Modèle de démogénétique environnementale

pour l'étude des processus d'invasion biologique

Arnaud Becheler

Chaire MMB

15 Février 2018

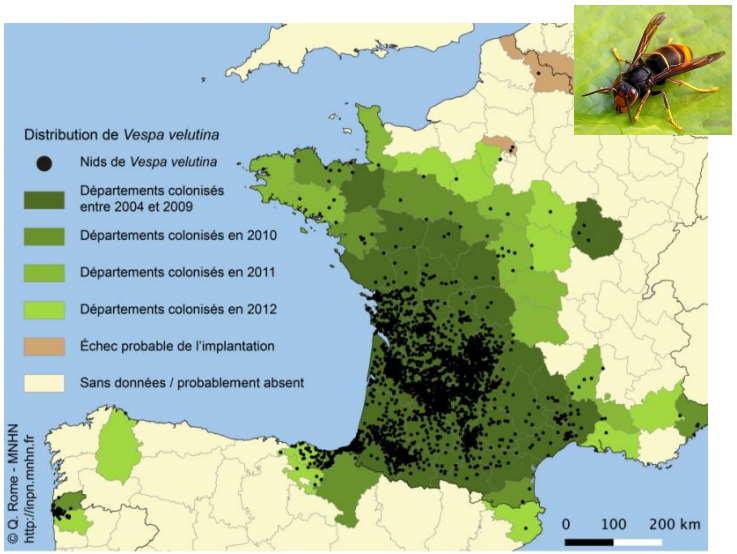
Invasion biologique

- Accroissement durable de l'aire de répartition d'un taxon/espèce.
- Processus faisant partie intégrante de l'évolution de la biodiversité.
- Forte accélaration du processus depuis quelques siècles/décennies.
- Seconde cause du déclin de la biodiversité.

Invasion du frelon à pattes jaunes

Vespa velutina nigrithorax apparaît dans une commune du Lot-et-Garonne en 2004 (Arca et al. 2015).
Il s'acclimate bien est commence à s'installer puis se répandre en France et au-delà.

- Problèmes économiques réels (prédation des abeilles domestiques)
- Problèmes écologiques encore à quantifier
- Risque sanitaire limité



Invasion du frelon à pattes jaunes

Questions d'intérêt écologiques encore mal explorées

- Modalités de dispersion (sans doute longue distance) ?
- Réaction à l'environnement (taux de croissance) ?
- Extrapolation géographique/temporelle de l'invasion ?

Etudier la génétique pour comprendre l'écologie ?

- On cherche à exploiter l'information contenue dans un jeu de données génétique.
- C'est *a priori* possible parce que l'écologie d'une espèce conditionne sa démographie, qui elle-même affecte les patrons génétiques des populations.

Données génétiques disponibles

	u_1	u_2	u_3	u_4
Lat	40	30	50	50
Lon	0	1	1	1
A_1	10	12	10	10
A_2	10	12	8	12
B_1	320	324	320	320
B_2	324	320	328	328

FIGURE – Exemple de jeu microsatellite : quatre individus diploïdes (colonnes) échantillonnés à différentes localisation (Lon, Lat) ont été génotypés à deux loci ($\{A_1, A_2\} et \{B_1, B_2\}$)

Modèle rendant compte de ces données ?

Hétérogénéité spatiale et temporelle

- Le paysage est discrétisé en I dèmes.
- On se donne L variables environnementales.
- On note $E_{l,i}$ la valeur de la variable environnementale l dans le dème i.

