

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.

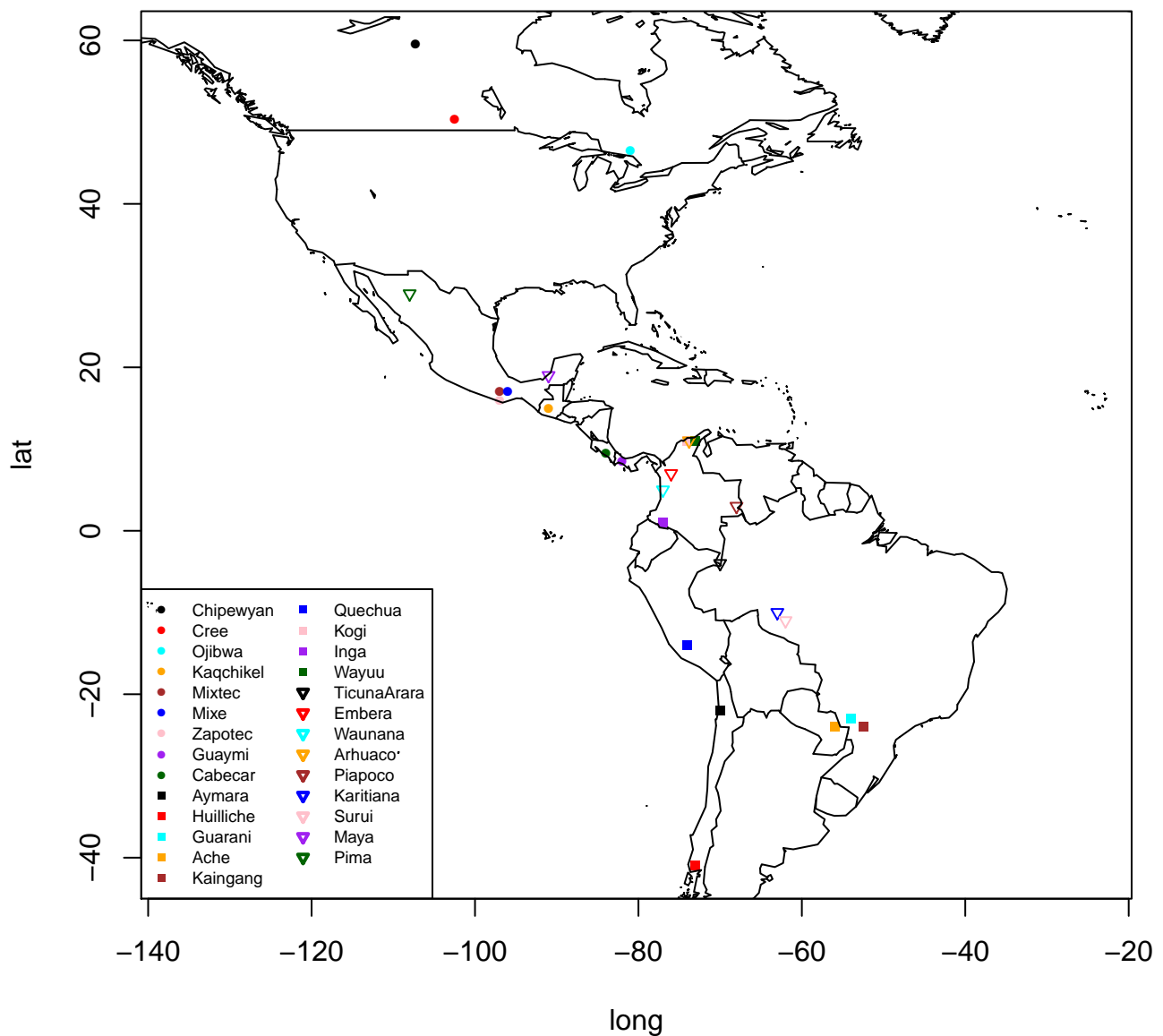


FIGURE 1 – Origine géographique des indiens d'Amérique

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.

Avec les 2 premiers axes de l'ACP, les individus qui sont facilement identifiables sont les Surui et les Aches.

