Solving Biological Problems that require Math 2010

« Speciation and extinction rate estimation using phylogenetic trees »

Semestre de printemps 2010

Supervisé par Micha Hersch et Nicolas Salamin

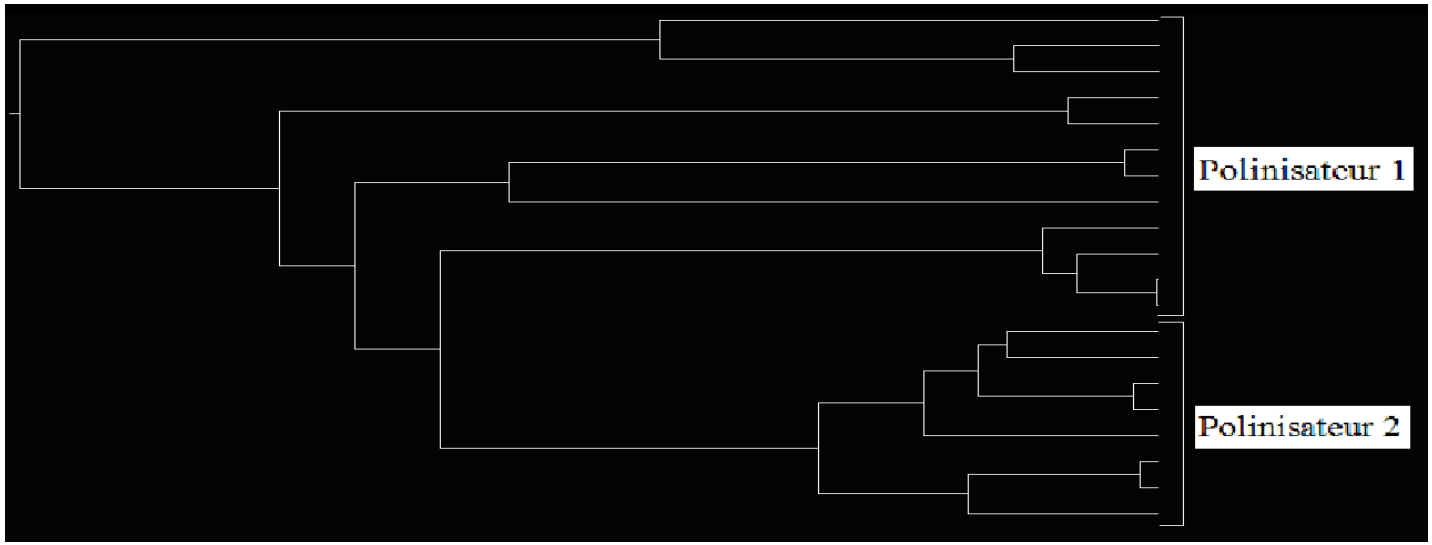
herve.cachin@unil.ch et trestan.pillonel@unil.ch

Solving Biological Problems that require Math 2010

Plan :

1. But
2. Background
3. Méthodologie
4. Résultats
5. Perspectives

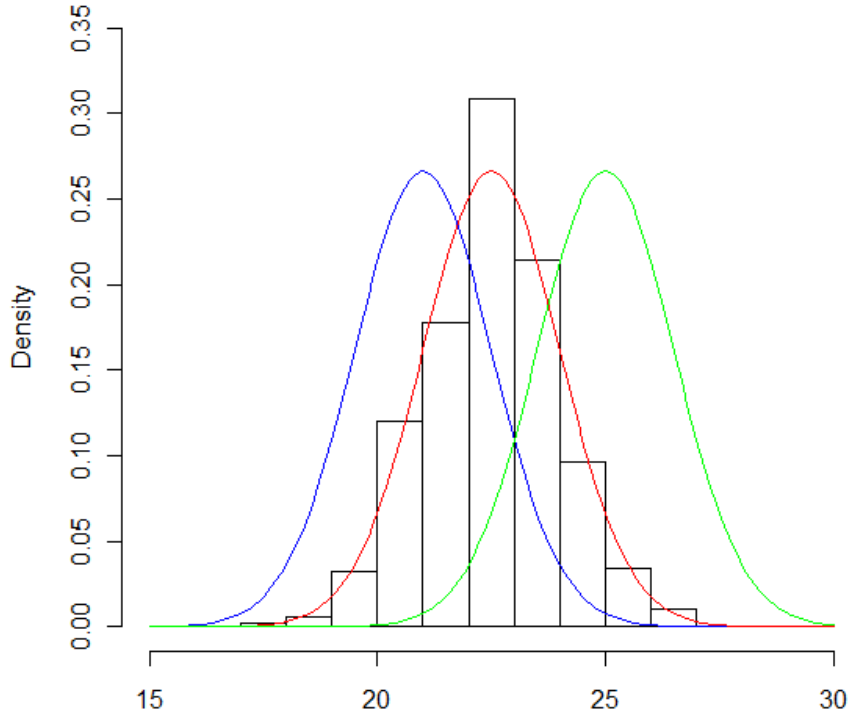
- Parvenir à estimer les paramètres les plus probables du taux d'extinction et d'apparition de nouvelles espèces connaissant un arbre phylogénétique
- Comparer et évaluer différentes méthodes.
- Découvrir si un trait spécifique (type de polinisateur p.ex) à un taxon peut influencer le taux d'extinction (β) et le taux de spéciation (μ)



- Compréhension de certains phénomènes évolutifs et des facteurs qui les influencent
- Utilisation et interprétation des données phylogénétiques
- Meilleure compréhension de l'origine de la biodiversité

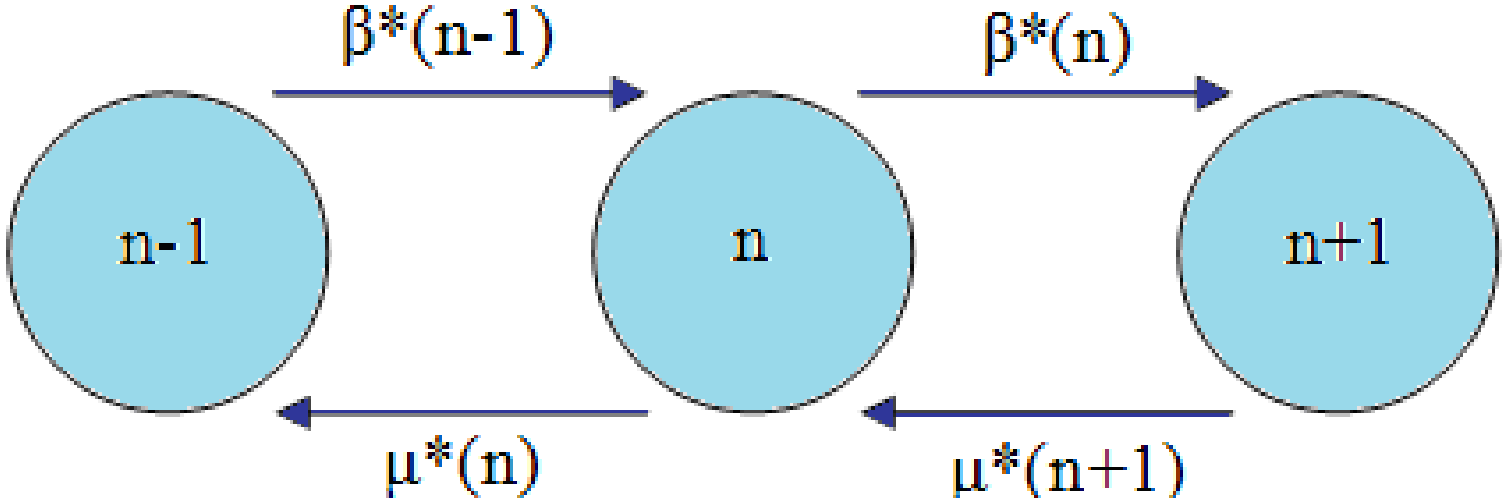
Le maximum de vraisemblance

- Méthode mathématique servant à déterminer les paramètres de la densité de probabilité (distribution) d'un échantillon.
- fonction de vraisemblance = probabilité d'observer ce que l'on observe selon différents paramètres
  - Exemple: On connaît la taille de 21 personnes :



- Connaissant la variance, on veut estimer la moyenne qui expliquerait au mieux nos 21 observations sachant que la taille suit une distribution normale

Modèle « birth and death » :



- On extrait la relation suivante :

$$P_n(t+\Delta t) = P_n(t)+P_{n-1}(t)*(n-1)*\beta + P_{n+1}(t)*(n+1)*\mu - n*P_n(t)*(\beta+\mu)$$

$$dP_n(t)/\delta t = P_{n-1}(t)*(n-1)*\beta + P_{n+1}(t)*(n+1)*\mu - n*P_n(t)*(\beta+\mu)$$

NB: $p_n(t)$ est la probabilité qu'un lignage ait exactement n descendant après t unités de temps.

