \beta^G)$

# Algorithme SOM-Clusterwise

## Présentation de l'algorithme

Algorithme batch
A $t$ fixée :

**Phase d'affectation :**
$\forall i \in [\![1, n]\!]\ A(Z_i) = \arg\min_{r \in [\![1,G]\!]} \sum_{c=1}^{G} K_t(\delta(r, c))(y_i - X_i^T \beta^c)^2$
*On tient compte des distances sur la carte*

**Phase de minimisation :**
$\forall c \in [\![1, G]\!]\ \beta^c = \left(X^T M_c X\right)^{-1} X^T M_c y$
*Régression pondérée de $Y$ par $X$ par les valeurs de la fonction de voisinage*

*Algorithme kreg avec carte topologique*

# Algorithme SOM-Clusterwise

## Déroulement de l'algorithme

On fait décroître la valeur de la température $t$ au fur et à mesure de l'exécution (réduction du voisinage considéré)

Pour chaque valeur de $t$ : atteinte d'un minimum local de la fonction de coût

*A t fixée, on atteint un minimum local de la fonction de coût en un temps fini*

# Algorithme SOM-Clusterwise

## Illustration

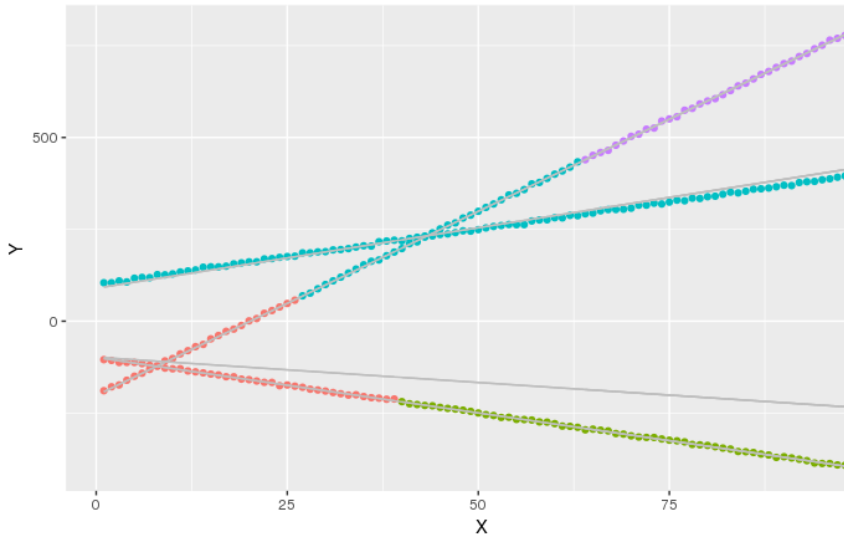



FIGURE 3 – 4-moyennes puis régression linéaire

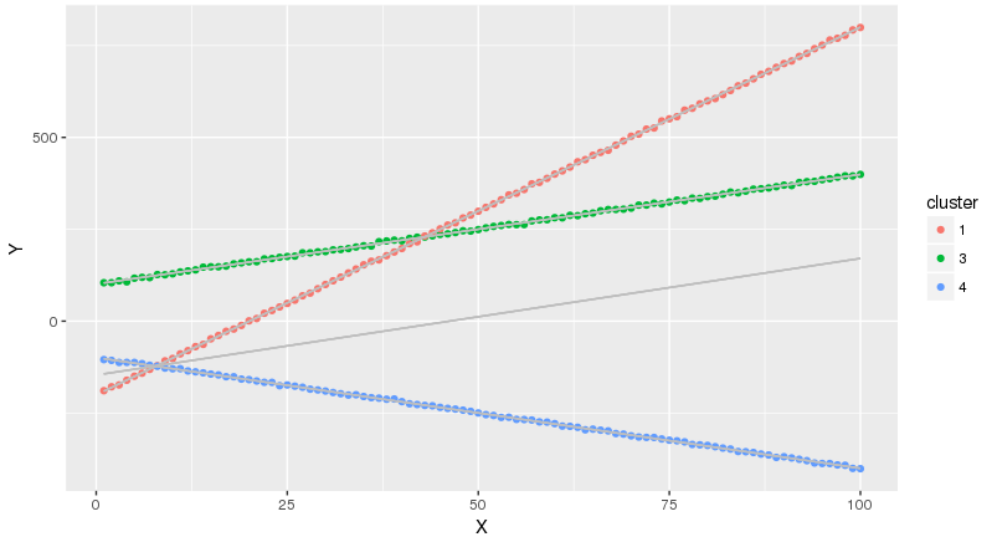



FIGURE 4 – Algorithme SOM-Clusterwise

# Algorithme SOM-Clusterwise

## Illustration

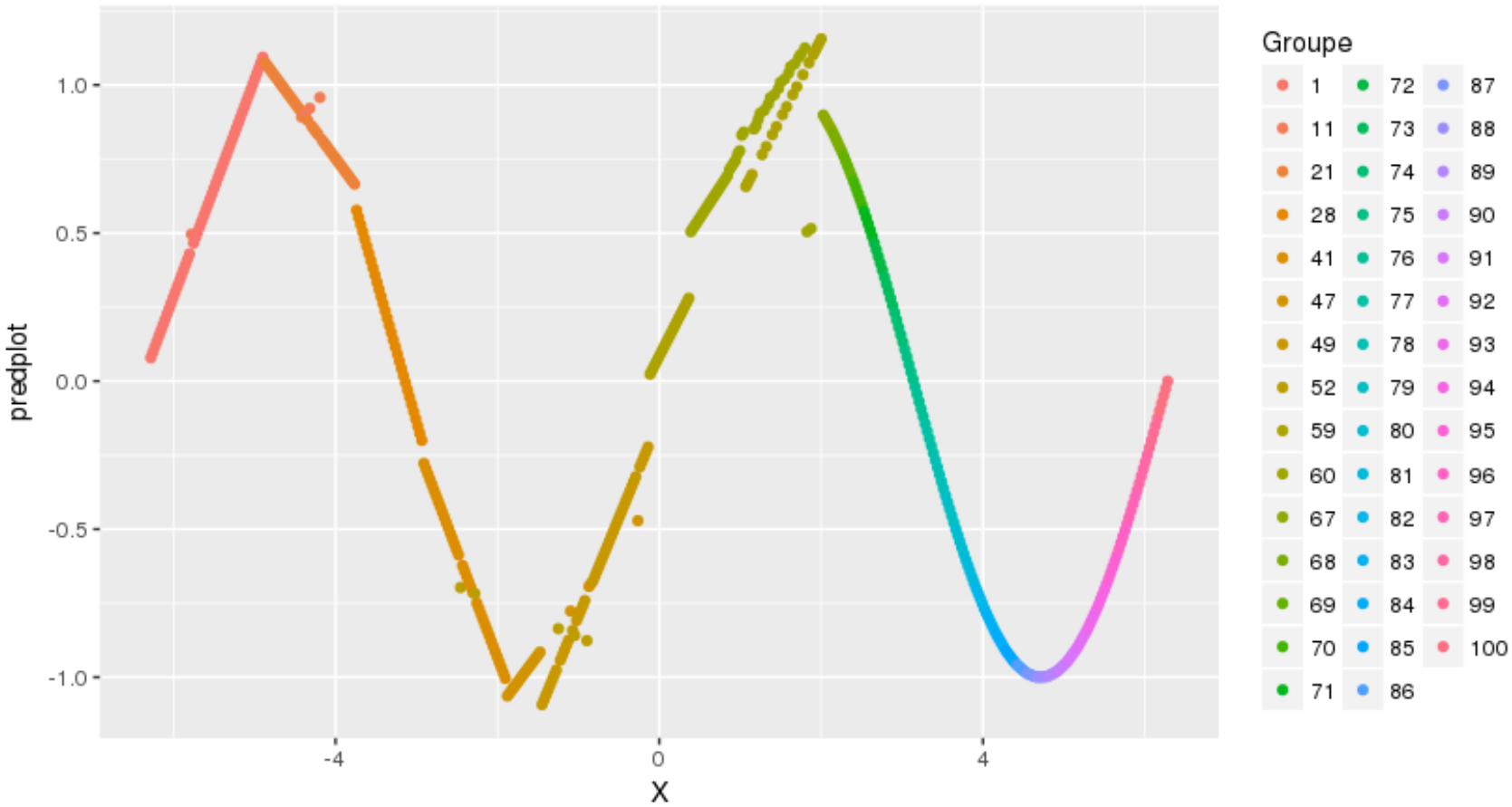



FIGURE 5 – Illustration de la notion d'ordre

# Algorithme SOM-Clusterwise

## Robustesse

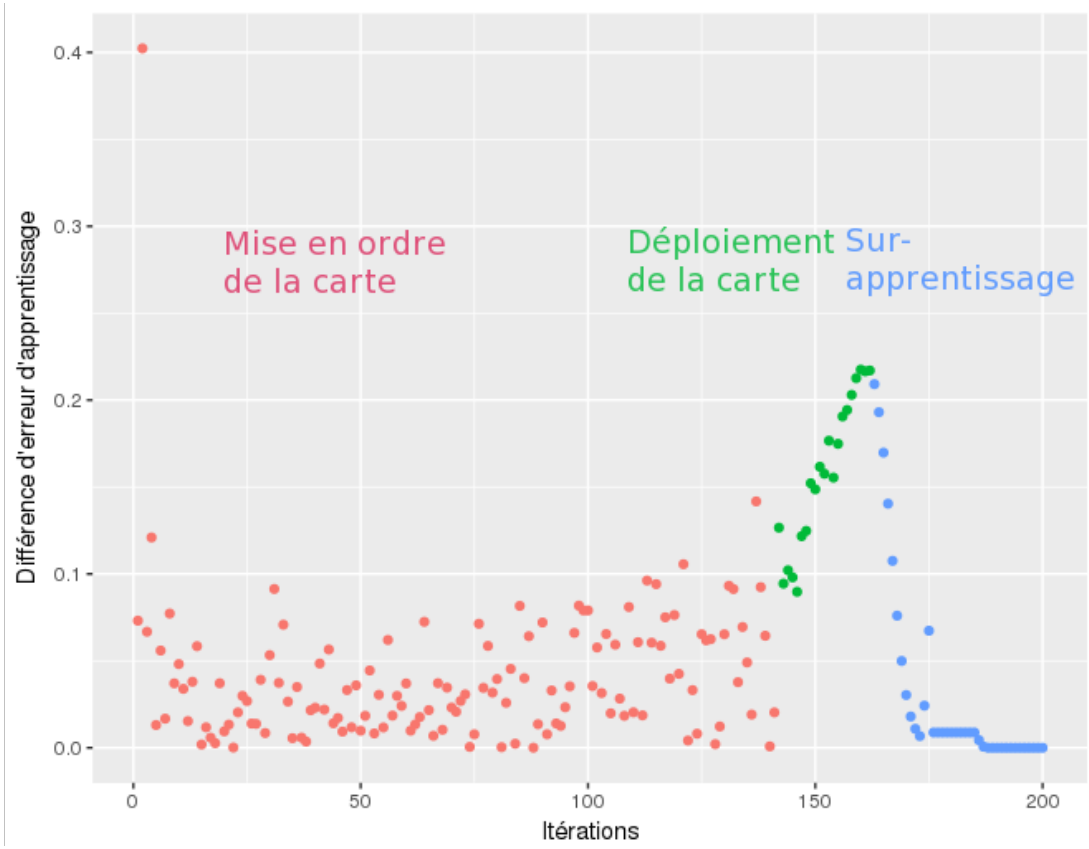
**Motivation :**
Si $p$ et $G$ sont élevés par rapport à $n$, la régression linéaire au sein d'un groupe peut devenir impossible à basse température.

**Solution :**
Définir un critère d'*early-stopping*, basé sur l'évolution de l'erreur sur l'ensemble d'apprentissage en fonction de la température

# Algorithme SOM-Clusterwise

## Robustesse

**Erreur d'apprentissage :** Trois phases bien distinctes




*Early-stopping* : arrêter l'algorithme avant le sur-apprentissage

# Méthodologie des comparaisons

**Données simulées** :

- chaque variable explicative $X^j$ : lois normales $(\mu_1, \mu_2, ..., \mu_G)$, $\sigma = 1$
- $Y^c = \sum_{j=1}^{p} \alpha_c X^j + \epsilon$ avec $\epsilon \sim \mathcal{N}(0, 0.1)$, $\alpha_c \in [\![-1, 1]\!]$

**SOM-Clusterwise et kreg** :
Pour chaque cas envisagé, 5 simulations
Pour chaque simulation, 5 exécutions des algorithmes
*On garde les résultats relatifs à la plus petite valeur de la fonction de coût*

