2

Statistiques pour la biologie évolutive

Jean-Dominique LEBRETON [1], *Étienne KLEIN* [2], *Olivier GIMENEZ* [1] *& François ROUSSET* [3]

1 INTRODUCTION

Le développement des ordinateurs a fait définitivement basculer pratiques et concepts de la science vers le quantitatif. L'écologie évolutive n'y échappe pas : à côté de modèles théoriques (Maynard Smith 1974), de nombreuses approches visent à confronter des données à des attendus ou à estimer des quantités d'intérêt biologique et relèvent donc de la statistique au sens large. Statistique et modélisation, servant par essence à comprendre et analyser des résultats quantitatifs, font donc nécessairement partie de la culture de l'évolutionniste.

La vision moderne de la statistique s'appuie largement sur la notion de « modèle statistique ». Un exemple simple est le modèle qui sous-tend l'estimation de la moyenne : des observations quantitatives (y_1, y_2, ..., y_n) sont supposées provenir de n variables aléatoires indépendantes (Y_1, Y_2, ..., Y_n), obéissant à la relation

$$Y_i = \mu + \varepsilon_i \qquad [1]$$

dans laquelle apparaissent une part systématique, ou paramétrique, μ, et une part aléatoire, ε_i (avec $E(\varepsilon_i) = 0$, Var $(\varepsilon_i) = \sigma^2$, indépendance des ε_i). La part systématique peut être diverse, sans pouvoir être compliquée à l'infini : modèles linéaires

[1] *CEFE ,UMR 5175, CNRS, 1919 Route de Mende, 34 293 Montpellier cedex 5*
[2] *BioSP, UR 546, INRA, Domaine St-Paul, 84 914 Avignon cedex*
[3] *ISEM, UMR 5554, UM II, Place Eugène Bataillon, 34 095 Montpellier cedex 5*

(Dagnélie 2006, Tomassone *et al.* 1983) ; modèles linéaires généralisés (Aitkin *et al.* 1988, Crawley 1993) ; modèles non-linéaires (Seber et Wild 1988, Huet *et al.* 1992). La part aléatoire peut elle aussi être compliquée, avec plusieurs « composantes de la variance » (Searle *et al.* 1992) considérant comme aléatoires des variations que l'on ne souhaite pas représenter de manière explicite dans la partie systématique du modèle. Dans tous ces modèles, des paramètres représentent des quantités d'intérêt et l'on parle de statistique paramétrique. Nous n'évoquerons que brièvement la statistique non paramétrique.

Dans tout modèle statistique apparaît une tension inévitable entre complexité du réel et choix simplificateurs, extrêmes bien sûr dans l'exemple du modèle [1] : le modèle n'est qu'un outil (Legay 1973) et modéliser reste un art de la simplification, adaptée aux objets et aux objectifs. La culture statistique de l'évolutionniste inclut dès lors :

- une compréhension des modèles statistiques et de quelques approches générales ;
- une connaissance plus approfondie de quelques modèles pertinents pour l'écologie évolutive.

L'appui des logiciels tend à faire négliger le second niveau. Le risque est de réduire la statistique à l'utilisation de quelques pratiques dominantes, propres à des sous-domaines et jamais réévaluées, engendrant des blocages scientifiques. Pour ne citer qu'un seul exemple, l'emploi généralisé des proportions d'individus réobservés, qui suppose à tort une détectabilité parfaite des individus, a engendré et entretenu une sous-estimation des probabilités de survie dans les populations animales, qui n'a été résolue qu'avec le développement des méthodes de capture-recapture (Gimenez *et al.* 2008, Lebreton 2006 ; cf. Paragraphe 4). Ce chapitre s'efforce donc de favoriser une attitude culturelle face à la statistique. Le Paragraphe 2 développe de façon concise les approches générales. Les paragraphes suivants (3 à 5) développent trois sujets retenus pour leur spécificité, leur complémentarité et leur importance pour l'évolutionniste :

- les modèles linéaires mixtes, utilisés dans l'analyse d'expériences mais aussi de pedigrees (Kruuk 2004) ;
- les modèles d'estimation de traits démographiques (modèles de capture-marquage-recapture), largement utilisés pour les populations animales (Lebreton *et al.* 1992) mais aussi végétales (Kéry *et al.* 2005) ;
- les principales approches statistiques de la génétique des populations, notamment pour l'analyse de données moléculaires (par ex. Rousset 1997).

Les termes les plus techniques sont introduits entre guillemets et autant que possible brièvement définis. Les notions d'échantillon, de variable aléatoire et de test statistique sont supposées connues (Dagnélie 2007). Un minimum de familiarité avec des méthodes de base comme la régression linéaire et l'analyse de variance à un facteur est requis du lecteur (voir Sokal et Rohlf 1981).

2 QUELQUES CONCEPTS GÉNÉRAUX DE LA STATISTIQUE

La distinction entre observations $(y_1, y_2, ..., y_n)$ et variables aléatoires $(Y_1, Y_2, ..., Y_n)$ est cruciale. Les observations sont des réalisations des variables aléatoires, elles-mêmes supposées obéir à certaines règles constituant le modèle statistique. Cette dualité observations/variables aléatoires sous-tend la notion de réplication. Si l'on obtient un second échantillon, les observations $(y_1', y_2', ..., y_n')$ différeront ; les variables aléatoires sous-jacentes $(Y_1, Y_2, ..., Y_n)$ ainsi que les propriétés et les paramètres du modèle restent inchangés. La notion de réplication est essentielle, même en l'absence de réplicat. En effet toute propriété déduite de $(y_1, y_2, ..., y_n)$ qui ne s'appliquerait pas à un réplicat n'aurait qu'une validité interne, restreinte à cet échantillon. Toute conclusion valide sur un réplicat, même s'il n'est pas réalisé, participe de la validité externe, la seule intéressante scientifiquement : la conclusion est valide dans le monde réel et n'attend que la réalisation d'un réplicat éventuel pour se manifester : on passe pour ainsi dire du fait brut au résultat scientifique. Le Tableau 1 donne un exemple de procédure à forte validité interne, mais sans aucune validité externe.

Estimations et tests vont de pair et constituent les fondements de la statistique paramétrique. Dans notre modèle, la variable aléatoire $Y_\cdot = \frac{Y_1 + Y_2 + ... + Y_n}{n}$ est un « estimateur » de μ. Cette appellation sous-entend que Y apporte une information pertinente sur μ, ce que confirme l'étude mathématique. L'estimateur Y est ainsi sans biais ($E(Y) = \mu$; Figure 1). Si l'on suppose en outre que les ε_i ont une distribution normale (ou « sont gaussiens », ce qu'on notera $\varepsilon_i \approx N(0, \sigma^2)$), $Y_\cdot$, combinaison linéaire de variables gaussiennes, sera lui aussi distribué normalement.

L'indépendance des observations est cruciale : en conduisant à $\mathrm{var}(Y_\cdot) = \frac{\sigma^2}{n}$, elle permet littéralement de transporter la variabilité de l'individu (σ^2) à la moyenne. Sans l'indépendance, assurée en pratique par un tirage aléatoire de l'échantillon dans la population cible, seules des réplications des « moyennes empiriques » $y_\cdot$, et donc plusieurs échantillons, permettraient d'estimer cette variance.

La statistique des observations dépendantes, relevant notamment du vaste champ des processus stochastiques, est difficile et complexe. En effet, la structure de dépendance du modèle doit, tout en restant suffisamment simple, permettre d'estimer les paramètres de la part aléatoire du modèle. C'est le cas par exemple pour les processus markoviens (Iosifescu et Tautu 1973) et pour les modèles mixtes du Paragraphe 3.

Dans tous les cas, des méthodes générales, dont au premier chef la méthode du maximum de vraisemblance (MV) (Mood *et al.* 1974 Chapitre VII), permettent d'obtenir des estimateurs des paramètres aux propriétés optimales. Le principe peut en être rappelé sur l'exemple simple de l'observation de variables indépendantes Y_i ($i = 1, ..., n$) distribuées selon une loi de Bernoulli, c'est-à-dire prenant les valeurs 0 et 1 avec les probabilités $1 - p$ et p. La probabilité d'observer Y_i, $\Pr(Y_i) = pY_i + (1 - p)(1 - Y_i)$, dépend bien sûr de la valeur de Y_i considérée mais aussi du paramètre p. La probabilité des Y_i ($i = 1, ..., n$) est égale à $p^k(1 - p)^{n-k}$ pour tous les échantillons menant à la même valeur $k = \sum_{i=1}^{n} y_i$, en nombre $C_n^k = \frac{n!}{k!(n-k)!}$. Le modèle s'écrit donc en fonction de la variable aléatoire « nombre de succès » $K = \sum_{i=1}^{n} Y_i$ dont la distribution est binomiale : pour $k = 0, 1, ..., n$, $\Pr(K = k) = C_n^k p^k (1 - p)^{n-k}$. Une fois l'échantillon $(y_1, y_2, ..., y_n)$ et donc une valeur k de K

Tableau 1 **Sélection de la variable X_i la plus corrélée avec une variable d'intérêt Y, parmi X_1, X_2, ..., X_{20}. Pour un échantillon de référence, la variable 2 est la plus corrélée (en valeur absolue) à Y. La validation croisée, examinant les corrélations sur un réplicat, indique l'absence totale de validité externe de cette procédure, la variable 9 étant désormais la plus corrélée. L'illusion provient du fait que $r = 0{,}53$ ne diffère pas significativement de 0 lorsqu'il est rapporté non pas à la distribution d'un coefficient de corrélation, mais à celle du plus grand coefficient de corrélation en valeur absolue parmi 20. L'ensemble des données de cet exemple est en fait formé de nombres au hasard.**

