Nous travaillons sous R / RSTUDIO durant cette séance.

Supports de référence :

- **Vidéo 1** – « Combinaison K-Means et CAH », https://www.youtube.com/watch?v=nik6O1DGykc
- **Vidéo 2** – « Combinaison Carte de Kohonen et CAH », https://www.youtube.com/watch?v=HIV1WJ6dg6A

1. Inspection et préparation des données

Nous travaillons avec le fichier « **covertype_100k.xlsx** » (le fichier est accessible sur notre drive : https://drive.google.com/drive/folders/1ZgmOZoC7HAqqpTF2neNIcMhN73-fVwr6 - Voir dans le dossier **TD 4**). Différents types de forêts sont décrites à l'aide de variables cartographiques. Des modifications ont été réalisées par rapport à la base originelle (source Kaggle) : seules les 10 premières variables ont été conservées, nous nous contentons de traiter 100.000 observations.

1. [**Préambule**] Pour que nous ayons tous le même environnement, après avoir démarré RSTUDIO, vous créez un nouveau notebook (FILE / NEW FILE / R NOTEBOOK – voir https://www.youtube.com/watch?v=u6pqsK8_vO4 à partir de **06:50**). Vous pourrez ainsi visualiser en temps réel vos sorties sous forme de fichier HTML (**08:30**).
2. Si ce n'est déjà fait, installez la librarie « **readxl** » (dans RStudio, voir le menu « TOOLS / INSTALL PACKAGES ; ou encore, utilisez la commande install.packages).
3. Importez la librarie « readxl » (library).
4. Chargez (read_excel -- https://readxl.tidyverse.org/reference/read_excel.html) la feuille « **var_actives** » de « **covertype_100k.xlsx** ». Affichez les caractéristiques de la base (str).
5. La colonne « Numéro » est un simple identifiant qui permet de repérer les individus. Les variables actives sont « X1 » … « X10 ». Créez une copie de la base excluant la colonne « Numéro ».
6. Calculez les statistiques descriptives des variables (summary). Que constatez-vous si l'on s'intéresse aux moyennes ? (les variables ne sont pas exprimées sur les mêmes échelles)
7. Centrez et réduisez les variables (scale) (**Vidéo 1, 09:44**)
8. Affichez de nouveau les statistiques descriptives. Qu'observez-vous ? (moyennes = 0)
9. Affichez les attributs de la structure issue du centrage-réduction (attributes).
10. Comment accéder aux propriétés spécifiques de l'objet ? (attr) Affichez les moyennes et écarts-type utilisés lors de l'opération de standardisation ci-dessus.

2. Pré-clustering avec les K-Means

11. Nous souhaitons opérer un premier regroupement en (**K = 10**) clusters des individus via la méthode des K-Means (kmeans). Avant de lancer les calculs, fixez la valeur de départ du générateur de nombres aléatoires pour que nous ayons les mêmes résultats [set.seed(0)] (**Vidéo 1, 11:02**).
12. Affichez les résultats (print). Quelle est la part d'inertie expliquée par la partition ? (55.7%)
13. Affichez les attributs de l'objet issu du clustering (attributes)
14. Quelle est la valeur de l'inertie totale T ? ($totss)
15. Essayez de la recalculer en vous aidant des différentes formules vues en cours, sachant que les variables ont été centrées et réduites [attention, R utilise « 1/(n-1) » pour le calcul de la variance].

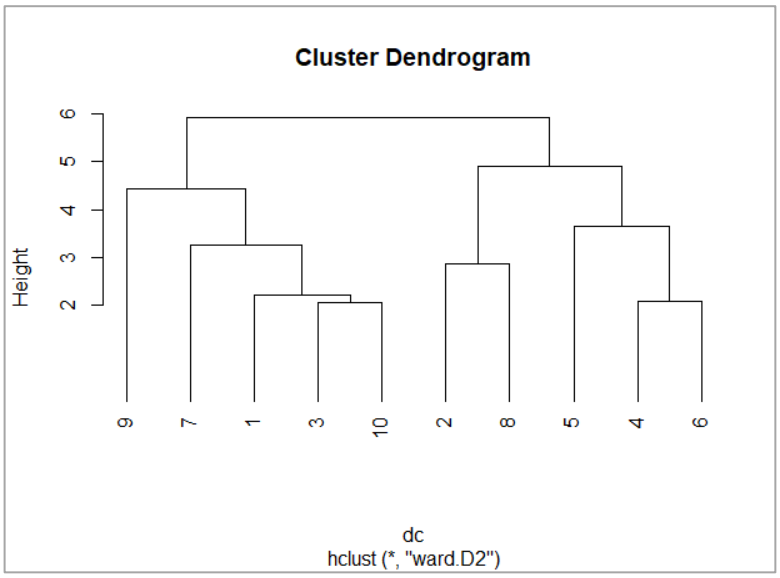
$$TSS = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2$$