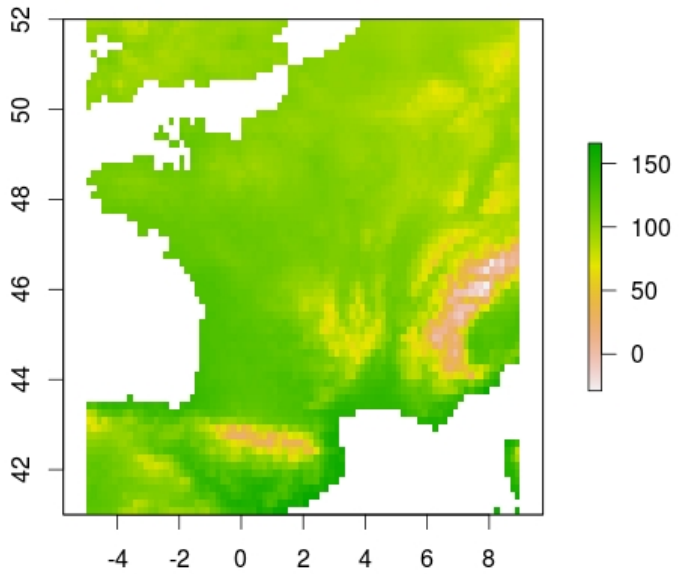


FIGURE – Température moyenne en degré Celsius * 10

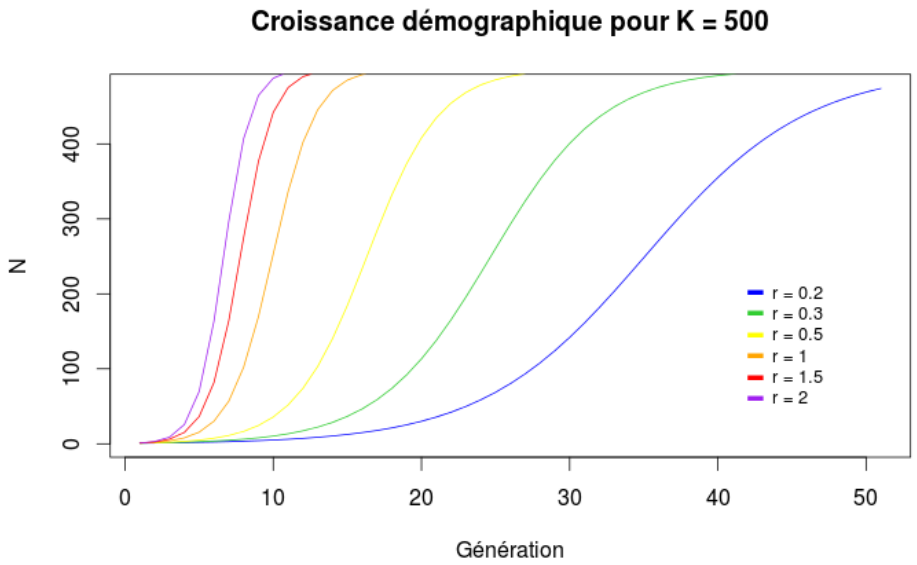
Initialisation de l'invasion

- Une femelle a vraisemblablement été introduite à Nérac en 2004 via des poteries chinoises.
- La femelle est habituellement fécondée par plusieurs mâles (poly-fécondation).
- (N_t^i) : taille de la population d'allèles (ou séquences) dans le dème i au temps t .
- distribution initiale est connue avec imprécision et devient un paramètre à estimer.

Evènement de reproduction

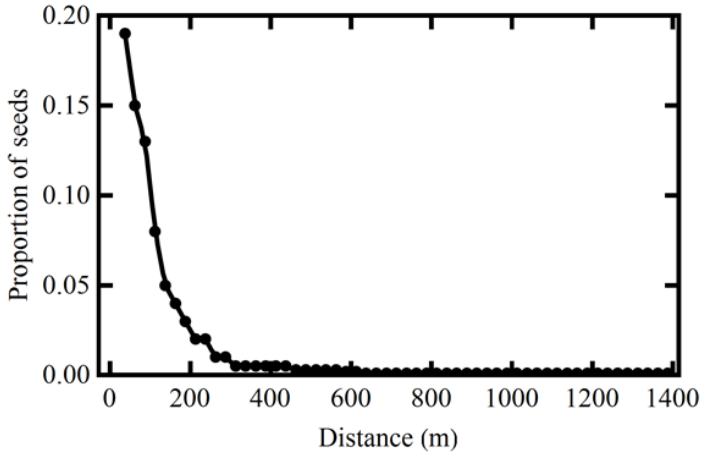
- On définit la taille de population après reproduction par :

$$\tilde{N}_t^i = \frac{N_t^i \times (1+R_t^i)}{1+\frac{R_t^i N_t^i}{K_t^i}}$$



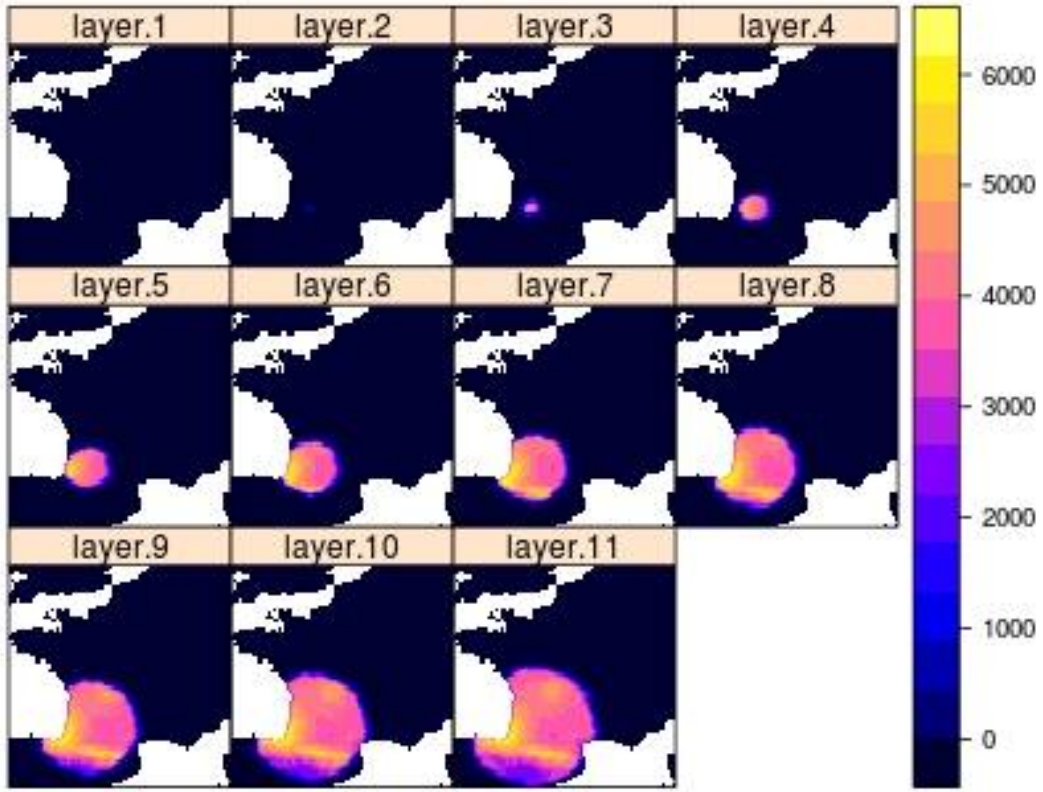
Evènement de dispersion

- Soit d_{ij} la distance géographique entre le dème i et le dème j.
- Soit un noyau de dispersion $f : R^+ \rightarrow [0, 1]$
- Probabilité m_{ij} de migrer du dème i vers le dème j : $m_{ij} = \frac{f(d_{ij})}{\sum_{k \in I} f(d_{ik})}$