Numéro i de la variable X_i	*Échantillon de référence*	*Réplicat de l'échantillon*
1	0,05	– 0,10
2	**0,53**	***– 0,24***
3	0,22	0,07
4	0,17	– 0,08
5	– 0,21	0,12
6	– 0,60	0,22
7	– 0,08	0,13
8	– 0,21	0,26
9	– 0,60	**0,36**
10	0,12	– 0,00
11	– 0,04	0,12
12	– 0,04	– 0,07
13	– 0,09	0,36
14	– 0,16	– 0,06
15	– 0,31	– 0,09
16	0,00	0,15
17	0,17	0,25
18	0,24	0,20
19	0,38	– 0,01
20	0,40	0,01

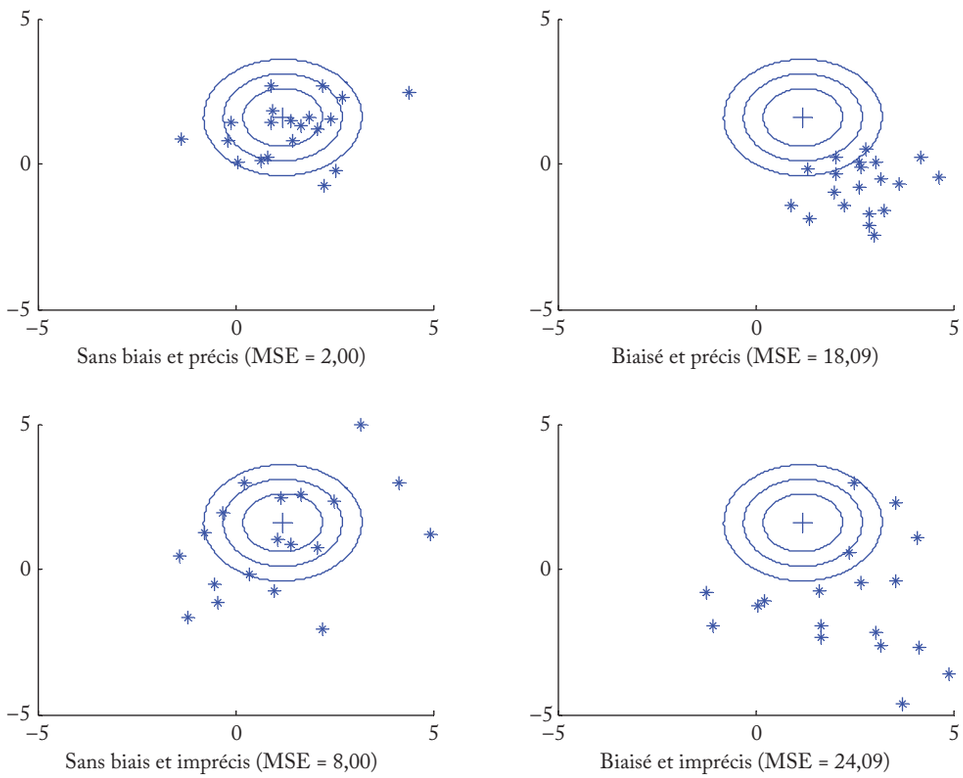

Figure 1 **Quatre types d'estimateurs : l'analogie de la cible et des fléchettes.** Le paramètre bidimensionnel inconnu est le centre de la cible, les étoiles sont des réalisations d'estimateurs. L'erreur quadratique moyenne (« *Mean Squared Error* », MSE), définie comme le carré du biais plus la variance, résume en un seul nombre les propriétés de chaque estimateur.

obtenus, cette probabilité est une fonction de p seulement, notée $L(p)$. L'estimation par MV de p consiste à prendre pour estimateur la quantité qui maximise cette fonction, appelée « vraisemblance », ou de façon équivalente, son logarithme $\log\left(L\left(p\right)\right) = \log\left(C_n^k\right) + k\log\left(p\right) + (n-k)\log\left(1-p\right)$. Ce maximum correspond au point d'annulation de la dérivée, solution de $\frac{d\log\left(L\left(p\right)\right)}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = 0$. En effet, la dérivée seconde $\frac{d^2\log\left(L\left(p\right)\right)}{dp^2} = -\frac{k}{p^2} - \frac{n-k}{\left(1-p\right)^2}$ est négative : la fonction $\log\left(L\left(p\right)\right)$ a sa concavité tournée vers le bas et admet bien un maximum au « zéro » de sa dérivée. La variable aléatoire $\widehat{p} = \frac{K}{n}$, racine de l'équation ci-dessus, est « l'estimateur du maximum de vraisemblance » (EMV) de p.

Quand le nombre d'observations augmente, la dérivée de la log-vraisemblance rapportée au nombre d'observations, $\frac{1}{n}\frac{d\log\left(L\left(p\right)\right)}{dp} = \frac{\widehat{p}-p}{p\left(1-p\right)}$, tend vers une loi normale d'espérance nulle (car $E(K) = np$) et de variance $\frac{1}{np\left(1-p\right)}$ (car var $(K) = np(1-p)$). Comme cette variance tend vers 0 lorsque $n \to \infty$, l'estimateur converge donc, dans un sens précisé par les mathématiques (Tassi 1985 p. 182), vers la vraie valeur du paramètre. Il suffit donc, si les données proviennent bien du modèle considéré, d'augmenter le nombre d'observations pour rapprocher en probabilité $\widehat{p}$ de la vraie valeur inconnue. L'exemple d'une partie de pile ou face est bien caractéristique de cette propriété (Figure 2). Dans le cas général, sous des conditions peu restrictives, les EMV ont asymptotiquement (c'est-à-dire quand $n \to \infty$) d'excellentes propriétés (Figure 3). Pour cette raison, la méthode du MV est largement utilisée mais ne dit rien en elle-même sur le modèle sous-jacent. Ainsi, quand on qualifie une approche, par exemple de reconstruction d'arbres phylogénétiques de « méthode du MV » (par ex. Sullivan 2006), on devrait dire qu'il s'agit de méthode basée sur un modèle probabiliste paramétrique traité par MV. Une telle approche est toujours préférable à une approche *ad hoc*, mais encore faut-il préciser et évaluer le modèle probabiliste.

La méthode du MV permet aussi d'obtenir des estimations de précision des estimateurs et donc des « intervalles de confiance », qui sont des intervalles de valeurs acceptables au sens d'un test. Il ne s'agit pas d'intervalles où le paramètre a une certaine probabilité de se trouver, puisque le paramètre est supposé avoir une valeur inconnue fixe, reflétant « l'état du monde ». Si l'on souhaite refléter l'incertitude sur les paramètres en les traitant comme des quantités aléatoires, il faut sortir de ce cadre classique, celui de la statistique dite « fréquentiste », pour se placer dans le cadre de la « statistique bayésienne » (Ellison 2004). Là aussi, une certaine confusion règne : on peut utiliser des méthodes bayésiennes sans pour autant entrer forcément dans la statistique bayésienne proprement dite. Ainsi, le calcul de la vraisemblance pour l'obtention d'EMV réclame parfois, notamment dans des modèles mixtes avec des distributions non normales (cf. Paragraphe 4), des intégrations multiples impraticables par les méthodes

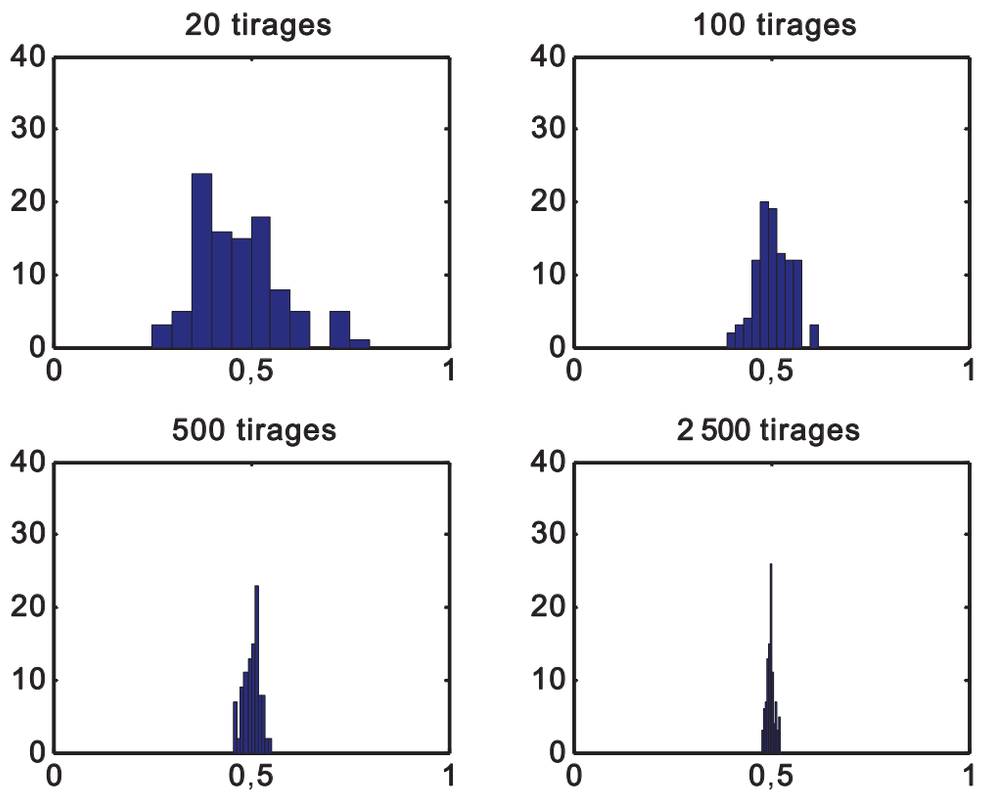

Figure 2 Proportion de « piles » dans des parties de pile ou face à $n = 20$, 100, 500 et 2 500 tirages. Dans chaque cas, l'histogramme de 100 répétitions est montré. La convergence de l'estimateur du maximum de vraisemblance vers 0,5 quand l'effectif n de l'échantillon augmente est clairement visible.