Modèle « birth and death »

- On résout et on obtient (Kendall, 1948):

$$P_n(t) = (1-P_0)*(1-u)*u^{n-1}$$

Où $u = \beta*(e^{(\beta-\mu)*t}-1) / \beta*e^{(\beta-\mu)*t} - \mu$

$P_0 = \mu*(e^{(\beta-\mu)*t}-1) / \beta*e^{(\beta-\mu)*t} - \mu$

(Pour $n > 0$)

- On cherche ensuite $P_1(t)$ qui est la probabilité qu'un lignage ait exactement n descendants après t unités de temps.

$$P_1(t) = (1-P_0)*(1-u)*1$$

- On résout et on obtient :

$$P_1(t) = e^{(\beta-\mu)*t}*(\beta-\mu)^2 / (\beta* e^{(\beta-\mu)*t} - \mu)^2$$

En résumé :

- Modèle birth and death d'où l'on extrait certaines relations
- On obtient une équation différentielle que l'on résout afin d'obtenir la formule permettant de calculer la probabilité qu'un lignage ait exactement n descendants après t unités de temps.
- Puis on la développe ce résultat pour $P_1(t)$ afin d'obtenir la probabilité qu'un lignage ait exactement 1 descendant après t unités de temps .

Explications notations

- On a un arbre avec des longueurs de branches, on connaît la formule donnant P1(t). On veut trouver une fonction de vraisemblance nous donnant une bonne estimation des paramètres de l'arbre.

- On a une formule donnant la probabilité qu'un lignage ait exactement 1 descendant après t unités de temps (On a donc une formule donnant la probabilité d'observer une branche d'arbre à un temps t) :

$$P_1(t) = e^{(\beta-\mu)*t}*(\beta-\mu)^2 / (\beta* e^{(\beta-\mu)*t} - \mu)^2$$

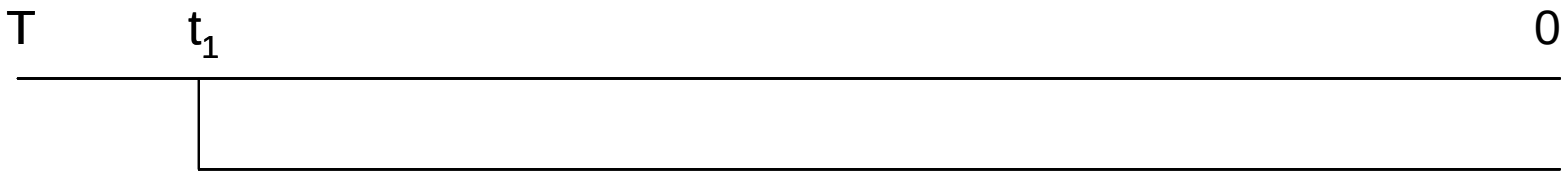
Mise en pratique

T 0

- On peut développer cela de la façon suivante :

$P_1(T)$

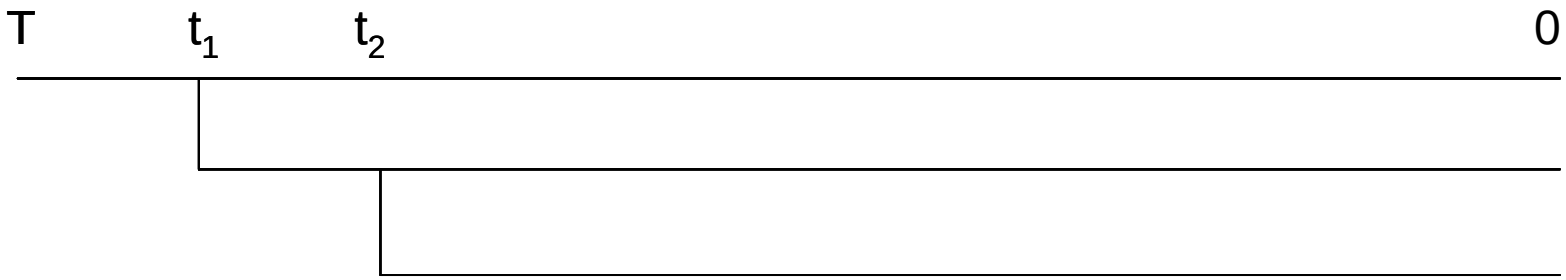
Mise en pratique



- On peut développer cela de la façon suivante :

$P_1(T) * \beta * \Delta t * P_1(t_1) * 1$

Mise en pratique



- On peut développer cela de la façon suivante :

$P_1(T) * \beta * \Delta t * P_1(t_1) * 1 * \beta * \Delta t * P_2(t_2) * 2$

Mise en pratique

- On peut développer cela de la façon suivante :

$P_1(T) * \beta * \Delta t * P_1(t_1) * 1 * \beta * \Delta t * P_2(t_2) * 2 * \beta * \Delta t * P_3(t_3) * 3$

= probabilité d'observer cet arbre aux temps t1, t2 et t3

Mise en pratique

- On peut développer cela de la façon suivante :

$P_1(T) * \beta * \Delta t * P_1(t_1) * 1 * \beta * \Delta t * P_2(t_2) * 2 * \beta * \Delta t * P_3(t_3) * 3$

= probabilité d'observer cet arbre aux temps t1, t2 et t3

- Généralisation (formule 1):

$$P_1(T) * (n-1)! * \beta^{n-1} * \prod P_1(ti)$$

Mise en pratique

- Il existe une seconde méthode qui permet d'éviter le problème du T en le faisant tendre vers l'infini.
- Parmi tous les arbres on va en quelque sorte enlever ceux qui n'ont pas n espèces. (auparavant on a travaillé avec tous les arbres mais maintenant on n'utilise que ceux qui ont n branches)→Formule 2.1 :

$$P(\{t_i\}) = \frac{(e^{(\beta-\mu)*ti}*(\beta-\mu)^2)}{(\beta* e^{(\beta-\mu)*ti} - \mu)^2} * \frac{(\beta*e^{(\beta-\mu)*T} - \mu)}{(e^{(\beta-\mu)*T} - 1)}$$

- En faisant tendre T vers l'infini, on peut se débarrasser du problème du T → Formule 2.2

En résumé :

- Formule 1 : Chance d'observer cet arbre avec les paramètres β et μ.

$$\text{Prob}\,(n, t_1, t_2, \ldots, t_{n-1} \mid \beta, \mu, T) \;=\; p_1(T)\,\beta^{n-1}\,(n-1)!\,\prod_{i=1}^{n-1} p_1(t_i)$$

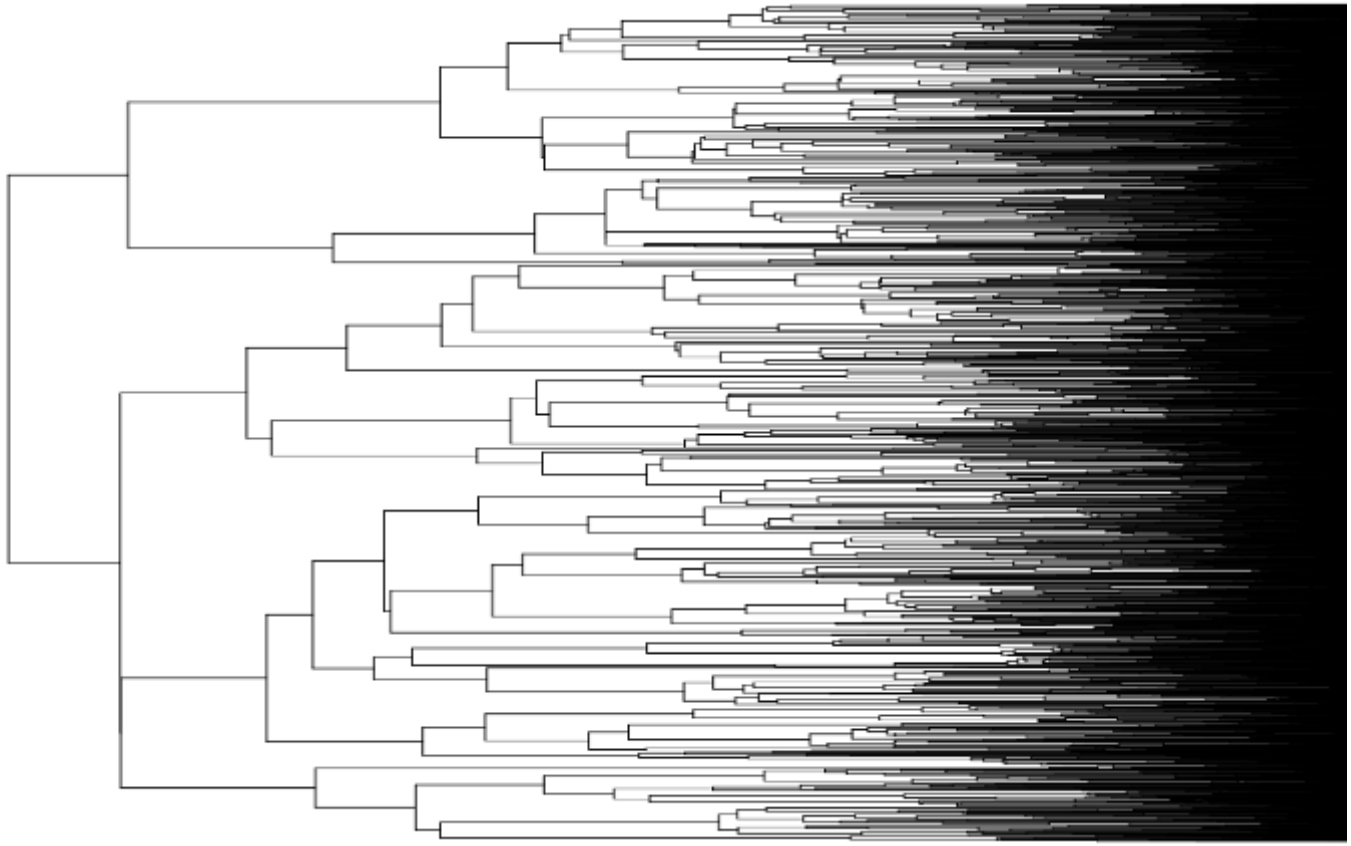
- Formule 2.1 : Chance d'observer cet arbre en sachant à l'avance qu'on va observer un arbre à n branches avec β et μ (on réduit le nombre de possibilité).