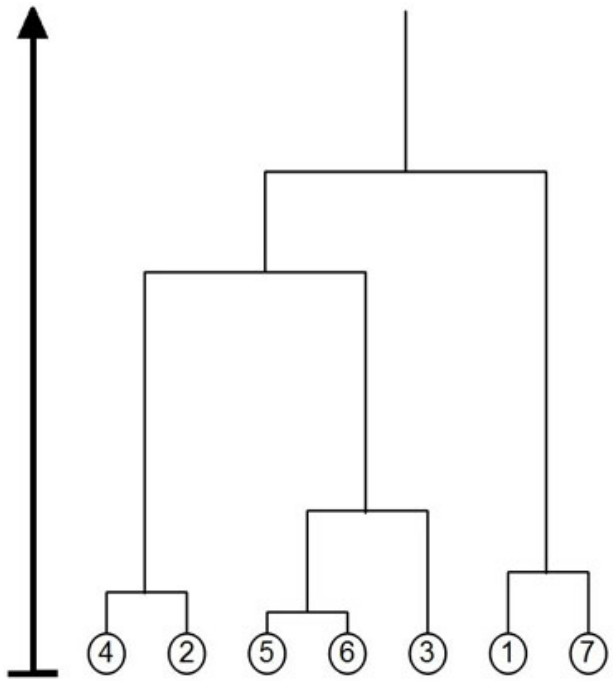
Processus démographique

$$N_{t+1}^{i} = \sum_{k \in I} m_{ki} \tilde{N}_{t}^{k}$$



Processus génétique en temps inverse

- La coalescence est une manière de voir la réplication de l'ADN en temps inverse.
- Conditionnellement à la démographie et à l'échantillon de séquences observées à la dernière génération, trouver les séquences parentes à la génération précédente.



Processus génétique en temps inverse

Soit e une séquence enfant et p sa séquence parente. La probabilité que le parent de e soit dans le dème i au temps $t-1$ sachant que l'enfant est dans le dème j au temps t est :

$$P(p_1 \in i \mid e_1 \in j) = \frac{m_{ij} N_{t-1}^i}{\sum_k m_{kj} N_{t-1}^k}$$

Soit e_1 (e_2) une séquence enfant et p_1 (p_2) sa séquence parente. La probabilité que les séquences e_1 et e_2 soient des copies de la même séquence parente sachant que leurs parents proviennent du même dème est :

$$P(p_1 = p_2 \mid p_1 \in i, p_2 \in i) = 1/N_{t-1}^i$$

Généalogies incomplètes

On ne peut pas forcément remonter jusqu'à la copie du gène MRCA :

- On ne sait pas grand chose de ce qui se passait *avant* en Chine, donc on ne peut pas continuer le processus
- Plusieurs copies différentes ont été introduites (femelle polyfécondée)
- On se retrouve donc en début d'invasion avec plusieurs généalogies non reliées