16. Quelle est la valeur de l'inertie inter-classes B ? ($betweenss)
17. Calculez explicitement le ratio B/T. Obtenez-vous la même valeur issue de print() ci-dessus ? (oui, sinon very big problème…)
18. Affichez les effectifs par cluster ($size). Effectuez-en la somme (sum). Quelle valeur obtenons-nous ? (100.000) Est-ce normal ? (oui, c'est l'effectif initial de la base, on n'a perdu personne, on n'en a pas gagné non plus, heureusement…)
19. Affichez le n° de groupe d'appartenance des 10 premiers individus ($cluster + head).
20. Comptabilisez (table) les effectifs par classe à l'aide des indicateurs de groupes d'appartenance « $cluster ». Les valeurs sont cohérentes avec celles de la propriété « $size » ? (il faudrait s'inquiéter si ce n'était pas le cas).
21. Affichez les coordonnées des centres de classes ($centers)

3. CAH à partir des pré-clusters des K-Means

22. Calculez et affichez les distances entre centres des pré-clusters issus des K-Means (dist) (**Vidéo 1, 14:50**)
23. Lancez la CAH (hclust) à partir de cette matrice de distance. Utilisez le critère de Ward comme méthode d'agrégation (method = 'ward.D2'). N'oubliez pas de pondérer les « individus » par l'effectif des pré-clusters (members). Affichez le dendrogramme (plot, avec l'option 'hang = -1' [testez avec et sans pour voir la différence]).

24. Combien de classes suggère le dendrogramme ? Effectuez le découpage correspondant (cutree) (**Vidéo 1, 16:38**) (bon, pour que nous soyons tous raccords, mettons que nous choisissons k = 4)



25. Affichez les groupes d'appartenance. Calculez les effectifs par groupe (table). Que constatez-vous ? (les effectifs correspondent en réalité au nombre de pré-clusters associés à chaque groupe, à savoir : 4, 2, 3, 1)
26. Affichez la table de correspondance entre les identifiants de pré-clusters et les identifiants de groupes finaux (issus de la CAH). Mettez-la en relation avec le dendrogramme ci-dessus. Les résultats sont-ils cohérents ? (il vaut mieux que oui) (**Vidéo 1, 18:23**)

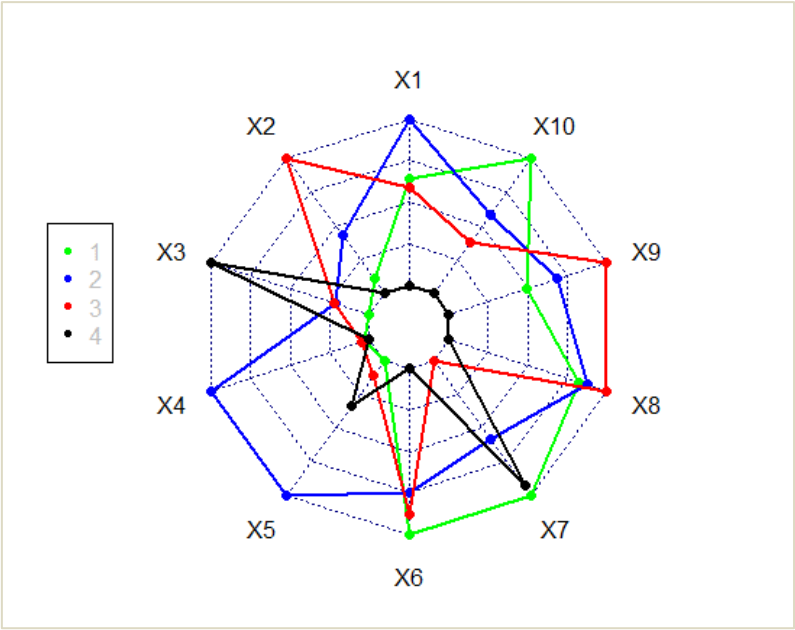
	[,1]	[,2]	[,3]	[,4]
[1,]	1	0	0	0
[2,]	0	1	0	0
[3,]	1	0	0	0
[4,]	0	0	1	0
[5,]	0	0	1	0
[6,]	0	0	1	0
[7,]	1	0	0	0
[8,]	0	1	0	0
[9,]	0	0	0	1
[10,]	1	0	0	0

