université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

Inférence démographique - Approximate Bayesian Computation

Gustavo Magaña López Théo Roncalli

20/12/2021

Contents

1 Introduction

Dans le cas présent, un jeu de données nous est fourni sur le chromosome 22 pour 54 individus, c'est-à-dire pour 108 haploïdes. Celui-ci provient du Panel Diversity de Complete Genomics. Nous avons deux fichiers : le premier renvoie diverses informations sur chaque haploïde (numéro du chromosome, id du snp, position sur le chromosome de référence, nucléotide de référence, variant, et le nucléotide effectivement observé pour chaque haploïde. Le second fichier renvoie des métadonnées sur chaque haploïde. Nous avons l'identifiant de l'individu, le code renseignant sur la population à laquelle il appartient et sa région de provenance. Parmi les populations recensées, nous en avons plusieurs, comme les portoricains (PUR), les Utah d'ascendance européenne du nord et de l'ouest (CEU), les Yoruba (YRI), les chinois Han (CHB), etc. Les scientifiques considèrent les CEU comme Européens alors que ces individus habitent aux Etats-Unis car il y a eu une migration des Européens de l'Ouest vers les Amériques, ce qui implique un bottleneck. Comme c'est une communauté très fermée, on le considère comme un échantillon européen étant donné qu'il n'y a pas d'échange de matériel génétique avec d'autres populations américaines.

Afin de reconstruire les tailles de population et histoires démographiques pour l'espèce humaine, nous pouvons commencer par étudier certaines statistiques telles que le D de Tajima:

$$D = \frac{\Theta_T - \Theta_W}{\sqrt{V(\Theta_T - \Theta_W)}}$$

La **figure 1** fournit le D de Tajima pour chaque population recensée. Un D de Tajima positif signifie qu'il y a un excès d'allèles en fréquence intermédiaire, alors qu'un D de Tajima négatif signifie qu'il y a plutôt de nombreux singletons. Dans le cas présent, on remarque que les populations africaines (YRI, ASW, LWK, MKK) ont un D de Tajima négatif (très élevé en valeur absolue). Nous pouvons donc supposer que ces populations africaines ont connu une expansion démographique. En revanche, les autres populations ont

un D de Tajima positif, ce qui laisse supposer qu'il y a eu un bottleneck pour celles-ci. On peut donc supposer qu'il y a eu une migration de l'espèce Homo Sapiens hors d'Afrique. Cette observation déduite de nos résultats semble être confirmée par la théorie "Origine africaine de l'Homme moderne." On appelle ce phénomène le Out Of Africa (OOA) qui désigne une très petite partie de la population africaine ayant migré vers les autres continents il y a environ 50000 à 60000 années.

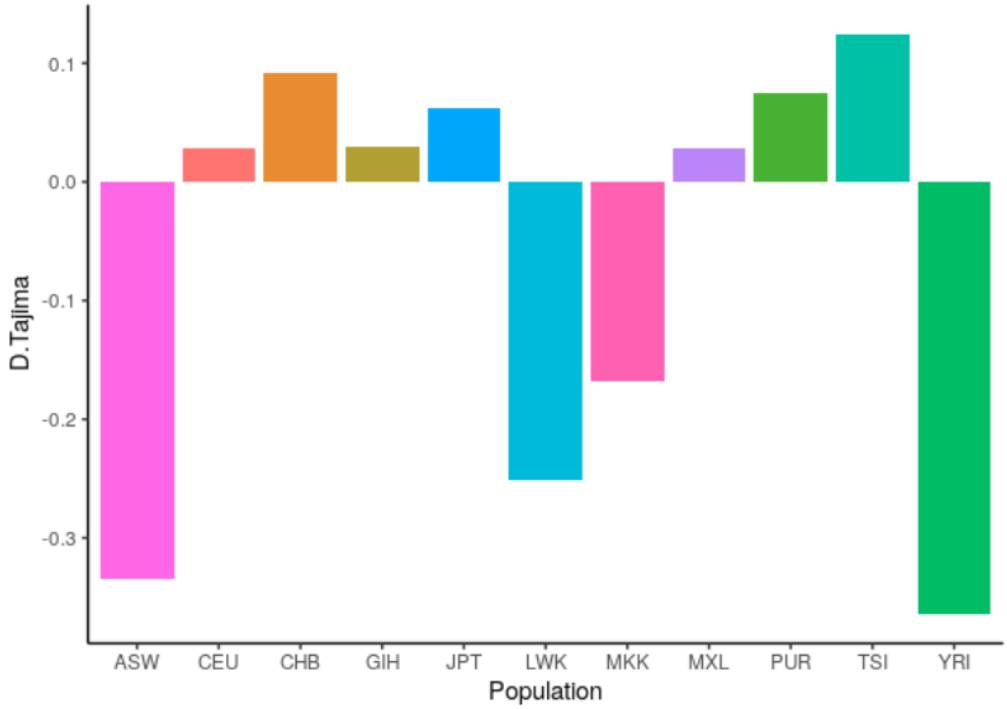

Figure 1: D de Tajima pour chaque population

2 Sélection du modèle démographie

Au regard des D de Tajima différents que nous avons obtenu précédemment, les populations humaines que nous étudions ne semblent pas avoir les mêmes caractéristiques de diversité génétique. Par conséquent, nous allons estimer des modèles démographiques plus complexes pour élaborer des analyses plus approfondies.