Généalogies incomplètes

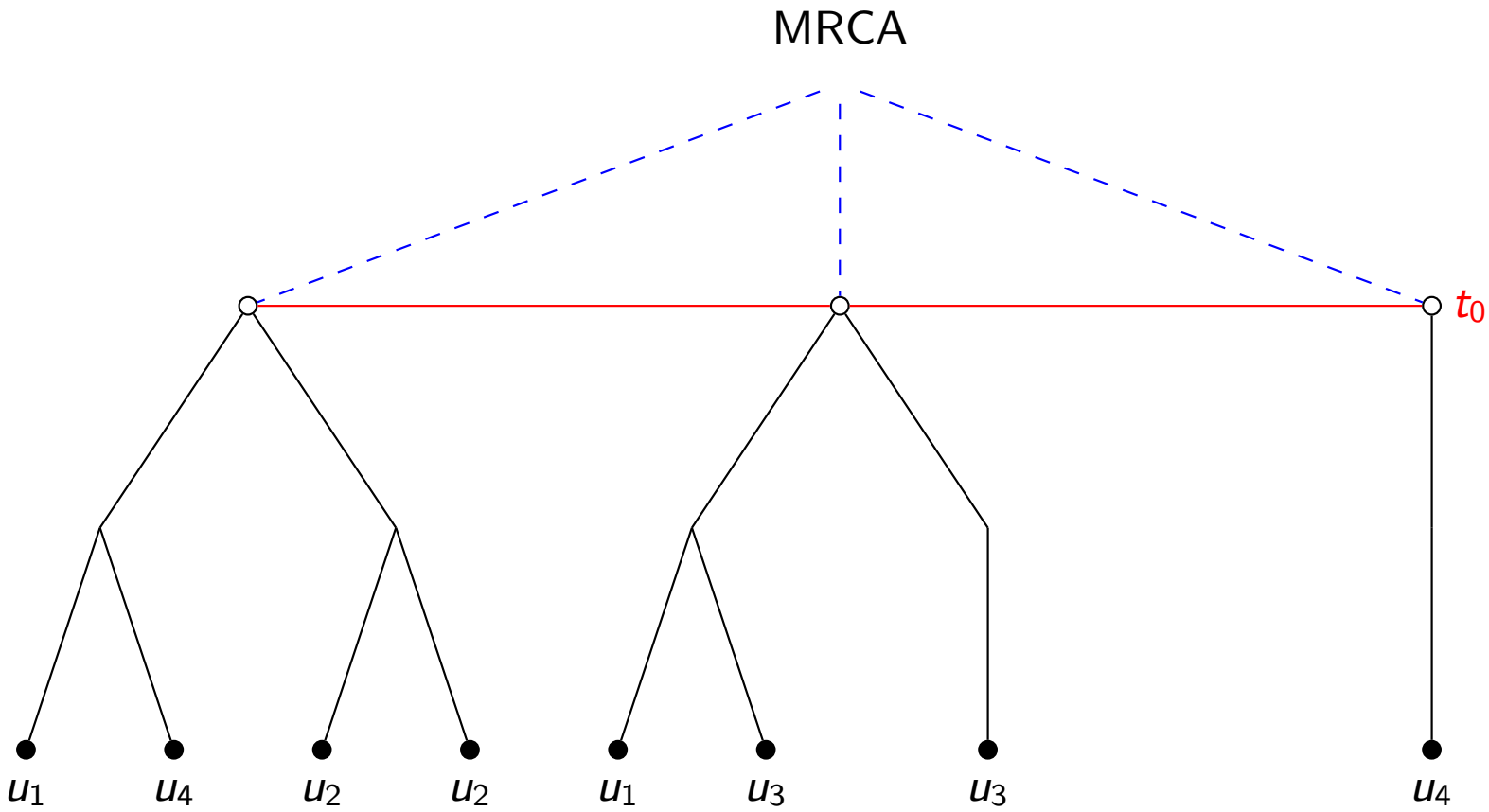


FIGURE – Coalescence incomplète : le processus de simulation des généalogies de gènes s'arrête à la date du début d'invasion t_0. On n'obtient donc pas *forcément* 1 généalogie comme dans le cas classique.

Choix de la méthode inférentielle

Problèmes

- Modèle compliqué : outils classiques d'estimation inopérants.
- Fonction de vraisemblance du modèle incalculable.

Solution : Calcul Bayésien Approché (ABC)

Utiliser des simulations massives pour explorer l'ensemble des possibles

Principe de l'ABC

Algorithme ABC

- On tire les paramètres dans la distribution a priori : $\theta' \sim p(\theta)$
- On simule les données selon le modèle : $y' \sim p(y|\theta')$
- On attribue un poids au couple (θ', y') en fonction de la distance $\|y' - y_{obs}\|$

Simuler : quand l'Humain parle à la machine

Une différence fondamentale à garder en tête :

> L'homme est lent, peu rigoureux et très intuitif. L'ordinateur est super rapide, très rigoureux et complètement con. On essaie de faire des programmes qui font une mitigation entre les deux.
>
> Gérard Berry, médaille d'or du CNRS 2014.

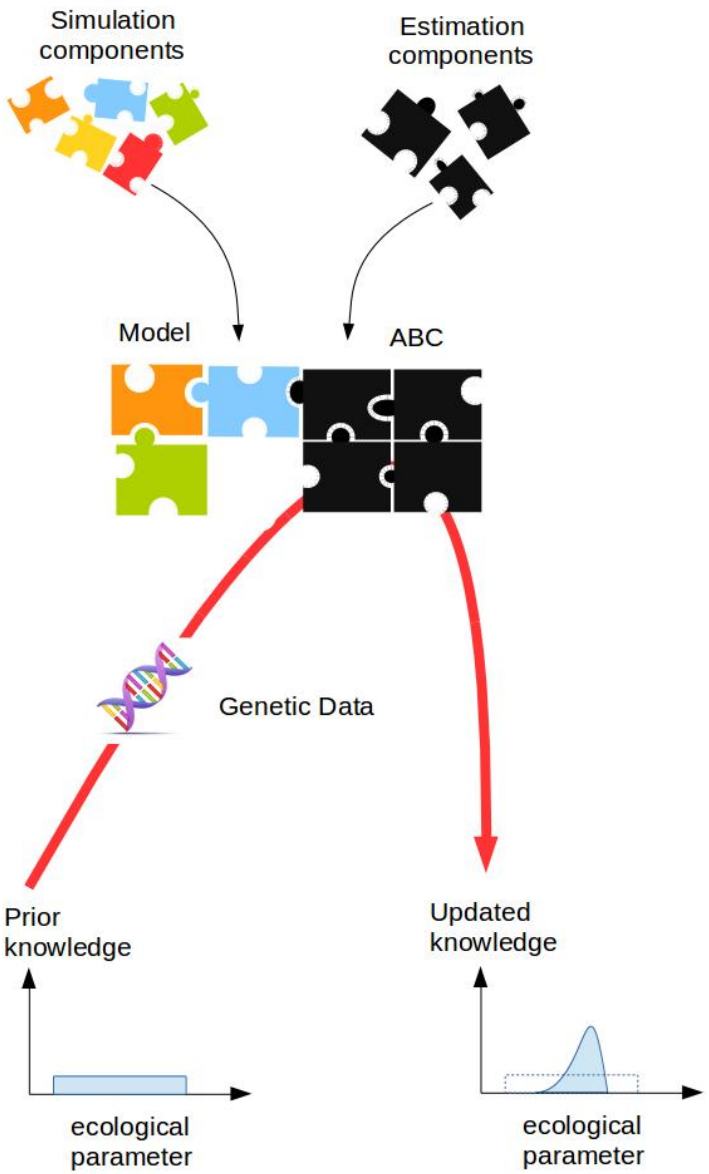
Pourquoi concevoir une bibliothèque ?