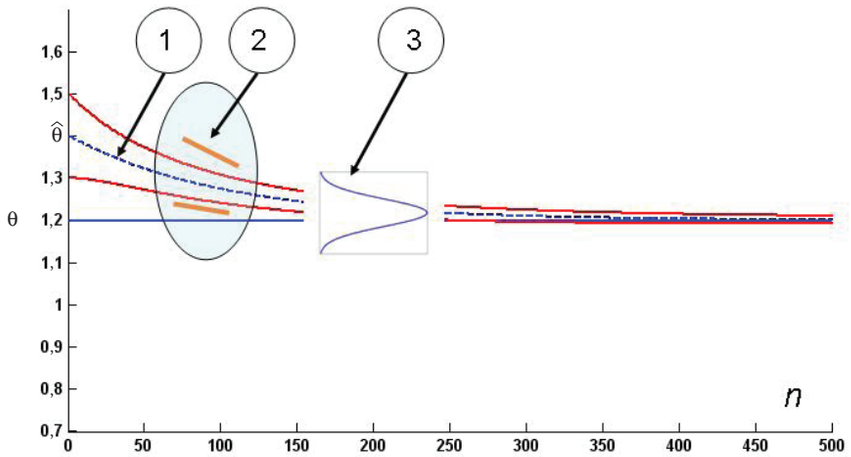

Figure 3 **Illustration des propriétés de l'estimateur du maximum de vraisemblance (EMV) $\hat{\theta}$ d'un paramètre θ en fonction du nombre d'individus n de l'échantillon :**
1. asymptotiquement sans biais ;
2. asymptotiquement de variance minimale parmi les estimateurs asymptotiquement sans biais ;
3. asymptotiquement distribué normalement.

usuelles. Des algorithmes performants et attractifs, relevant de la statistique bayésienne, comme les méthodes de chaînes de Markov par Monte-Carlo (appliquées à des données génétiques par Raymond et Rousset 1995) sont alors de plus en plus couramment utilisés (notamment grâce à la disponibilité de logiciels tels que WinBUGS, *e.g.* Gimenez *et al.* 2009), sans que l'on sorte pour autant d'une approche classique par MV ni donc de la statistique paramétrique fréquentiste.

Une part des propriétés des modèles dépend de leur structure profonde, le squelette relationnel du modèle donné dans l'Équation [1] ; une autre part d'une enveloppe extérieure, les hypothèses distributionnelles. On peut ainsi aisément démontrer que l'on peut estimer σ^2 sans biais par $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - Y_.\right)^2$ sans la moindre hypothèse de normalité, ni même sans supposer que les ε_i aient la même distribution. Il s'agit d'une propriété « sans hypothèses distributionnelles » ou, avec la concision de l'anglais, « *distribution free* ». Mais elle est tout sauf « *non-parametric* » puisqu'elle concerne un paramètre. L'absence d'hypothèses distributionnelles et l'absence de paramètres sont malheureusement souvent confondues en français sous l'appellation « non-paramétrique ». Les méthodes non-paramétriques *sensu stricto*, c'est-à-dire sans paramètres, sont fréquemment basées sur la conversion des données en rangs et sur des méthodes combinatoires, par exemple pour comparer deux distributions sans exprimer leurs différences éventuelles sous forme paramétrique (Conover 1980, Potvin et Roff 1993). Leur emploi est souvent basé sur un raisonnement partiellement circulaire : c'est le cas lorsqu'une différence de variance, impliquant nécessairement une différence des distributions, écarte de l'emploi de tests paramétriques simples de comparaisons de moyenne. On perd de toute façon l'interprétabilité attachée aux paramètres (Johnson 1995).

Quel que soit le cadre ou la méthode utilisée, il faut bien comprendre que les données n'ont jamais été générées par un modèle, sauf dans le cas de simulations. Le plus que l'on puisse demander à la statistique, c'est de nous permettre de dire « *avec les outils dont je dispose, les données ne sont pas distinguables de données issues de tel ou tel modèle* ». Les outils de diagnostic de la qualité de l'ajustement (« *goodness-of-fit* ») sont dès lors au moins aussi importants que l'obtention des estimations et des tests (pour la régression linéaire, Tomassone *et al.* 1983), ou que la notion de « robustesse » : comment se comporte tel estimateur ou tel test si l'on s'écarte des hypothèses du modèle ? La normalité de la distribution des Y_i est ainsi relativement secondaire dans les modèles linéaires (Ito 1980 p. 205), contrairement à ce qu'écrivent bien des ouvrages de statistique pour biologistes.

3 MODÈLES LINÉAIRES MIXTES

3.1 Modèles à effets fixes, modèles à effets aléatoires, modèles mixtes

Les modèles linéaires sont depuis l'origine intimement liés à la génétique, l'agronomie et la biologie évolutive : la dénomination « régression linéaire » provient des travaux de Galton (1886) sur l'hérédité de la taille corporelle ; l'analyse de variance a été formalisée par Fisher (1918) pour traiter des expériences agronomiques en plans contrôlés ; les modèles de composantes de la variance se sont développés avec les besoins de l'amélioration animale (Searle *et al.* 1992).

Le modèle standard (modèle à effets fixes) d'analyse de variance à un facteur (ANOVA, pour « *ANalysis Of VAriance* ») permet de comparer les moyennes de I groupes à partir de :

$$Y_{ij} = \mu + a_i + \varepsilon_{ij} \qquad [2]$$

où Y_{ij} est la variable aléatoire attachée à l'individu j ($j = 1, ..., n_i$) du groupe i ($i = 1, ..., I$). Ce modèle se distingue donc du modèle [1] par une part systématique plus compliquée,

qui suppose que les espérances de la variable étudiées dans les différents groupes i ($i = 1, ..., I$), $E(Y_{ij}) = \mu + a_i$, peuvent différer dès lors que l'on s'écarte de l'hypothèse $a_i = 0$, pour tout i.

Dans les notations traditionnelles, à partir des moyennes des observations par groupes $y_{i.}$ et de la moyenne générale $y_{..}$, le « test F » compare l'importance des effets estimés $\hat{a}_i = y_{i.} - y_{..}$ aux écarts résiduels $y_{ij} - y_{i.}$. Pour cela, on rapporte sous l'hypothèse H_0, $a_i = 0$ pour tout i, complétée de l'hypothèse de normalité, le rapport $\dfrac{\sum_{i=1}^{I} n_i \hat{a}_i^2 / (I-1)}{\sum_{i=1}^{I}\sum_{j}\left(y_{ij} - y_{i.}\right)^2 / (n-I)}$ à une distribution « F » de Fisher-Snedecor (Dagnélie 2006).

Mais, si les modalités du facteur ($i = 1, ..., I$) ne sont pas prédéfinies et proviennent par tirage au hasard d'une « population de modalités », il est plus judicieux de considérer les effets a_i comme des réalisations d'une variable aléatoire A_i (avec $E(A_i) = 0$, $\mathrm{var}(A_i) = \sigma_A^2$, et indépendance des A_i entre eux et avec les ε_{ij}). C'est ainsi le cas dans l'étude de la taille de cèdres à 20 ans dans une plantation comparative de différentes provenances (Figure 4). Les paramètres a_i ($i = 1, ..., k$) du modèle « fixe » disparaissent dans le modèle « aléatoire » :

$$Y_{ij} = \mu + A_i + \varepsilon_{ij} \qquad [3]$$

au profit du seul paramètre $\mathrm{var}(A_i) = \sigma_A^2$, qui vient se ranger aux côtés de $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$ comme une seconde « composante de la variance ». On parle de modèles (à effets) aléatoires, et lorsque effets fixes et aléatoires sont simultanément présents, de modèles mixtes. C'est la part aléatoire du modèle [1] qui est modifiée quand on passe au modèle [3], et non plus la part systématique comme c'était le cas avec le modèle [2]. Dans le modèle aléatoire d'ANOVA, reconnu depuis longtemps comme proche mais différent du modèle fixe (Eisenhart 1947), on teste l'effet *provenance* en testant la nullité de la variance inter-provenance (H_0 : $\sigma_A^2 = 0$ vs. H_1 : $\sigma_A^2 > 0$). Le test utilisé est identique au test F du modèle fixe, mais la logique en diffère profondément et considérer un facteur fixe ou aléatoire n'est pas anodin (Bennington et Thayne 1994). Tout d'abord, la portée des résultats diffère : dans l'exemple du cèdre, le modèle fixe permet de tirer des conclusions sur les I provenances étudiées, sans extrapolation possible à d'autres populations, alors que dans le modèle aléatoire le test porte sur l'ensemble de toutes les provenances qui auraient pu être utilisées ; il a donc une portée plus générale. Le test F indique un effet significatif de la provenance sur la hauteur et donc la croissance des cèdres ($F_{5,645} = 18$, $P < 0{,}0001$, Tableau 2A). Dans le modèle fixe, ce test montre seulement que les six provenances étudiées ont des croissances significativement différentes. Dans le modèle aléatoire, il indique une différence significative entre provenances de cèdres choisies au hasard. Mais cette plus grande généralité des conclusions repose sur le choix des modalités du facteur aléatoire (ici les provenances) : les 6 provenances de cèdres étudiées, pour être représentatives d'un plus grand ensemble de populations, doivent y avoir été échantillonnées de manière sensiblement aléatoire.

