

ASM-TP2 : ACP-régression en génétique

Pierre Petitbon

Florian Privé

Xinrui Xu

1 Préparation des données

1.a) OK

1.b) Le script place toutes les origines des individus des données sur une carte du continent américain, suivant la latitude et la longitude (figure 1). La carte obtenue correspond bien à celle du sujet.

2 Régression

La régression n'est pas effectuée par R : les coefficients $(\hat{\beta}_j)_{0 \leq j \leq p}$ renvoyés valent *NA* (Not available). Le problème est la non-unicité de la régression.

Lorsque $n \geq p + 1$, alors $X^T X$ est inversible, et l'unique solution de la régression est :

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Par contre, ici, $n < p + 1$. D'où :

$$\begin{aligned} \text{rang}(X^T X) &\leq \text{rang}(X) \\ &\leq n \\ &< p + 1 \end{aligned}$$

$X^T X$ est une matrice carrée de taille $p + 1$, de rang strictement inférieur à $p + 1$. Donc $X^T X$ n'est pas inversible, et la solution de la régression n'est pas unique.

3 ACP

3.a) Le principe de l'Analyse en Composantes Principales est de tirer parti de la corrélation entre les variables du problème pour expliciter un nombre réduit de nouvelles variables, combinaison linéaires des anciennes, qui permettent de modéliser correctement la variance de l'échantillon. Son intérêt majeur est donc de limiter le nombre de variables à étudier. Ici, réaliser une ACP permet ensuite de régresser les coordonnées géographiques.

