TP 1: Quelques manipulations élémentaires autour de l'inertie (des peuplements forestiers)

NB. Pour ceux d'entre vous qui ne sont pas encore familiers avec R, la première partie de ce premier TP est entièrement faisable en Python, et même... sur tableur. Prenez soin de visualiser la matrice des données après chaque transformation, ainsi que toute quantité ou vecteur calculé, et d'exercer votre esprit critique devant chacun de ces objets.

Partie I

Charger dans le logiciel les données relatives au peuplement arboré de la forêt du bassin du Congo (**Datagenus.csv**). Inspectez le fichier (attention: *une* anomalie ruine *toute* l'analyse !).

Ces données fournissent sur 1000 parcelles de cette forêt: les variables de comptage de 27 espèces d'arbres (*gen1*, ..., *gen27*), la surface de la parcelle, le type forestier (*forest*) et le type géologique (*geology*) tels qu'identifiés par les écologues . On ne tiendra pas compte des autres variables.

Les parcelles seront traduits en nuage dans l'espace des 27 densités de peuplement, $\mathbb{R}^{27}$.

1) Calculer la densité de peuplement de chaque espèce par unité de surface pour les 1000 parcelles. Justifiez l'emploi de cette variable au lieu de la variable de comptage originelle pour toutes les analyses.

Centrer-réduire les 27 densités de peuplement. Montrez théoriquement, puis vérifiez informatiquement qu'alors:

- le barycentre du nuage se trouve à l'origine;
- l'inertie totale du nuage est égale au nombre des variables (27).

2) Calculer les poids et les barycentres des sept types forestiers. Puis, calculer les normes euclidiennes carrées de ces sept barycentres.

En déduire l'inertie inter-types forestiers, puis le R^2 de la partition des parcelles en types forestiers.

Quel est le pourcentage d'information (variabilité du peuplement) dont cette partition rend compte?

3) On voudrait savoir quelles sont les espèces qui sont les plus liées au type. Calculez le $R^2 = \frac{variance\ inter-types\ forestiers}{variance\ totale}$ de chaque variable densité de peuplement. Quelles sont les densités qui sont les plus (respectivement les moins) liées au type?

Montrez mathématiquement, puis vérifiez informatiquement, que le R^2 de la partition est égal à la moyenne arithmétique des R^2 des variables.

Partie II

*NB. Cette partie est à faire en R, et l'interface **R-studio** de R est fortement conseillée!*

On notera X la matrice dont les colonnes sont les 27 densités centrées-réduites, Y la matrice dont les colonnes sont les indicatrices de types forestiers, et Z celle dont les colonnes sont les indicatrices de sols (*geology*).

On notera $W=\frac{1}{n}I_n$ la matrice des poids des individus et $M=\frac{1}{p}I_p$ celle des poids des variables. Bref, ici, tout est équipondéré.

1/ a) Rappeler pourquoi $\forall j, \Pi_Y x^j=\Pi_{Y^c} x^j$. Rappeler l'interprétation statistique de $\|\Pi_Y x^j\|_W^2$.

b) Programmer et calculer Π_Y , puis, pour chaque x^j : Π_{x^j} , $tr(\Pi_{x^j}\Pi_Y)$. Rappelez l'interprétation statistique de cette dernière quantité.

c) On note $R=XMX'W$. Programmer et calculer $tr(R\Pi_Y)$. Interprétez statistiquement cette quantité.

d) Rapprochez les résultats obtenus de ceux de la première partie.

2/ Programmez puis calculer chaque $tr(\Pi_{x^j}\Pi_Z)$, $tr(R\Pi_Z)$, et interprétez ces résultats statistiquement.