2.1 Hypothèses

D'après les statistiques obtenues précédemment, nous pouvons supposer que les populations humaines ont subies des processus démographiques différents. Ces différences de processus démographiques seraient surtout importantes entre les populations africaines (YRI, ASW, LWK, MKK) et les autres populations. Pour la suite du devoir, nous nous intéresserons à trois scénarios démographiques pour chacune des populations étudiées :

- constant (pas de changement significatif de la taille de la population au cours du temps)
- bottleneck (population connaissant un déclin démographique important puis, après une période de plusieurs générations, une croissance démographique très importante ramenant soudainement l'effectif de la population à son état initial)
- expansion (population connaissant une forte croissance démographique à partir d'une certaine période)

Figure 2 fournit une représentation graphique dans le temps de la taille d'une population et de ses coalescences pour chaque scénario démographique. Pour simuler ces modèles, différents paramètres peuvent être

utilisés, tels que la taille effective de la population Ne ou le taux de mutation . Ces valeurs sont souvent inconnues et nécessite toutefois des simulations.

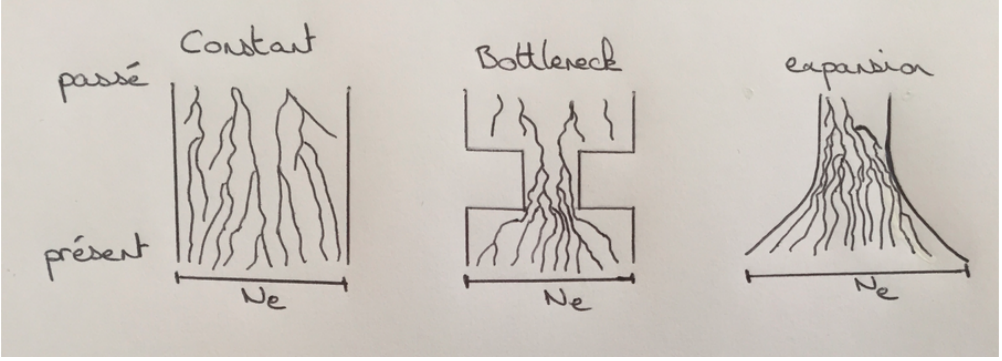

Figure 2: Schématisation des modèles démographiques

Notons que le spectre de fréquence allélique SFS fournit des informations importantes sur les scénarios démographiques passés. Par exemple, la proportion de fréquences alléliques dérivés pour des comptages élevés est plus importante pour le scénario de population constante en comparaison des deux autres scénarios. Observons également que Le D Tajima est une statistique résumant la densité du SFS. Le D Tajima espéré pour chaque modèle démographique est le suivant:

- $D_{Tajima} < 0$: expansion car excès d'allèles en fréquence intermédiaire (Theta Tajima supérieur au Theta Watterson)
- $D_{Tajima} = 0$: constant (population à l'équilibre neutre sous le modèle de Wright-Fisher)
- $D_{Tajima} > 0$: bottleneck car pauvreté d'allèles en fréquence intermédiaire (Theta Tajima inférieur au Theta Watterson)

3.2 Construire une base de simulation de référence

Pour retracer l'histoire démographique humaine, le modèle ABC peut être utilisé. Différentes statistiques résumées peuvent être utilisées : comptage des allèles, diversité génétique, spectre des fréquences alléliques, etc. Ces statistiques résumées peuvent être simulées à partir de différents paramètres tels que qu'énoncé précédemment (comme par exemple le taux de mutation). La distribution à priori de chacun de ces paramètres suivent une loi uniforme. Nous pouvons proposer des bornes faibles afin d'être parcimonieux dans nos résultats. Si nous fixons un taux de tolérance trop élevé, nos trois modèles vont tendre vers l'équiprobabilité. En revanche, il ne faut pas fixer un taux de tolérance trop faible non plus, afin d'éviter d'augmenter la variance de nos résultats qui seraient trop dépendantes d'une très faible quantité de simulations conservées. Un taux de tolérance de 5% semble donc intéressant et raisonnable.

Dans le cas présent, un jeu de statistiques résumées déjà simulées pour 50000 valeurs différentes de paramètres pour les trois scénarios démographiques est déjà proposé. Les statistiques résumées simulées sont le Theta Tajima, le D Tajima et la variance du D Tajima. Nous rappelons que le D Tajima est une statistique résumée du SFS et permet donc de retracer l'évolution démographique d'une population. **Figure 3** fournit les boxplot de ces trois statistiques résumées pour les trois scénarios démographiques étudiés : population constante, goulot d'étranglement et expansion démographique. On remarque que la médiane du D Tajima est positif en cas de bottleneck, nul pour une population constante et fortement négative pour une population en expansion. Ces résultats sont en cohérence avec la théorie établie. Également, on remarque que la variance du D Tajima est plus élevée en cas de bottleneck et plus faible en cas de population constante et en expansion.

Nous allons maintenant nous intéresser à la sélection de modèle démographique avec la méthode Approximate Bayesian Computation (ABC) en récupérant les probabilités bayésiennes a posteriori pour chacun des scénarios. Pour commencer, nous pouvons récupérer une unique simulation et ses statistiques résumées. Par exemple, prenons la première simulation associée à un goulot d'étranglement. Pour cette simulation, nous avons $\Theta_{Tajima} = 0.001053111$, $D_{Tajima} = 0.05764936$ et sa variance $Var(D_{Tajima}) = 1.302449$. Si nous