> Doing anything significant using only basic language features – that is, without the use of libraries – is unpleasant and unproductive. The key to elegant code and productivity lies in the production and use of libraries.
>
> Bjarne Stroustrup

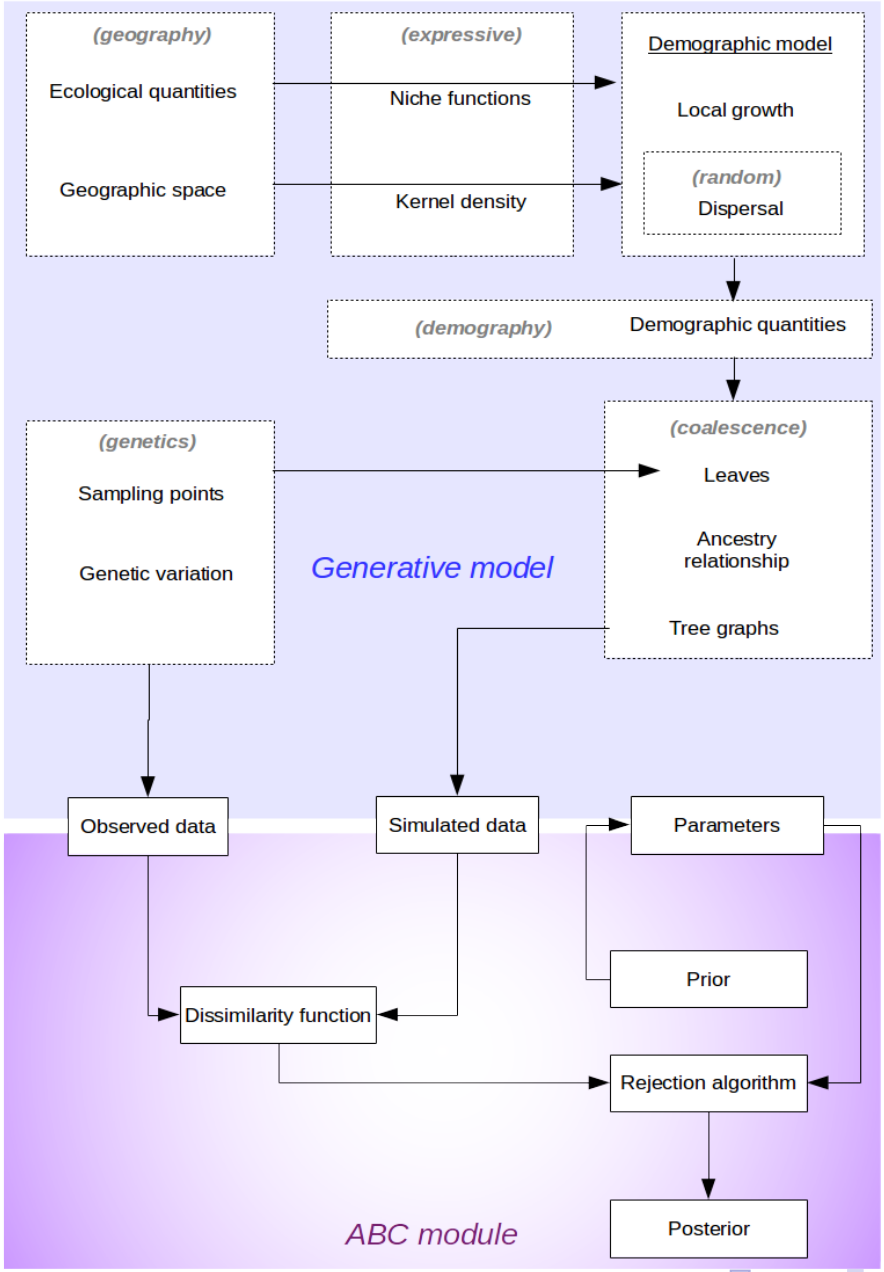
Définition

Bibliothèque : un ensemble d'implémentations de comportements (routines), écrites dans un langage de programmation, qui a une interface bien définie. Chaque routine est relativement indépendante et peut donc être réutilisée dans différents programmes.

Objectifs : modularité, efficacité, extensibilité



Résultat : Quetzal-CoalTL v1.0.0



Outils et partage

- Templates (oublier les types de données)
- Métaprogrammation (automatiser l'écriture de code)
- Move semantics (éviter les copies coûteuses)
- Classes de politiques (injecter du code contexte-spécifique dans un algorithme général)

Dépôt de code

https ://github.com/Becheler/quetzal

Site internet pour la documentation :

https ://becheler.github.io/quetzalAPI/html/

Internet Relay Chat

Channel #quetzal sur Freenode

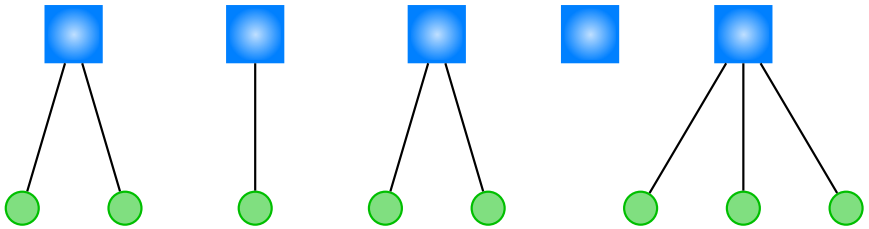
Stratégies pour la simulation de coalescents

- Hypothèse courante : $n << N$: arbres binaires, simulations rapides.
- Pas toujours raisonnable (invasion biologique, ou peu de contrôle sur N) : arbres n-aires, simulation lentes.
- Dans tous les cas, peu de code standard et réutilisable.