$$\text{Prob}\,(t_1, t_2, \ldots, t_n \mid n, \beta, \mu, T) \;=\; \prod_{i=1}^{n-1} \left[\frac{p_1(t_i)\,dt_i}{\int_0^T p_1(u)\,du} \right]$$

- Formule 2.2 : Formule 2.1 en faisant tendre T vers l'infini

Simulations :

- Nous avons testé la capacité des 3 différentes formules d'estimer les taux de spéciation et d'extinction sur la base d'arbres générés avec des paramètres définis.



Résultats

- Les premiers tests ont montré qu'il est difficile d'estimer les taux sur la base de petits arbres, nous avons donc fait de nouveaux tests sur de plus grands arbres.

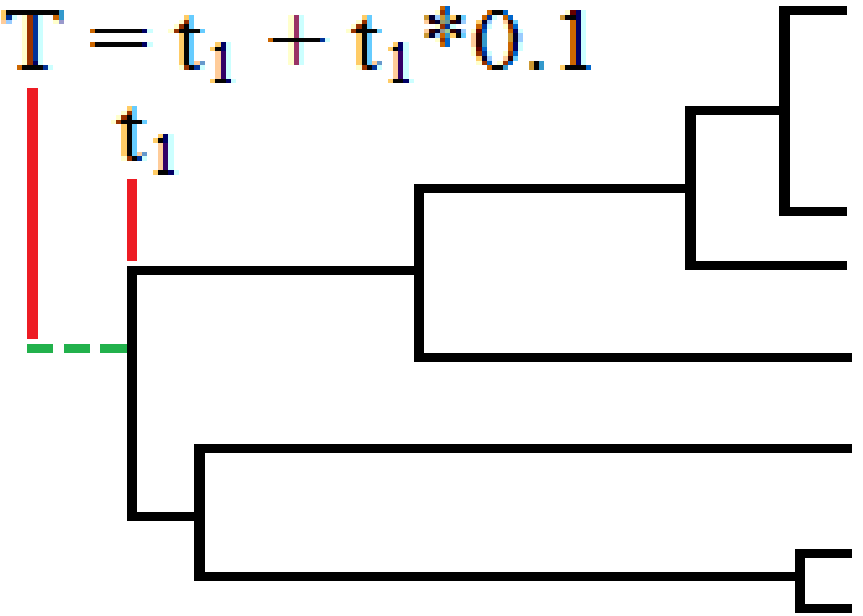
Résultats

- Paramètres pour la génération de plus grands arbres (100 arbres en tout) :

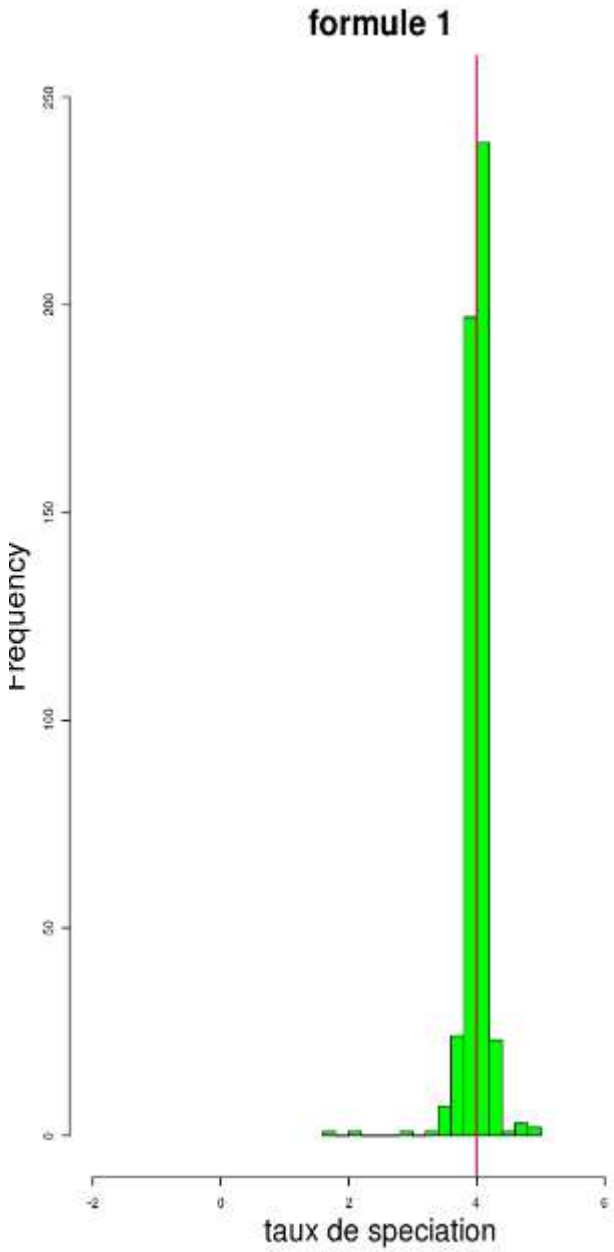
  - Taux de spéciation (β): 4
  - Taux d'extinction (μ): 1

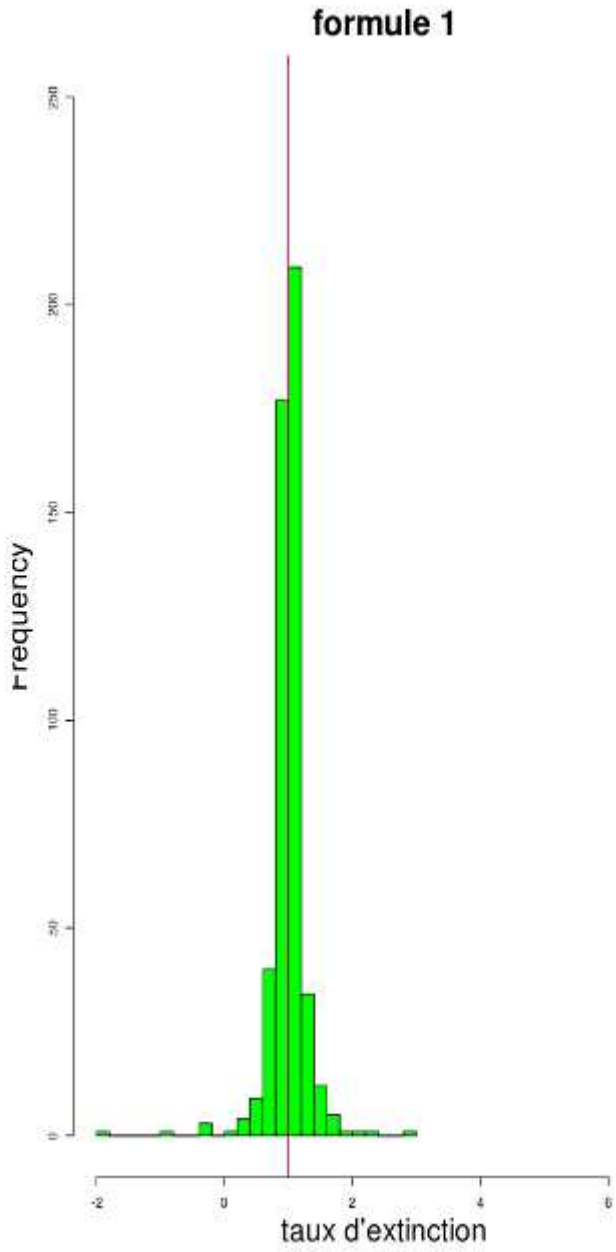
- De cette manière, nous avons obtenu des arbres ayant entre 135 et 29749 espèces (plus de 7000 en moyenne).