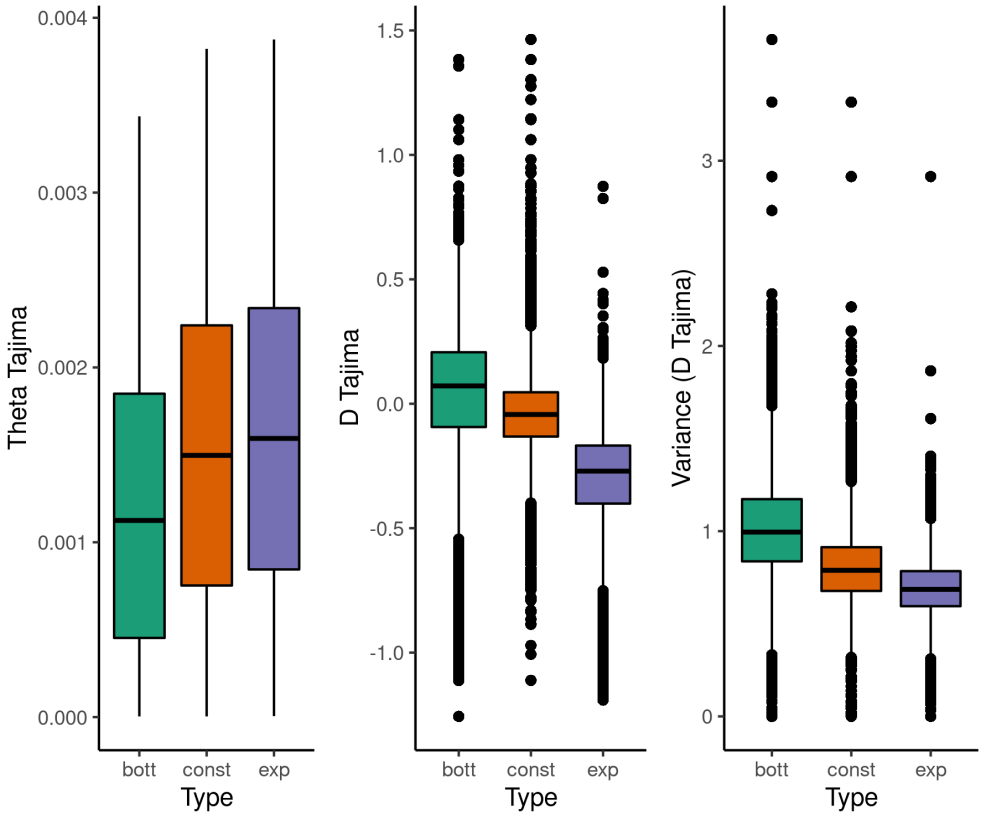

Figure 3: Statistiques résumées simulées pour chaque scénario démographique

appliquons la méthode ABC avec un seuil de tolérance fixé à 5%, nous obtenons les probabilités a posteriori suivantes : 13.20% pour une population constante, 1.33% pour un goulot d'étranglement et 86.67% pour une population en expansion. En revanche, si nous avons un seuil de tolérance égal à 100, nous obtenons des probabilités a posteriori équiprobable, c'est-à-dire à 33.3% chacun. Cela résulte du fait que nous conservons toutes les simulations lorsque le seuil de tolérance est à 100% et il n'est donc pas possible de prédire le scénario démographique auquel elle appartient. En revanche, lorsque nous fixons un seuil de tolérance à 5%, nous prenons uniquement les 5% de simulations les plus proches de l'observation et nous utilisons les proportions de chaque scénario simulé récupéré pour estimer la probabilité bayésienne a posteriori. Ajoutons que nous avons retiré l'observation pour estimer son postérieur. Nous faisons cela afin d'avoir une estimation out-of-sample, et donc corriger le biais de l'erreur d'apprentissage.

Nous avons 3 modèles démographiques. Nous définissons le taux de mal classés comme suit:

$$MCR = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \neq \hat{y}(x_i)\}$$

Si l'on prédisait un modèle au hasard, le taux de mauvaise classification serait de deux tiers. Avec le modèle computationnel ABC, nous estimons le scénario démographique en fonction des autres observations. Toutefois, deux hyperparamètres sont requis pour notre modèle ABC :

- Seuil de tolérance ε
- Méthode utilisée α

Un hyperparamètre est un paramètre sélectionné par l'utilisateur, en amont de l'exécution du modèle. Lorsqu'un modèle est nouveau, la sélection de l'hyperparamètre est arbitraire. Néanmoins, dans le cas

de prédiction ou classification, il est important de procéder à des tests computationnels pour sélectionner une valeur pour l'hyperparamètre. Dans notre cas par exemple, nous avons un problème de classification à trois modalités. Notre objectif est d'utiliser un modèle qui prédit un scénario démographique en minimisant le taux d'erreur. Pour évaluer la qualité de prédiction associée à un hyperparamètre, différentes méthodes sont possibles. Dans le cas présent, nous allons utiliser une méthode de cross-validation pour sélectionner la valeur des hyperparamètres. La validation croisée consiste à diviser les données en deux sous jeux de données : le premier est appelé jeu de données d'apprentissage (servant à calculer les estimateurs pour la prédiction du modèle) et le second est appelé jeu de données test (servant à calculer l'erreur de prédiction). Ainsi, nous chercherons à minimiser l'erreur test comme suit :

$$\{\hat{\varepsilon}, \hat{\alpha}\} = \arg\min \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{y_i \neq \hat{y}^{-i}(x_i)\}$$

Pour rechercher la valeur de nos deux hyperparamètres (seuil de tolérance ε et méthode utilisée α), nous utiliserons une méthode basée sur le *leave-one-out cross validation*. Cette dernière consiste à considérer, pour une observation donnée, un échantillon d'apprentissage composé de toutes les observations sauf celle donnée, et d'un échantillon test composé de cette unique observation, puis de calculer le modèle sur l'échantillon d'apprentissage et la prédiction sur l'observation donnée. Cette opération est réitérée sur l'ensemble des observations disponibles afin que nous puissions calculer le taux de mal classés.