27. Connaissant les effectifs dans les 10 pré-clusters issus des K-Means, déduisez le nombre d'observations que l'on peut rattacher aux 4 groupes (50381, 13450, 30214, 5955). La somme de ces valeurs est égale à combien ? Ce dernier résultat est-il cohérent avec notre analyse ? (100000, oui cohérent, nous avons bien le nombre initial d'individus à traiter).

28. Affecter chaque individu à son groupe final mis au jour par la CAH (4 groupes en tout). Calculez les effectifs par groupe. Retrouvons-nous les valeurs de la question précédente ? (il vaut mieux…) (**Vidéo 1, 19:19**)
29. **Interprétation des groupes**. Sur les variables centrées et réduites, calculez les moyennes conditionnellement aux groupes d'appartenance (j'ai arrondi les valeurs à la 3ème décimale dans la copie d'écran ci-dessous).

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	0.011	-0.613	-0.308	-0.263	-0.381	0.124	0.542	-0.029	-0.414	0.272
2	0.689	0.038	0.115	1.668	1.655	-0.112	-0.081	0.105	0.117	-0.104
3	-0.086	1.168	0.131	-0.240	-0.160	0.005	-0.955	0.419	0.992	-0.284
4	-1.213	-0.827	1.676	-0.329	0.300	-0.824	0.440	-2.114	-1.794	-0.624

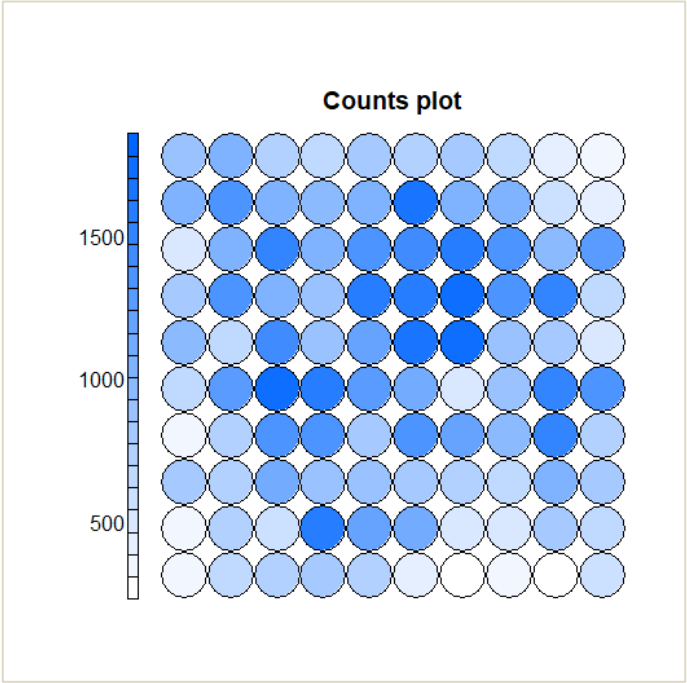
30. Faites afficher ces valeurs dans un graphique « radar » (voir https://r-graph-gallery.com/143-spider-chart-with-saveral-individuals.html, il faudra éventuellement installer la libraire « **fmsb** » avant de pouvoir la charger et l'utiliser ; pour la commande radarchart, spécifiez l'option maxmin = FALSE pour que l'outil s'ajuste automatiquement à vos valeurs extrêmes). Votre graphique devrait ressembler à ceci (j'ai rajouté une légende pour que l'on identifie correctement les groupes d'appartenance). Qu'est-ce qui caractérise le cluster n°1 [celui en vert] ? (par rapport aux autres groupes, valeurs les plus faibles pour X3, X4 et X5 ; les plus élevées pour X10, X6, X7). Le groupe n°2 ? Etc.