# Comparaisons SOM-Clusterwise et kreg

## Résultats

| **Cas numéro** | **n** | **p** | **G** | **Séparation** | $\mathbf{G_{algo}}$ |
|---|---|---|---|---|---|
| 1 | 540 | 2 | 4 | moyennement séparés | 9 |
| 2 | 540 | 2 | 4 | bien séparés | 9 |
| 3 | 180 | 2 | 4 | bien séparés | 16 |

| **Cas numéro** | **algorithme** | **rand** | **CH** | $\mathbf{err_{train}}$ | $\mathbf{err_{test}}$ |
|---|---|---|---|---|---|
| 1 | SOM-clus | 0.88 | 137 | 0.08 | 4.25 |
| | kreg | 0.5 | 47 | 30 | 7.13 |
| | SOM-clust-d | 0.84 | 58 | 0.09 | 3.31 |
| 2 | SOM-clus | 0.81 | 55 | 0.03 | 0.34 |
| | kreg | X | X | X | X |
| | SOM-clust-d | 0.8 | 22 | 0.05 | 0.28 |
| 3 | SOM-clus | X | X | X | X |
| | kreg | X | X | X | X |
| | SOM-clust-d | 0.8 | 5.23 | 0.01 | 0.25 |

# Comparaisons SOM-Clusterwise et kreg

## Résultats

- Critères d'*early-stopping* efficaces en termes de robustesse et de performances en prédiction
- Justification de la pertinence de la carte topologique dans un algorithme de régression clusterwise

# Données
## Description

1960 observations (1960 relevés à Banizoumbou, Niger)

1. *Variables explicatives* : 16 variables locales et 936 variables de grande échelle
   - Date (heure, jour, mois, année) (non utilisées dans cette étude)
   - Vitesse du vent (en $m.s^{-1}$)
   - Direction du vent (de 0 à 360 degrés)
   - Température (en °C)
   - Humidité (en pourcentage)
   - Pression (en Bar)
   - Quatre mesures de $AOT$ (Epaisseur Optique en Aérosols) (sans unité)
   - Trois mesures du coefficient d'Angström (sans unité)
   - 936 variables de grande échelle (non utilisées dans cette étude)
2. *Variable réponse* : Pollution en particules PM10 ($\mu g.m^{-3}$)

# Données

## Méthodes linéaires simples

Ecart-type de $y$ : 288 $\mu g.m^{-3}$

| Régression linéaire | 12 variables | 7 variables |
|---|---|---|
| Coefficient de détermination $R^2$ | 0.66 | 0.62 |
| RMSE (10-cross validation) en $\mu g.m^{-3}$ | 440 | 343 |

TABLE 1 – Résultats pour la régression linéaire classique

Régression PLS avec 12 composantes :

