Estimation non paramétrique dans des modèles de cladogenèse

Jean Velluet

Sorbonne Université & Ecole Normale Supérieure

May 22, 2024

Sommaire

Espèce

Ensemble d'individus isolés d'un point de vue reproductif

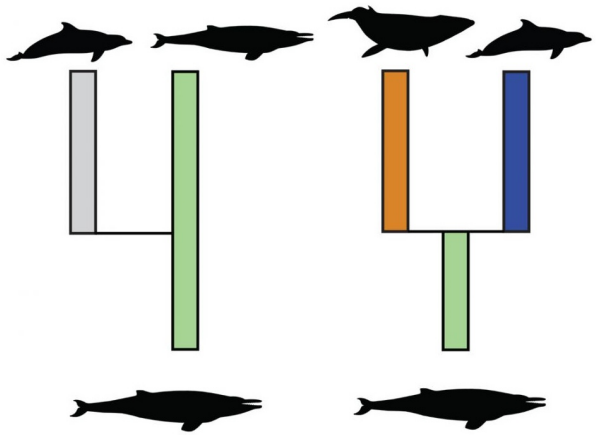


Figure: Cladogenèse

Clagogénèse = Processus de création de nouvelles espèces

- mutations, sélections, ségrégations
- causes externes (glaciation, fermeture d'un isthme...) ou internes (spécialisation de niche, innovation phénotypique...)

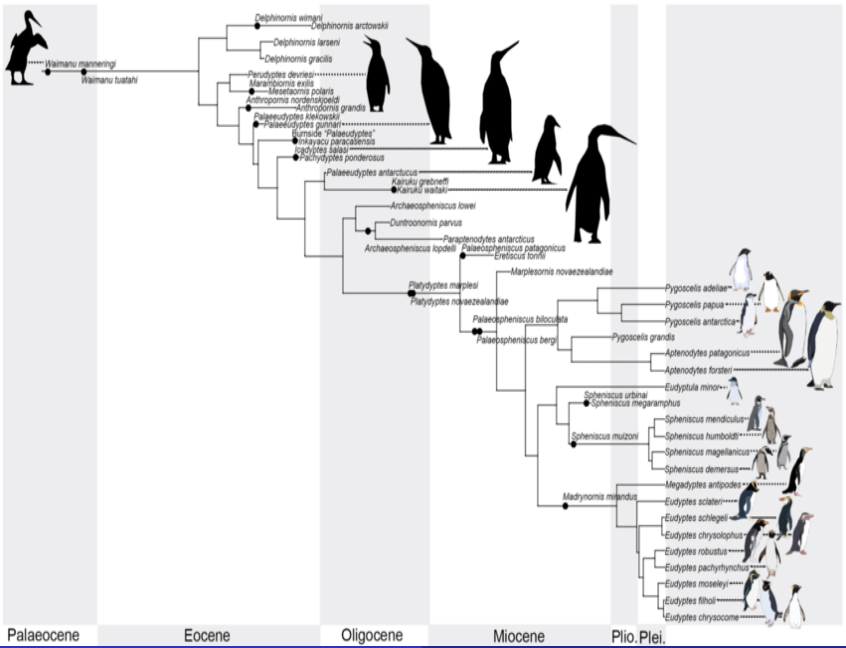
$\Rightarrow$ spéciation par cladogenèse

Taux de spéciation et d'extinction

Definitions

- λ : taux de spéciation (en Myrs^{-1})
- μ : taux d'extinction (en Myrs^{-1})

Dépendent à priori du temps, du nombre d'expèces co-existantes, de traits héritables ou non [1]



Question

Biologiste : connaitre évolution de ces taux au cours du temps

- mieux comprendre l'hisoire évolutive des espèces.
- en particulier : liens avec environnement **externe** (climat, altitude, migration...)

$\Rightarrow$ Compter les espèces qui sont apparues et qui ont disparu ?

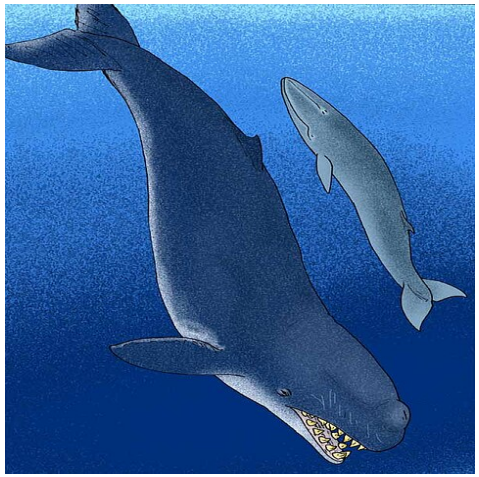


Figure: Espèce ayant vécue il y a environ 9,9 à 8,9 Myrs

Figure: Fossile du crâne découvert en 2008

Données disponibles pour l'inférence de ces taux

- On a "aucune" trace des espèces disparues
- En pratique, on n'a seulement accès à une fraction ρ des espèces actuelles
- Sequençage génétique des espèces actuelles $\Rightarrow$ reconstruire des arbres phylogénétiques

