MESURES D'EWENS

L'objectif de ce devoir est d'étudier la répartition en diverses espèces des individus d'un écosystème ; par exemple, la répartition par espèces des arbres d'une forêt tropicale. Dans la première partie, on introduit un modèle markovien qui prend en compte les phénomèmes de spéciation et d'extinction des espèces, et qui permet de decrire l'évolution d'un écosystème. Dans la seconde partie du devoir, on s'intéresse à l'évolution d'une espèce fixée au sein d'un écosystème, et on montre une borne sur le temps d'extinction de cette espèce. Enfin, dans la dernière partie, on s'intéresse au régime stationnaire de la chaîne de Markov introduite dans la première partie.

1. Évolution d'un écosystème

Dans ce qui suit, on fixe un entier $N \geq 1$, qui représentera le nombre total d'individus dans l'écosystème que l'on considère, toutes espèces confondues. La répartition des individus en différentes espèces est encodée par une *composition* de taille N, c'est-à-dire une suite

$$c = (N_1, N_2, \ldots, N_R)$$

d'entiers telle que $N_i \geq 0$ pour tout i, et telle que $\sum_{i=1}^{R} N_i = N$. Le nombre N_i représente le nombre d'individus appartenant à l'espèce i. Par exemple, si l'on considère une forêt avec 10 arbres dont 5 chênes, 2 peupliers et 3 sapins, alors on pourra représenter cette répartition par la composition $c = (5, 2, 3)$.

On note $\mathfrak{C}_N$ l'ensemble des compositions de taille N :

$$\mathfrak{C}_N = \left\{ (N_1, \ldots, N_R) \mid R \geq 1,\ N_i \geq 0 \text{ pour tout } i \in [\![1, R]\!],\ \sum_{i=1}^{R} N_i = N \right\}$$

On fixe également un paramètre $\nu \in (0, 1)$. On va introduire une chaîne de Markov $(c(t))_{t\geq 0}$ à valeurs dans $\mathfrak{C}_N$ et qui modélise l'évolution de la répartition entre espèces. La taille N de l'écosystème restera fixe au cours du temps. On fixe une composition de départ $c(0) \in \mathfrak{C}_N$. Si $c(t) = (N_1(t), N_2(t), \ldots, N_{R(t)}(t))$ est construite, pour obtenir $c(t+1)$, on procède comme suit :

(E1) On choisit au hasard un individu de la population qui meurt. La probabilité que cet individu soit issu de l'espèce i avec $i \in [\![1, R(t)]\!]$ est égale à $\frac{N_i(t)}{N}$.

(E2) Avec probabilité ν, l'individu de type i qui est mort à l'étape (E1) est remplacé par un nouvel individu d'une nouvelle espèce (phénomène de spéciation). Dans ce cas, la nouvelle composition $c(t+1)$ s'écrit :

$$c(t+1) = (N_1(t), \ldots, N_{i-1}(t), N_i(t) - 1, N_{i+1}(t), \ldots, N_{R(t)}(t), 1).$$

(E3) Avec probabilité $1-\nu$, l'individu de type i qui est mort à l'étape (E1) est remplacé par un individu d'une espèce j avec $j \in [\![1, R(t)]\!]$. La probabilité de choix de j est $\frac{N_j(t)}{N}$. Les indices i et j étant choisis, la nouvelle composition $c(t+1)$ s'écrit :

$$c(t+1) = \begin{cases} (N_1(t), \ldots, N_i(t) - 1, \ldots, N_j(t) + 1, \ldots, N_{R(t)}(t)) & \text{si } j \neq i, \\ c(t) & \text{si } j = i. \end{cases}$$

On suppose que les choix faits dans les étapes (E1), (E2) et (E3) sont indépendants entre eux, et indépendants pour des temps distincts. L'algorithme écrit ci-dessus définit alors une chaîne de Markov à valeurs dans $\mathfrak{C}_N$. Notons que l'ensemble $\mathfrak{C}_N$ est infini, car il contient toutes les compositions $c = (0, 0, \ldots, 0, N)$ avec un nombre arbitraire de 0. Ceci n'empêche pas de travailler avec la chaîne $(c(t))_{t\in\mathbb{N}}$.