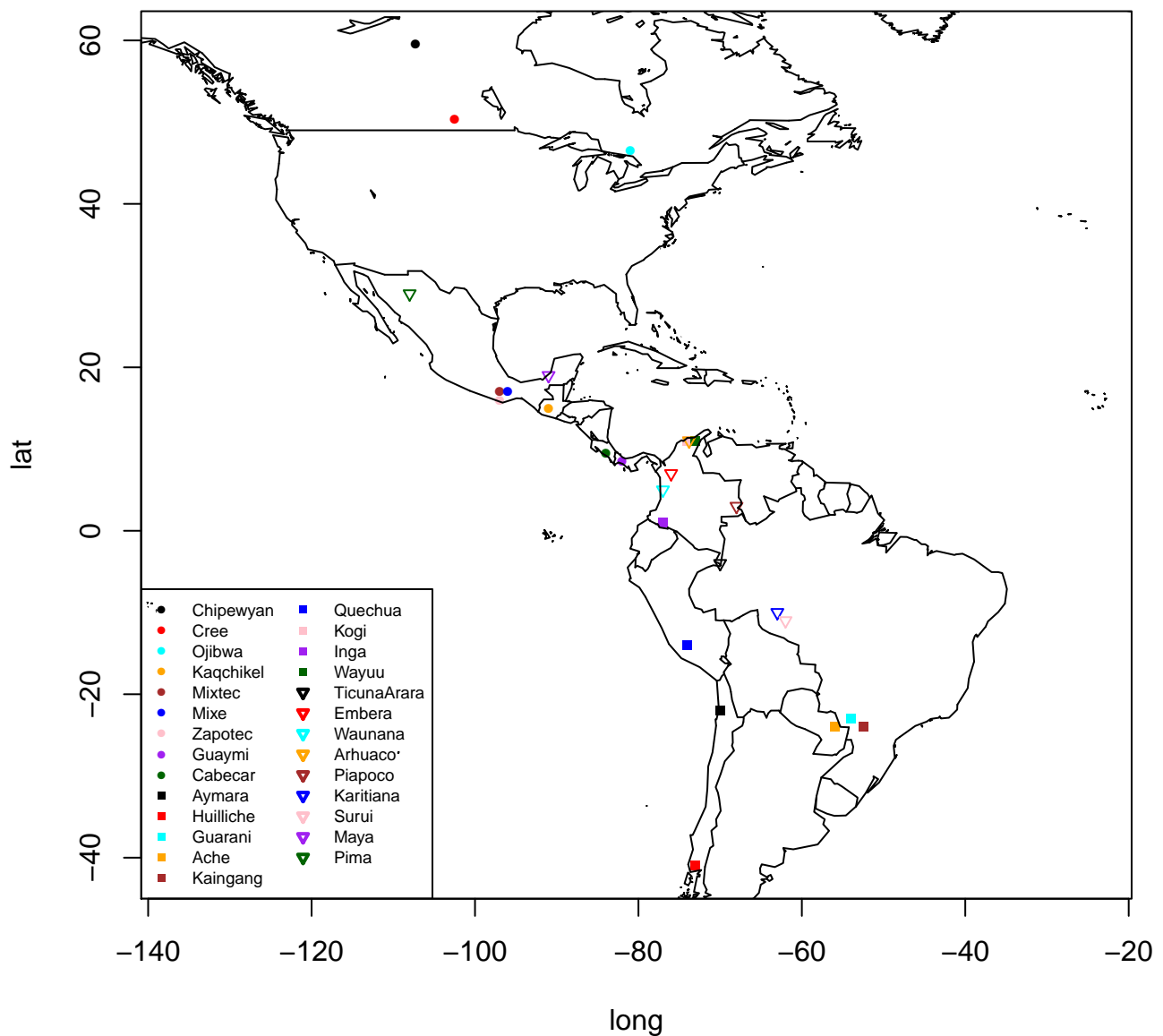


FIGURE 1 – Origine géographique des indiens d'Amérique

3.b) On réalise une ACP sur les données génétiques avec tous les individus. On n'a pas besoin d'utiliser l'argument scale car les variables (les marqueurs génétiques) n'ont pas d'unités de mesures : les valeurs sont binaires. Il n'y a donc pas besoin de mettre ces variables sur une même échelle.

En considérant les 2 premiers axes de l'ACP, les individus qui sont facilement identifiables sont les Surui et les Aches. En effet, ils sont relativement éloignés de l'origine, et les vecteurs associés ne sont pas colinéaires et de même sens (cf. figure 2). Cependant, en observant la figure 1, on s'aperçoit que les localisations associées sont relativement

