Analyse Statistique Multidimensionnelle – TP

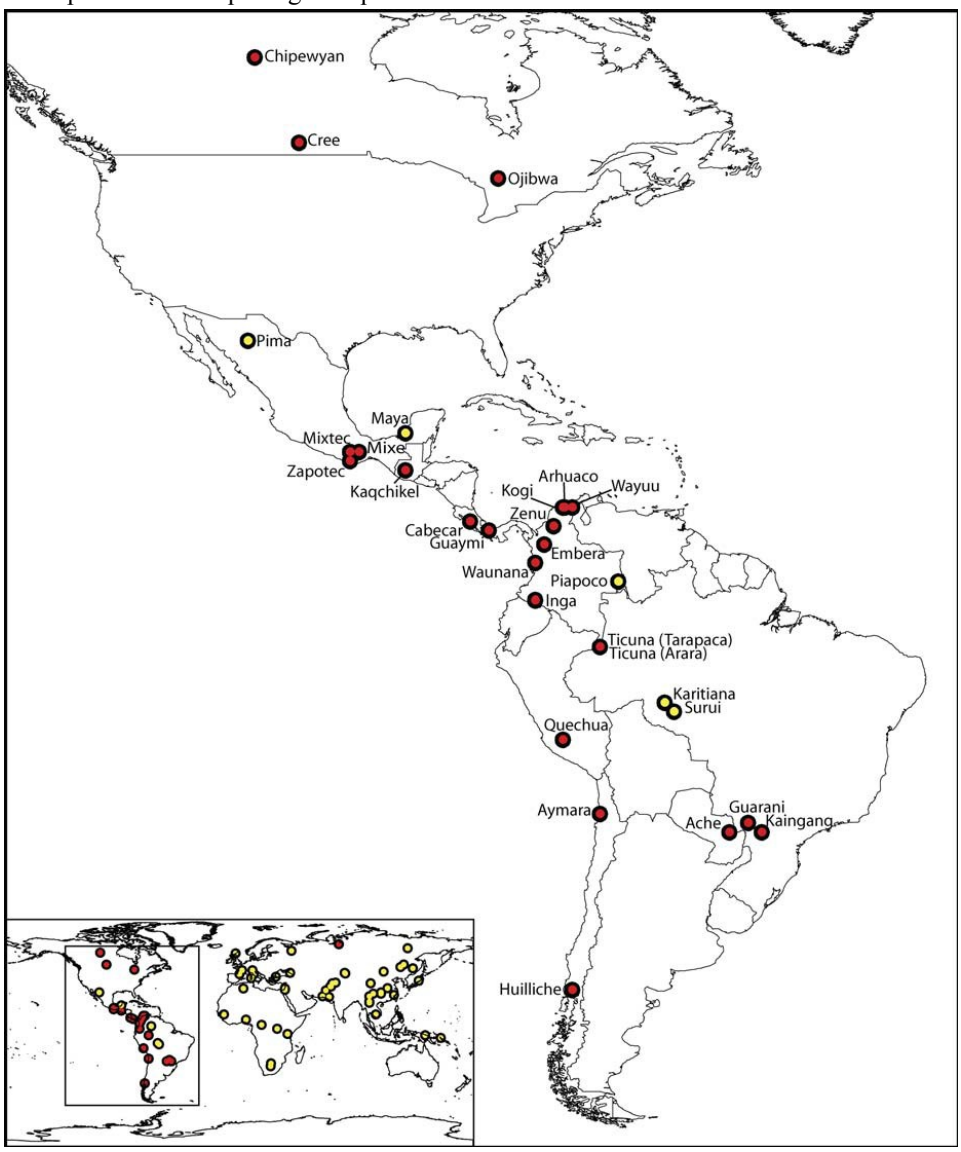
Flora Jay flora.jay@lri.fr

Les calculs, les graphiques et la programmation se feront avec le logiciel **R**.
La représentation graphique est très importante en statistique. Chaque graphe doit être accompagné au minimum d'un titre et de noms d'axes. Il doit toujours être expliqué par quelques phrases dans le texte (Qu'a-t-on tracé ? Qu'en déduire ?).

TP : Différenciation génétique.

1. Préparation données

a) Chargez le fichier Nam.txt avec `read.table(.., header=T,stringsAsFactor=F)`, appelez le `NAm`.
Chaque ligne correspond à un individu.
Vérifiez que les colonnes aient des noms explicites. La colonne 3 contient la population d'origine de l'individu. Les colonnes 7 et 8 contiennent la latitude et la longitude. Chaque colonne à partir de la 9ième correspond à un marqueur génétique.



b)
```
names=unique(NAm$population.name)
npop=length(names)
coord=unique(NAm[,c("population.name","long","lat")]) #coordinates for each pop

colPalette=rep(c("black","red","cyan","orange","brown","blue","pink","purple","
darkgreen"),3)
pch=rep(c(16,15,25),each=9)

plot(coord[,c("long","lat")],pch=pch,col=colPalette,asp=1)
# asp permet d'avoir le rapport correct entre les axes  longitude et latitude
# ainsi la carte n'est pas déformée
legend("bottomleft",names,col=colPalette,lty=-1,pch=pch,cex=.75,ncol=2,lwd=2)
library(maps);map("world",add=T)
```

Rq : la dernière ligne fonctionne uniquement si vous pouvez installer le package maps.
Que fait ce script ? Vérifiez que cela correspond à la carte ci-dessus.