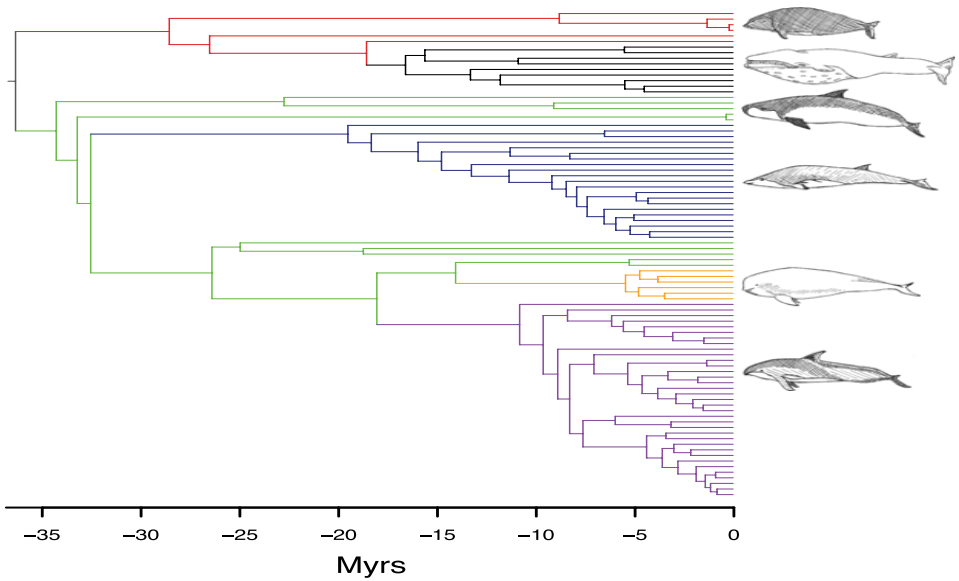


Figure: Phylogénie reconstruite des Cétacés [3]

Modèle mathématique

Processus de naissance-mort

- Processus de markov à temps continu
- Transitions : de n à $n+1$ individus "naissance" avec un taux λ ou de n à $n-1$ individus "mort" avec un taux μ

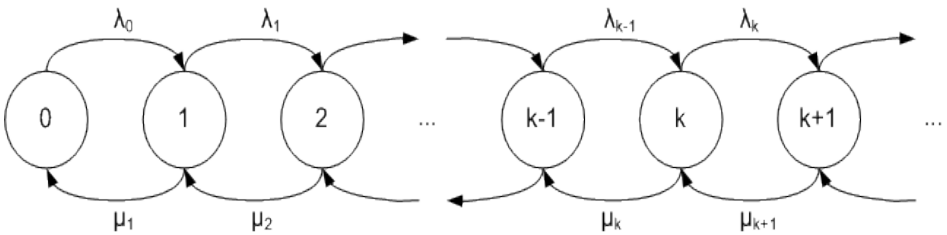


Figure: Chaine processus de naissance-mort

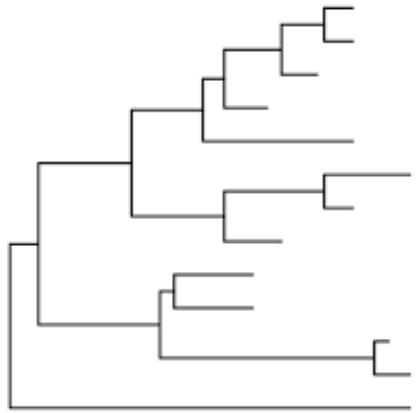


Figure: Arbre résultant

Hypothèses et but

Hypothèse :

- L'évolution se fait selon un processus de naissance mort.
- L'arbre résultant est constitué des espèces existantes et éteintes $\Rightarrow$ L'**élagage** espèces éteintes et fraction $(1 - \rho)$ des espèces existantes : un arbre "reconstruit".

But

Inférer les vitesses de spéciation λ et de mort μ des espèces à partir de l'abre phylogénétique reconstruit.

Quelques résultats préliminaires

identifiabilité "asymptotique"

On dit qu'un modèle est "asymptotiquement" identifiable si à partir d'un nombre infini d'observations, on peut retrouver les vrais paramètres.

Stadler 2009

Même lorsque λ et μ sont supposés constants, les paramètres (λ, μ, ρ) ne sont pas identifiables, mais (λ, μ) le sont.
$\Rightarrow$ On suppose ρ connu.

Louca & Pennell (2020) [4]

Pour un arbre reconstruit T_b, il existe une infinité de fonctions dérivables λ et μ qui ont la même probabilité d'avoir généré cet arbre. On note $\mathcal{C}(T_b)$ cet ensemble appelé "classe de congruence".

Comment approcher les "vrais" paramètres ?

Considérer des formes paramétriques

λ et μ évoluant de manière exponentielle ou linéaire avec le temps (ex. $\lambda(t) = \lambda_0 e^{at}, \mu(t) = \mu_0 + bt$)

Legried & Terhorst (2022)

Les taux de speciation et de mort constants par morceaux sont identifiables lorsque : $n > 8K$. (Où n est le nombre de feuilles et K est le nombre d'intervalles de temps.)

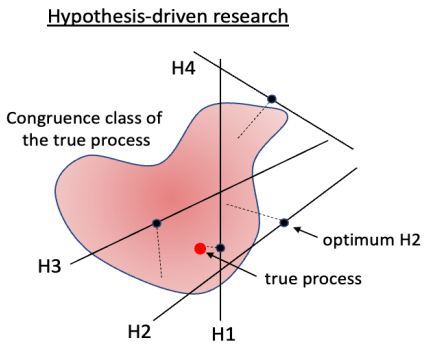


Figure: Classe de congruence et hypothèse biologique [3]

Comment approcher les "vrais" paramètres ?

Filtration

- Echantillonner la classe de congruence
- Filtrer en ne gardant que les taux "biologiquement" possibles.