Trois questions peuvent guider le choix de considérer un facteur comme fixe ou aléatoire :

1. Les modalités du facteur ont-elles un intérêt pour elles-mêmes ou ont-elles été choisies au hasard ?

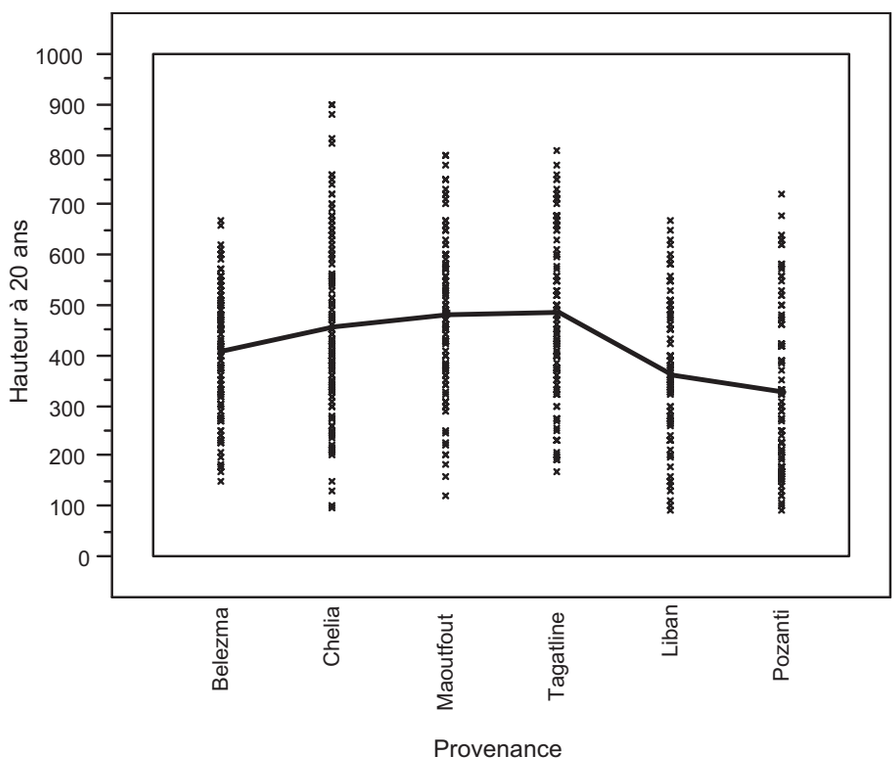

Figure 4 Dans un test de provenances de cèdres, 36 à 141 individus par provenance (pour 4 provenances de *C. atlantica* [Belezma, Chelia, Maoutfout et Tagatline] et 2 provenances de *C. libani* [Liban et Pozanti]) ont été mesurés à l'âge de 20 ans.

Tableau 2 **Effet de l'espèce sur la croissance des cèdres en test de provenances. Les résultats de l'analyse de variance à deux effets hiérarchisés déclarés fixes sont présentés dans les tableaux A et B. L'analyse de variance avec le facteur *espèce* fixe et le facteur *provenance* aléatoire est conduite avec la méthode de l'espérance des carrés moyens (CM). Dans ce cas, le dénominateur de la statistique du test F de l'effet *espèce* égale 0,88 × CM(provenance[espèce]) + 0,12 × CM(Résiduelle), ces pondérations étant calculées à partir de l'expression des espérances des carrés moyens *espèce* et *provenance[espèce]* (Dagnélie 2006). On utiliserait CM(provenance[espèce]) si le plan d'expérience était équilibré.**

A. Analyse de variance

Source	DDL	Somme des carrés	Carrés moyens	F	Prob > F
Modèle	5	2 157 342	431 468	18,0	<,0001
Résiduelle	645	15 449 092	23 952		
C. Total	650	17 606 434			

B. Tests dans le modèle à effets fixes

Source	DDL	Somme des carrés	Carrés moyens	F	Prob > F
Espèce	1	1 670 893	1 670 893	69,8	<,0001
Provenance[espèce]	4	431 468	107 602	4,5	0,0014

C. Tests dans le modèle à effets aléatoires

Source	DDL Num.	DDL Dénom.	Carrés moyen numérateur	Carrés moyen dénominateur	F	Prob > F
Espèce	1	4,25	1 670 893	97 327	17,2	0,0126
Provenance[espèce] et Random	4	645	107 602	23 952	4,5	0,0014

2. Les conclusions de l'analyse concerneront-elles les modalités étudiées ou un ensemble plus grand de modalités ?
3. Si l'expérimentation devait être refaite, les modalités étudiées seraient-elles les mêmes ou d'autres tirées dans un plus grand ensemble de modalités ?

3.2 Modèle mixte pour tester des effets fixes à partir d'observations non indépendantes

La présence d'un effet aléatoire induit une dépendance entre variables aléatoires, contrairement au cas des modèles fixes. En effet, si deux individus (j et l) de groupes différents (i et k) sont bien indépendants, les variables aléatoires de $Y_{ij} = \mu + A_i + \varepsilon_{ij}$ et $Y_{kl} = \mu + A_k + \varepsilon_{kl}$ étant toutes indépendantes, ce n'est plus le cas pour deux individus (j et l) d'un même groupe i car le même terme aléatoire A_i intervient dans $Y_{ij} = \mu + A_i + \varepsilon_{ij}$ et $Y_{il} = \mu + A_i + \varepsilon_{il}$. Une corrélation entre deux individus d'un même groupe apparait (Searle *et al*. 1992) :

$$\text{corr}\left(Y_{ij}, Y_{il}\right) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2} \qquad [4]$$

Dans l'exemple des cèdres, $\hat{\sigma}_A^2 = 3\,771$ et $\hat{\sigma}^2 = 23\,952$ conduisent à une corrélation intra-groupe estimée de 0,14 : la variance inter-groupe représente 14 % de la variance totale.

L'effet aléatoire engendre donc une dépendance entre observations. Le modèle est effectivement équivalent à un modèle sans l'effet aléatoire A_i, semblable au modèle [1], mais dans lequel on représenterait les dépendances entre individus dans les variances et les covariances des ε_{ij}. Cette approche, consistant à modéliser paramétriquement une « structure de covariance » entre variables aléatoires du modèle sans utiliser d'effets aléatoires, est un deuxième regard, plus récent, sur le modèle mixte. Quand le plan d'expérience et/ou le mode de recueil des données induisent une corrélation entre observations, on peut donc la modéliser et obtenir des tests adaptés en s'affranchissant de l'hypothèse d'indépendance des variables aléatoires ε, cruciale pour la validité des tests F dans les modèles à effets fixes. À côté des cas classiques de données répétées sur une même unité statistique au cours du temps et de données spatialisées (corrélation spatiale, Ives et Zhu 2006), la biologie évolutive s'intéresse notamment au cas de données concernant des espèces liées par une phylogénie (Blomberg *et al.* 2003) et aux observations sur des individus liés par des relations de parenté (pedigree) (Milner *et al.* 2000, Kruuk 2004).

Les modèles d'analyse de variance à effets hiérarchisés aléatoires présentent typiquement une structure de corrélation entre observations. Dans l'exemple de la Figure 4, les 6 provenances de cèdre proviennent de deux espèces et la *provenance* est considérée comme un facteur aléatoire hiérarchisé dans le facteur fixe *espèce* :

$$Y_{ijk} = \mu + a_i + B_{i(j)} + \varepsilon_{ijk} \qquad [5]$$

où Y_{ijk} est la $k^{\text{ème}}$ observation de la $j^{\text{ème}}$ provenance de l'espèce i. La « hiérarchisation », marquée par les parenthèses dans $B_{i(j)}$, indique que les provenances sont propres à chaque espèce : on ne peut parler d'un facteur « provenance » sans préciser d'abord l'espèce concernée. Les 651 arbres étudiés ne sont plus indépendants puisque les données d'une même provenance sont désormais corrélées (cf. [4]). La présence du facteur aléatoire, et donc de ces corrélations entre individus, modifie le test de l'effet fixe *espèce*. Dans le modèle avec effet *provenance* fixe, la variation résiduelle inter-individus (ε), seule source d'aléa, était utilisée comme référence (ou « terme d'erreur ») pour tester l'effet *espèce*, alors

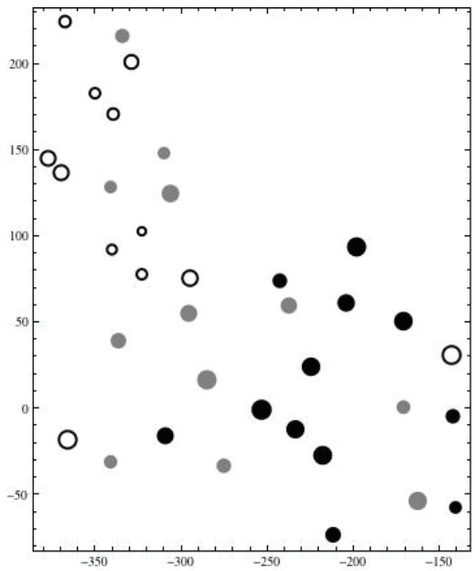

Figure 5 Position spatiale de 36 individus de cèdres issus de trois générations (G0 = ● = introduction en 1960, G1 = ● = issue de G0, G2 = ○ = issue de G0 et G1) dans une parcelle du Luberon. On représente leur croissance radiale sur leurs 30 premières années (proportionnelle au rayon du point). Celle-ci peut dépendre de la génération ou de l'environnement local (corrélation spatiale).

très significatif (P < 0,0001, Tableau 2B). Par contre, le modèle avec effet *provenance* aléatoire [5] contient deux sources d'aléa : la variation inter-provenance (B) et la variation inter-individu (ε). Le terme d'erreur utilisé pour tester l'effet fixe *espèce* est maintenant essentiellement composé de la variation inter-provenance (Tableau 2C). Dans ce test, valide même en présence d'une variabilité non-nulle entre provenances (var $B_{i(j)} > 0$)), l'effet *espèce* reste significatif, mais à un moindre degré (P = 0,0126, Tableau 2C).

Un autre exemple, dans lequel on spécifie directement une structure de covariance des variables ε, est celui de données corrélées spatialement (Ives et Zhu 2006), illustré par une étude de la croissance de 36 cèdres de trois générations successives (Figure 5). Les individus d'une même génération sont proches dans l'espace, signe d'une expansion de la population après son introduction. Cette proximité induit une confusion d'effets entre l'effet *génération* (adaptation génétique à un nouvel environnement) et un effet « micro-environnement » potentiellement dû à des zones plus favorables que d'autres (caractéristiques du sol, exposition, compétition interspécifique, etc.) : l'ANOVA classique, considérant les observations indépendantes, détecte un effet *génération*. L'utilisation du modèle mixte comprenant une structure de covariance spatiale pour les variables résiduelles permet d'estimer un paramètre de structure spatiale (δ, échelle à laquelle la corrélation spatiale est sensible) et de corriger notablement le test de l'effet fixe *génération* qui n'est plus significatif (P = 0,3, Tableau 3). Lorsque l'on considère des observations comme indépendantes de manière abusive et, d'une manière générale, quand la structure de covariance des termes aléatoires n'est pas correctement spécifiée vis-à-vis des contraintes d'échantillonnage, on parle souvent de « pseudo-réplication » (Hurlbert 1984).

