Université d'Orléans 2021–2022
CoST Miage M1

Analyse de données – TP n° 4
ACP avec individu(s) supplémentaires
données du "Canidé de Jussac"

Il s'agit de données réelles d'origine paléontologiques. Le crâne d'un animal préhistorique appartenant à la famille des canidés a été découvert il y a quelques années, dans la région de Jussac (Auvergne).

La question que se posait les scientifiques était de savoir si cet animal se rapprochait plus d'un chien ou d'un loup.

On a donc mesuré six grandeurs caractéristiques sur des crânes chiens de même taille que celle de l'animal inconnu (berger allemand, lévrier, doberman,...), et sur des crânes de loups. Les variables mesurées sont (cf. Fig.1) :

— X_1 : longueur condylo-basale (LCB)
— X_2 : longueur de la mâchoire supérieure (LMS)
— X_3 : largeur bi-maxilaire (LBM)
— X_4 : longueur de la carnassière supérieure (LP)
— X_5 : longueur de la première molaire supérieure (LM)
— X_6 : largeur de la première molaire supérieure (LAM)

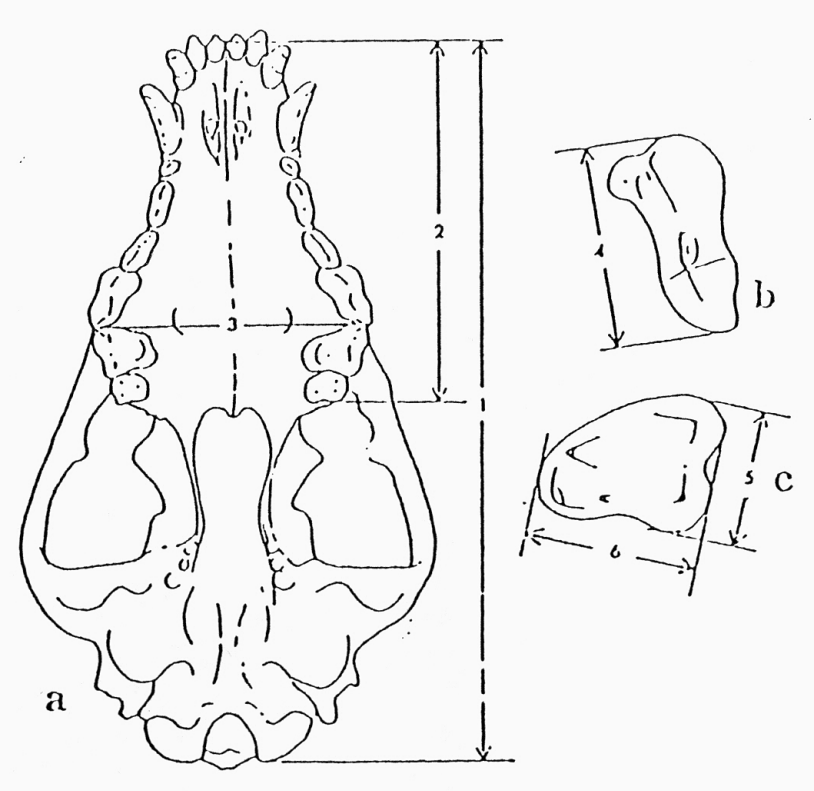

FIGURE 1 – *Signification des variables.*

Le fichier des données `Jussac.txt` contient, en plus des 6 variables ci-dessus, une colonne d'indices, `no` inutile, une variable `Type` donnant l'identifiant (chien, loup, Jussac), et une colonne précisant en plus la race des chiens qui pourra être utilisée plus tard éventuellement.

Méthodologie : ACP avec individu(s) supplémentaire(s)

L'objectif est :

1. d'analyser ce qui distingue les populations de chiens et de loups au mieux
2. de voir si l'individu de race inconnu est plutôt dans la population des chiens ou dans celle des loups.

Il existe plusieurs approches possibles en data mining (analyse discriminante, clustering supervisé,...). Mais on se propose ici de répondre à la question (1) par une ACP, en espérant que le plan principal (ou un autre) sépare bien les populations chiens/loups.

Dans ce contexte, l'individu `Jussac` a un statut particulier : **il ne doit pas participer à la construction des axes**, mais doit être projeté *a posteriori* sur les plans individus. C'est la notion d'individu supplémentaire(s) vue en cours.

Questions

(1) Chargez les données dans R, puis séparez la table en deux sous-tables : celle des individus actifs, et celle des individus supplémentaires (réduite à 1 individu ici, mais on souhaite un programme le plus générique possible). Avec le fichier décrit ci-dessus, il faudra manipuler un peu sous R afin de construire les `data.frames` adaptés.

(2) Analyse préliminaire de la table `actifs`, distributions, etc., puis choix de métrique : l'ACP doit-elle être normée ou non ?

(3) Adaptez le programme R d'ACP du TP n°3 pour réaliser l'ACP des individus actifs.

(4) Interprétez les résultats de cette ACP, en lien avec l'objectif 1.

(5) Lire l'aide `?suprow` puis utilisez cette fonction pour calculer les coordonnées des individus suppplémentaires (1 seul ici).

(6) Lire l'aide `?s.label` puis utilisez cette fonction pour projeter l' individu suppplémentaire sur le plan principal. Interprétez les résultats de l'analyse.

(7) Compléments : étude avec les races de chiens, comparaison des métriques...