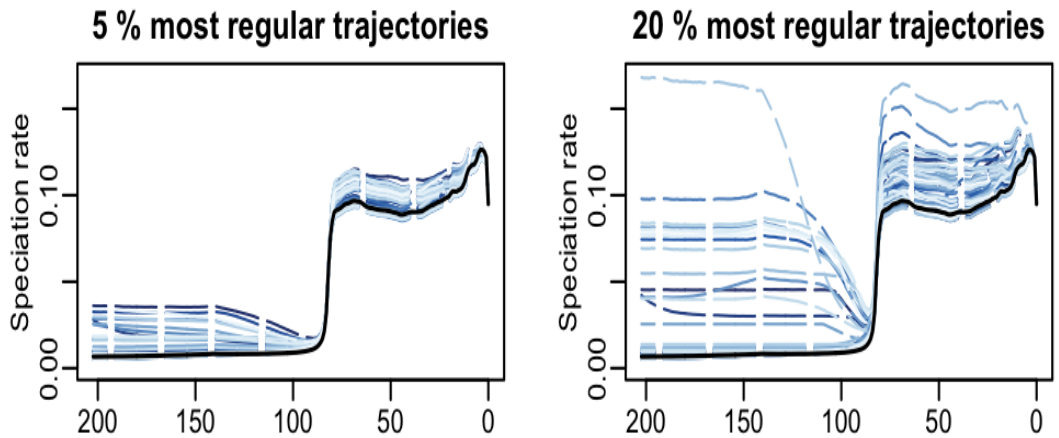


Figure: Filtration des éléments de la classe de congruence par régularité de la trajectoire [**Andreoletti2023**]

Comment approcher les "vrais" paramètres ?

L'approche choisie pour le stage :

- Taux de spéciation et d'extinction plus flexibles(abus de langage "non paramétrique")
- Pénaliser les trajectoires peu probables biologiquement $\Rightarrow$ contraindre la classe de congruence vers les "vrais" paramètres.

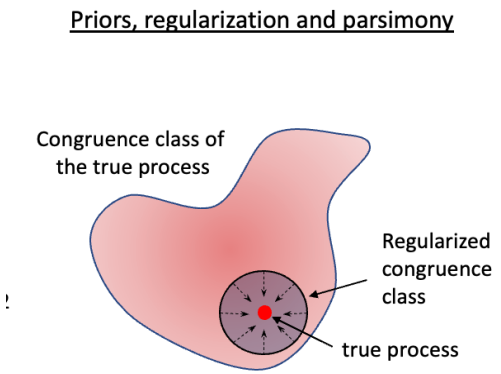


Figure: Classe de congruence régularisée [2]

Méthode d'estimation

Statistique à "fitter"

Etant donné un arbre reconstruit Y, d'instants de branchements $T_b = \{t_2 \geq, \ldots, \geq t_n\}$, il existes des statistiques (la vraisemblance, le nombre d'espèces présente à chaque instant, le taux de diversification net...) qui sont invariantes dans la classe de congruence.

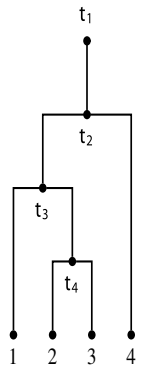


Figure: Exemple d'un arbre reconstruit

Maximimum de vraisemblance

- $T_b = \{t_2 > \cdots > t_n\}$: Temps de branchements
- On note $\theta := (\lambda, \mu)$, vraisemblance sans conditionnement [1] :

$$\mathcal{L}(\theta; T_b) = \rho^n \Psi(0, t_2)^2 \prod_{i=3}^{n} \lambda(t_i)\Psi(0, t_i)$$

$$\Psi(0,t) = e^{\int_0^t (\lambda(u)-\mu(u))du}\left(1 + \rho \int_0^t e^{\int_0^\tau (\lambda(\sigma)-\mu(\sigma))d\sigma}\lambda(\tau)d\tau\right)^{-2}$$

Remarque

Les statistiques ne dépendent que des instants de branchements.

Réduction de la classe de congruence

Rappel

La classe de congruence n'est pas réduite à un singleton $\Rightarrow$ le maximum de vraisemblance n'est pas unique.

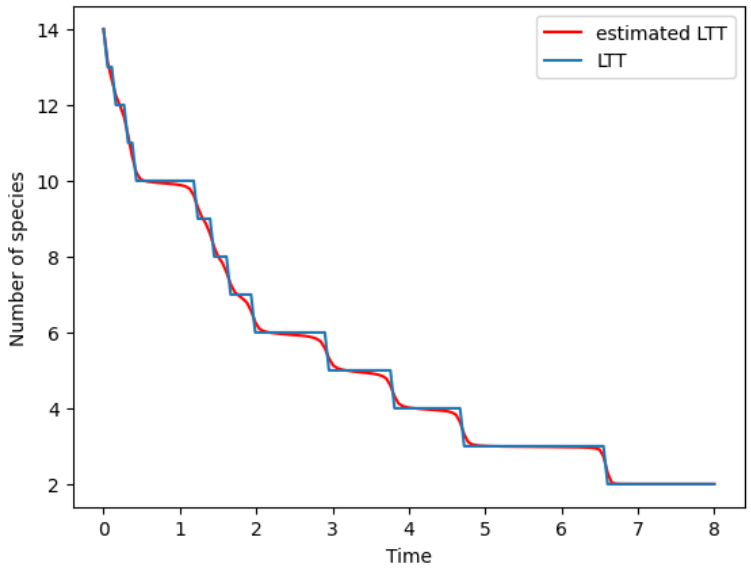


Figure: dLTT calculé avec les paramètres estimés (en rouge) et une réalisation du LTT (en bleu)