Q1. On note $\mathfrak{C}_N^*$ l'ensemble des compositions de taille N sans part égale à 0 ; par exemple, $\mathfrak{C}_4^* = \{(4), (3,1), (2,2), (1,3), (2,1,1), (1,2,1), (1,1,2), (1,1,1,1)\}$. Montrer que pour tout $N \geq 1$, $\mathfrak{C}_N^*$ contient 2^{N-1} compositions. On pourra procéder par récurrence sur N.

Q2. Dans ce qui suit, on suppose que $c(0)$ appartient à $\mathfrak{C}_N^*$. Les autres compositions $c(t \geq 1)$ pourront en revanche avoir des parts $N_i(t)$ égales à 0, qui représenteront les espèces i éteintes au temps t. Écrire un programme qui simule $(c(0), c(1), \ldots, c(T))$, avec par exemple $N = 10$, $\nu = 0.1$, $c(0) = (5, 2, 3)$ et $T = 30$. Donner un résultat de cette simulation.

Q3. On note $R(t)$ le nombre total de parts de $c(t)$. Montrer qu'on peut écrire $R(t)$ sous la forme

$$R(t) = R(0) + V_1 + V_2 + \cdots + V_t,$$

où les V_i sont des variables indépendantes et de même loi, que l'on précisera. Décrire le comportement des variables aléatoires $R(t)$ et $\frac{R(t)}{t}$ lorsque t tend vers l'infini. Que peut-on dire du nombre

$$E(t) = \text{card}\,\{i \in [\![1, R(t)]\!] \mid N_i(t) = 0\}$$

d'espèces qui sont éteintes au temps t, en particulier lorsque t tend vers l'infini ?

2. Évolution d'une espèce fixée

Dans cette partie, on s'intéresse à l'évolution d'une espèce fixée dans l'écosystème, par exemple celle d'indice $i = 1$. Les paramètres N et ν sont fixés comme précédemment, avec $\nu > 0$.

Q4. Montrer que $(N_1(t))_{t\geq 0}$ est une chaîne de Markov sur $[\![0, N]\!]$, dont les probabilités de transition sont

$$P(k, k+1) = \frac{(N-k)k}{N^2}\,(1-\nu);$$
$$P(k, k) = \frac{N-k}{N}\,\nu + \frac{k^2 + (N-k)^2}{N^2}\,(1-\nu);$$
$$P(k, k-1) = \frac{k}{N}\,\nu + \frac{(N-k)k}{N^2}\,(1-\nu).$$

Q5. Montrer que 0 est l'unique état absorbant de la chaîne $(N_1(t))_{t\geq 0}$. En déduire que $N_1(t) \to_{t\to\infty} 0$ presque sûrement.

Q6. On cherche à déterminer à quelle vitesse l'espèce d'indice $i = 1$ s'éteint. Pour $k \in [\![0, N]\!]$, calculer l'espérance conditionnelle $\mathbb{E}[N_1(t+1) \mid N_1(t) = k]$, et en déduire par récurrence sur t que

$$\mathbb{E}[N_1(t)] = N_1(0)\,\left(1 - \frac{\nu}{N}\right)^t.$$

Q7. Soit $T = \inf\{t \geq 0 \mid N_1(t) = 0\}$. Montrer que $\mathbb{P}[T > t] \leq \mathbb{E}[N_1(t)]$. En déduire que

$$\mathbb{E}[T] \leq \sum_{t=0}^{\infty} \mathbb{E}[N_1(t)] = \frac{N\, N_1(0)}{\nu}.$$

Commenter ce résultat.