- Nous avons fait les 3 tests avec un T 10% supérieur au 1er nœud de l'arbre.

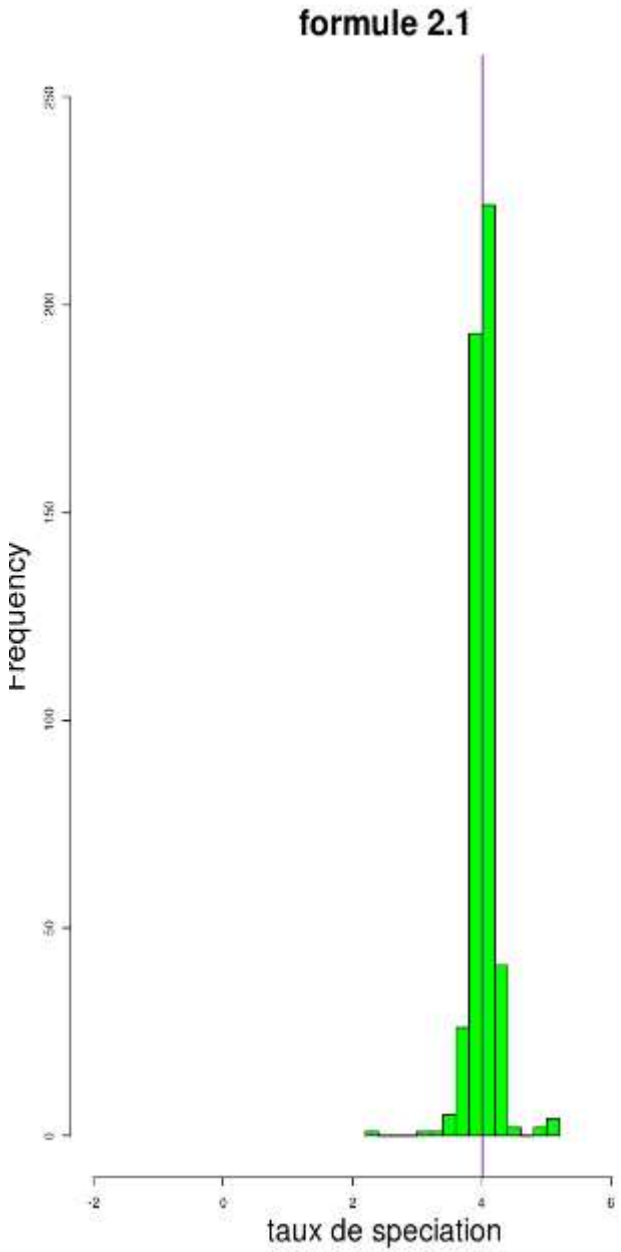


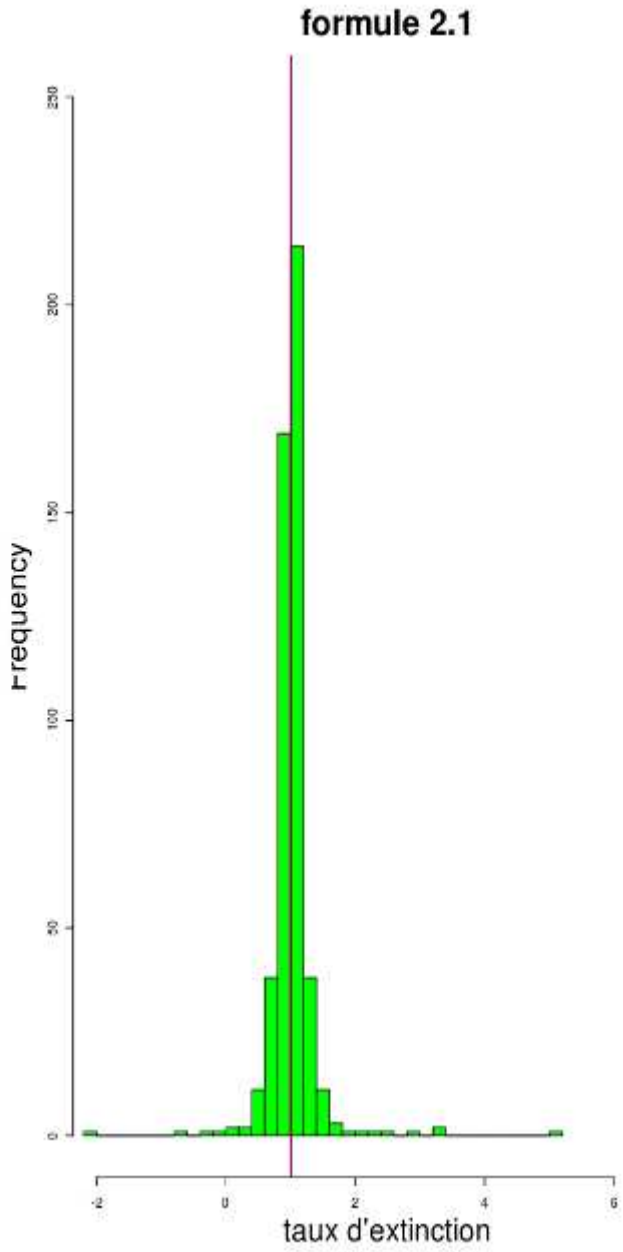
- Formule 1
- Résultats comparables à la formule 2.2
- Variance :
  - Taux de spéciation: 0.216207
  - Taux d'extinction: 0.2979145