Dans le présent travail, une fonction nous a été fournie pour effectuer le leave-one-out cross validation. Comme le modèle ABC est consommateur en temps, l'énoncé nous indique d'utiliser directement la méthode de base, i.e. rejet. En revanche, nous allons évaluer le taux de mal classés en fonction du seuil de tolérance epsilon. Le leave-one-out cross validation est une opération coûteuse en temps. Afin de pallier ce problème, nous avons parallélisé la méthode de cross validation. Nous avons donc lancé le (pseudo) leave-one-out cross-validation sur 150 observations (50 pour chaque scénario) pour différents seuils de tolérance. **Figure 4** reporte le taux d'erreur test pour un seuil de tolérance epsilon allant de 0.05% à 10%. On remarque que le seuil de tolérance est croissant à mesure que le seuil de tolérance augmente. Le taux d'erreur tend vers deux tiers à mesure que nous approchons de un. Il convient donc de fixer un seuil de tolérance faible pour minimiser le taux d'erreur. Toutefois, on remarque également du bruit pour des seuils de tolérance très faibles (entre 0.05% et 2.5%). En effet, si notre modèle ne prend qu'un très faible nombre de simulations proches de l'observation, le poids d'une simulation est très important, ce qui augmente la variance de nos prédictions. Il convient donc de fixer un seuil de tolérance faible mais suffisamment élevé également pour éviter de capturer du bruit. Un seuil de tolérance fixé à 2.5% semble raisonnable.

2.2 Sélection du modèle démographique pour les données réelles

Maintenant que nous avons déterminé la valeur de nos hyperparamètres grâce au cross-validation sur données simulées (seuil de tolérance $\varepsilon = 2.5\%$ et méthode utilisée $\alpha =$ rejetction), nous pouvons appliquer le modèle ABC sur nos données réelles. TODO : reference table Table 1 fournit, pour chaque population, la probabilité bayésienne a posteriori de chaque scénario démographique et la prédiction du scénario. Nous remarquons que pour quatre populations, le scénario d'expansion est prédit. Cela concerne les quatre populations africaines (YRI, ASW, LWK, MKK). Parmi les populations pour lesquelles le modèle ABC prédit un goulot d'étranglement, nous avons les résidents de l'Utah d'ascendance européenne du Nord et de l'Ouest (CEU), les chinois Han à Pékin (CHB), les japonais à Tokyo (JPT) et les Toscans en Italie (TSI). Pour toutes les autres populations, le modèle ABC prédit une démographique constante dans le temps. Pour certaines populations, le scénario démographique prédit est très vraisemblable. C'est le cas par exemple pour les Yoruba à Ibadan au Nigéria (YRI) pour lesquels nous estimons le scénario d'expansion à 93.3% ou les individus d'ascendance africaine dans le sud-ouest des États-Unis (ASW) pour lesquels nous estimons le même scénario à 90.7%. Pour d'autres populations, le scénario démographique prédit est plus incertain. Par exemple, les portoricains (PUR), les individus d'ascendance mexicaine à Los Angeles (MXL) et les indiens Gujarati à Houston au Texas (GIH) ont une probabilité d'avoir une population constante dans le temps à 55% environ et d'avoir connu un goulot d'étranglement à 45% environ. Nous pouvons donc nous demander si le scénario de prédiction pour ces populations est correct. Si, pour ces populations, le scénario de goulot

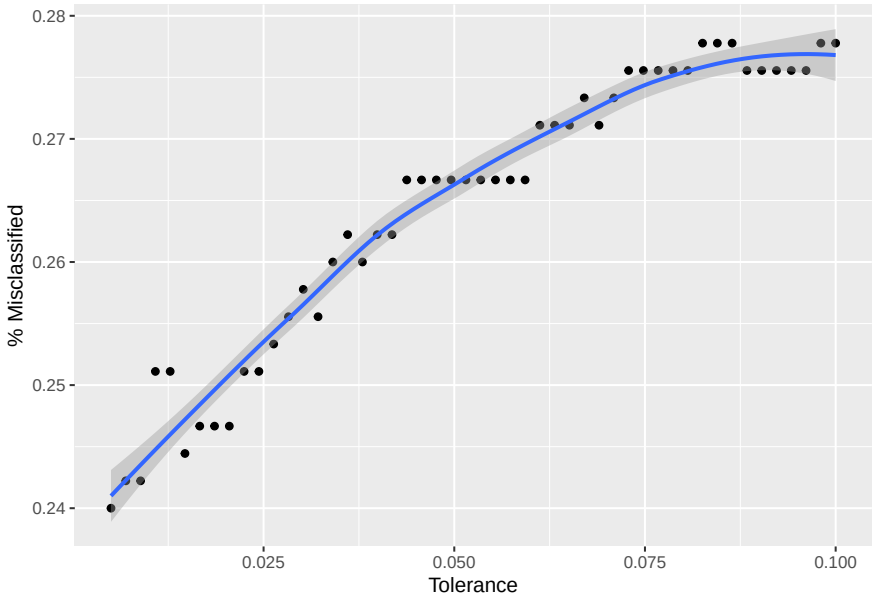

Figure 4: Taux d'erreur test en leave-one-out cross-validation

Table 1: Prédiction du scénario démographique pour chaque population

Population	P.bott	P.const	P.exp	Scenario
PUR	0.434	0.556	0.0104	const
CEU	0.57	0.419	0.0107	bott
YRI	0.0016	0.0653	0.933	exp
CHB	0.68	0.318	0.00267	bott
JPT	0.666	0.331	0.0032	bott
LWK	0.00587	0.2	0.794	exp
MXL	0.446	0.539	0.0144	const
ASW	0.00267	0.0901	0.907	exp
TSI	0.57	0.427	0.00213	bott
GIH	0.433	0.554	0.0136	const
MKK	0.0285	0.422	0.549	exp

d'étranglement était vrai, alors la théorie de l'origine africaine de l'Homme moderne se confirmerait davantage, avec un scénario d'expansion pour les populations africaines et un scénario de bottleneck pour toutes les autres populations.

Pour vérifier le goodness-of-fit de nos résultats, nous pouvons utiliser l'ACP. Figure 5 fournit les ACP pour les populations YRI et PUR. Nous remarquons que pour les Yoruba à Ibadan au Nigéria (YRI), le scénario d'expansion est bien prédit. En revanche, pour les portoricains (PUR), nous remarquons qu'il est plus difficile de prédire correctement le scénario entre population constante et bottleneck (comme nous l'avons souligné précédemment). En effet, l'observation se situe à l'intérieur du cercle pour la population constante et du cercle pour le bottleneck.