2. **Régression**
En utilisant comme prédicteurs les 678 marqueurs génétiques régressez la latitude. Quel est le problème ?

3. ACP

a) Rappelez brièvement le principe d'une ACP.

b) Réalisez une ACP sur les données génétiques (attention données génétiques uniquement !) des individus issus des populations Kogi, Waunana, Arhuaco, Quechua et Ache. Qui y a-t-il dans l'attribut x ? Et dans l'attribut `rotation ?`
Aide : Pour sélectionner les individus vous pouvez utiliser la fonction `%in%` **:**
`NAm[,3] %in% c("Kogi","Waunana","Arhuaco","Quechua","Ache")`

c) Utilisez la fonction biplot pour observer les résultats de l'ACP, expliquez ce qui est tracé. Vous pouvez tracer uniquement les valeurs de l'attribut x pour y voir plus clair :
```
plot(resultat$x[,1:2])
text(resultat$x[,1:2],rownames(resultat$x))
```
Interprétez les résultats, le rôle des différentes composantes principales (PC), etc. Tracer les 3ème et 4ème PC pour voir si vous captez un autre signal. Quel pourcentage de variance est expliqué par les 2 premières PC ?

d) Réalisez maintenant une ACP sur les données génétiques de tous les individus. Stockez le résultat dans l'objet `pcaNAm`. *Pensez aux arguments scale et center de l'ACP.*
Ce script vous permettra d'y voir plus clair :
```
caxes=c(1,2)
plot(pcaNAm$x[,caxes],col="white")
for (i in 1:npop) {
       browser()
       print(names[i])
       lines(pcaNAm$x[which(NAm[,3]==names[i]),caxes],
       type="p",col=colPalette[i],pch=pch[i])
}
legend("top",legend=names,col=colPalette,lty=-1,pch=pch,cex=.75,ncol=3,lwd=2)
```
Expliquez.

e) Quel pourcentage de variance est expliqué par les 2 premières PC ?

4. PCR Principal Components Regression

a) Régressez la latitude puis la longitude en utilisant comme prédicteurs les 250 premiers axes de l'ACP.

b) Affichez sur un graphe les coordonnées spatiales prédites (comme dans 3d rendez distinguables les populations d'origine). Que constatez-vous ?

c) On choisit comme erreur la distance moyenne entre les vraies coordonnées (des populations d'origine) et les coordonnées prédites (attention, utilisez la « distance du grand cercle » pour cela). Calculez l'erreur pour le modèle précédent (avec les 250 axes).
Aide : `??rdist.earth` (pensez à l'option `miles=F`).

5. PCR et validation croisée

Notre but est de construire le meilleur modèle prédictif. Pour sélectionner le nombre d'axes à conserver (`naxes`), on va appliquer une méthode de validation croisée 10-fold.

a) Rappelez brièvement le principe de la méthode de validation croisée 10-fold. Expliquer pourquoi il est bien d'utiliser cette méthode pour construire un modèle prédictif.

Les données doivent être « partagées » en 10 jeux qui serviront tour à tour de jeu de validation. Construisez un vecteur `set` de taille égale au nombre d'individus dans lequel vous stockerez l'indice du jeu de validation auquel l'individu appartient.

Exemple pour 9 individus et une validation croisée 3-fold :
`set = c(1,2,3,1,2,3,1,2,3)` ou bien `set=c(1,3,1,3,3,2,1,2,2)`
Vous pouvez construire ce vecteur aléatoirement en étant sûr d'avoir le même nombre d'individus dans chaque jeu de validation :

```
labels=rep(1:3,each=3)
set=sample(labels,9)
```

b) On étudiera les modèles avec naxes variant de 2 au maximum possible (de 10 en 10 par exemple). Mais commençons d'abord avec `naxes=4`.

i. Construire une matrice vide predictedCoord à 2 colonnes ("longitude", "latitude") et autant de lignes que d'individus.

ii. Avec comme prédicteurs les axes 1 à 4 de l'ACP, régressez les variables latitude et longitude en utilisant les individus n'étant pas dans le jeu de validation n°1

```
Aide : Vous devez régresser Nam[set!=1,]$lat par ...
Attention, pcaNAm$x devra sûrement être converti en un objet de classe
data.frame avec as.data.frame
```

iii. A partir des régressions de ii. prédire la latitude et la longitude des individus **du jeu de validation n°1.** Stockez les coordonnées prédites dans `predictCoord` (dans les lignes correspondant aux index des individus, pour pouvoir comparer aux vraies coordonnées ensuite).

iv. Recommencez pour les jeux de validation de 1 à 10. A la fin la matrice `predictCoord` doit être remplie.
Calculez l'erreur de prédiction. Utilisez pour cela la distance entre vraies coordonnées et coordonnées prédites (cf. 4.c).

c) Refaire les étapes du b) en faisant varier `naxes` de 2 au maximum possible. Tracer l'erreur de prédiction et l'erreur d'apprentissage.
Aide : `seq(2,ncol(pcaNAm$x),by=10)`

d) Quel modèle sélectionnez-vous ? Quelle est l'erreur de prédiction pour ce modèle ? Quelle est l'erreur en longitude et en latitude séparément ? Affichez les coordonnées prédites sur la carte.

6. **Résumez/Concluez** (on attend de vous au moins une tentative d'interprétation, pas seulement des lignes de R)