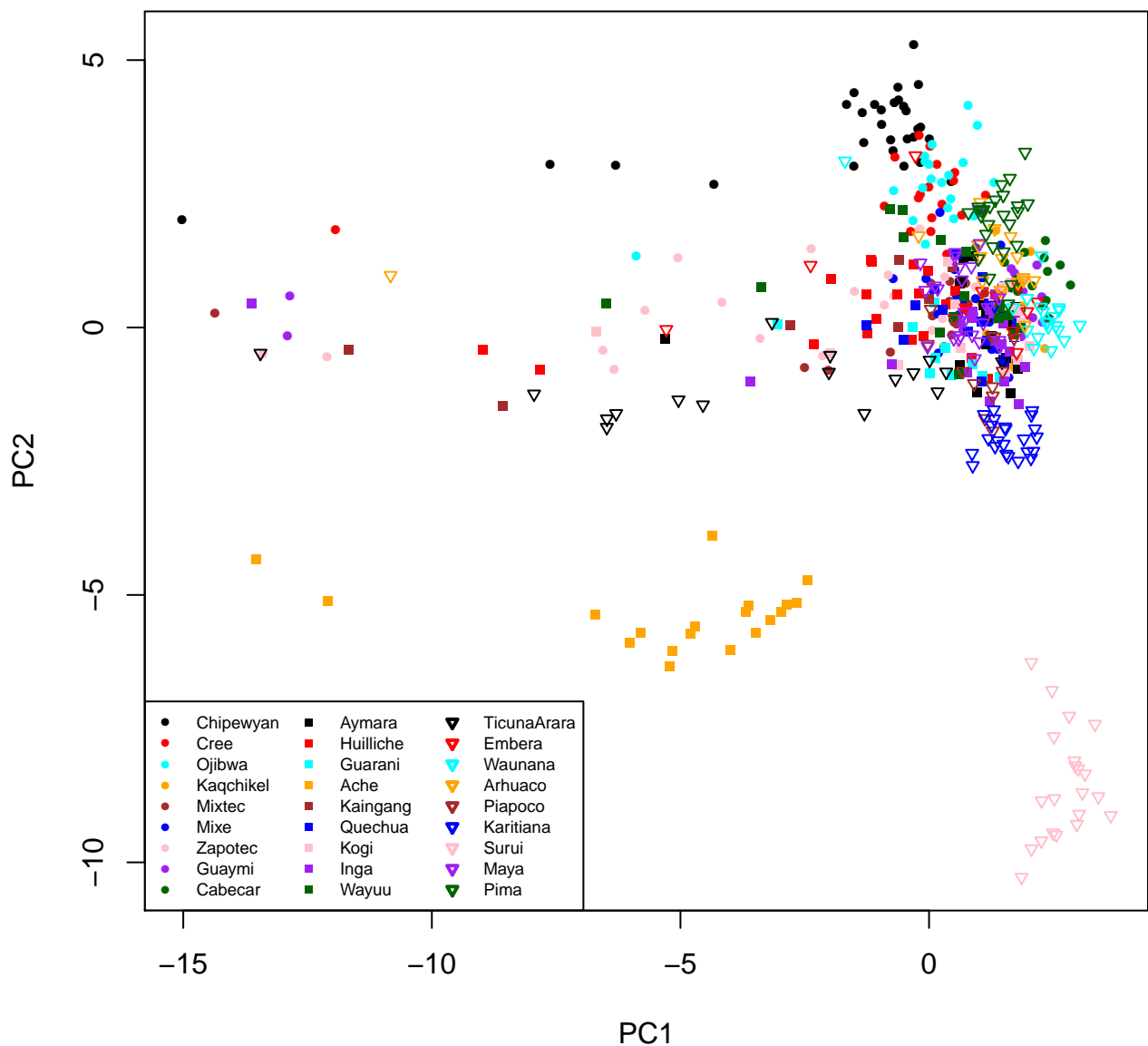


FIGURE 2 – Les populations projetées sur les 2 premières PC

Avec les axes 5 et 6, ce sont les Karitiana et les Pima qui sont facilement identifiables.

3.c) Le pourcentage cumulé de variance expliqué par les 2 premières PC est de 3,57%. Elle est très faible donc l'ACP n'est pas réalisable : en retenant que les 2 premières PC, on ne peut pas bien expliquer le modèle. Pour avoir un pourcentage de variance correct (70%) on va devoir garder 210 PC, ainsi on perdrait tout l'intérêt d'un ACP.