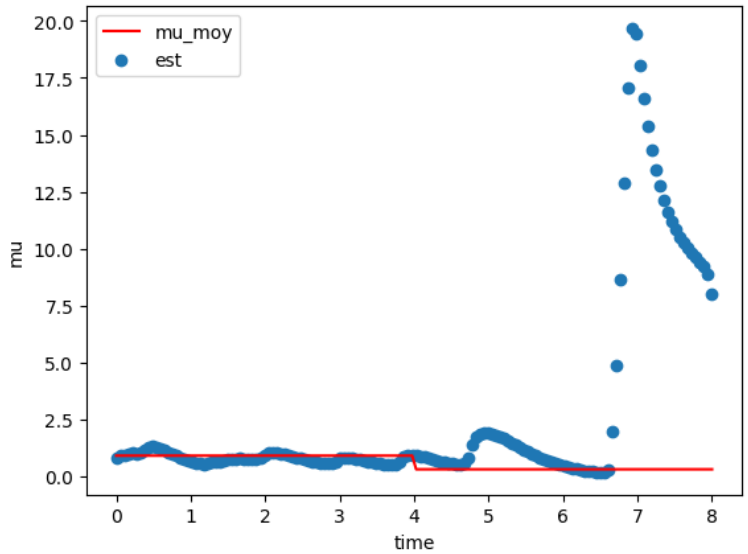


Figure: Taux d'extinction estimé μ (en bleu) et valeur réelle du paramètre (en rouge)

Régularisation

Choix des pénalisations

λ : variations assez lisses

μ : constant par morceau avec de possibles sauts

$$p(\lambda) = ||\lambda''||_2 \simeq \frac{1}{\Delta_t^2} \sum_{i=1}^{n} (\lambda_{i+1} - 2\lambda_i + \lambda_{i-1})^2$$

$$q(\mu) = ||\mu'||_1 \simeq \frac{1}{\Delta_t} \sum_{i=1}^{n} |\mu_{i+1} - \mu_i|$$

Remarque

D'autres pénalisations possibles (qui annulent certaines formes paramétriques classiques)

Procédure d'optimisation

fonction objectif régularisée

$$J_\gamma : \theta \to -\log(\mathcal{L}(\theta; T_b)) + \alpha p(\lambda) + \beta q(\mu)$$

avec $\gamma := (\alpha, \beta)$

- Descente de gradient avec Pytorch (Adam).
- Moment de Nesterov pour accélérer la procédure

quelques difficultés rencontrées

- Trouver un bon taux d'apprentissage initial (dépend de l'ensemble de données, des hyperparamètres α, β, de la vitesse de convergence)
- Non-différentiabilité de la pénalité q

Approximation différentiable [2]

$|x| \simeq \frac{2x}{s} \int_0^{x/s} e^{-t^2} dt$ pour s petit.

quelques propriétés

- différentiable
- converge exponentiellement rapidement

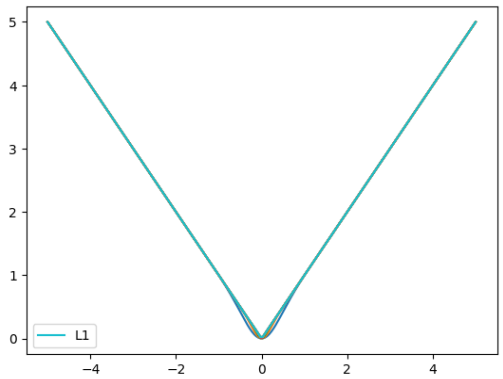


Figure: Boule unité $L1$ et son approximation

$\Rightarrow$ Evite descente de gradient proximale (problème dual + KTT).