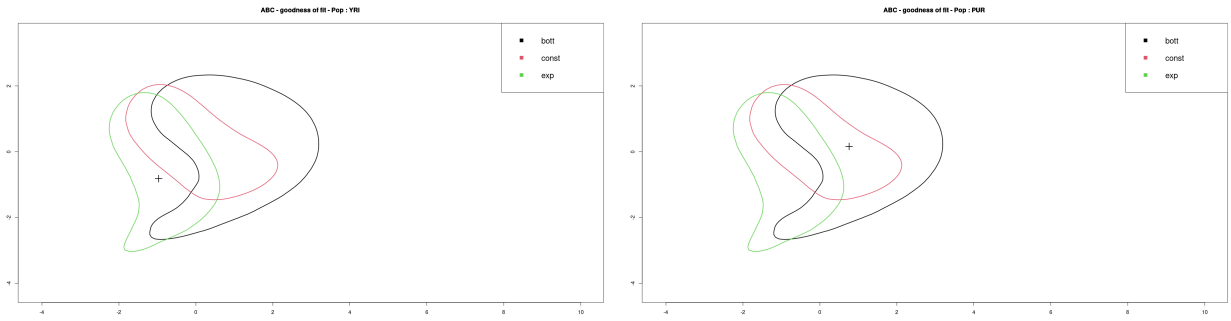

Figure 5: ACP des statistiques résumées (populations YRI et PUR)

3 Estimer les paramètres d'un modèle donné

Ayant choisi des modèles démographiques maintenant nous nous intéressons à l'estimation des paramètres utilisés précédemment. Ce sont les suivants :

- N_e, la taille efficace avant et après le goulot d'étranglement.
- a, le facteur de réduction, i.e. N_e divisée par la taille pendant le goulot d'étranglement ($a = N_e/N_{bottleneck}$).
- $duration$, la durée du goulot (en années).
- $start$, temps depuis le début du goulot (en années).

3.1 Exploration, intuitions

Toute analyse bayésienne repose sur les mêmes principes : la formulation d'un modèle, ajuster le modèle aux données pour finalement améliorer le modèle en vérifiant la qualité de l'ajustement et en le comparant à d'autres modèles.

Le Théorème de Bayes permet de calculer le posterior $P(\theta|X)$ en fonction de la vraisemblance $P(X|\theta)$, le prior $P(\theta)$ et l'évidence $P(X)$ fournie par nos données. Ce théorème peut être vu comme une façon de "mettre à jour nous croyances" sur la valeur de notre paramètre d'intérêt θ au fur et à mesure que l'on obtient plus d'informations sur le phénomène étudié $P(X)$.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

En sachant que les paramètres sont tirés indépendamment, à partir du théorème, on fait la substitution de $\theta = (N_e, a, duration, start)$ dans le prior, ce qui nous permet de calculer le posterior:

$$P(\theta|X) = \frac{P(X|\theta)\prod_{i=1}^{n}\theta_i}{P(X)} = \frac{P(X|\theta)P(N_e)P(a)P(duration)P(start)}{P(X)}$$

Les distributions à priori des paramètres de la **figure 6** sont toutes uniformes sauf une, celle de a, le facteur de réduction. Cela s'explique par l'interprétation de chacun des paramètres dans le contexte de l'inférence démographique. A priori, nous ne savons pas quelle aurait pu être la taille efficace (N_e) avant un boulot d'étranglement. Le même raisonnement s'applique à la durée du goulot ($duration$) et le temps depuis le goulot ($start$). Pour cette raison l'hypothèse la plus pertinente est celle d'une équiprobabilité entre les valeurs minimum et maximum que l'on peut imaginer. Or, quant au facteur de réduction, si un sous-ensemble trop réduit migre ou survie à une catastrophe naturelle ($min(N_{bott}) \rightarrow max(a)$ grand) la probabilité de survie et

puis de reproduction devient trop faible. Cela justifie la loi avec une densité plus importante pour faibles valeurs de a que l'on observe.

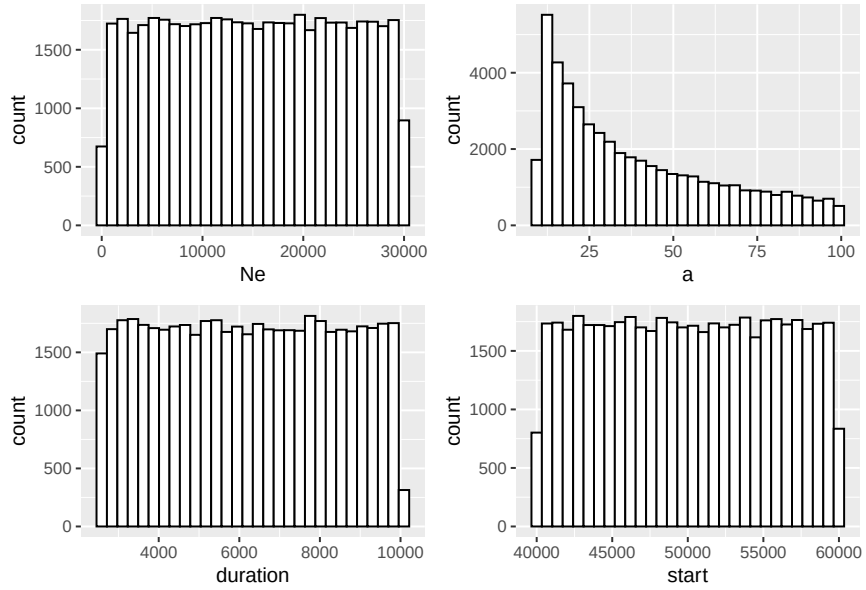

Figure 6: Distribution empirique des paramètres

En analysant le comportement des statistiques résumées en fonction de la taille efficace, nous constatons que θ_{Tajima} (pi) montre une claire corrélation positive. Le D_{Tajima} ainsi que sa variance montrent d'abord une corrélation négative et puis positive dans la région $N_e \leq 1000$. Après les deux statistiques semblent atteindre un plateau.