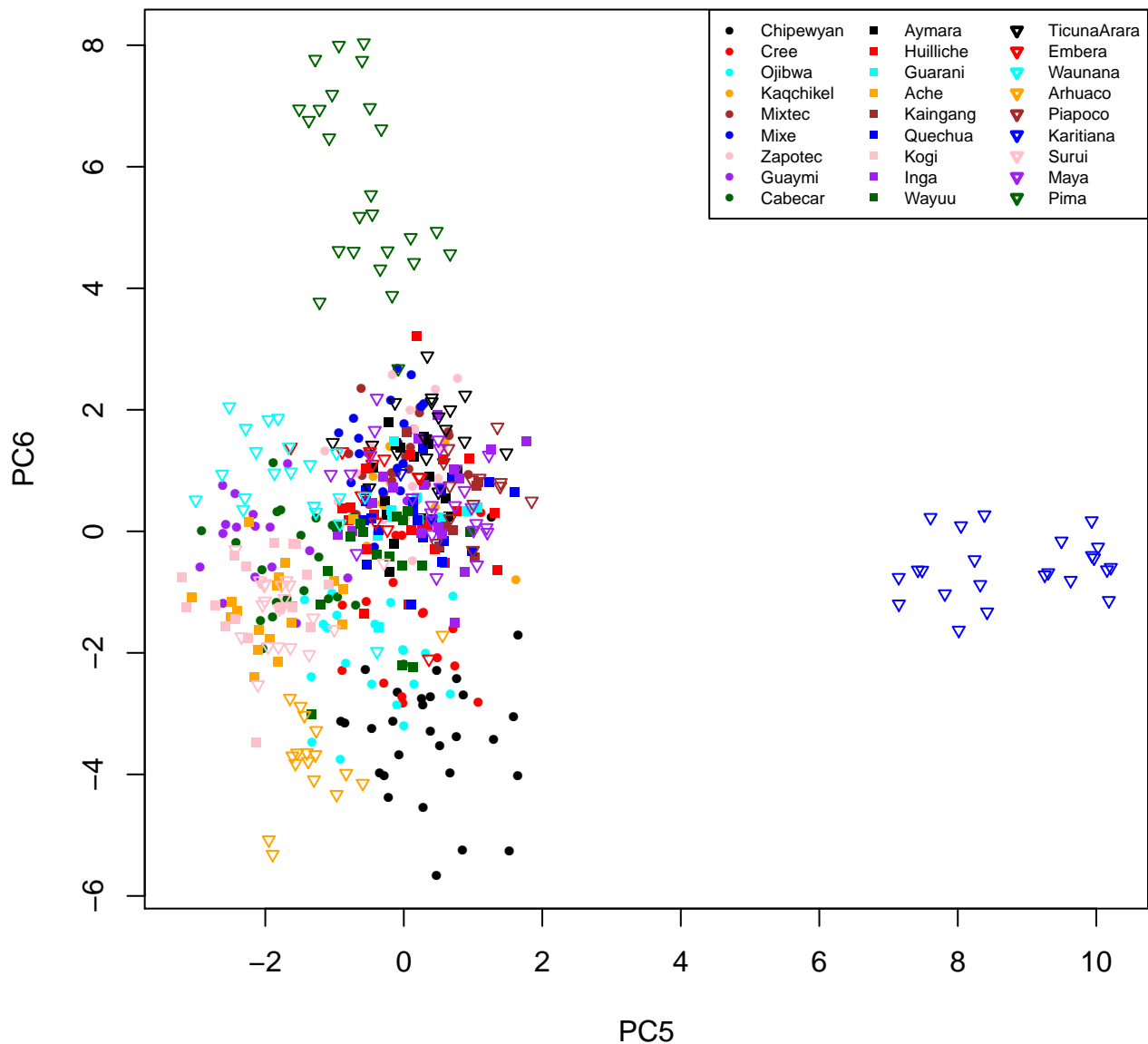


FIGURE 3 – Les populations projetées sur les PC 5 et 6

4 PCR Principal Components Regression

- 4.a) La régression de la latitude et de la longitude en utilisant comme prédicteurs les scores des 250 premiers axes de l'ACP nous donne que l'ACP1 n'influe pas sur la localisation même si elle maximise la variance des variables. Dans les deux régressions, la p-valeur de l'ACP1 est très grande (respectivement 0.44% et 0.23% pour la régression de la longitude et de la latitude) par rapport aux autres p-valeurs des autres ACP. Donc il y a de forte chance que le coefficient de régression associé à ACP1 soit nul.

4.b) Les coordonnées spatiales prédites pour chaque population sont localisées dans la même zone que les coordonnées réelles de chaque population. Cependant, on observe des erreurs : les coordonnées prédites peuvent être parfois très dispersées (comme les Huilliches). De plus, à cause de cette dispersion, on a beaucoup de mal à distinguer les populations qui sont situées très près les unes des autres (comme toutes celles d'Amérique centrale) : les points prédits se recoupent trop.