Choix des hyper-paramètres

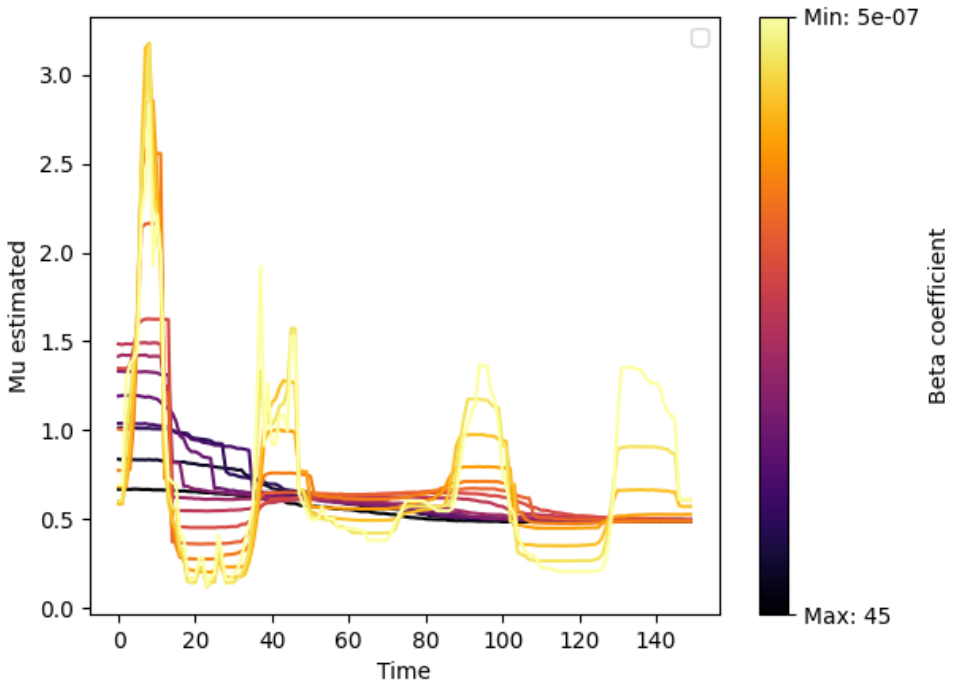


Figure: relaxation de la pénalité sur μ

Validation croisée

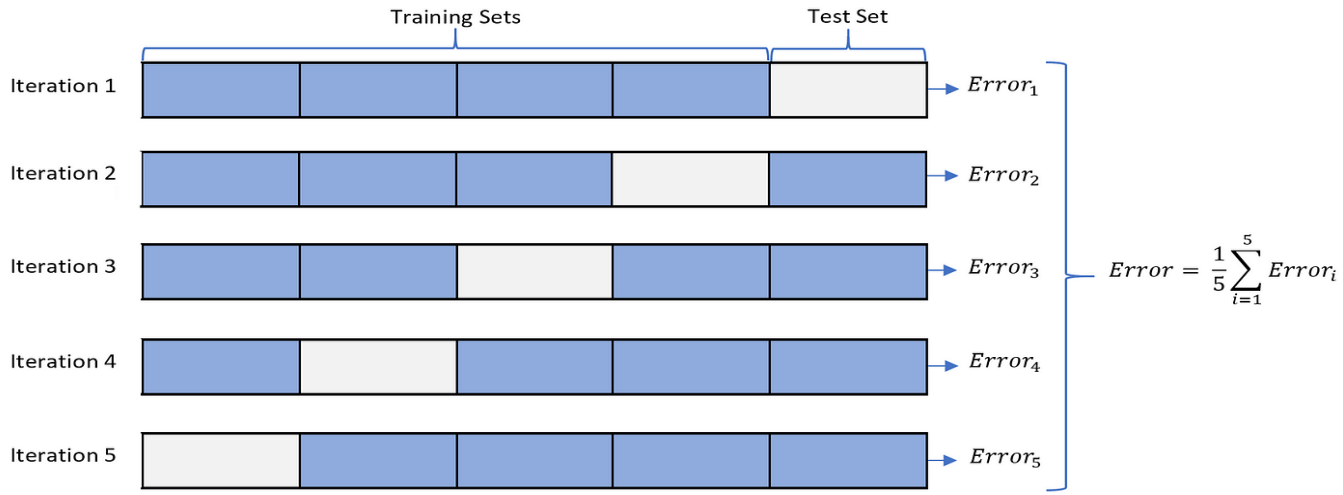


Figure: Validation croisée à 5 blocs

$$\underset{\gamma}{\text{Argmin}} \left\{ \frac{1}{B} \sum_{b=1}^{B} -\log(\mathcal{L}(\widehat{\theta}^b(\gamma), Y^{T,b})) \right\}$$

$$\text{s.t. } \widehat{\theta}^b(\gamma) \in \underset{\theta}{\text{Argmin}}\{J_\gamma(\theta, Y^{L,b})\}$$

Validation croisée : données i.i.d.