3.3 L'exemple de données phénotypiques sur un pedigree

Enfin, dans les modèles mixtes, les effets aléatoires (A_i dans [3]) peuvent être considérés comme dépendants, liés par une première structure de covariance, avec une deuxième structure de covariance pour les variables aléatoires résiduelles (les ε_{ij} dans [3]).

Le modèle animal en génétique quantitative qui utilise un pedigree pour estimer l'héritabilité d'un caractère individuel, par exemple le nombre de descendants, est particulièrement pertinent pour la biologie évolutive. Avec une seule observation par individu du pedigree, il s'écrit (Lynch et Walsh 1998, Kruuk 2004) :

$$Y_i = \mu + A_i + \varepsilon_i \qquad [6]$$

où les A_i sont les valeurs génétiques additives des n individus du pedigree, avec $\mathrm{Var}(A_i) = V_A$, la variance génétique additive (cf. aussi Chapitre 13), et des covariances entre individus i et j proportionnelles aux coefficients de parenté ϕ_{ij} (Gallais 1990, Lynch et Walsh 1998) : $\mathrm{cov}(A_i, A_j) = 2V_A\phi_{ij}$. Les variables ε, avec $\mathrm{Var}(\varepsilon_i) = V_R$, modélisent les effets environnementaux et génétiques non additifs. Dans ce modèle, les effets aléatoires sont liés par des corrélations mesurant les ressemblances entre apparentés mais les erreurs résiduelles sont indépendantes. Connaissant le pedigree, et donc tous les coefficients de parenté, on estime la variance additive V_A et la variance phénotypique totale $V_P = V_A + V_R$ pour en déduire l'héritabilité du caractère.

Tableau 3 **Test de l'effet *génération* sur la croissance des cèdres du Luberon. Deux tests sont présentés : (i) avec une ANOVA à un facteur et (ii) dans un modèle mixte comprenant une structure de corrélation spatiale dans les variables résiduelles.**

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \text{ avec } \mathrm{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = \sigma^2 \exp\left(-\frac{d(ij, i'j')}{\delta}\right)$$

où Y_{ij} est le $j^{\text{ème}}$ arbre de la $i^{\text{ème}}$ génération et $d(ij,i'j')$ la distance entre les arbres ij et $i'j'$.

Modèle utilisé	*Valeurs estimées des paramètres de variance*	*Statistique F pour l'effet génération*	*P-valeur pour l'effet génération*
(i) Pas de prise en compte de la structure spatiale	$\hat{\sigma}^2 = 5\,210$	$F_{2,33} = 8,80$	P = 0,0009
(ii) Prise en compte de la structure spatiale	$\hat{\sigma}^2 = 6133$ $\hat{\delta} = 30,2$ m	$F_{2,29} = 1,08$	P = 0,315

Charmantier *et al.* (2006) analysent ainsi 2 141 tailles de couvées de cygnes tuberculés *Cygnus olor*, provenant de 502 mères issues d'un pedigree de 960 individus. Des estimations de la variance additive de 0,41 ± 0,13 et de la variance phénotypique de 2,09, on déduit une héritabilité estimée de la taille de couvée de 19,5 % ± 5,8. En s'appuyant sur une moyenne de 4,3 observations par mère, Charmantier *et al.* (2006) ajoutent une structure de covariance pour la résiduelle ε et distinguent un effet environnemental permanent de variance V_{EP} (commun à toutes les portées d'une mère mais indépendant de la valeur génétique additive) ainsi qu'un effet environnemental spécifique à chaque portée, de variance V_R. Avec V_{EP} estimé à 0,29 ± 0,11, une fraction égale à 14 % de la variance phénotypique est attribuée à des effets permanents (effets à long terme des conditions juvéniles, territoire de ponte, effets génétiques non additifs, etc.). Finalement, on peut aussi prédire les valeurs des effets aléatoires (« *Best Linear Unbiased Predictors* » ou BLUP). Charmantier *et al.* (2006) prédisent ainsi les valeurs génétiques des 502 mères pour la taille de couvée et montrent une augmentation significative avec l'année de naissance des mères, indiquant donc les effets d'une sélection directionnelle sur ce caractère.

En conclusion, les modèles mixtes peuvent donc être vus comme différentes manières de modéliser des corrélations entre observations, parfois pour mieux tester les effets fixes, parfois pour estimer les composantes de la variance. Malgré leur technicité et leur diversité, ils bénéficient d'une écriture mathématique synthétique (Searle *et al.* 1992, Kruuk 2004) qui permet leur traitement mathématique (Searle *et al.* 1992) et informatique (Littell *et al.* 2006). À côté de logiciels spécialement dévolus aux modèles mixtes (*e.g.* ASReml, Gilmour *et al.* 2006), la procédure « PROC MIXED » du logiciel SAS (Littell *et al.* 2006) propose une gamme très riche de méthodes d'estimation et de structures de covariances. Le logiciel R était plus centré sur les modèles à effets hiérarchisés jusqu'au développement du package lme4 (Pinheiro et Bates 2000).

Malgré les variétés des utilisations des modèles mixtes sous leur forme générale et leur apport potentiel aux questions de biologie évolutive, ils restent sous-utilisés (Ives et Zhu 2006). Cette sous-utilisation tient peut-être au poids de la méthode d'estimation historique, basée sur l'espérance des carrés moyens, peu adaptée aux structures complexes et aux données déséquilibrées. La généralisation de méthodes d'estimation comme le MV restreint (REML) devrait permettre de populariser ces modèles.

4 STATISTIQUE ET DÉMOGRAPHIE : MODÈLES DE CAPTURE-MARQUAGE-RECAPTURE

Les tests des forces évolutives (Kingsolver *et al.* 2001) s'appuient de plus en plus sur des suivis temporels de populations naturelles qui permettent d'aborder des processus difficiles à traiter en laboratoire. Leur analyse, du fait de dépendances complexes au cours du temps, réclame des modèles statistiques spécifiques et pose des problèmes méthodologiques récurrents. Il est ainsi presqu'impossible de mesurer de façon directe la valeur sélective sur le terrain : il faudrait en effet suivre de façon continue tous les individus d'une population de leur naissance à leur mort. En pratique, les individus d'une population ne peuvent être détectés ou capturés que de façon imparfaite. Pour accéder aux composantes de la valeur sélective, il faut alors utiliser des modèles permettant d'estimer des paramètres démographiques tels que des probabilités de survie tout en considérant des probabilités de détection potentiellement inférieures à 1. On parle de modèles de « capture-marquage-recapture » (CMR, Lebreton *et al.* 1992). Il s'agit de modèles probabilistes dont les paramètres seront le plus souvent estimés par MV. Ces modèles sont plus largement utilisés en écologie et en biologie de la conservation (Williams *et al.* 2002), qu'en biologie évolutive (Clobert 1995, Cam 2009) où l'on a souvent supposé de façon abusive (voir Gimenez et al. 2008) que la détection était exhaustive pour se ramener aux méthodes statistiques classiques plus simples utilisées en épidémiologie humaine (Skalski *et al.* 1993).

4.1 Les modèles de CMR uni-états

Un protocole de CMR consiste en $J + 1$ occasions d'échantillonnage au cours desquelles I individus au total sont capturés ou observés (nous ne distinguerons pas ces deux cas, les modèles utilisés étant identiques). À chaque occasion, les individus non marqués reçoivent des marques uniques puis sont relâchés. L'identité des individus précédemment marqués est relevée avant leur relâché. Les données se ramènent à un ensemble de I histoires de détection individuelles faites de $J + 1$ « 1 » et « 0 » selon que l'individu a été capturé ou pas (Lebreton *et al.* 1992, Williams *et al.* 2002). Par exemple, pour 3 occasions de capture, l'individu i avec l'histoire $h_i = (1, 0, 1)$ a été capturé pour la première fois à la première occasion, marqué puis relâché, non capturé à la deuxième occasion, puis capturé à la troisième et dernière occasion.

L'estimation des paramètres démographiques s'appuie sur le modèle probabiliste de base suivant. Notons ϕ_{ij} la probabilité qu'un individu i ($i = 1, ..., I$) vivant à l'occasion j ($j = 1, ..., J$) survive jusqu'a l'occasion $j + 1$ et p_j la probabilité de détection à l'occasion j d'un individu vivant ($j = 2, ..., J + 1$). Par simplicité, on suppose que les probabilités de détection ne diffèrent pas entre individus (pas d'indice i) mais varient au cours du temps. La probabilité h_i de l'histoire ci-dessus est alors $\phi_{i2}\phi_{i3}(1 - p_2)p_3$. Sous l'hypothèse d'indépendance des individus, la vraisemblance totale est proportionnelle au produit $\prod_{i=1}^{I} h_i$. En pratique, les données de CMR ne permettent pas d'estimer une survie différente pour chaque individu et l'on s'appuiera sur diverses hypothèses d'homogénéité. Le modèle de base (voir Lebreton *et al.* 1992) suppose que la probabilité de survie est identique pour tous les individus et ne varie qu'au cours du temps. Diverses généralisations et particularisations ont été proposées : effets groupes (mâle *vs.* femelle par exemple), âge (jeune *vs.* adulte), contraintes sur le temps (survie constante ou variant linéairement avec des covariables environnementales) (Lebreton *et al.* 1992). Cette flexibilité conduit à une stratégie de modélisation basée sur un modèle « parapluie » contenant tous les effets biologiquement plausibles, à partir duquel on obtient d'autres modèles plus restrictifs par des contraintes linéaires, de façon analogue aux modèles linéaires généralisés (Lebreton *et al.*

1992). La vraisemblance est en général suffisamment complexe pour que l'obtention des EMV réclame des méthodes numériques itératives. Divers logiciels facilitent considérablement la mise en œuvre de ces méthodes pour les biologistes : les plus performants sont MARK (White et Burnham 1999), M-SURGE (Choquet *et al.* 2004) et E-SURGE (Choquet *et al.* 2009). La qualité d'ajustement du modèle de départ peut être évaluée à partir de tests de tables de contingence (logiciel U-CARE, Choquet *et al.* 2005), et la sélection de modèles s'effectue ensuite en général par minimisation d'un critère global comme l'AIC (Burnham et Anderson 2002) pour éviter de multiplier les tests.