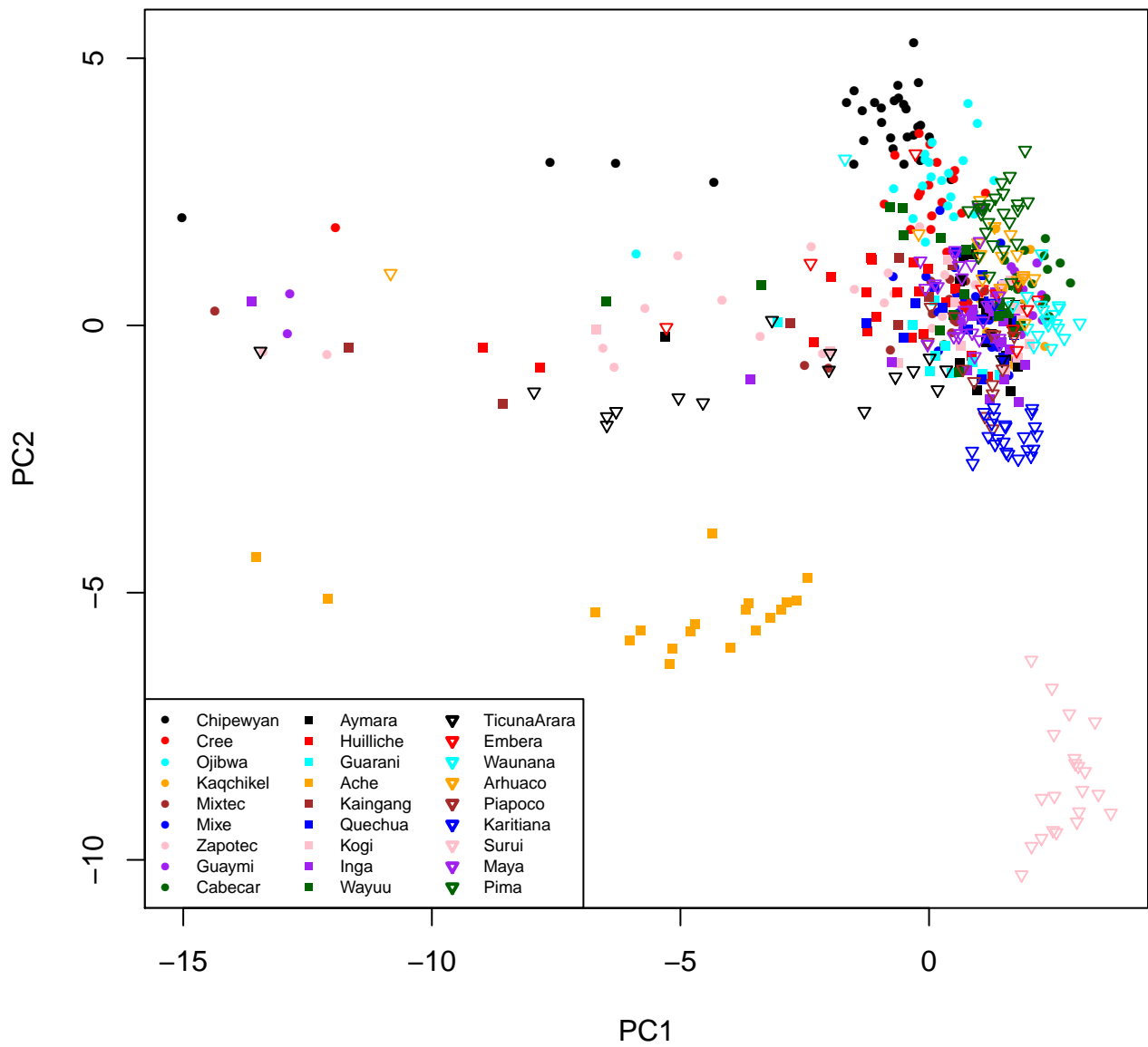


FIGURE 2 – Les populations projetées sur les 2 premières PC

proches. Ainsi, les Surui et les Aches se différencient bien génétiquement, mais ces différences ne sont pas corrélées à une origine géographique différente. En régressant les coordonnées géographiques selon les axes de l'ACP, on s'aperçoit effectivement que le premier axe est indépendant de la localisation (cf. question 4.a)).