- RMSE : 282 $\mu g.m^{-3}$
- *R squared* : 0.66

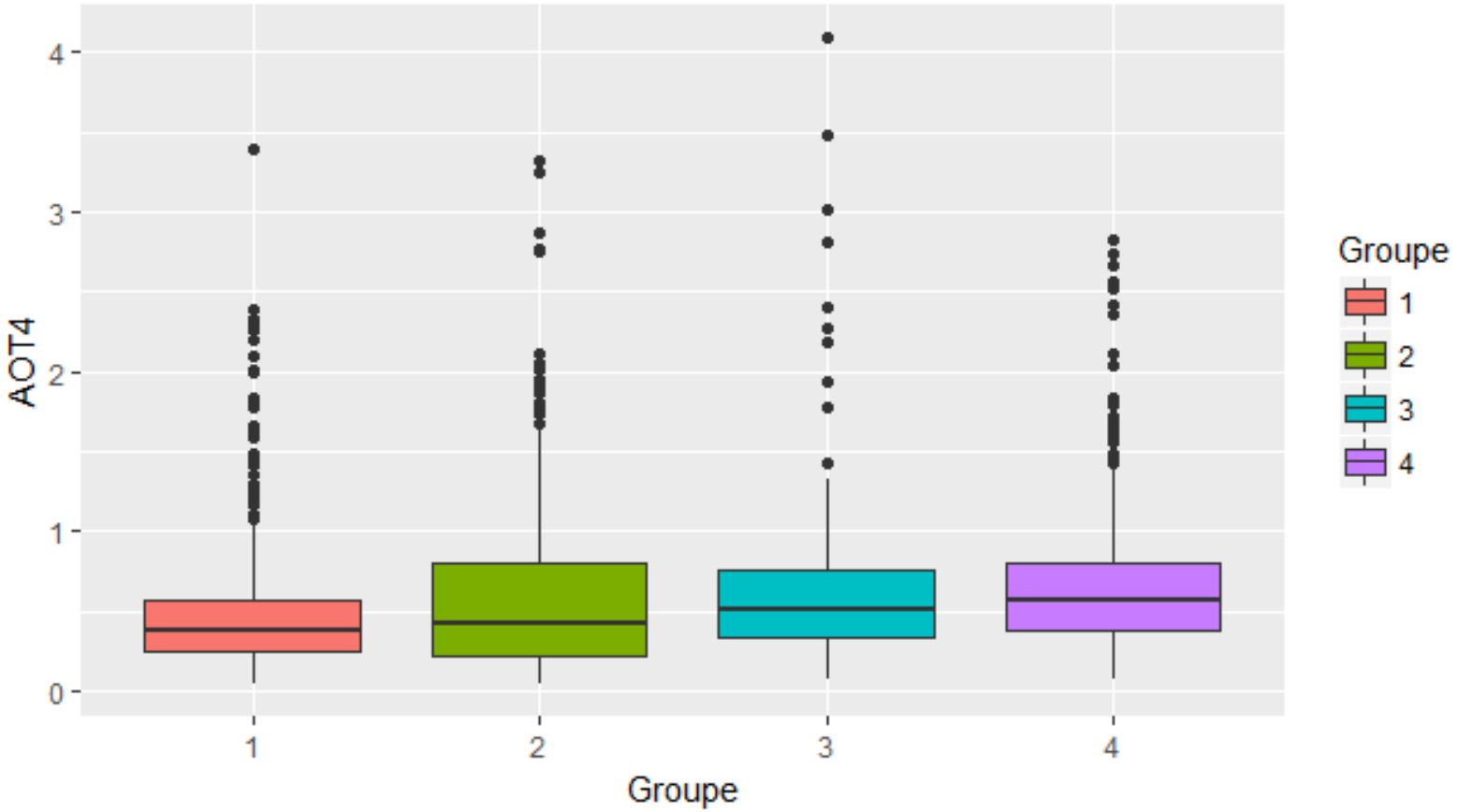
# Données
## Algorithme SOM-Clusterwise

| RMSE(ensemble de test) | | Algorithme | | |
|---|---|---|---|---|
| | | kreg | Spaeth | SOM-clus |
| $G_{algo}$ | 4 | 177.1 | 174.9 | 149.42 |
| | 6 | 183.39 | 172.81 | 168.44 |

TABLE 2 – Comparaison de l'algorithme SOM-Clusterwise aux algorithmes kreg et Spaeth

*Spaeth : version stochastique de l'algorithme kreg*

# Données
## Interprétation




*AOT : bonne séparation suivant les groupes*

# Données
## Interprétation

| | |
|---|---|
| Groupe 1 | AOT |
| Groupe 2 | Pression |
| Groupe 3 | AOT |
| Groupe 4 | Angström, Direction du vent |

TABLE 3 – $G_{algo} = 4$ : Variables importantes dans chaque groupe

| | |
|---|---|
| Groupe 1 | Humidité |
| Groupe 2 | Humidité, Pression, Direction du vent |
| Groupe 3 | AOT |
| Groupe 4 | Pression, Direction du vent |
| Groupe 5 | |
| Groupe 6 | Température |

TABLE 4 – $G_{algo} = 6$ : Variables importantes dans chaque groupe

# Conclusion et perspectives

## Bilan

- **SOM-Clusterwise** :
  1. bonnes performances en généralisation
  2. robustesse assurée
  3. rapidité de convergence
- **SOM-PLS-Clusterwise** :
  1. plus performant que *mbpls* dans certains cas ($G_{algo} > G$)

## Perspectives :

1. Algorithmes SOM-Clusterwise et PLS-SOM-Clusterwise
   $\rightarrow$ parallélisation des tâches (Spark,...)
2. Critère pour le choix du nombre de groupes

# Régressions SOM-Clusterwise et SOM-PLS-Clusterwise

**Encadrants :**
Ndeye Niang
Sylvie Thiria
**Stagiaire :**
Rémi Bernhard

CNRS/Locean

Mars-Juin 2018

**Références :**

Hervé Abdi. Partial least squares regression and projection on latent structure regression (pls regression). Wiley Interdisciplinary Reviews : Computational Statistics, 2(1) :97–106, 2010.

Stéphanie Bougeard, Hervé Abdi, Gilbert Saporta, and Ndèye Niang. Clusterwise analysis for multiblock component methods. Advances in Data Analysis and Classification, 2017.

Helmuth Späth. Algorithm 39 clusterwise linear regression. Computing, 22(4) :367–373, 1979.