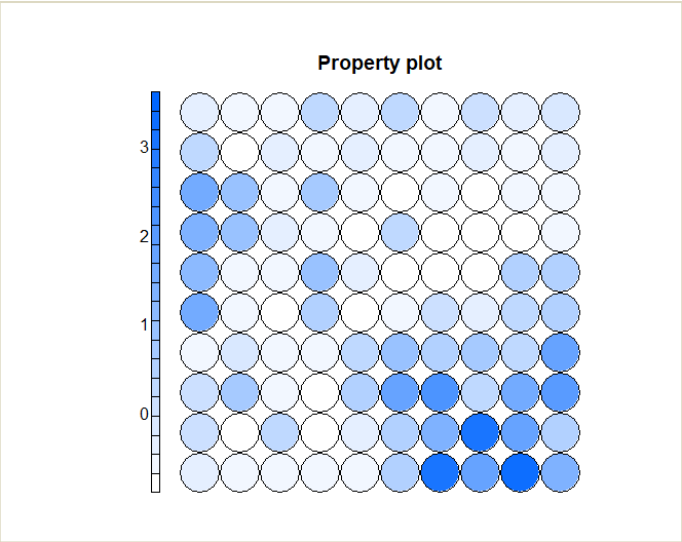
4. Combinaison Kohonen – CAH

Nous souhaitons mettre en œuvre la stratégie mixte en combinant les cartes de Kohonen avec la CAH dans cette section.

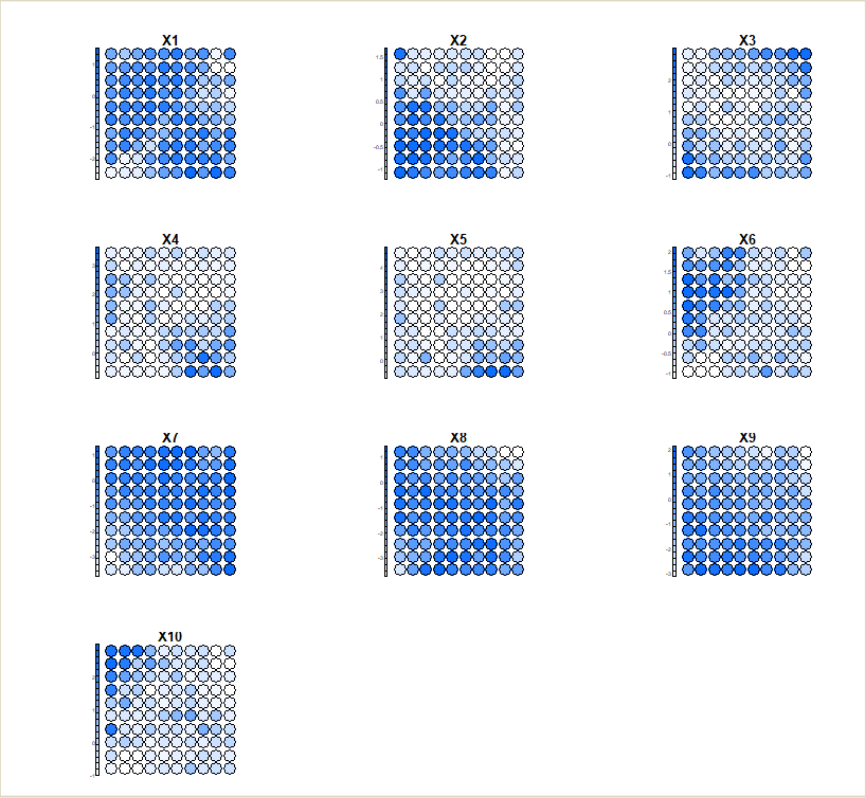
31. Installez puis chargez la librairie « **kohonen** ».
32. Initialisez le générateur de nombres aléatoires pour que nous ayons tous les mêmes résultats [set.seed(0)] (**Vidéo 2, 12:52**)
33. Lancez la procédure som() sur les données centrées et réduites. Nous désirons produire une carte **rectangulaire** de dimension « 10 x 10 ».
34. Affichez les nœuds d'appartenance des 20 premiers individus ($unit.classif)
35. Calculez les effectifs par nœuds (table) (**Vidéo 2, 17:30**). Quel est le nœud qui présente l'effectif le plus élevé ? (max, which.max) (le nœud n°67, avec 1857 obs.) Le plus faible ? (min, which.min) (n°7 avec 230 obs.)
36. Affichez la carte avec les effectifs. Utilisez un dégradé de bleu pour discerner le nombre d'observations par nœud. Le graphique obtenu est-il en cohérence avec les réponses de la question précédente ? (oui, identifiez par exemple le nœud n°67, le plus bleu parmi les bleus) (**Vidéo 2, 18:07**)



