Université d'Aix-Marseille **Année 2024-2025**
Master 1 Mathématiques Appliquées, Statistique
Cours de classification, fiche 2

Objectif : savoir mettre en œuvre l'algorithme des moyennes mobiles (k-means)

La table 1 présente[1] pour l'année 2023 et pour 5 départements français les valeurs prises par les variables :

- EqSport : "Nombre d'équipements sportifs (piscines, tennis, cyclisme, athlétisme, sports de glace) pour 100 000 habitants";
- EqCult : "Nombre d'équipement culturels (cinémas et théâtres) pour 100 000 habitants".

Département	EqSport	EqCult
Alpes-de-Haute-Provence	450.41	72.84
Dordogne	412.48	66.13
Landes	328.18	39.94
Mayenne	500.12	56.91
Pyrénées-Orientales	355.29	41.52

Table 1: Pour chaque département, nombre d'équipements sportifs pour 100000 habitants et nombre d'équipements culturels pour 100000 habitants.

1. On donne $s(\text{EqSport})=69,73$. Quelle est la valeur de $s(\text{EqCult})=?$
2. Quelle distance a été considérée pour construire la table 2 ? Comment cette distance peut-elle être justifiée ? Compléter la valeur de la distance entre les départements des Pyrénées-Orientales et de la Mayenne.

	Alpes-de-Haute-Provence	Dordogne	Landes	Mayenne
Dordogne	0.71			
Landes	2.85	2.16		
Mayenne	1.30	1.41	2.73	
Pyrénées-Orientales	2.54	1.87	0.40	**?**

Table 2: Valeurs des distances entre chaque paire d'individus.

3. On décide de réaliser une partition des individus en deux groupes, en utilisant l'algorithme des centres mobiles. Pour cela on choisit comme centres initiaux des classes les individus Alpes-de-Haute-Provence et Mayenne.
   (a) Définir les partitions engendrées par les centres initiaux.
   (b) Donner l'inertie intra-classes associée à cette partition.
   (c) Déterminer les coordonnées des nouveaux centres de classes c_1 et c_2.
   (d) En vous aidant des résultats présentés table 3, construire la partition induite par les nouveaux centres de classes.
   (e) En déduire la variance intra-classes associée à cette partition.

[1] Source : Base permanente des équipements 2023

individu	c_1	c_2
Alpes-de-Haute-Provence	0.36	2.00
Dordogne	0.36	1.39
Landes	2.51	?
Mayenne	1.31	1.68
Pyrénées-Orientales	2.20	0.65

Table 3: Distances entre chaque individu et les centres des classes de la partition initiale.

(f) Décrire la suite de l'algorithme jusqu'à sa fin.

4. Créer un dossier au nom de `TP-CLASSIF`. Le dossier `TP-CLASSIF` contiendra les dossiers des différents TP de l'UE Classification. Créer dans `TP-CLASSIF` un nouveau dossier au nom de `TPKmeans`. Retrouver le résultat précédent sous `R` en utilisant la fonction `kmeans`. Pour cela on pourra sous `R` :

   (a) télécharger l'objet `equip5.rds` depuis AMeTICE et l'enregistrer sous `TPKmeans/data`

   (b) définir le répertoire `data` comme étant le répertoire de travail
   `setwd(dir=" /F/MASS/2024-2025/M1/classification/TP-CLASSIF/TPkmeans/data")`

   (c) charger sous R l'objet `equip5` :
   `>load("equip5.rds")`

   (d) visualiser l'objet equip5 :
   `>View(equip5)`
   A quoi correspond l'objet equip5 ?

   (e) Quelle est la conséquence des opérations suivantes ?
   ```
   >equip5[,1]<-equip5[,1]/sd(equip5[,1])
   >equip5[,2]<-equip5[,2]/sd(equip5[,2])
   >dist(equip5)
   >cl<-kmeans(x=equip5,centers=equip5[c(1,5),],trace=F)
   ```

   (f) Décrire les éléments de `cl`

   (g) Décrire le graphique obtenu en exécutant les lignes suivantes
   ```
   plot(equip5,col=cl$cluster,pch=16)
   abline(v=mean(equip5[,1]),h=mean(equip5[,2]),pch=2,col="grey")
   ```

On considère maintenant le jeu de données équivalent au précédent mais portant sur l'ensemble des départements français. Ce jeu de données est disponible sur AMeTICE via le fichier `equip.rds`.

5. En s'inspirant de la question 4 réaliser une partition des départements suivant l'algorithme des centres mobiles.

6. On s'intéresse maintenant au choix du nombre k d'éléments de la partition : quel critère pourrait-on proposer pour choisir k ?

   Après avoir exécuté le programme ci-dessous, proposer un choix pour k. Vous expliquerez comment les centres de classes ont été choisis.

```
K=30
W<-rep(NA,K)
W[1]<-ncol(equip)*(nrow(equip-1))
for(k in (2:K)){
cl<-kmeans(x=equip,centers=equip[sample((1:nrow(equip)),k),])
W[k]<-cl$tot.withinss
}
barplot(W,names.arg=c(1:30))
```

7. En utilisant la valeur de k déterminée à l'étape précédente, exécuter le programme suivant :

```
cl<-kmeans(equip,equip[sample((1:nrow(equip)),k),])
plot(equip,col=(cl$cluster),pch=16)
abline(v=mean(equip[,1]),h=mean(equip[,2]),pch=2,col="grey")
cl$tot.withinss
```

```
cl<-kmeans(equip,equip[sample((1:nrow(equip)),k),])
plot(equip,col=(cl$cluster),pch=16)
abline(v=mean(equip[,1]),h=mean(equip[,2]),pch=2,col="grey")
cl$tot.withinss
```

   En déduire les facteurs dont dépendent la partition obtenue pour une réalisation de l'algorithme des centres mobiles.