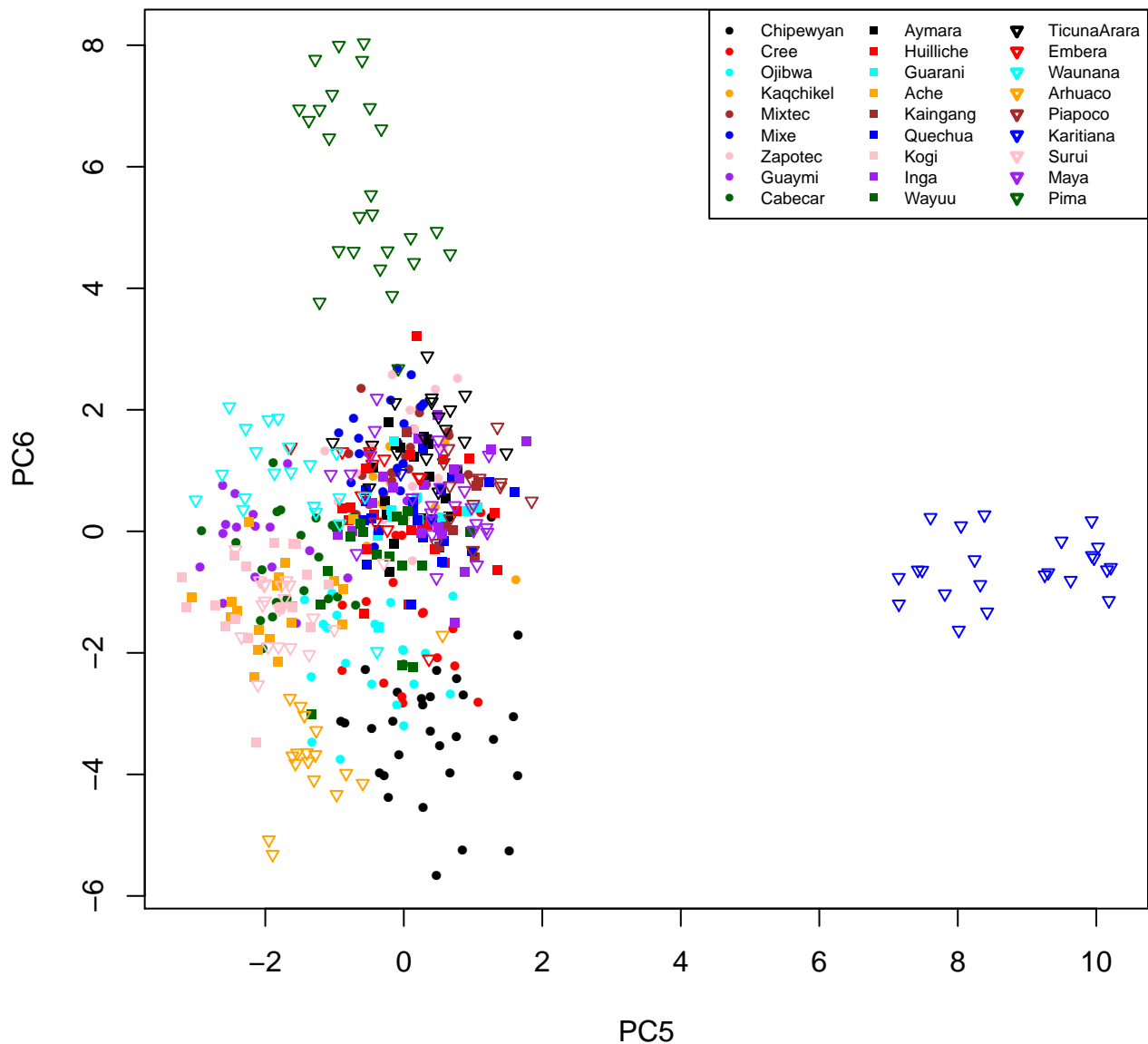


FIGURE 3 – Les populations projetées sur les PC 5 et 6

Avec les axes 5 et 6, ce sont les Karitiana et les Pima qui sont facilement identifiables, d'autant plus que chacun des vecteurs associés est aligné sur un axe différent (cf. figure 3).

- 3.c) Le pourcentage de variance cumulée expliquée par les 2 premières PC est de 3,57%. Ce pourcentage paraît relativement faible, mais il ne faut pas oublier que l'on part d'un très grand nombre de variables ($p = 5709$ marqueurs génétiques). En prenant deux variables de départ quelconques, on expliquerait en moyenne $2/p = 0.035\%$ de variance. On a donc réussi à expliquer 100 fois plus de variance.

En gardant les 250 premiers axes de l'ACP, on exprime 77% de la variance du jeu de données, en divisant par 20 le nombre de variables.

4 PCR Principal Components Regression

4.a) On régresse la latitude et la longitude en utilisant comme prédicteurs les scores des 250 premiers axes de l'ACP. On observe que le premier axe n'influe pas sur la localisation en dépit du fait qu'il maximise la variance des variables. Dans les deux régressions, la p-valeur associée est très grande (respectivement 44% et 23% pour la régression de la longitude et de la latitude) par rapport aux p-valeurs des autres axes. Donc il y a de forte chance que le coefficient de régression associé au premier axe soit nul.

4.b) On affiche sur une carte les coordonnées spatiales régressées pour chaque individu (cf. figure 4). Cette carte donne une image trop optimiste de la capacité à retrouver l'origine géographique d'un individu à partir de ses marqueurs génétiques. En effet, on a considéré uniquement l'erreur de régression, et pas l'erreur de prédiction. On fait donc face à un problème d'overfitting : on cherche trop à "coller" aux données, autrement dit on régresse aussi le bruit.

Par ailleurs, il est illusoire de croire que tous les marqueurs génétiques (ou tous les axes de l'ACP) sont corrélés à l'origine géographique. On l'a vu : le premier axe de l'ACP est indépendant de l'origine géographique.

4.c) Pour le modèle précédent (250 axes), l'erreur moyenne en termes de distance est de 641 km. Etant donnée la complexité de l'histoire humaine (migrations, brassages génétiques), et à l'échelle du continent américain, c'est une erreur relativement faible. Mais n'oublions pas que le problème d'overfitting n'est pas encore réglé, et que l'erreur finale sera donc plus grande.