3. Régime stationnaire de la chaîne de Markov

Si $c \in \mathfrak{C}_N$ est une composition de taille N, on note $p = \pi(c)$ la *partition* de taille N qui est obtenue en retirant de c les parts de taille 0, et en réordonnant ses parts de manière décroissante. Par exemple, si $c = (1, 5, 1, 0, 2, 1, 0)$, alors $p = \pi(c) = (5, 2, 1, 1, 1)$. On notera $\mathfrak{P}_N$ l'ensemble des partitions de taille N (compositions sans part égale à 0, et avec des parts classées par ordre décroissant) :

$$\mathfrak{P}_N = \left\{ (N_1 \geq N_2 \geq \cdots \geq N_R) \mid R \geq 1,\ N_i \geq 1 \text{ pour tout } i \in [\![1, R]\!],\ \sum_{i=1}^{R} N_i = N \right\}.$$

Par exemple,

$$\mathfrak{P}_4 = \{(4), (3, 1), (2, 2), (2, 1, 1), (1, 1, 1, 1)\}.$$

On voit facilement que si $(c(t))_{t\geq 0}$ est la chaîne de Markov introduite dans la première partie, alors $p(t) = \pi(c(t))$ définit encore une chaîne de Markov, cette fois-ci à valeurs dans l'ensemble fini $\mathfrak{P}_N$. De plus, pour construire la chaîne $(p(t))_{t\geq 0}$, on peut utiliser les mêmes étapes (E1), (E2) et (E3) que pour la chaîne $(c(t))_{t\geq 0}$; la seule différence est qu'il faut à chaque fois supprimer les éventuelles parts 0 qui ont été créées, et reclasser par ordre décroissant les parts de $p(t)$.

Q8. Si $N = 4$, écrire en fonction de ν la matrice de transition de la chaîne de Markov $(p(t))_{t\in\mathbb{N}}$ (comme $\operatorname{card} \mathfrak{P}_4 = 5$, cette matrice est de taille 5×5).

Q9. Pour tout entier $N \geq 1$, montrer que la chaîne $(p(t))_{t\in\mathbb{N}}$ est irréductible sur $\mathfrak{P}_N$, et qu'elle est apériodique.

Q10. Montrer qu'il existe une mesure de probabilité $\mu_{N,\nu}$ sur $\mathfrak{P}_N$ telle que

$$\lim_{t\to\infty} \mathbb{P}[p(t) = p] = \mu_{N,\nu}(p)$$

pour toute partition $p \in \mathfrak{P}_N$.

Étant donnée une partition $p \in \mathfrak{P}_N$ et un entier $i[\![1, N]\!]$, on note $m_i(p)$ le nombre de parts de p égales à i. Par exemple, si $p = (5, 2, 1, 1, 1) \in \mathfrak{P}_{10}$, alors la suite de ses multiplicités est :

$$(m_1, m_2, \ldots, m_{10}) = (3, 1, 0, 0, 1, 0, 0, 0, 0, 0).$$

Si θ est un paramètre positif, la mesure d'Ewens de paramètre θ sur $\mathfrak{P}_N$ est la mesure positive définie par

$$\rho_{N,\theta}(p) = \frac{N!}{\prod_{i=1}^{N} i^{m_i}\,(m_i)!}\, \frac{\theta^{m_1+m_2+\cdots+m_p}}{\theta(\theta+1)(\theta+2)\cdots(\theta+N-1)}$$

où $(m_1, m_2, \ldots, m_N)$ sont les multiplicités de la partition p.

Q11. Montrer que pour tout $\theta > 0$, $\rho_{4,\theta}$ est une mesure de probabilité sur $\mathfrak{P}_4$. Si $\theta = \frac{4\nu}{1-\nu}$, montrer que $\rho_{4,\theta}$ est laissée invariante par la matrice de transition calculée à la question Q8, et donc que $\rho_{4,\theta} = \mu_{4,\nu}$.

On admet dans la suite que pour tout entier $N \geq 1$, si $\theta = \frac{N\nu}{1-\nu}$, alors la mesure d'Ewens $\rho_{N,\theta}$ est la mesure invariante $\mu_{N,\nu}$ introduite à la question Q10. Les dernières questions vont permettre de simuler la loi $\rho_{N,\theta}$ sur l'ensemble des partitions $\mathfrak{P}_N$.