Spectre d'occupation

- n balles indistinguables sont placées uniformément au hasard dans m urnes, chaque urne ayant une probabilité d'affectation de $1/m$.
- Si une urne contient r balles : r est le *nombre d'occupation* de l'urne (voir Johnson and Kotz, 1977, p. 115).
- On désigne par M_r le nombre d'urnes ayant r balles.
- On appelle spectre d'occupation le vecteur $M_0, M_1, \ldots, M_n$.

Spectre d'occupation



- $M_0 = 1$: nombre de parents sans enfants
- $M_1 = 1$: nombre de lignées qui ne coalescent pas
- $M_2 = 2$: nombre de coalescences binaires
- $M_3 = 1$: nombre de coalescences ternaires.
- $M = 1121$: spectre d'occupation.

Spectre d'occupation agit comme une *interface* car il peut être généré de différentes manières, appelées stratégies.

Stratégies

- A la volée : à chaque lignée un parent (OTF, on-the-fly).
- Par échantillonnage direct dans sa distribution de probabilité pré-calculée (MEM, mémoïsation)

Soit D_n^m la distribution jointe des spectres d'occupation émergeants lorsque n balles sont lancées dans m urns. Expression de D_n^m (von Mises, 1939) :

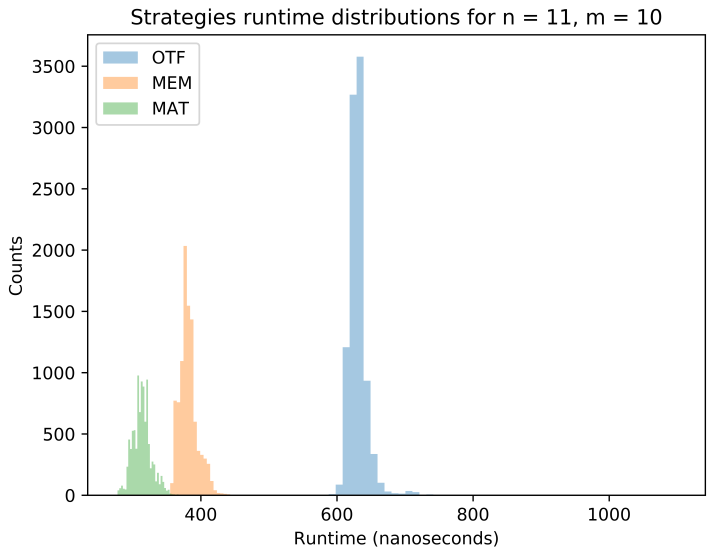
$$\Pr[\bigcap_{j=0}^{n}(M_j = m_j)] = \frac{m!n!}{m^n \prod[(j!)^{m_j} m_j!]} \quad (1)$$

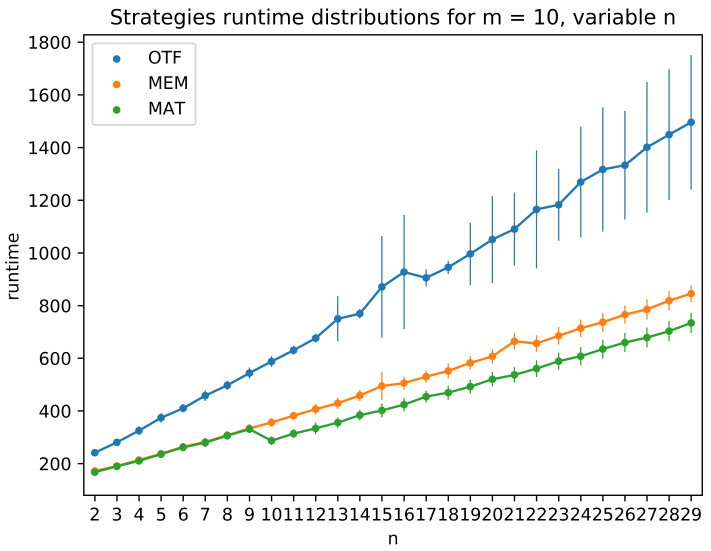
$$\text{avec} \quad \sum_{j=0}^{n} m_j = m \quad \text{(conservation du nombre d'urnes)} \quad (2)$$

$$\text{et} \quad \sum_{j=0}^{n} j m_j = n \quad \text{(conservation du nombre de balles)} \quad (3)$$

Distribution mémoïsée : réduction de l'empreinte mémoire

- Un algorithme maison permet de construire le support de D_n^m, qui grossit vite avec des spectres de très faible probabilité.
- Stratégie MAT permet d'effectuer de l'approximer et de tronquer certains spectres pour éviter des itérations inutiles.





Idée générale : forêts et partitions

- Soit un ensemble S de n de noeuds (copies de gènes) échantillonnés au temps t_s.
- Une forêt d'arbres aléatoires est construite en temps inverse de t_s à t_0 par le processus de coalescence.
- On définit la relation d'équivalence $\sim$ sur S «*sont feuilles d'un même sous-arbre*».
- Au temps t, les classes d'équivalence forment une partition P_t de S.
- Sous l'hypothèse de non mutation, P_t est un raffinement de la partition observée O_{t_s}.
- Introduire une distance mesurant cette proximité entre partition observée et simulée.