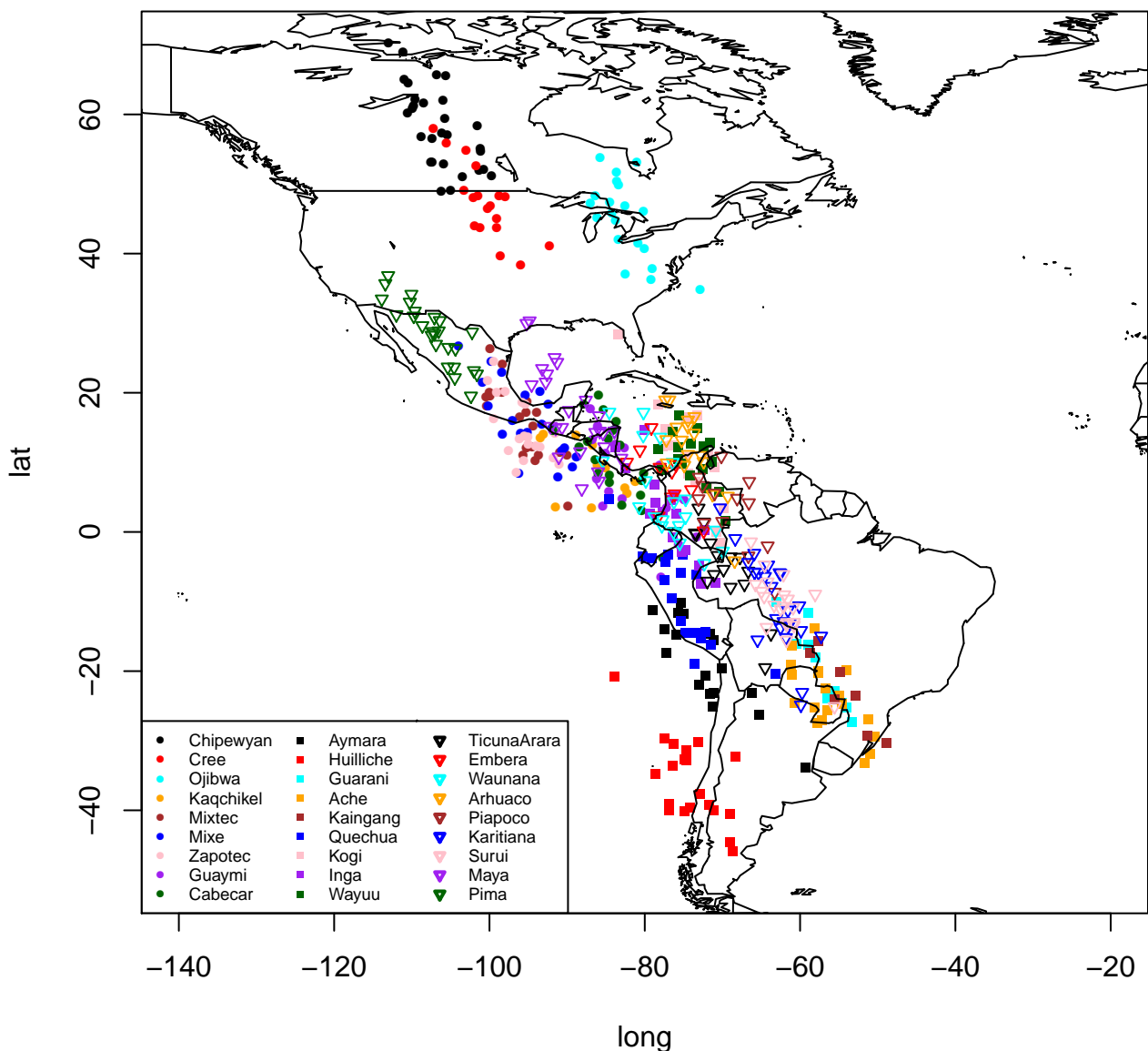


FIGURE 4 – Régression des coordonnées spatiales avec les 250 premiers axes de l'ACP

4.c) Au final sur tous les écarts entre les valeurs prédites et valeurs réelles mesurés avec la

distance du grand cercle on obtient 641km. Cette valeur a été obtenue en prenant la moyenne des écarts pour chaque population. C'est une valeur à relativiser. A l'échelle de l'Amérique ce n'est pas très grand et ce n'est pas si mal. Pour certaines populations cela se voit d'ailleurs, mais dès que les populations sont plus proche que cette distance on peine à les distinguer et donc notre modèle prédictif n'est pas très concluant pour ces populations.