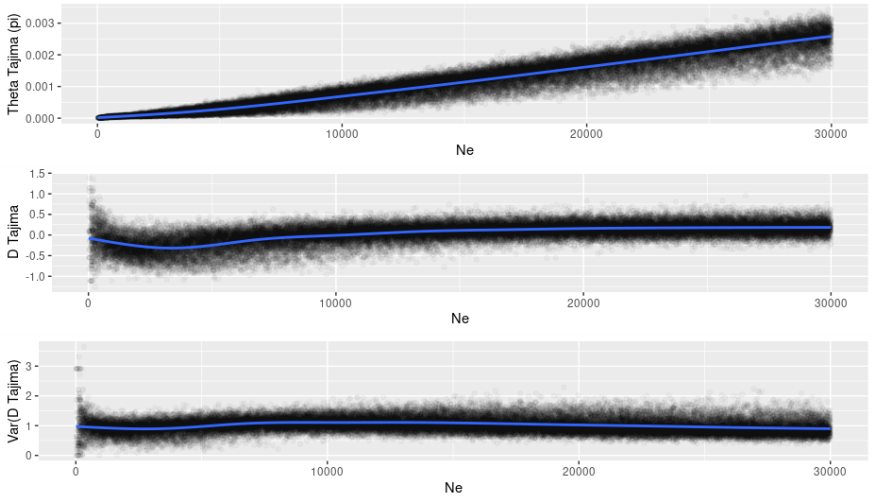

Figure 7: Statistiques vs Ne

3.2 Exemple d'inférence

Afin de voir comment la fonction `abc` nous permet d'échantillonner la loi à posteriori de nos paramètres ce qui souvent ne peut pas se faire de façon analytique. La **figure 8** montre l'échantillonnage du posterior de chacun de nos paramètres d'intérêt pour une observation tirée aléatoirement. La fonction `abc` a été appelée en fixant `tol=0.05` et `method='loclinear'`. Nous pouvons constater que le comportement des trois lois est bien différent pour chacun de nos quatre paramètres.

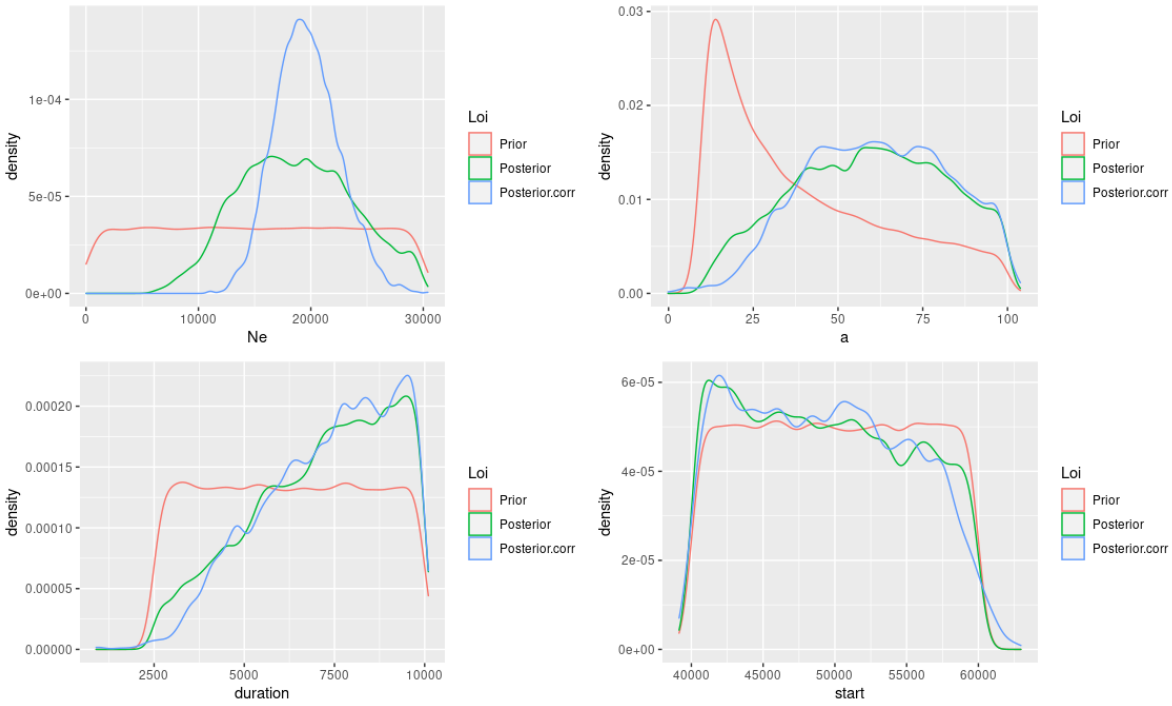

Figure 8: Comparaison des lois pour chaque paramètre

Nous observons que N_e est près d'être un cas idéal d'estimation du posterior. En partant d'une loi uniforme la méthode de rejet nous permet d'arriver à une loi qui s'approche d'une gaussienne qui serait centrée sur la vraie valeur de la taille effective de la population. En appliquant la correction par régression linéaire locale, notre estimation devient beaucoup plus précise ce qui est mis en évidence par la courbe bleue.