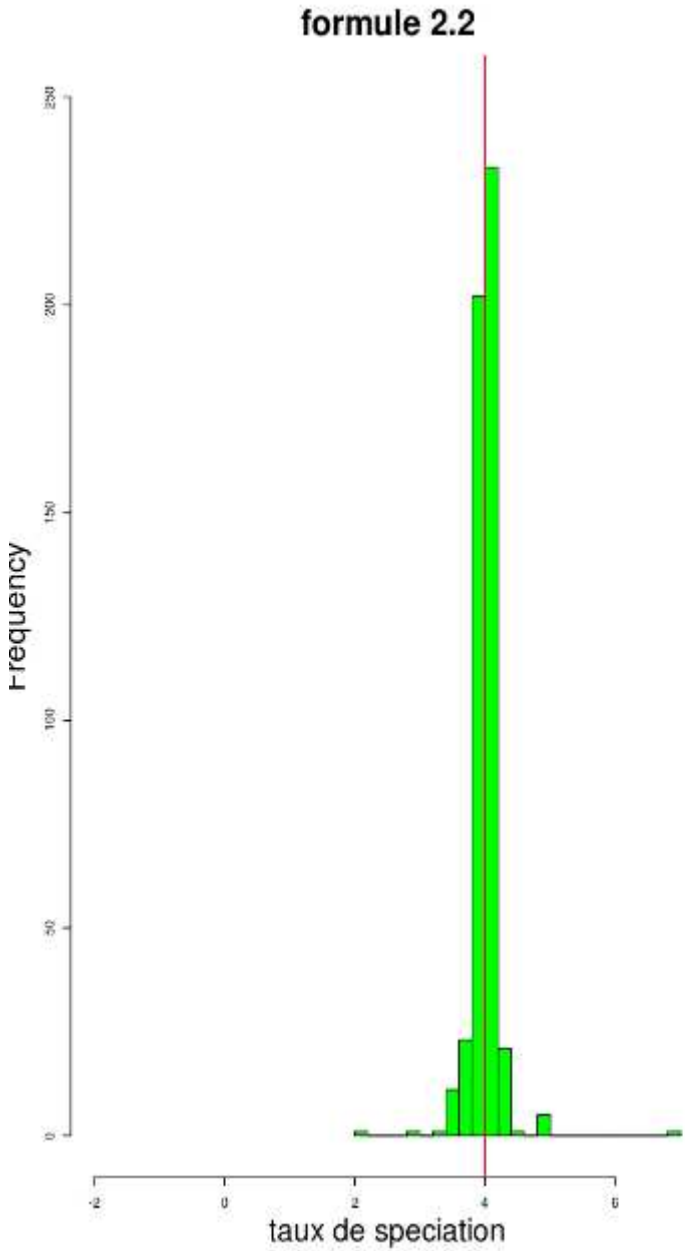


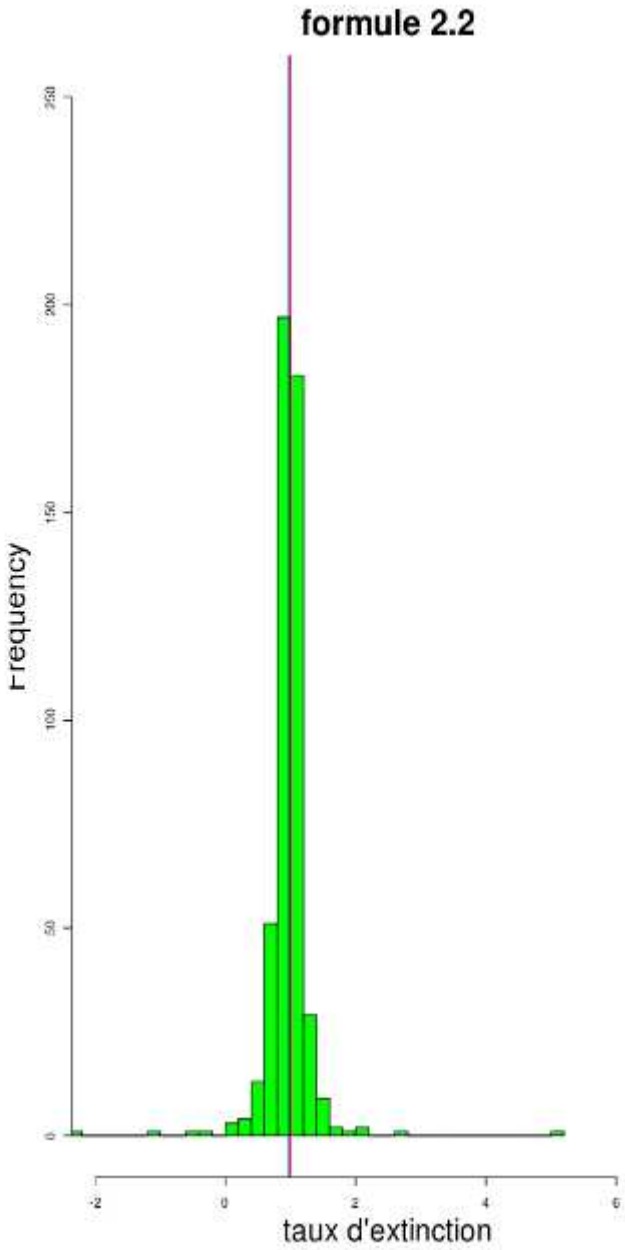
- Formule 2.1
- Résultats assez bons
- Variance :
  - Taux de spéciation: 0.2033684
  - Taux d'extinction: 0.3755941



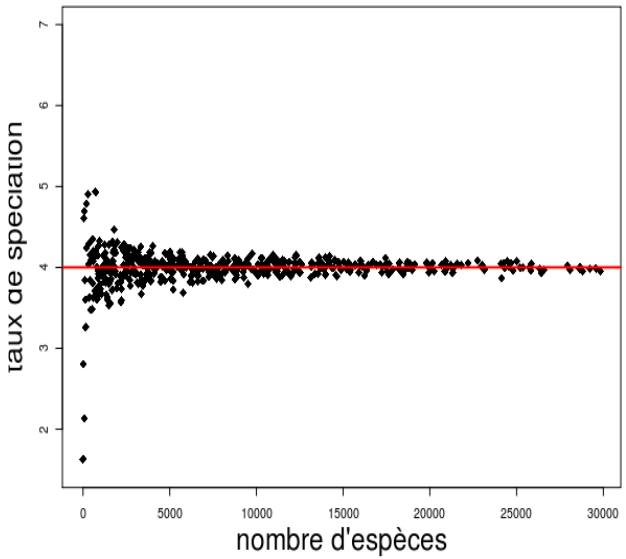


- Formule 2.2
- Résultats assez bons
- Variance :
  - Taux de spéciation: 0.2322773
  - Taux d'extinction: 0.3537552