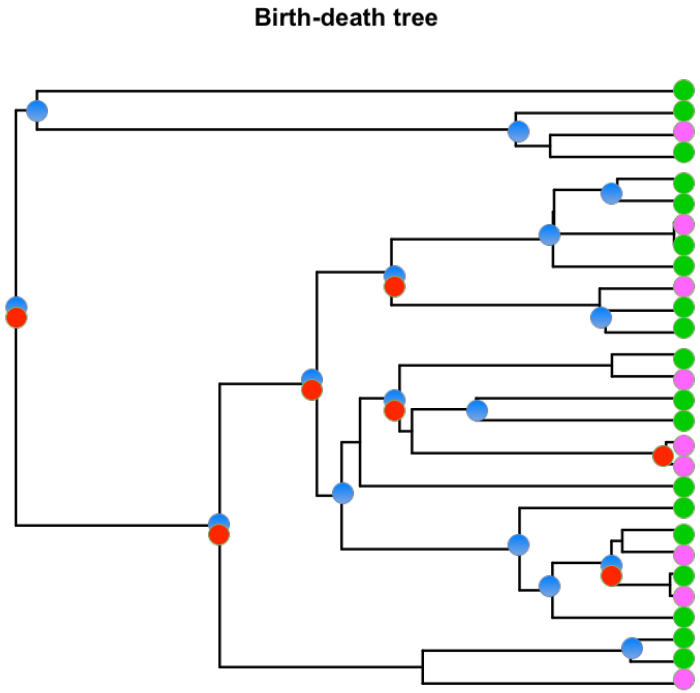


Figure: à partir des feuilles

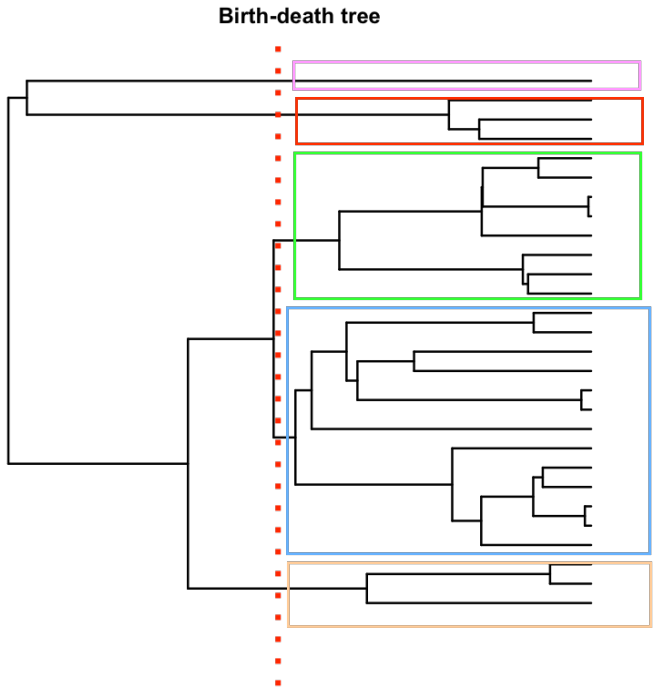


Figure: à partir des sous arbres

Validation croisée

effet de "saturation"

- au dela d'une certaine valeur, augmenter β n'améliore plus le score.
  $\Rightarrow$ favorise μ constant.
- Pas surprenant (μ plus dur à estimer que λ)

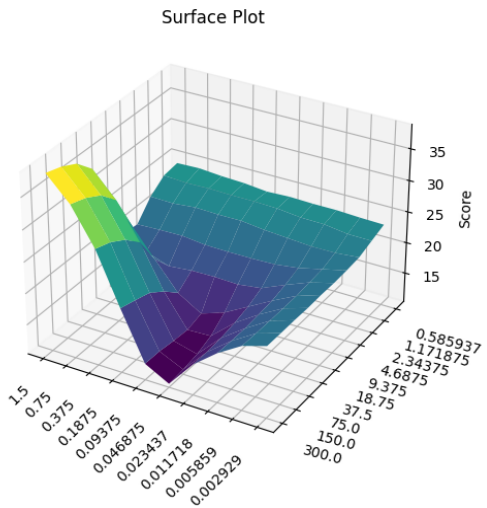


Figure: Score moyen sur les données test pour différentes valeurs de γ

Simulations

- Simulation arrêtée à $T = 12$ Myrs
- $\alpha = 0.2$, $\beta = 30$
- intervalles de confiances approximatifs : Jacknife

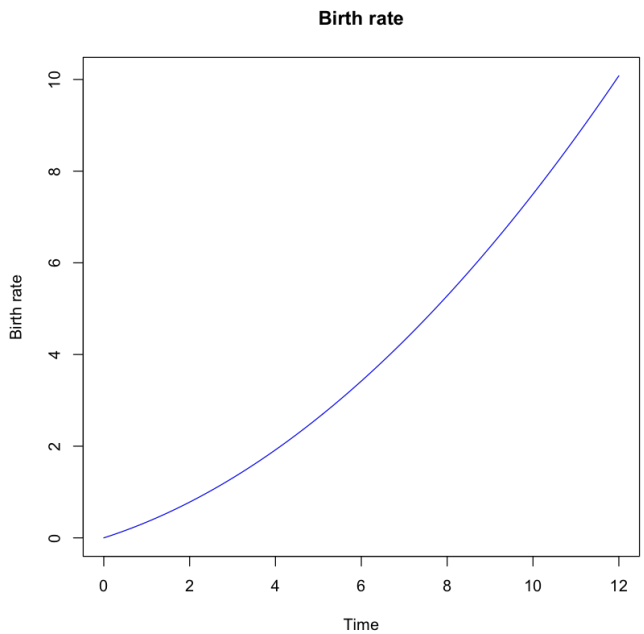


Figure: Taux de spéciation

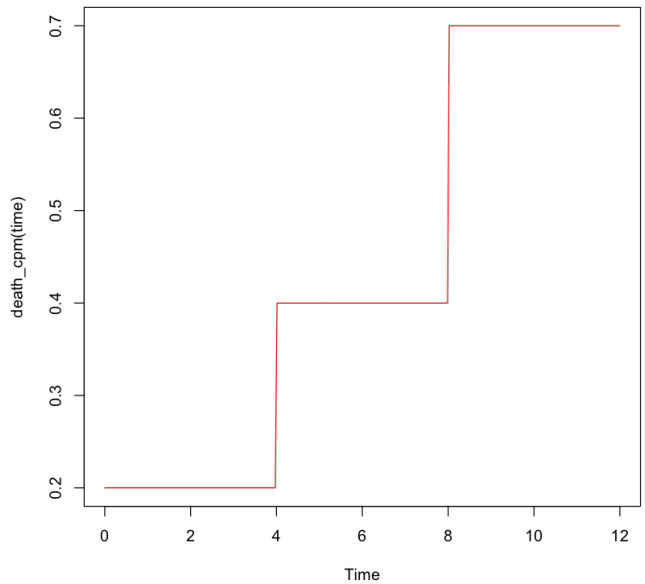


Figure: Taux d'extinction

estimation

Fonction objective non convexe $\Rightarrow$ point de départ de l'optimisation très important