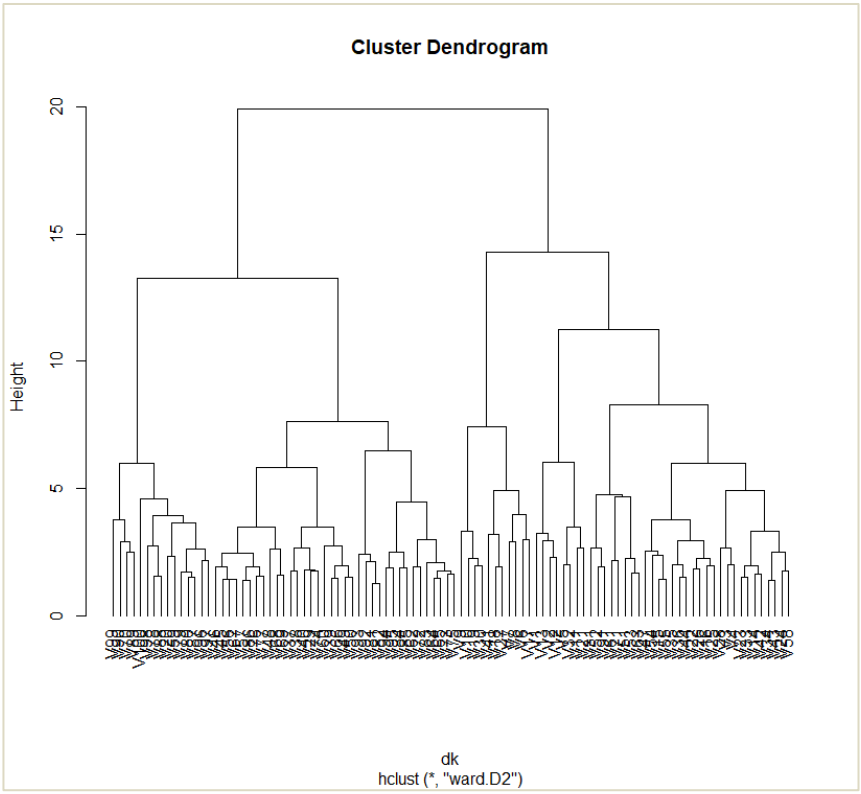
37. Affichez les codebooks de la carte ($codes). Puis, plus particulièrement, celui des 2 premiers nœuds (**Vidéo 2, 19:55**)
38. Affichez la carte en mettant en évidence les contrastes portées par la variable **X4**. Dans quelle zone, la variable **X4** prend-elle des valeurs élevées ? (sud-est) (**Vidéo 2, 22:58 ;** à la différence que je réalise les cartes de l'ensemble des variables dans la vidéo, à ce stade, on ne vous demande que la carte de la variable X4)



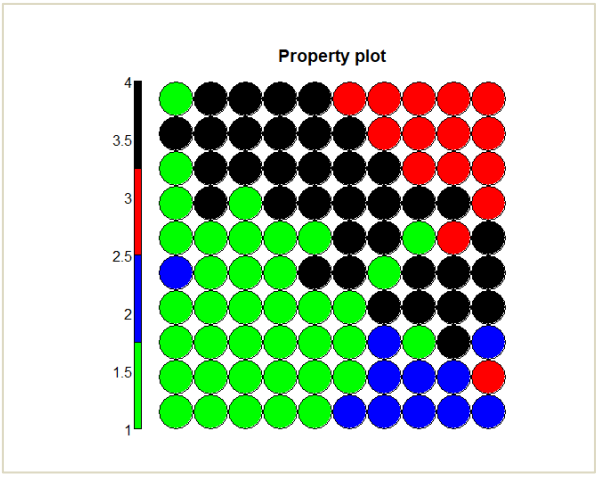
39. Affichez le même graphique pour les 10 variables. Quelles sont les variables qui apportent les plus forts contrastes ? (X2, X3, X4, etc.) Quelles sont les variables peu discriminantes ? (X7, X8, X9)