La qualité des estimations du posterior des autres paramètres est loin d'être idéal. Pour `start`, les lois du posterior naïf (rejet) et posterior corrigé (rég. lin.) se superposent avec le prior. Quant à `duration`, les posteriors se sont écartés du prior mais on n'a pas une estimation très confiable. Un résultat intéressant est celui du paramètre `a` pour lequel l'estimation n'est pas optimale mais on voit que le prior et posterieur sont contrairement biaisés. Ceci semble contredire l'hypothèse que nous avions précédemment émise. Cependant le résultat n'est pas conclusif car l'estimation pourrait encore être améliorée et elle n'a été faite que sur une observation tirée aléatoirement. De plus, les hyperparamètres utilisés ont été choisis arbitrairement et non pas par validation croisée.

3.3 Performances

Pour une (ou plusieurs) variable continue, plusieurs métriques existent. L'une des plus fréquentes est la racine de l'erreur quadratique moyenne appelée $RMSE$ pour son nom en anglais. Elle est définie comme suit :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_t - y_t)^2}{n}}$$

Cette métrique est facilement utilisable pour comparer les performances de différents modèles. Elle a l'avantage de se trouver dans le même ordre de grandeur que la variable d'intérêt. Cette métrique permet d'estimer l'écart en moyenne entre les estimations $\hat{y}_t$ et la valeur observée y_t pour une variable y, mais que devons-nous faire quand plusieurs paramètres doivent être considérés ? Par exemple, pour comparer $RMSE(N_e)$ et $RMSE(a)$ nous avons besoin que les erreurs soient du même ordre de grandeur. Pour ce faire, nous avons utilisé la fonction R `scale(x, center=F, scale=T)` afin de réduire chacun des $RMSE_i$.

Une option est de ne regarder qu'une parmi les quatre valeurs de RMSE, mais ceci représenterait une perte d'information. Une solution considérant toutes les variables est de prendre la moyenne de leurs erreurs. Cette moyenne peut être naïve ou pondérée. En tout cas le choix de poids est crucial car il déterminera le classement de méthodes de correction et valeurs de tolérance, et donc tous les résultats obtenus.

La pondération peut être justifiée par une hypothèse biologique ou bien par un critère numérique pertinent. Nous avons décidé de prendre l'écart type du vecteur d'erreur de chaque variable comme son poids. Si l'on fait une moyenne simple, on risque de ne pas représenter la sensibilité de chaque paramètre au seuil de tolérance ε.

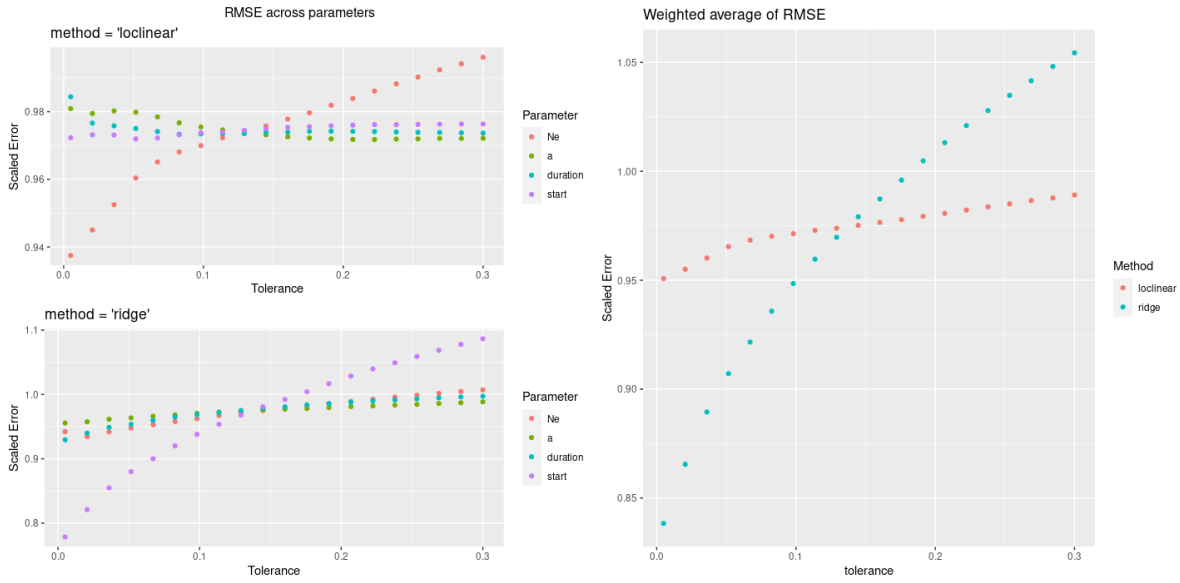

Figure 9: A gauche : valeurs réduits pour chaque paramètre, à droite : moyenne pondérée

Dans la **figure 9** nous observons à gauche que le paramètre le plus sensible au changements de ε est N_e quand la méthode de correction est `loclinear` alors que pour `ridge`, c'est *start*. A droite nous observons que la méthode `ridge` se montre plus sensible au choix de tolérance et permet d'obtenir globalement un RSME réduit plus faible que `loclinear`. Pour cette raison nous avons fixé nos paramètres à `method='ridge'` et `tol=0.025`.

Table 2: Statistiques Résumées pour la pop. CHB (bottleneck)

Type	pi	tajimas_d	tajimas_d_var
Estimate	0.0009316	0.0893783	1.077206
Observed	0.0009436	0.0911399	1.079177

3.4 Application aux données réelles et conclusion

Maintenant que nous avons bien calibré les hyperparamètres par validation croisés en prenant en compte la métrique que nous avons proposée (moyenne des valeurs $RMSE_i$ pondérée par leurs écarts type), nous pouvons appliquer la méthode `abc` pour estimer les paramètres $\theta = (N_e, a, duration, start)$. La **table 2** montre aussi la valeur observée et l'estimation par `abc` qui est très proche pour toutes les statistiques résumées.