Ce résultat nous permet néanmoins d'affirmer qu'il n'est en général pas possible de prédire très précisément l'origine géographique d'un individu à partir de ses marqueurs génétiques. Sur la carte (figure 4), on remarque en particulier que les coordonnées régressées des individus en provenance de tribus d'Amérique Centrale sont fortement mélangées, on ne peut donc pas reconstituer les tribus d'origine de ces individus.

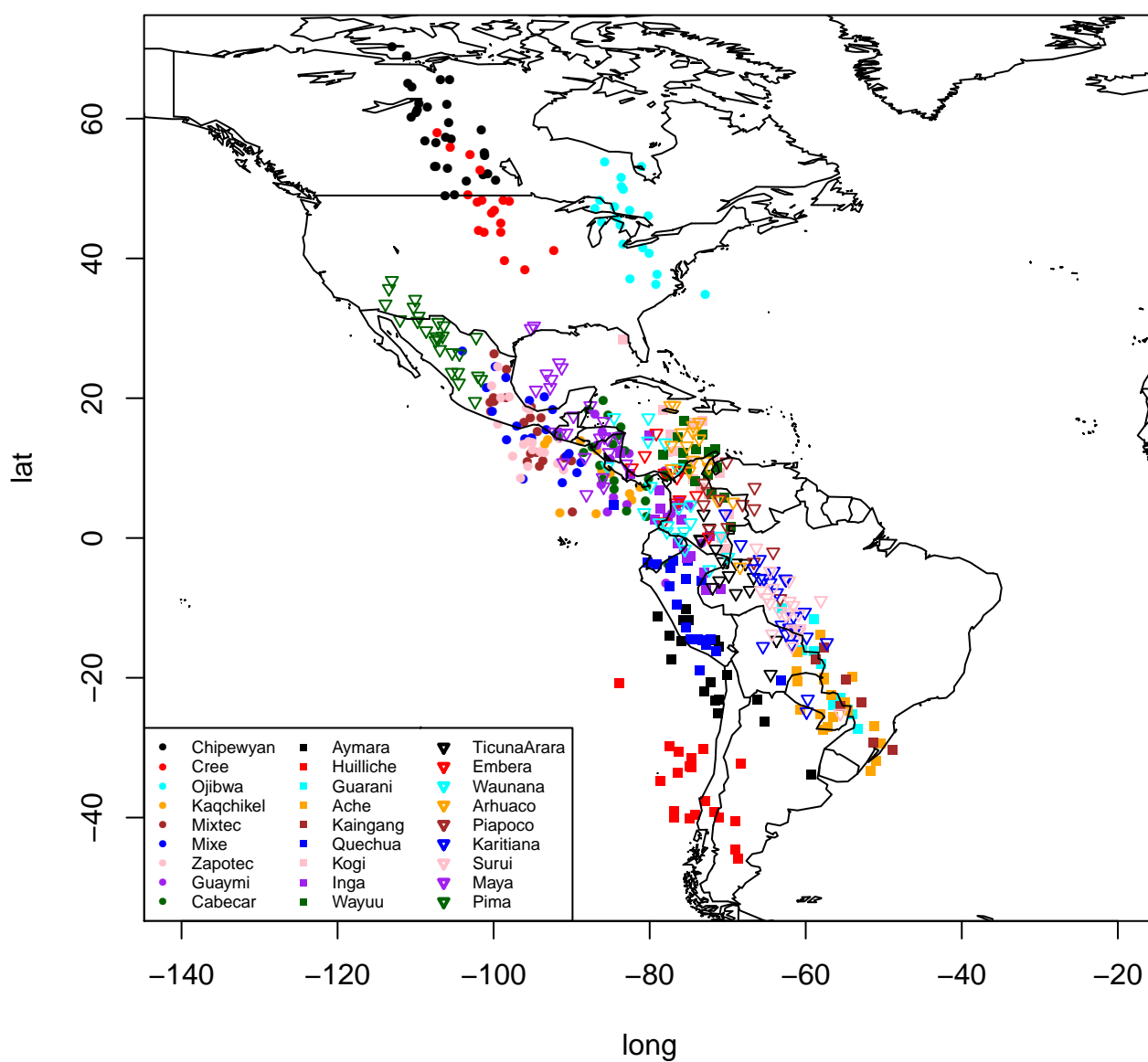


FIGURE 4 – Régression des coordonnées spatiales avec les 250 premiers axes de l'ACP