# TP 1: Quelques manipulations élémentaires autour de l'inertie (des vins de Loire)

*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 aux vins de Loire (**wine.csv**).

Elles contiennent deux variables qualitatives (Appellation: "*Label*" = {*Bourgueil, Chinon, Saumur*} et Sol: "*Soil*" = {*Env1, Env2, Env3=référence, Env4*}) et 29 variables quantitatives décrivant diverses intensités sensorielles (odeur, arôme, goût, couleur etc).

Les vins seront traduits en nuage dans l'espace des 29 variables quantitatives, $\mathbb{R}^{29}$ .

1) Centrer-réduire les variables quantitatives. 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 (29).

2) Calculer les poids et les barycentres des trois appellations (Bourgueil, Chinon, Saumur). Puis, calculer les normes euclidiennes carrées de ces trois barycentres.

En déduire l'inertie inter-appellations, puis le $R^2$ de la partition des vins en appellations.

En déduire que l'appellation n'explique qu'environ 11% des disparités sensorielles entre les vins de Loire.

3) On voudrait savoir quelles sont les variables qui sont les plus liées à l'appellation. Calculez l e $R^2 = \frac{variance\, inter-appellation}{variance\, totale}$ de chaque variable sensorielle. Quelles sont les variables qui sont les plus (respectivement les moins) liées à l'appellation?

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 29 variables quantitatives centrées-réduites, $Y$ la matrice dont les colonnes sont les indicatrices d'appellations, et $Z$ celle dont les colonnes sont les indicatrices de sols.

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 $\Pi_Y$ , puis, pour chaque $x^j$ : $\Pi_{x^j}$ , $tr(\Pi_{x^j}\Pi_Y)$ . 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.