Des approches bayésiennes (algorithme de type MCMC) permettent par exemple de considérer des effets aléatoires (Gimenez *et al.* 2009) en contournant les problèmes d'intégration multiple qu'ils soulèvent. Le point de vue bayésien en tant que tel, notamment l'injection d'information *a priori*, reste encore peu répandu (McCarthy et Masters 2005).

4.2 Quantifier la sélection sur des traits phénotypiques

Quantifier l'action de la sélection sur des traits phénotypiques dans des populations naturelles permet, si ces traits sont héritables, de prédire leur réponse à l'évolution (Kingsolver *et al.* 2001). La méthode de Lande et Arnold (1983), qui relie la valeur sélective à des traits par régression multiple, est facilement transposable au contexte des CMR (Kingsolver et Smith 1995). La probabilité de survie, composante importante de la valeur sélective, peut ainsi être exprimée comme fonction d'une ou plusieurs variables explicatives :

$$\text{logit}(\phi_{ij}) = f(x_{ij}),\ i = 1, ..., I \text{ et } j = 1, ..., J$$

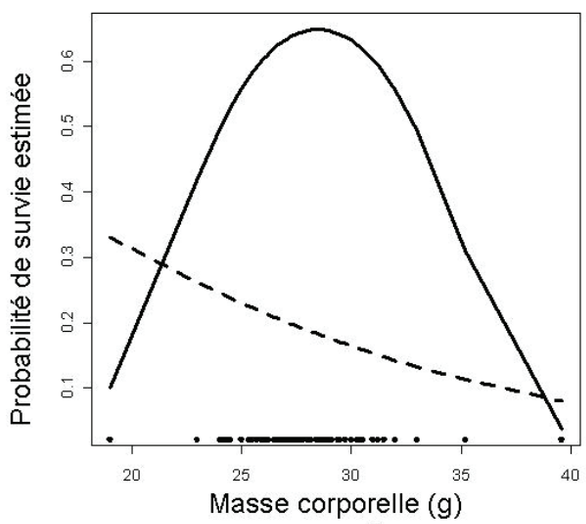

Figure 6 Relation entre survie et masse corporelle chez le tisserin social *Philetairus socius* par une analyse de CMR (trait plein) et une analyse naïve supposant une détection parfaite des individus (trait pointillé). Les valeurs observées de la masse sont représentées par des cercles pleins. Comme les probabilités de détection varient au cours du temps (entre 0,124 ± 0,045 et 0,829 ± 0,085) au cours des 7 années d'étude, l'analyse naïve sous-estime les probabilités de survie, mais conclut également à une variation linéaire plutôt que quadratique avec la masse corporelle.

où x_{ij} est la valeur de la variable explicative pour l'individu i à l'occasion j. Une fonction de lien, $\text{logit}(x) = \log(x/(1-x))$, permet de maintenir l'estimateur de la probabilité de survie entre 0 et 1. Le prédicteur $f(x_{ij})$ peut être linéaire ($f(x) = \beta_0 + \beta_1 x$) ou bien quadratique ($f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$), ou basé sur une forme plus souple sans *a priori* (Gimenez *et al.* 2006). Covas *et al.* (2002) étudient ainsi la survie du tisserin social *Philetairus socius* en Afrique du Sud de 1993-2000 à partir du marquage de 435 jeunes oiseaux. La probabilité de détection varie au cours du temps. En comparant les modèles reliant la probabilité de survie annuelle de façon respectivement linéaire et quadratique avec la masse corporelle, on peut discriminer entre sélection directionnelle (relation linéaire) et sélection stabilisante (pic de survie pour des valeurs intermédiaires de la masse : relation quadratique). On se ramène à la méthode de Lande et Arnold (1983) si l'on suppose que la détection est complète. Alors que cette approche naïve favorise le modèle linéaire et indiquerait donc une sélection directionnelle en faveur des individus les plus légers, l'approche par CMR sélectionne le modèle quadratique et plaide donc pour une sélection stabilisante, avec une survie optimale autour de la masse moyenne (Figure 6 ; Gimenez *et al.* 2008). Cet exemple illustre bien les risques d'inférence erronée sur la forme de la sélection quand l'imperfection de la détectabilité n'est pas prise en compte.

4.3 Modèles de CMR multi-états

Dans les modèles multi-états, les individus peuvent non seulement survivre au cours du temps, mais aussi se déplacer entre différents sites ou états (Hestbeck *et al.* 1991). Concrètement, si l'on considère 2 sites A et B, les histoires seront faites de 0, A et B pour spécifier des individus respectivement non-détectés, détectés sur le site A et détectés sur le site B. En termes de paramètres, les probabilités de survie ou de détection pourront dépendre du site de départ et/ou d'arrivée. Mais il faut aussi introduire des probabilités de transition, ψ_j^{pq}, qu'un individu vivant à l'occasion j ($j = 1, ..., J$) dans l'état p = A ou B se déplace vers l'état q = A ou B entre les occasions j et $j + 1$. Les états peuvent être des sites géographiques mais aussi des états définis au niveau individuel, comme l'état reproducteur. Ces modèles permettent d'aborder une grande variété de questions biologiques (Lebreton et Pradel 2002) : dispersion (Hestbeck *et al.* 1991), compromis entre traits d'histoire de vie (Nichols *et al.* 1994), accession à la reproduction (Lebreton *et al.* 2003).

Pradel *et al.* (1997) étudient ainsi l'âge de première reproduction chez le flamant rose *Phoenicopterus ruber roseus* en Camargue à partir de 7 822 individus bagués comme poussins (âge 0) entre 1977 et 1988 et revus comme reproducteurs entre 1983 et 1994. Une première reproduction précoce ne procurera un gain de valeur sélective que si elle n'est pas contrebalancée par un coût en termes de survie ou de reproduction future, c'est-à-dire par un « compromis démographique » (Stearns 1992, voir aussi Chapitre 10). Y-a-t-il un âge optimal de première reproduction ? On considère un modèle dit « de recrutement » à deux états, « non-reproducteur » et « reproducteur » (Nichols *et al.* 1994). Les non-reproducteurs n'étant pas observables, on fixe leur probabilité de détection à 0, tandis que celle des reproducteurs varie au cours du temps. Dans de tels modèles avec un ou plusieurs

états non-observables, l'estimation des paramètres exige en général des hypothèses complémentaires qui doivent être les plus réalistes possibles. Dans les modèles de recrutement, on suppose en général que :

- passé l'âge de première reproduction au niveau de l'espèce, la survie des non-reproducteurs est identique à celle des reproducteurs ;
- la transition de reproducteur à non-reproducteur est impossible : une fois reproducteur, un flamant le reste.

La probabilité de transition de l'état non-reproducteur à reproducteur peut alors être estimée. Les modèles de CMR multi-états permettent donc de tester des hypothèses directement sur le paramètre d'intérêt, ici la probabilité d'accession à la reproduction. On peut ainsi examiner à quel âge elle se stabilise, en comparant les modèles correspondant à différentes valeurs, cet exercice conduisant ici à l'âge de 10 ans. Sur la base de ce modèle, l'âge le plus probable d'accession à la reproduction est voisin de 7 ans (Figure 7).

Les exemples choisis n'illustrent qu'une part du rôle et du potentiel des modèles de CMR pour la biologie évolutive. On peut citer notamment leur apport aux sujets suivants :

- sénescence et probabilités de survie (par ex. Nichols *et al.* 1997, Loison *et al.* 1999) ;
- évolution de la dispersion (par ex. Casula 2006) ;
- mise en évidence de compromis évolutifs (voir Moyes *et al.* 2006, Townsend et Anderson 2007 pour deux exemples récents).

Récemment, une généralisation des modèles de CMR multi-états, les modèles multi-événements (Pradel 2005), a été proposée pour pallier les incertitudes sur l'assignation d'un individu observé à un état. L'observation d'un oiseau non accompagné d'un poussin mais au sein d'une colonie de reproduction peut ainsi laisser un doute sur son statut reproducteur ou non-reproducteur. Le principe est de séparer processus démographique et états sous-jacents d'une part, et détection par des événements liés de façon probabiliste aux états d'autre part (voir aussi Gimenez *et al.* 2007). Ces modèles laissent entrevoir la possibilité d'estimer les états successifs de chaque individu, la séquence des événements de reproduction par exemple, et d'accéder ainsi à des composantes essentielles de la valeur sélective (McGraw et Caswell 1996, Link *et al.* 2002) malgré les divers niveaux d'incertitude (détection incomplète, incertitude sur l'état reproducteur) présents dans les données. Ce serait répondre de façon correcte à une question clé abordée de façon *ad hoc* par l'école du « *Lifetime Reproductive Success* » (Clutton-Brock 1988, Newton 1989) qui considère le nombre de descendants observés cumulé sur plusieurs années comme connu sans incertitude ni biais.