Partition floue observée

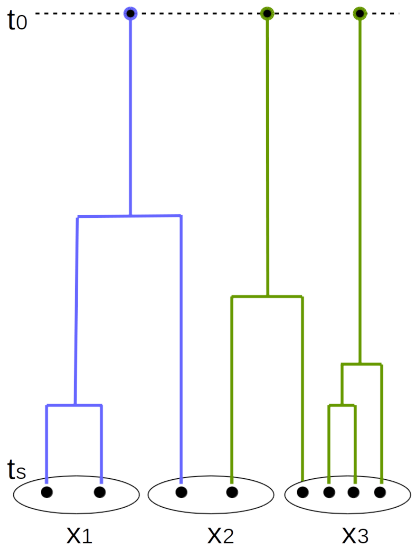
Les gènes échantillonnés dans un même dème ne sont pas distinguables. On change donc la représentation des données : chaque dème échantillonné appartient pour une certaine part (fréquence observée) à chaque état allélique.

Coordonnées	A1	A2
x_1	10	10
x_2	20	12
x_3	20	10
x_3	10	12

$$O_{t_s} = \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 0/2 & 1/2 & 1/4 \\ 2/2 & 0/2 & 2/4 \\ 0/2 & 1/2 & 1/4 \end{pmatrix} \end{array}$$

Partition floue simulée

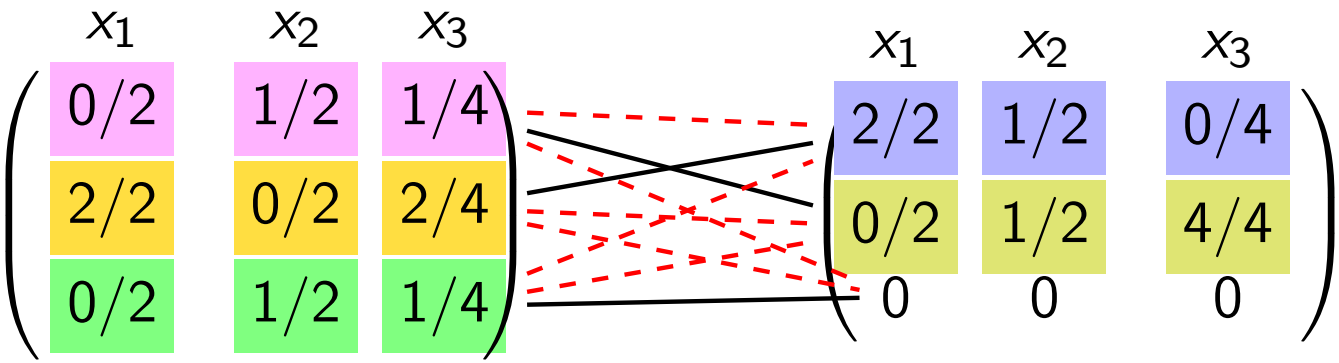
Certains gènes introduits à t_0 peuvent avoir le même état allélique, ce qui fusionne certains clusters de la partition généalogique



$$P_t = \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 2/2 & 1/2 & 0/4 \\ 0/2 & 1/2 & 4/4 \end{pmatrix} \end{array}$$

Distance de Transfert Floue

On contruit le graphe biparti pour trouver la meilleure affectation des clusters de O_{t_s} à ceux de P_t.
Chaque arrête est valuée par $w_{ij} = \sum_{k=1}^{n} |u_{ik} - v_{jk}|$.
L'algorithme de Kuhn-Munkres permet de trouver le couplage parfait de poids minimum.
Le poids minimal (arrêtes noires) définit la FTD (Campello, 2010). Ici FTD=2.25.



Test de la méthode - intégration à Quetzal

On cherche à estimer la taille d'une population de WF sur 50 générations, à partir d'un échantillon de taille $n = 50$:

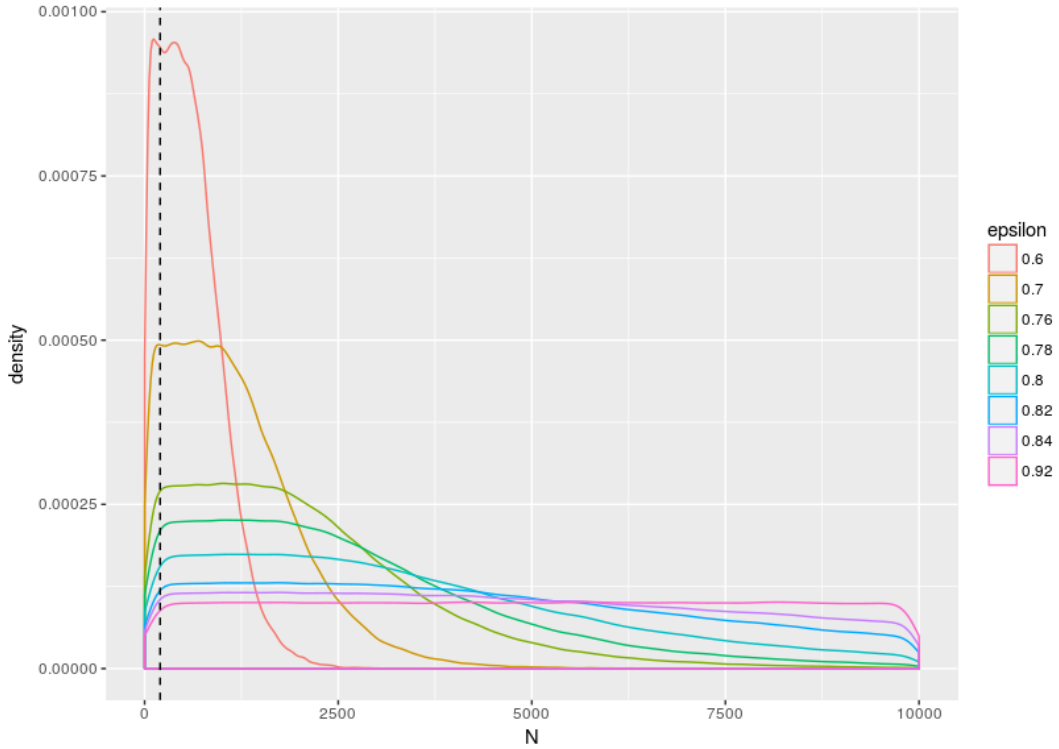


FIGURE – Densités postérieures pour différents ϵ pris comme les déciles de la distribution des FTD calculées (1 donnée pseudo-observée, 10^5 simulations). La vraie taille de population ($N = 200$) est indiquée par la droite verticale en pointillé.