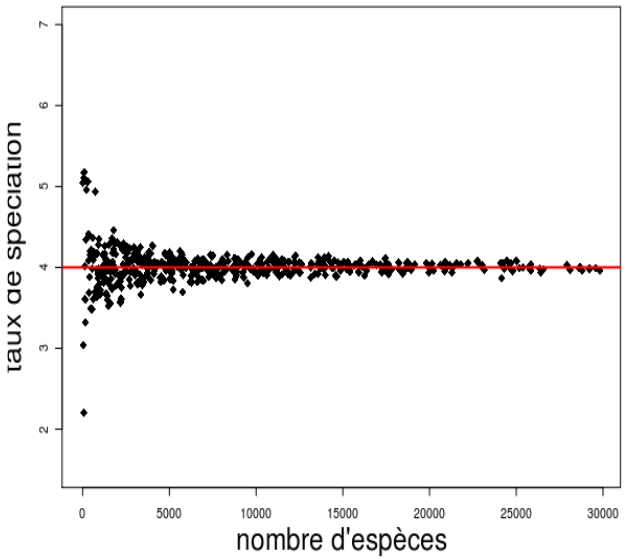




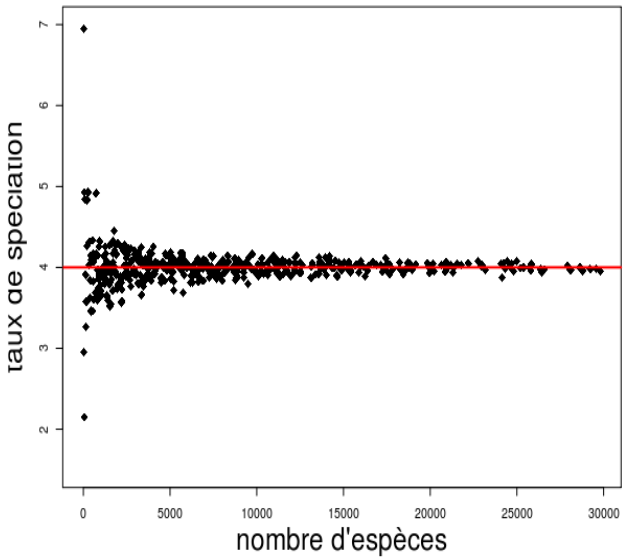
formule 1



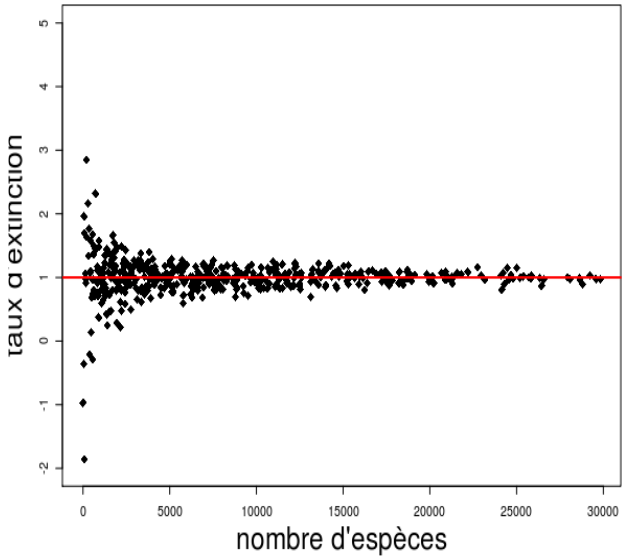
formule 2.1



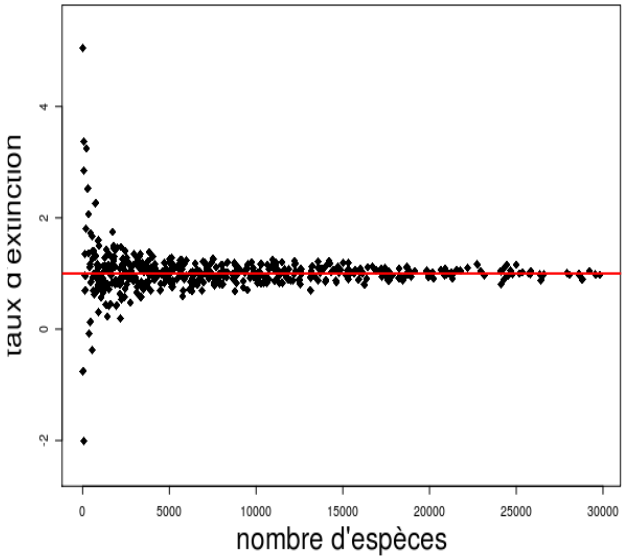
formule 2.2



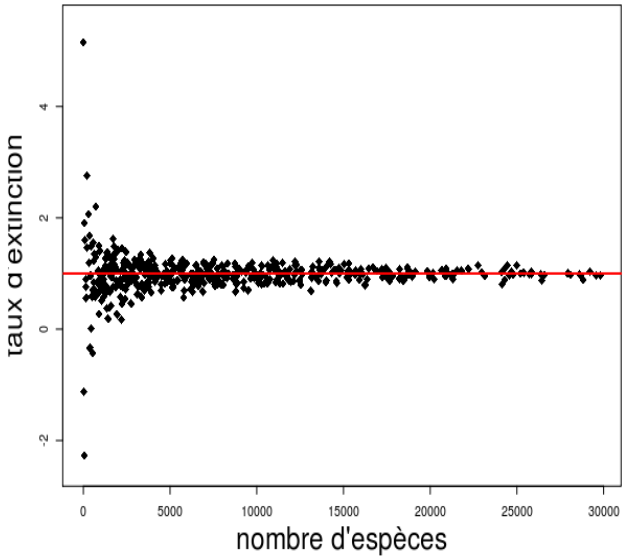
formule 1

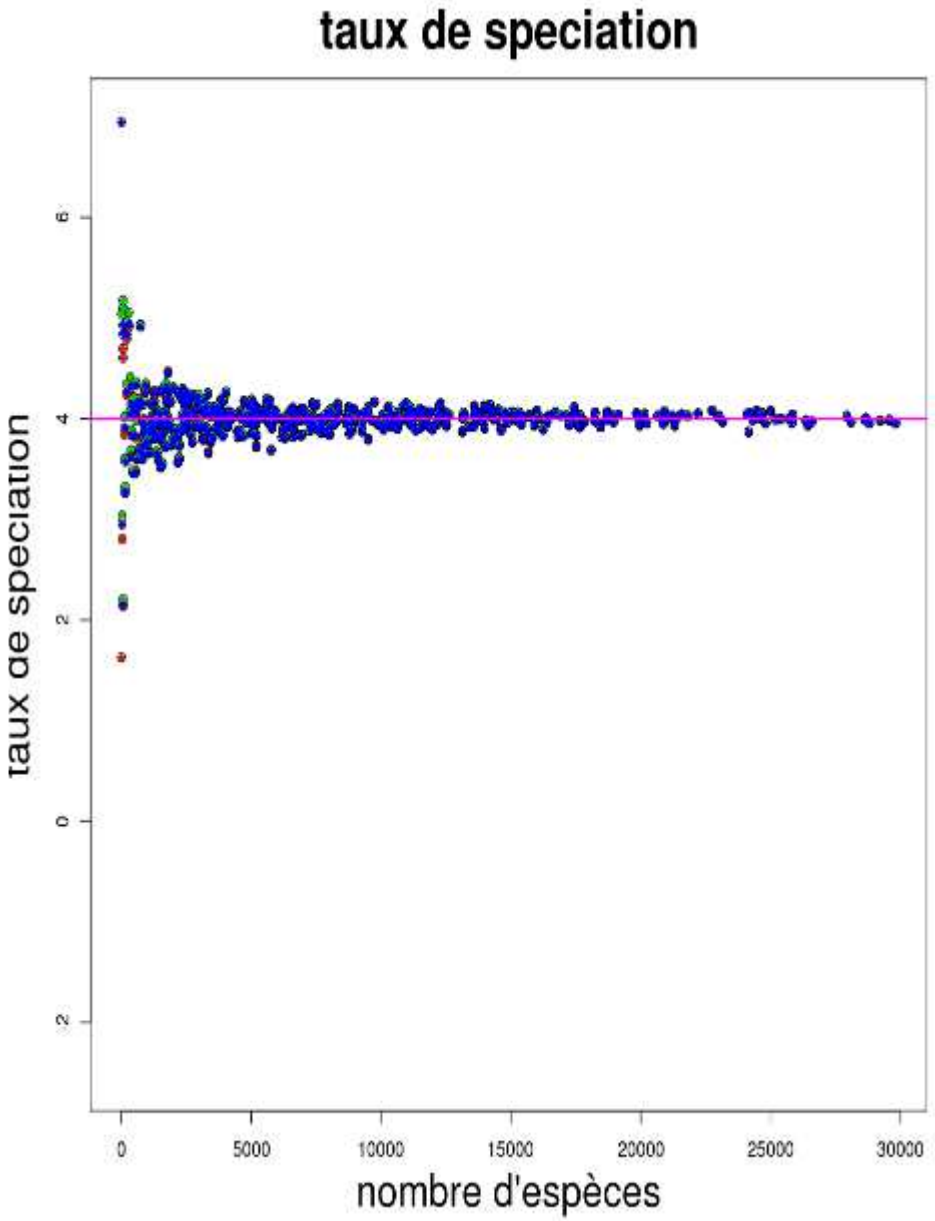


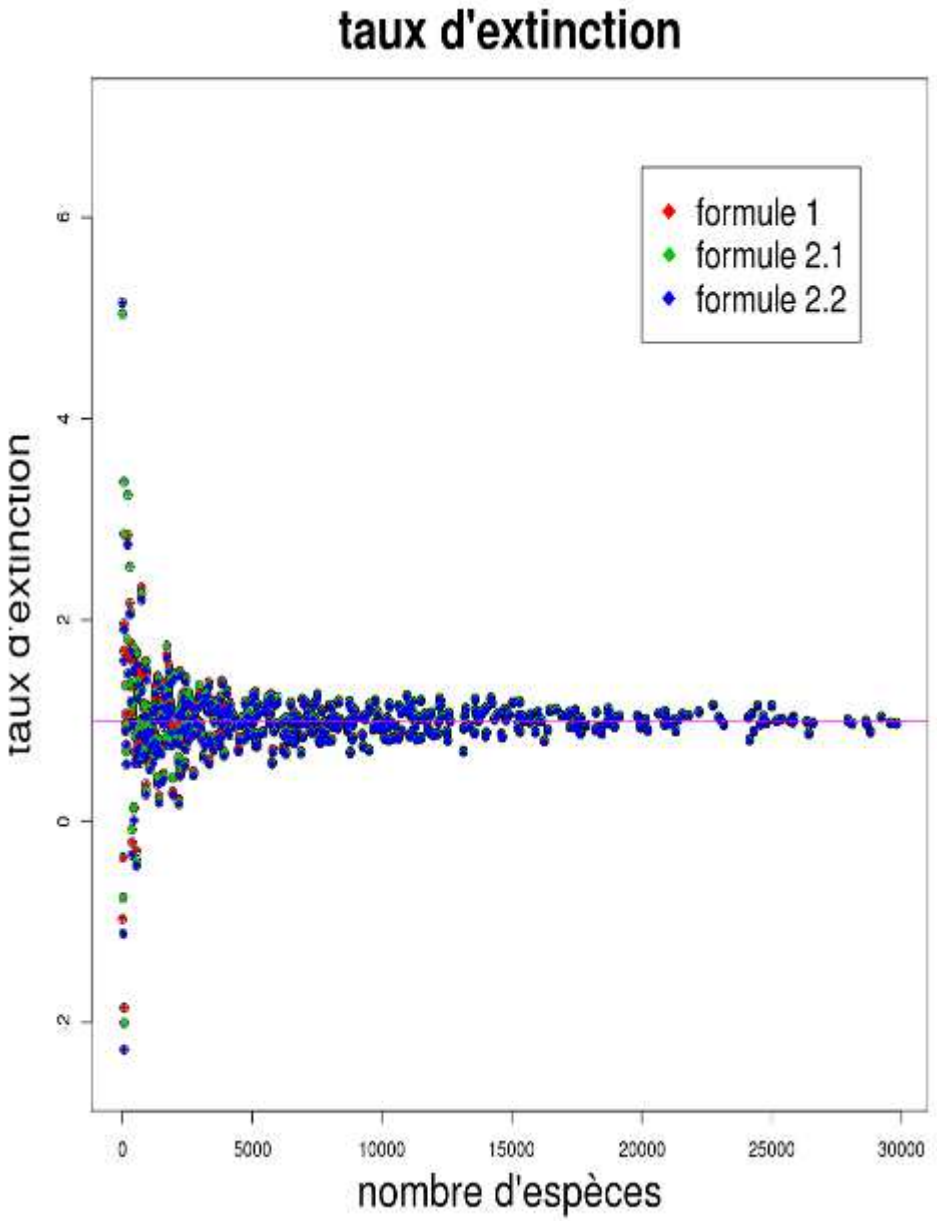
formule 2.1

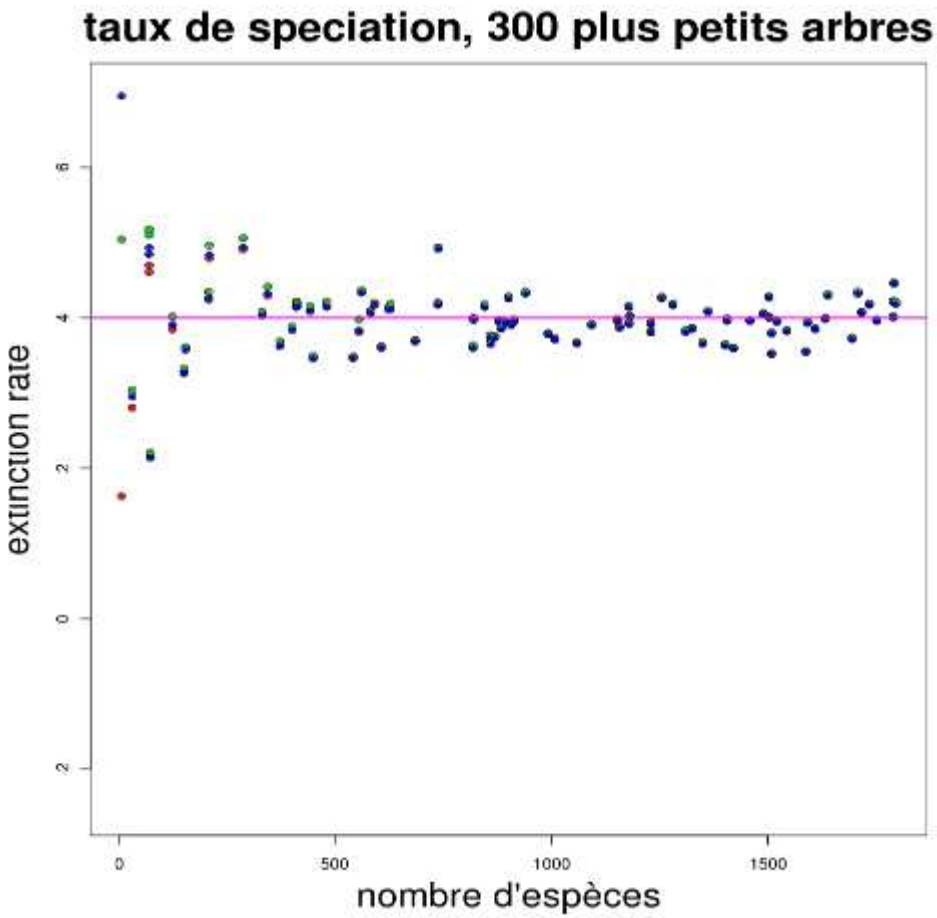