On note $\mathfrak{S}_N$ l'ensemble des $N!$ permutations de taille N, c'est-à-dire les bijections $\sigma : [\![1, N]\!] \to [\![1, N]\!]$. Un cycle $(a_1, a_2, \ldots, a_p)$ est une permutation qui envoie a_1 sur a_2, a_2 sur a_3, *etc.* jusqu'à a_p qui est envoyé sur a_1 ; et tous les autres éléments de $[\![1, N]\!]$ sont laissés invariants. On rappelle que toute permutation $\sigma \in \mathfrak{S}_N$ peut s'écrire sous la forme

$$\sigma = (a_{1,1}, \ldots, a_{1,p_1}) \circ (a_{2,1}, \ldots, a_{2,p_2}) \circ \cdots \circ (a_{\ell,1}, \ldots, a_{\ell,p_\ell}),$$

les cycles $(a_{i,1}, \ldots, a_{i,p_i})$ étant disjoints et recouvrant $[\![1, N]\!]$. Par exemple, la permutation $\sigma = 914362857$ qui envoie 1 sur 9, 2 sur 1, 3 sur 4, *etc.* se décompose en le produit de cycles $(1, 9, 7, 8, 5, 6, 2) \circ (3, 4)$. Le type cyclique d'une permutation $\sigma \in \mathfrak{S}_N$ est la partition $p(\sigma) \in \mathfrak{P}_N$ dont les parts sont les tailles des cycles de σ. Par exemple, la permutation 914362857, qui est de taille 9, a pour type cyclique $(7, 2)$.

Q12. Montrer que si $p \in \mathfrak{P}_N$, alors le nombre de permutations $\sigma \in \mathfrak{S}_N$ dont le type cyclique est p est égal à

$$\frac{N!}{\prod_{i=1}^{N} i^{m_i}\,(m_i)!},$$

où $(m_1, m_2, \ldots, m_N)$ sont les multiplicités de la partition p.

Q13. Montrer par récurrence sur N que toute permutation $\sigma \in \mathfrak{S}_N$ s'écrit de manière unique sous la forme

$$\sigma = (1, n_1) \circ (2, n_2) \circ (3, n_3) \circ \cdots \circ (N, n_N),$$

où chaque n_i appartient à $[\![1, i]\!]$, et où par convention, si $n_i = i$, alors (i, i) est la permutation identité. On pourra dans la récurrence poser $n_N = \sigma^{-1}(N)$. Montrer que de plus, le nombre de cycles disjoints de σ est

$$\ell(\sigma) = \operatorname{card}\{i \in [\![1, N]\!] \mid n_i = i\}.$$

Un paramètre $\theta > 0$ étant fixé, pour chaque $i \in [\![1, N]\!]$, on note m_i la variable aléatoire de loi

$$\mathbb{P}[m_i = j] = \frac{1}{\theta + i - 1} \text{ si } j \in [\![1, i-1]\!] \quad ; \quad \mathbb{P}[m_i = i] = \frac{\theta}{\theta + i - 1}.$$

Q14. Si $\sigma_N = (1, m_1) \circ (2, m_2) \circ \cdots \circ (N, m_N)$, montrer que la permutation aléatoire σ_N a pour loi

$$\mathbb{P}[\sigma_N = \sigma] = \frac{\theta^{\ell(\sigma)}}{\theta(\theta+1)(\theta+2)\cdots(\theta+N-1)},$$

où $\ell(\sigma)$ est le nombre de cycles de σ (les points fixes étant comptés comme des cycles de longueur 1). En déduire que $p_N = p(\sigma_N)$ a pour loi la mesure d'Ewens $\rho_{N,\theta}$.

Q15. Utiliser la question précédente pour écrire un programme qui simule une partition aléatoire de loi $\rho_{N,\theta}$, par exemple avec $N = 1000$ et $\theta = 10$. Dans `Sage`, on pourra utiliser la classe `Permutation` et ses méthodes, en particulier la méthode `.cycle_type()` qui affiche le type cyclique d'une permutation. Donner un résultat de cette simulation.

On obtient ainsi une méthode simple pour simuler la répartition par espèces des individus d'un écosystème. Le paramètre θ est appelé constante de biodiversité du modèle, et il régit le nombre moyen d'espèces distinctes et non éteintes dans une population.