Test de la méthode - intégration à Quetzal

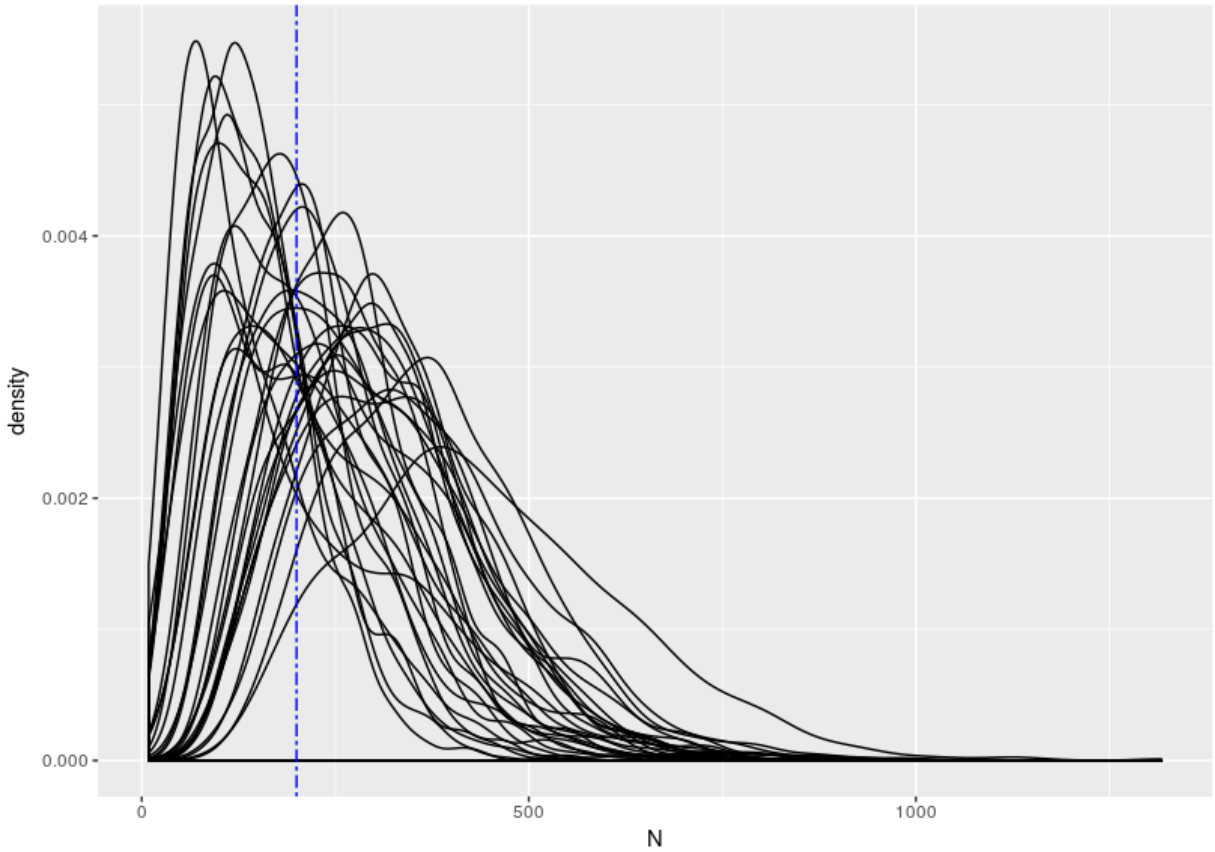


FIGURE – Densités postérieures pour 30 données pseudo-observées, avec ϵ pris comme le premier percentile de la distribution des FTD (10^5 simulations). La vraie taille de population ($N = 200$) est indiquée par la droite verticale.

Conclusion

- Quetzal : une ressource grandissante pour la simulation et l'ABC.
- Des stratégies de simulations flexibles pour plus d'efficacité et de généralité.
- Une distance pour l'ABC permettant de ne s'intéresser qu'à l'histoire généalogique récente (intéressant pour les processus d'invasion).

Perspectives

- mettre à profit ces avancées techniques et méthodologiques pour analyser le jeu de données du frelon asiatique.
- vérifier que la FTD puisse s'appliquer à un échantillonnage spatio-temporel pour pouvoir utiliser l'intégralité du jeu de données.
- définir des fonctions de niche pertinentes.

Remerciements

- Stéphane Dupas et Camille Coron.
- Ambre Marques.
- Florence Jornod.
- Loïc Joly, Philippe Dunsky et les nombreux autres membres de la communauté developpez.com .

Merci à tous pour votre attention

Penser c'est oublier des différences, c'est généraliser, c'est abstraire.

Jorge Luis Borges

Mettons en commun ce que nous avons de meilleur et enrichissons-nous de nos mutuelles differences.

Paul Valéry