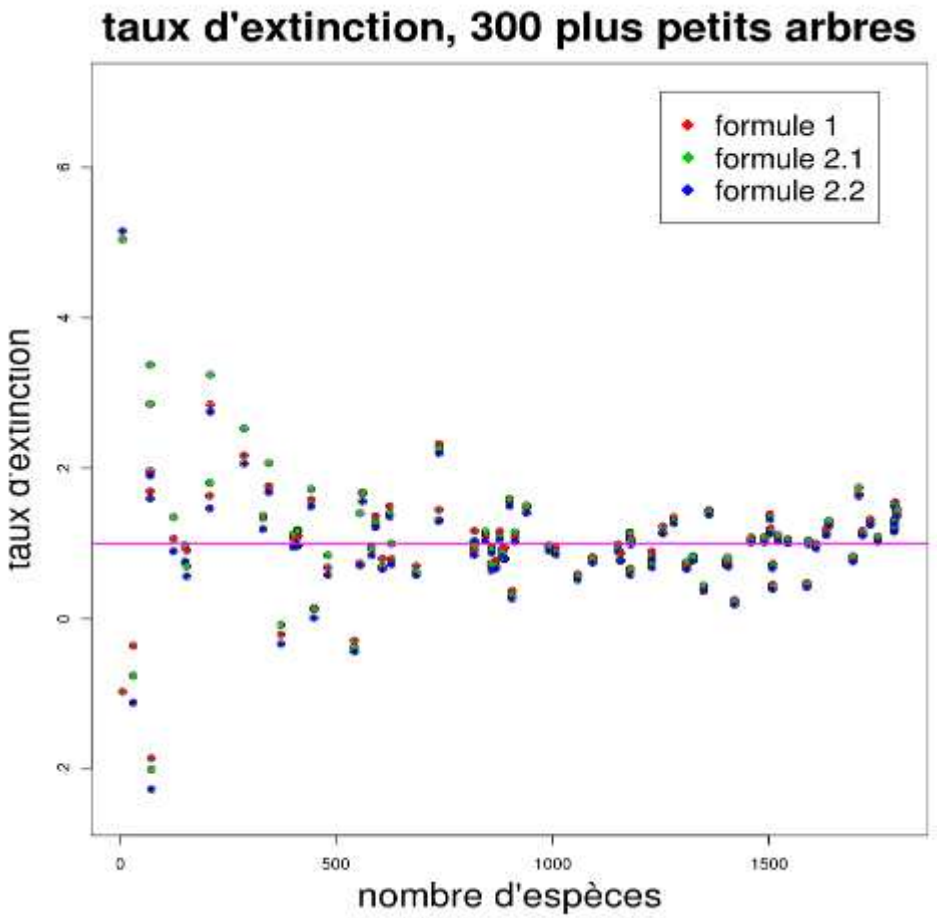
formule 2.2



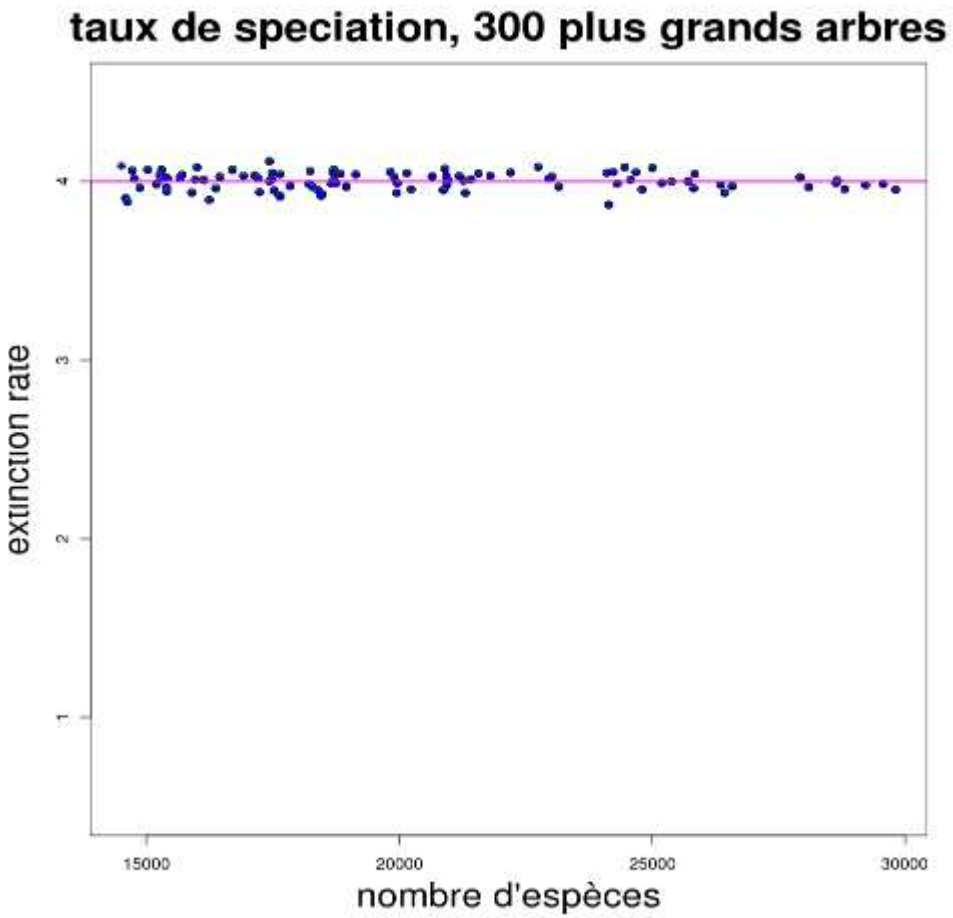

taux de speciation
taux de speciation
nombre d'espèces
0
5000
10000
15000
20000
25000
30000
2
0
2
4
6

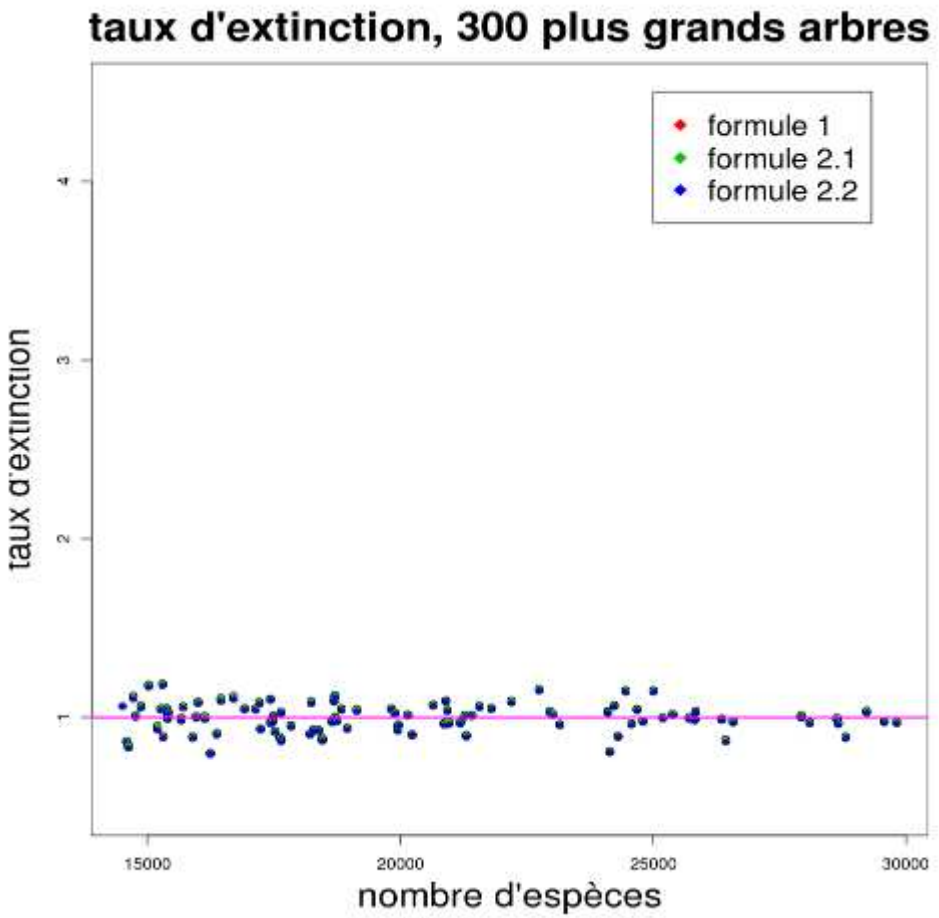

taux d'extinction
taux d'extinction
nombre d'espèces
formule 1
formule 2.1
formule 2.2
0
5000
10000
15000
20000
25000
30000
2
0
2
4
6