40. A l'instar de la section précédente (K-Means + CAH), calculez les distances entre les nœuds de carte. (**Vidéo 2, 26:09**)
41. Réalisez la CAH à partir de cette matrice de distances. N'oubliez pas que les nœuds sont pondérés par les individus qu'ils comportent. Affichez le dendrogramme (**Vidéo 2, 26:56**).



42. Ici aussi (comme pour K-Means + CAH), nous optons pour une partition en K = 4 groupes. Effectuez le découpage. Affichez les groupes d'appartenance issus de la CAH.
43. Calculez les effectifs par groupe (38, 11, 15, 36 ; la somme = 100 puisque nous sommes partis initialement de 10 x 10 = 100 nœuds dans la carte).
44. Faites apparaître les groupes de la CAH dans la carte de Kohonen (**Vidéo 2, 29:14**).



45. On s'intéresse au groupe situé dans la partie sud-est de la carte (groupe n°2, en bleu pour moi, tout dépend des couleurs que vous avez choisies pour vous). Enumérez les sommets correspondants (V6, V7, V8, V9, V10, V17, etc.)
46. Si on se réfère aux cartes des contrastes des variables ci-dessus (page 6), cette zone correspondrait aux valeurs élevées de X4 et X5. Essayons de vérifier cela.

a. Isolez dans un vecteur les effectifs associés aux nœuds « bleus » (V6 : 435, V7 : 230, V8 : 359, etc.)
b. Isolez dans une matrice les codebooks (moyennes conditionnelles) de ces nœuds.
c. Calculez la moyenne pondérée de la variable « X5 » (Z5 en réalité puisque les variables ont été centrées et réduites). Quelle valeur obtenez-vous ? (2.55) A quelle référence pourrait-on comparer cette valeur ? (à 0, puisque les variables présentées à l'algorithme ont été initialement centrées et réduites) Est-ce que notre intuition est confirmée (oui, le groupe bleu correspond aux valeurs élevées de X5)
d. Refaites la même manipulation pour X4 (2.04).
e. Et si on s'intéresse à X10, qu'obtenons-nous ? (-0.206). Est-ce cohérent avec les cartes des variables ? (oui, X10 prend ses valeurs élevées plutôt dans la zone nord-ouest)

47. Rattachez les individus à l'un des 4 groupes finaux issus de la CAH (**Vidéo 2, 30:44**). Quels sont les effectifs par groupe ? (36223, 6025, 12215, 45537)
48. Dans quelle mesure la partition issue de la première stratégie (K-Means + CAH) est-elle cohérente avec celle de la seconde (SOM + CAH) ?

	cluster_som			
cluster_final	1	2	3	4
1	3559	2	5785	41035
2	3534	5573	582	3761
3	29130	450	1	633
4	0	0	5847	108

5. Mesures d'évaluation des partitions

Pour aller plus loin : Laquelle des 2 partitions est la meilleure ? Il faudrait les confronter avec les vraies classes d'appartenance (7 catégories) recensées dans la feuille « étiquettes » de notre classeur, en utilisant les mesures externes d'évaluation des partitions par exemple (v de Cramer, Information Mutuelle normalisée, Indice de Rand, Homogénéité, Complétude, v-Mesure…. Voir https://www.youtube.com/watch?v=I6__VDxBH_M qui montre comment les calculer sous Python / Scikit-Learn). Une stratégie possible serait d'exporter les vecteurs d'affectation aux groupes générées par les deux approches dans un fichier Excel. Importez les sous Python ainsi que les « vraies » classes d'appartenance. Mettez en œuvre alors les commandes décrites dans la vidéo. Finalement, quelle approche fournit les meilleurs résultats ?