L'application des modèles de CMR en biologie évolutive s'oriente ainsi vers une intégration explicite d'aspects individuels dans les analyses. Une direction essentielle en ce sens, s'appuyant sur la parenté des modèles de CMR avec les modèles linéaires généralisés, concerne les effets aléatoires. Il s'agit de développer des modèles de CMR mixtes (cf. Paragraphe 3) pour considérer des effets aléatoires individuels et des structures de dépendance complexes. Dans le premier cas, on pourra prendre en compte l'hétérogénéité individuelle de survie, susceptible de masquer la sénescence (Vaupel et Yashin 1985, Cam *et al.* 2002) ; dans le second, la prise en compte des relations de parenté entre individus *via* l'intégration du pedigree devrait permettre de combiner « modèle animal » et modèles de CMR pour explorer la question de l'héritabilité de la survie et de traits phénotypiques.

Plusieurs publications (Clobert 1995, Nichols et Kendall 1995, Lebreton et Pradel 2002, Cam 2009) discutent les utilisations et le potentiel des méthodes de CMR pour la biologie évolutive.

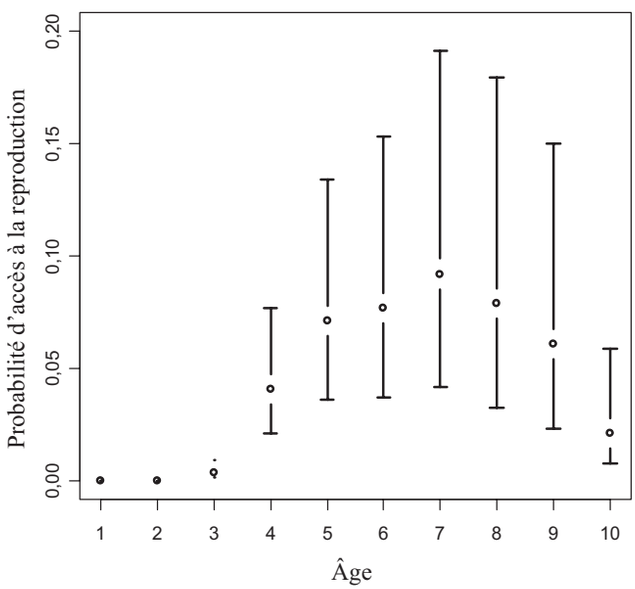

Figure 7 Probabilité d'accession à la reproduction en fonction de l'âge pour les flamants roses (barres verticales : intervalles de confiance à 95 %). Le modèle d'AIC minimal indique une stabilisation du recrutement à 10 ans. Les probabilités de survie des jeunes et des adultes sont estimées à 0,842 (± 0,257) et 0,963 (± 0,004) respectivement.

5 STATISTIQUE EN GÉNÉTIQUE ÉVOLUTIVE

L'analyse de données génétiques de populations naturelles couvre des questions très diverses : histoire sélective de mutations, histoire des populations (phylogéographie) et de phénomènes épidémiologiques, démographiques (effectifs de populations, taux de dispersion, etc.), et génétique proprement dite (taux de recombinaison, taux de mutation, héritabilité, etc.). Nous illustrons ici quelques spécificités et questions récurrentes : dépendance entre individus d'une population, rôle central des modèles d'évolution, critères de choix des méthodes.

5.1 Estimation d'un taux de mutation

Il s'agit d'estimer le taux de mutation à un locus à partir de génotypes d'individus prélevés simultanément dans une population. *A priori*, plus le taux de mutation est élevé, plus

deux gènes d'individus pris au hasard ont de chances de différer. En supposant que la population d'effectif N est haploïde, sans structure démographique, et que les mutations ne créent que des allèles nouveaux, la probabilité d'identité de deux gènes est au bout d'un temps infini $q = \frac{1}{1+2N\mu}$, où μ est le taux de mutation par gamète à ce locus (Malécot 1948). La probabilité d'identité décroît donc bien avec le taux de mutation, mais aussi avec N, l'effectif de la population. On peut alors, à partir de la formule précédente, estimer le produit $N\mu$ par :

$$\frac{\left(1-\frac{1}{\hat{q}}\right)}{2} \qquad [7]$$

où $\hat{q}$ est la proportion de paires de gènes identiques dans un échantillon qui estime donc la probabilité d'identité. Cet estimateur intuitif, dans lequel les deux paramètres N et μ ne sont pas séparément estimables, n'utilise que l'information contenue dans les paires de gènes.

Mais on peut aussi utiliser toute l'information de l'échantillon, c'est-à-dire les effectifs de différents allèles, rassemblés dans un vecteur noté E. La probabilité d'obtenir les effectifs E peut s'écrire (Ewens 1972) :

$$\Pr(E\,;N\mu) = \Pr(k \text{ allèles}\,;N\mu)\times\Pr(E/k \text{ allèles}) \quad [8]$$

où $\Pr(k \text{ allèles}\,;N\mu)$ est la probabilité d'observer k allèles dans l'échantillon, uniquement fonction du produit $N\mu$, et où la probabilité des effectifs alléliques pour k donné $\Pr(E/k \text{ allèles})$ ne dépend ni de μ ni de N. En l'absence d'informations supplémentaires, les deux paramètres ne sont donc toujours pas séparément estimables et $N\mu$ doit être traité comme un seul paramètre.

Dans la factorisation [8], le second terme ne contient pas le paramètre $N\mu$. Toute l'information sur $N\mu$ est donc contenue dans le nombre d'allèles observé k, appelé pour cette raison « statistique exhaustive ». L'EMV est alors la valeur de $N\mu$ qui maximise $\Pr(k \text{ allèles}\,;N\mu)$: il bénéficie des propriétés optimales usuelles mais doit être calculé numériquement. L'erreur quadratique moyenne (MSE, Figure 1) de l'estimateur intuitif [7] est effectivement plus élevée que celle de l'EMV, du fait de l'optimalité de ce dernier, pour des tailles d'échantillon modérées (par exemple, pour 100 individus, d'un facteur 5 quand $N\mu = 0{,}05$ et 2,47 quand $N\mu = 5$; Tavaré 2004). Mais le degré de « robustesse » des estimateurs aux hypothèses des modèles est aussi crucial (Staudte et Sheather 1990) : un estimateur sensible à des hypothèses irréalistes aura par exemple un fort biais, et un estimateur de plus grande variance mais robuste peut être préférable à un EMV fondé sur un modèle trop restrictif. Par exemple, la formulation d'Ewens (1972, [8]) ignore les effets de fluctuations passées de l'effectif ou ceux d'une structuration spatiale, qui peuvent avoir des conséquences marquées sur les estimations. Le problème de robustesse est particulièrement critique pour tout ce qui dépend des effectifs des populations (comme le produit $N\mu$). Pour prendre en compte la complexité biologique non représentée dans les modèles on remplace l'effectif N par une « taille efficace » N_e dans les formules. Cette approche *ad hoc* conduit à différentes définitions de la taille efficace dont le domaine d'application doit être précisé (Ewens 2004, Rousset 2004). Dans le cas de fluctuation des effectifs passés, il faut soit pouvoir inférer les aspects les plus cruciaux de l'histoire ancestrale, par exemple l'occurrence de goulots d'étranglement, soit définir des méthodes qui y sont peu sensibles, comme nous l'illustrons ci-dessous.

5.2 Structuration génétique

La « structuration génétique » peut être spatiale, suite à une dispersion limitée, ou résulter de l'apparentement entre individus (structuration inter-famille par exemple). Dans ce dernier cas, au sein d'une grande population haploïde sexuée se croisant au hasard, la probabilité que deux sœurs partagent un allèle A de fréquence p dans la population s'écrit : $\frac{p}{2}+\frac{p^2}{2}$. En effet, soit avec une probabilité $\frac{1}{2}$, les deux gènes sont copies du même gène parental, qui est A avec une probabilité p ; soit, avec une probabilité $\frac{1}{2}$, les deux gènes sont copies de deux gènes parentaux distincts, qui sont tous deux A avec une probabilité p^2. Plus généralement, par exemple en présence de dispersion (voir Chapitres 1 et 3), on peut souvent écrire une telle probabilité comme $Q = Fp + (1-F)p^2$ (Crow et Kimura 1970). F, égal donc à :

$$\frac{(Q-p^2)}{(p-p^2)} \qquad [9]$$

est appelé apparentement ou coefficient de consanguinité. Sa valeur résulte des relations généalogiques entre individus, partiellement accessibles en estimant F. La vraisemblance, pour plusieurs loci supposés indépendants, s'écrit :

$$L(F,p) = \prod_{\text{loci } l}\left(p_{i(l)}F + (1-F)\,p_{i(l)}p_{j(l)}\right)$$

où p désigne maintenant le vecteur des fréquences des allèles des différents loci. Pour le locus l, quand les deux gènes descendent d'un même ancêtre, $p_{i(l)}$ est la fréquence de l'unique allèle observé et, dans le cas contraire, $p_{i(l)}$ et $p_{j(l)}$ sont les fréquences des deux allèles observés, qui peuvent être identiques. Généralement, les fréquences alléliques sont inconnues et doivent elles-mêmes être estimées, avec des conséquences à la fois sur les calculs et sur les performances des estimateurs : on doit déterminer à la fois la valeur $\widehat{F}$ de F et le vecteur $\widehat{p}$ des fréquences alléliques conduisant au maximum absolu $L(\widehat{F},\widehat{p})$.

Diverses alternatives au MV peuvent donc être conçues. L'Équation [9] peut également s'écrire $F = \frac{Q-Y}{1-Y}$ où $Y = \sum_k p_k^2$ est la probabilité que deux gènes pris au hasard dans la population totale soient identiques. Q peut être interprété comme l'espérance d'une variable de Bernoulli X prenant la valeur 1 quand les deux gènes sont identiques, ou sinon 0. Un estimateur empirique de F est alors $\widetilde{F} = \frac{x-y}{1-y}$ où x est la proportion de paires de gènes identiques dans le sous-échantillon dont on cherche à estimer l'apparentement et y est la proportion de paires de gènes identiques dans l'échantillon total.