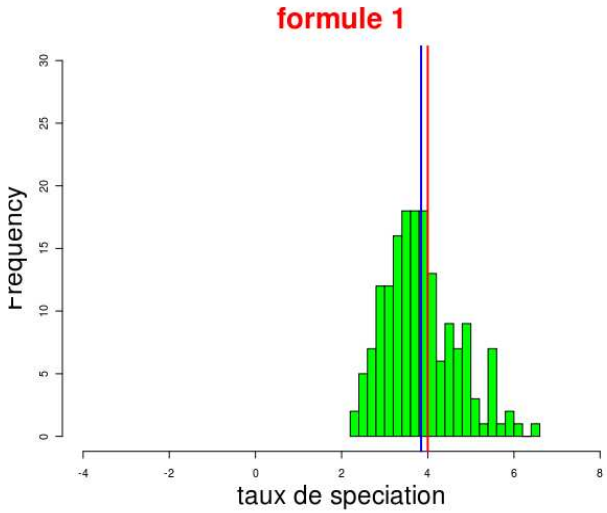


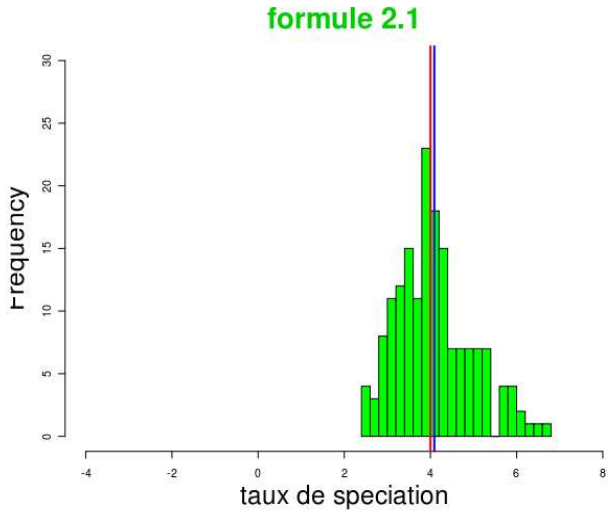
	Formule 1	Formule 2.1	Formule 2.2
Var. spéciation	0.4357404	0.4038708	0.4765981
Var. extinction	0.5858463	0.7716354	0.7251895

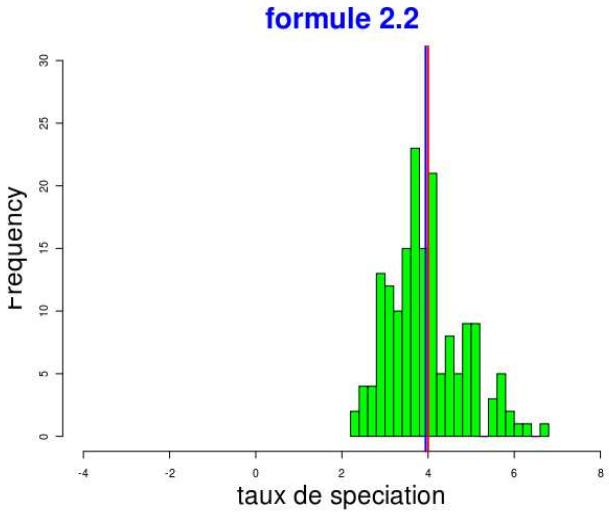


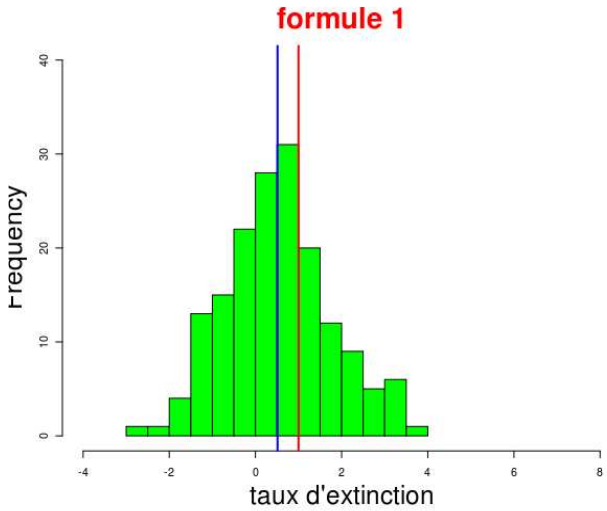


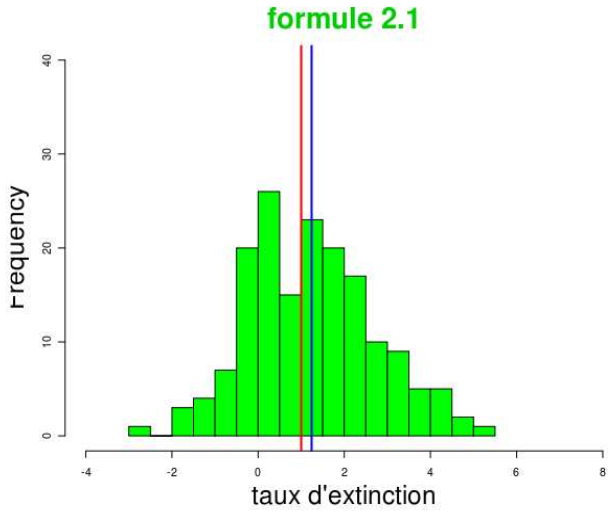
	Formule 1	Formule 2.1	Formule 2.2
Var. spéciation	0.0486401	0.04864382	0.0491386
Var. extinction	0.07917239	0.07966314	0.0790816

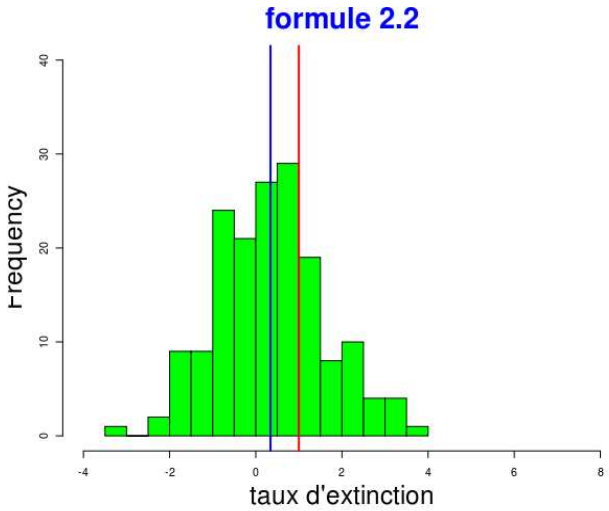












	Formule 1	Formule 2.1	Formule 2.2
Var. spéciation	0.845547	0.8787054	0.867201
Var. extinction	1.225961	1.474013	1.251919

Synthèse :

- Apparemment les trois méthodes semblent efficaces pour traiter des grands arbres mais la variance de l'estimation est beaucoup plus forte pour les arbres plus petits (limite→2500 espèces).
- La formule 1 semble tout de même supérieure
- Le taux de spéciation est toujours mieux estimé

- Il serait intéressant de refaire les simulations avec d'autres taux de spéciation et d'extinction → ordinateur puissant recommandé!
- Il serait intéressant de voir si certaines études s'étant basées sur les formules de vraisemblance que nous avons testées ont utilisé des arbres trop petits ce qui a très bien pu les mener à des conclusions erronées.
- Chercher une formule de vraisemblance fonctionnant bien pour des arbres phylogénétiques de taille normale

Merci de votre attention

Questions ?

trestan.pillonel@unil.ch et herve.cachin@unil.ch