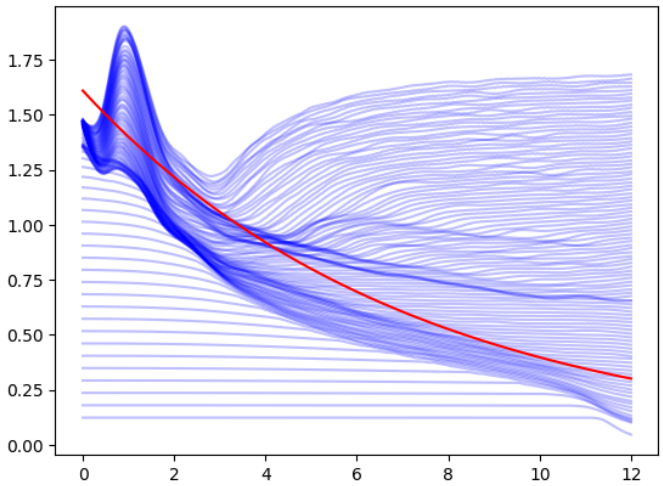


Figure: Estimations de λ en partant de différents points lors de l'optimisation

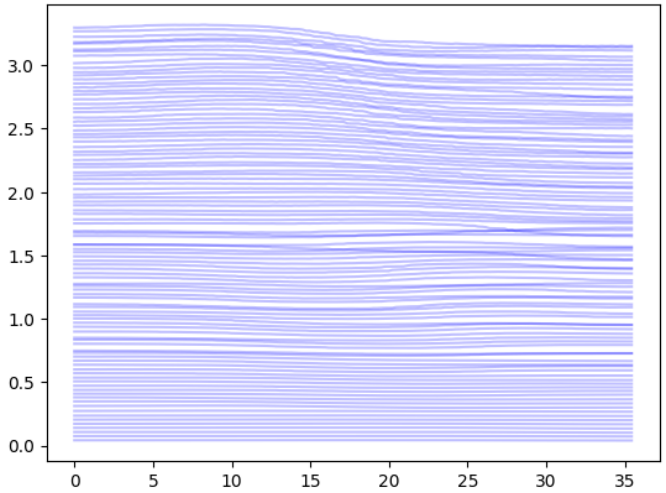


Figure: Estimations de μ en partant de différents points lors de l'optimisation

estimation

$\Rightarrow$ On garde la meilleure estimation

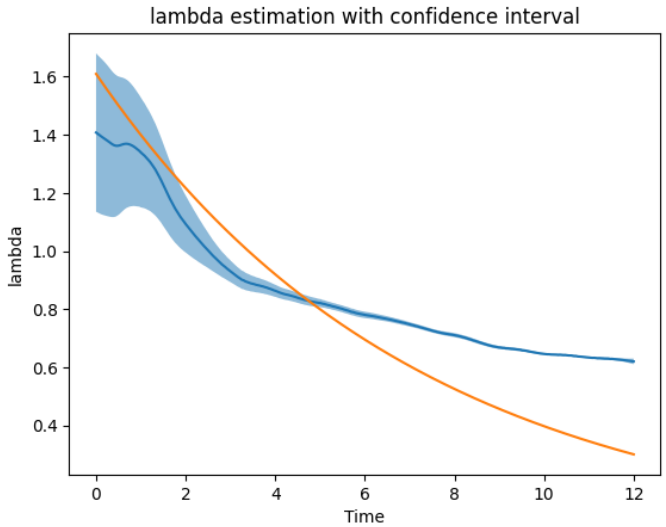


Figure: Estimation (bleu), vrai paramètres (orange) et intervalle de confiance à 95%

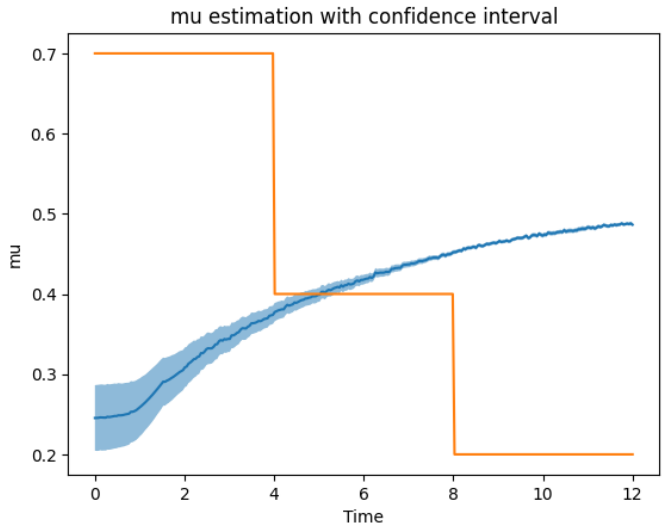


Figure: Estimation (bleu), vrai paramètres (orange) et intervalle de confiance à 95%

Cétacés

validation croisée

$p_1(\lambda) = ||\lambda''||_2$; $q_1(\mu) = ||\mu'||_1$, on trouve $\alpha = 0.04$, $\beta = 300$