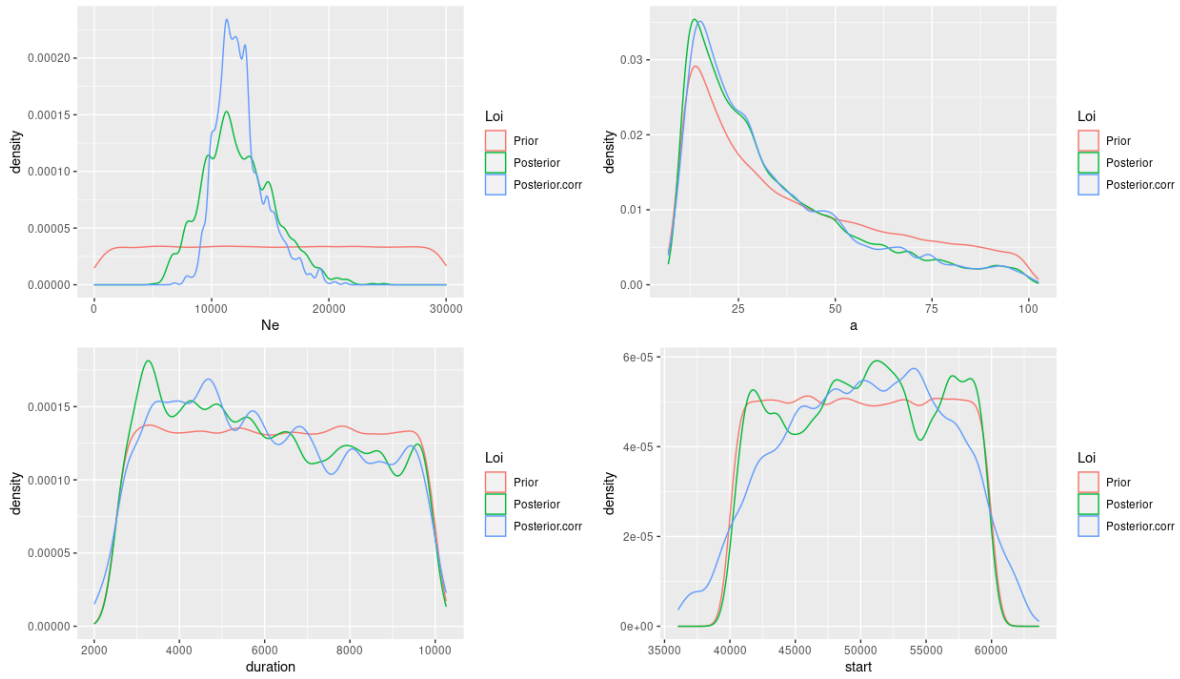

Figure 10: Distributions des paramètres, pop. CHB

Comme précédemment discuté pour l'exemple aléatoire, on voit que l'estimation de N_e est très différente de son prior. La variance de sa loi corrigé (corrected posterior) en utilisant la penalité `ridge` (l^2-norm) est moins importante que celle de l'estimation par simple rejet. Or, cette fois les deux estimations posterior pour a ne contredissent pas le prior. Ceci nous mène à croire que probablement l'observation prise au hasard pour produire les premières estimations de densité des lois des paramètre ne correspondait pas à une simulation du scénario bottleneck. Nous ne pouvons faire cette inférence que pour des données de simulations bottleneck car nous n'avons que les statistiques résumées pour ce scénario. Ceci est une amélioration à prendre en compte: ce serait intéressant d'avoir des données pour d'autres scénarios et analyser si cette variable a une influence sur la valeur des hyperparamètres optimaux cherchés par validation croisée. D'autres statistiques (sans nécessairement être exhaustives) pourraient être utiles pour améliorer les performances du modèle.

Références et Annexe

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r.* https://CRAN.R-project.org/package=rmarkdown.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics.* https://CRAN.R-project.org/package=gridExtra.

Bache, Stefan Milton, and Hadley Wickham. 2020. *Magrittr: A Forward-Pipe Operator for r.* https://CRAN.R-project.org/package=magrittr.

Csillery, Katalin, Olivier Francois, and Michael G. B. Blum. 2012. "Abc: An r Package for Approximate Bayesian Computation (ABC)." *Methods in Ecology and Evolution.* https://doi.org/http://dx.doi.org/10.1111/j.2041-210X.2011.00179.x.

Gaujoux, Renaud. 2020. *doRNG: Generic Reproducible Parallel Backend for Foreach Loops.* https://renozao.github.io/doRNG.

Katalin, Csillery, Lemaire Louisiane, Francois Olivier, and Blum Michael. 2015. *Abc: Tools for Approximate Bayesian Computation (ABC).* https://CRAN.R-project.org/package=abc.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

Müller, Kirill, and Hadley Wickham. 2021. *Tibble: Simple Data Frames.* https://CRAN.R-project.org/package=tibble.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes.* https://CRAN.R-project.org/package=RColorBrewer.

Ooms, Jeroen. 2021. *Magick: Advanced Graphics and Image-Processing in r.* https://CRAN.R-project.org/package=magick.

Paul R. Staab, and Dirk Metzler. 2016. "Coala: An R Framework for Coalescent Simulation." *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btw098.

Revolution Analytics, and Steve Weston. n.d. *Foreach: Provides Foreach Looping Construct.*

Staab, Paul, and Dirk Metzler. 2020. *Coala: A Framework for Coalescent Simulation.* https://github.com/statgenlmu/coala.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations.* https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for Ggplot2.* https://wilkelab.org/cowplot/.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr.* 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Repo github

L'intégralité du code utilisé pour cet analyse est disponible dans le lien suivant : https://github.com/gmagannaDevelop/MetaGenomique/tree/ABC. N'hésitez pas à utiliser et modifier les versions parallélisées des fonctions `cv4abc` et `cv4postpr` !