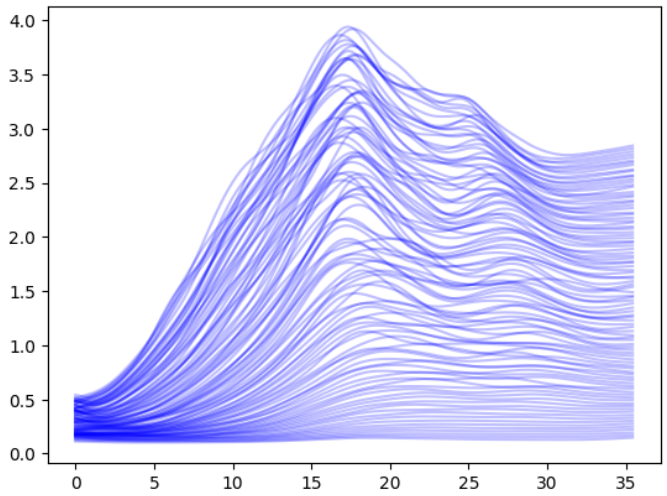


Figure: Estimations de λ en partant de différents points lors de l'optimisation

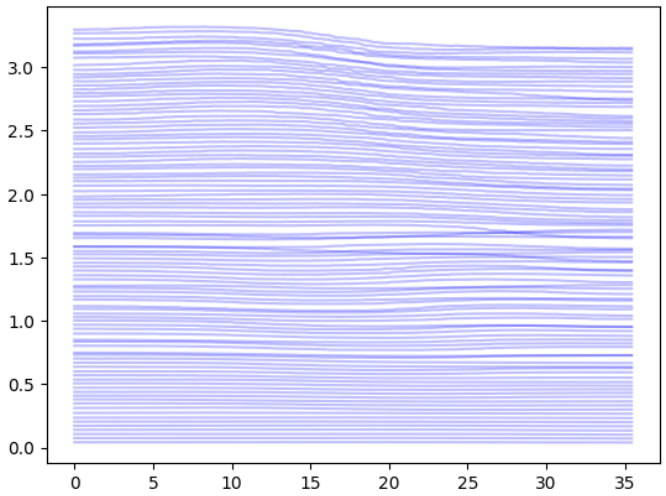


Figure: Estimations de μ en partant de différents points lors de l'optimisation

Comparaison temps géologiques et résultats existants

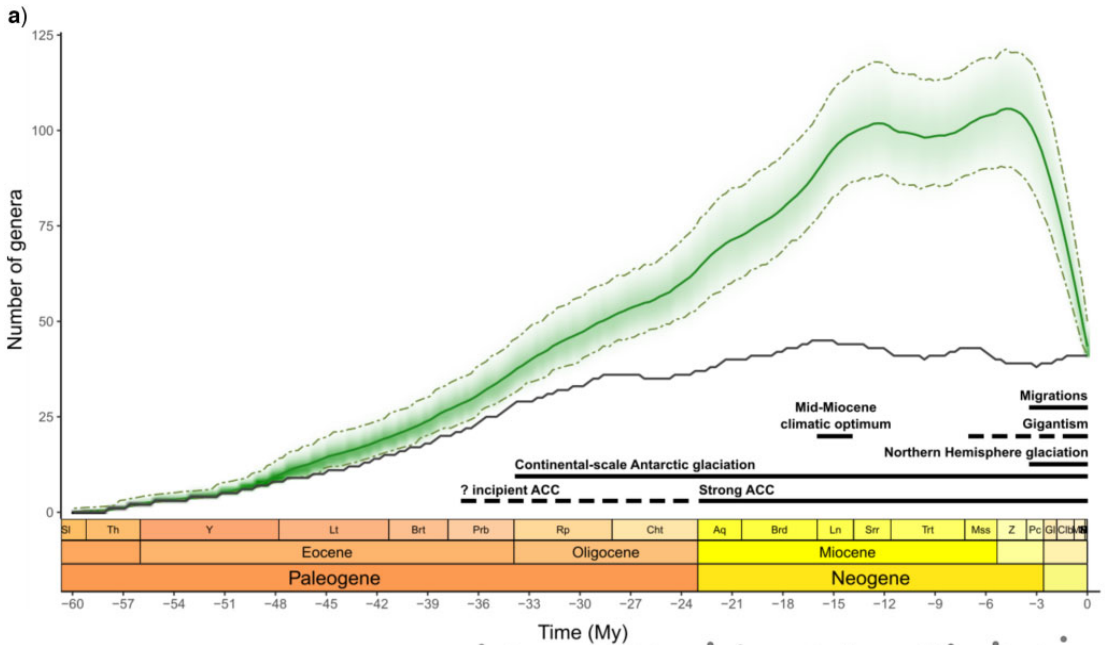


Figure: Diversification inférée (RevBayes) des cétacés. (Nombre total de genres au cours du temps avec les intervalles de confiance à 95%)

Comparaisons

Remarque

- Hypothèse forte (l'évolution a lieu selon un HBD) $\Rightarrow$ difficile de s'assurer de la plausibilité de ces résultats.
- Concordant avec les occurences fossiles

Conclusion

- Inférence purement statistique de ce type de modèle : pas possible $\Rightarrow$ utiliser des hypothèses biologiques (pénalisation, prior, formes paramétriques...)
- D'autres modèles : données fossiles (FBD), taux différents sur chaque branche etc.

Figure: Fossile d'un lézard mammalien herbivore et préhistorique

Bibliographie

[1] Tanja Stadler Amaury Lambert. "Birth-death models and coalescent point processes: the shape and probability of reconstructed phylogenies". In: *Theoretical Population Biology* 90 (2013), pp. 113–128. DOI: 10.1016/j.tpb.2013.10.002.

[2] Veronica Vinciotti Hamed Haselimashhadi. "A Differentiable Alternative to the Lasso Penalty". In: *ArXiv* (2016). DOI: arXiv:1609.04985.

[3] Todd L. Parsons Hélène Morlon and Joshua B. Plotkin. "Reconciling molecular phylogenies with the fossil record". In: *PNAS* 39 (2011), pp. 16327–16332. DOI: https://doi.org/10.1073/pnas.110254310.

[4] Stilianos Louca and Matthew W. Pennell. "Extant timetrees are consistent with a myriad of diversification histories". In: *Nature* 580 (2020), pp. 502–505. DOI: https://doi.org/10.1038/s41586-020-2176-1