On arrive à un estimateur similaire en écrivant l'état allélique d'un gène de l'individu j dans le groupe (famille ou sous-population) i sous la forme d'un modèle linéaire mixte (cf. Paragraphe 3) :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad [10]$$

Ici Y_{ij} prend la valeur 1 si le gène est d'un type allélique donné et 0 s'il est d'un autre type. α_i sera par exemple un effet aléatoire affectant la famille i et ε_{ij} un effet aléatoire affectant indépendamment chaque individu. Estimer l'apparentement entre les individus de la famille i revient alors à estimer la covariance intra-famille entre les états alléliques, induite par la composante commune α_i (voir [4]). Mais les hypothèses de normalité et d'égalité des variances ne s'appliquent pas. Dans le cas de l'apparentement, divers estimateurs ont été considérés selon la nature exacte des données (Wang 2002, Milligan 2003, Konovalov et Heg 2008). Pour une structuration spatiale, cette approche a été formalisée par Weir et Cockerham (1984). Les relations entre les différentes approches sont discutées par Rousset (2007) et par Konovalov et Heg (2008).

5.3 Indépendance statistique en génétique des populations

Le modèle linéaire mixte ci-dessus montre bien qu'un échantillon n'est pas forcément formé d'individus indépendants. Si les probabilités (p_1, p_2, ..., p_k) de k allèles sont déterminées, un tirage de n individus indépendants induira une distribution multinomiale des effectifs alléliques dans l'échantillon Mult (n ; p_1, p_2, ..., p_k). Mais les fréquences alléliques étant souvent elles-mêmes des variables aléatoires, des dépendances entre individus apparaîtront, dont il faudra tenir compte par exemple pour déterminer la variance d'un estimateur de $N\mu$.

Une façon de mettre en évidence cette dépendance est de considérer une distribution particulière des fréquences alléliques. Dans le cas le plus simple, on suppose ainsi qu'il peut y avoir K allèles à un locus, avec des taux de mutation identiques d'un allèle vers un autre pour toutes les paires d'allèles (modèle KAM pour « *K-allele model* »). Pour une population panmictique suivant le KAM, conditionnellement aux fréquences alléliques, la distribution des effectifs alléliques $\mathbf{n} = (n_1, ..., n_K)$ sera comme ci-dessus multinomiale. Mais on obtient une distribution non conditionnelle des effectifs alléliques plus dispersée qu'une loi multinomiale, dite Dirichlet-multinomiale.

La variance de l'estimation de la fréquence d'un allèle donné, $\frac{2N\mu+n}{2N\mu+1}\frac{P(1-P)}{n}$, est alors inévitablement plus élevée que celle donnée par le modèle multinomial, $\frac{P(1-P)}{n}$, en conséquence directe du caractère aléatoire des fréquences alléliques p_k et de la dépendance induite entre individus.

Dans les études de structure génétique, on utilise parfois des modèles où les fréquences alléliques sont considérées comme fixées dans la population totale (comme dans les estimateurs d'apparentement) et variables dans les sous-populations (ou dans les familles, pour le cas de l'apparentement) : c'est précisément cette variation aléatoire qui contient l'information concernant les paramètres que l'on cherche à estimer. La performance d'estimateurs sous l'hypothèse de tirages multinomiaux (Chakraborty et Danker-Hopfe 1991, Pons et Petit 1995) est alors sans intérêt. Les intervalles de confiance des paramètres tels que $N\mu$ (ci-dessus) ou des taux de migration doivent plutôt tenir compte de la variation aléatoire des fréquences alléliques entre populations. C'est le cas dans le modèle de migration « en île», dans lequel on suppose qu'il y a un grand nombre de sous-populations et qu'un émigrant a la même probabilité d'immigrer dans n'importe laquelle des autres sous-populations. Si le taux d'immigration m est très supérieur au taux de mutation, on peut alors approcher la vraisemblance d'un échantillon tiré de plusieurs sous-populations comme le produit de plusieurs termes de type Dirichlet-multinomiale, où $N\mu$ est remplacé par Nm (Wright 1931), et construire des intervalles de confiance pour Nm numériquement à partir de cette vraisemblance.

5.4 Tendances actuelles

De nombreux développements récents reposent sur la théorie de la « coalescence » qui modélise la probabilité d'un échantillon en considérant les événements affectant les lignées ancestrales. Ainsi, pour que deux séquences d'ADN à un locus donné soient identiques, il faut qu'il n'y ait pas eu de mutation depuis leur ancêtre commun ; comme précédemment, on ignore la possibilité de mutations identiques dans différentes lignées. Pour calculer la probabilité correspondante, on peut d'abord considérer la distribution de probabilité des arbres généalogiques des copies de gènes considérés. Cette distribution dépend des effectifs, des événements de migration et d'autres processus démographiques, mais est indépendante de l'occurrence de mutations si celles-ci sont sélectivement neutres. On peut ensuite considérer la probabilité des événements mutationnels conditionnellement à une généalogie donnée. Si on arrêtait cette logique aux paires de gènes, on retrouverait au mieux les méthodes précédentes basées sur les probabilités d'identité de paires de gènes. Mais l'accroissement des moyens de calculs et les progrès théoriques (Tavaré 2004, Ewens 2004, Hein *et al.* 2005, Nordborg 2007) ont permis d'implémenter des algorithmes calculant la vraisemblance de jeux de données de grande taille, sous des scénarios démographiques variés. Ces méthodes permettent l'analyse par MV de problèmes allant de la divergence entre paires de populations (Beerli et Felsenstein 1999, Nielsen et Wakeley 2001) à l'isolement par la distance, c'est-à-dire la migration de proche en proche dans un réseau de populations (Rousset et Leblois 2007), situation illustrée par l'exemple suivant.

Watts *et al.* (2007) ont génotypé 14 loci chez 240 libellules (*Coenagrion mercuriale*) échantillonnées sur un segment de ruisseau long de 10 km, habitat approximativement linéaire. On peut comparer des méthodes d'inférences démographiques fondées sur les probabilités d'identité de paires de gènes et sur le traitement par MV d'un modèle probabiliste. La première méthode utilise la relation à l'équilibre entre distance géographique linéaire r dans un habitat linéaire et probabilités d'identité Q_r :

$$\frac{Q_0 - Q_r}{1 - Q_0} \approx a + \frac{r}{2D\sigma^2} \qquad [11]$$

où D est la densité linéaire de population, σ^2 le carré moyen de la distance parent-descendant (Rousset 1997). On peut donc estimer $D\sigma^2$ à partir de la variation spatiale d'estimateurs de $\frac{(Q_0 - Q_r)}{(1-Q_0)}$ et en obtenir un intervalle de confiance approché (Davison et Hinkley 1997).

Le calcul de la vraisemblance s'appuie sur un modèle mutationnel des loci, une description de la totalité des populations connectées par migration et une forme particulière de la distribution de dispersion. Dans le calcul suivant, on considère ainsi que l'on a n sous-populations de même taille N et que tous les loci évoluent sous un modèle KAM avec le même taux de mutation μ. On suppose qu'une fraction m des individus migre et que la distance parcourue suit une loi géométrique. On peut alors estimer un paramètre de mutation, $N\mu$, deux paramètres de dispersion, le produit Nm et le paramètre de la loi géométrique, et déduire de ces deux estimations un estimateur de $D\sigma^2$. On ne peut toujours pas séparer les paramètres N et μ (ou N et m). Alors qu'une étude par simulation montre que les estimations sont peu robustes vis-à-vis d'une mauvaise spécification du nombre de sous-populations, l'estimation de $D\sigma^2$ semble relativement robuste (Rousset et Leblois 2007). L'habitat de cette libellule étant essentiellement continu, le découpage en sous-populations est un artifice de calcul. Les mêmes simulations suggèrent que l'estimation de $D\sigma^2$ est d'autant plus précise que le nombre de sous-populations est élevé, mais des contraintes de calcul limitent ce nombre. Pour $n = 20$ et $n = 80$ sous-populations, les EMV de $2D\sigma^2$ sont ainsi de 113 000 et 92 000 individus.m, avec un intervalle de confiance à 95 % de 50 000 à 140 000 pour cette dernière valeur, alors que l'estimation *via* la relation [10] est de 223 000 individus.m (intervalle de confiance 66 000 – 393 000). Ces valeurs sont à comparer à une estimation démographique de $2D\sigma^2$, qui néglige divers facteurs pouvant réduire les effectifs efficaces des sous-populations, de 278 000 individus.m (Watts *et al.* 2007).

Les différences entre estimateurs sont ici très modérées par rapport aux biais souvent attribués aux estimations de taux de dispersion et de tailles efficaces (Whitlock et McCauley 1999, Frankham 1995), vraisemblablement parce que seuls les estimateurs les plus robustes ont été considérés.

Plusieurs de nos exemples doivent cependant être vus plus comme des exercices conceptuels illustrant des principes généraux, que comme des exemples d'inférence fiables. Le modèle démographique sous-jacent à l'estimateur du taux de mutation fondé sur la formule d'Ewens [8], aussi bien que le modèle de migration « en île » sont rarement applicables. La complexité technique, déjà marquée, s'accroît quand on considère des modèles plus réalistes. Malgré le potentiel de ces méthodes, le domaine de validité des inférences à partir des données génétiques reste à cerner et l'application de ces méthodes doit toujours s'appuyer sur une compréhension des modèles génétiques sous-jacents.

6 CONCLUSION

Au-delà de la diversité des méthodes présentées dans ce chapitre et de la complexité propre à chacune, le rôle central de la statistique paramétrique pour un champ scientifique aussi riche que la biologie évolutive apparaît clairement. Les possibilités de diagnostics de la qualité d'ajustement, la robustesse souvent élevée et l'interprétabilité des paramètres doivent pousser à utiliser pleinement la puissance modélisatrice de la statistique paramétrique. À côté de l'estimation linéaire et de la méthode du maximum de vraisemblance, les méthodes utilisant plus intensivement l'ordinateur, comme simulations, permutations et bootstrap, permettent d'ailleurs de s'affranchir assez largement des hypothèses distributionnelles traditionnelles, souvent vues comme un obstacle plus important qu'il n'est. Un domaine qui n'est pas présenté ici, l'étude des liaisons entre tableaux de données, est ainsi sorti du champ traditionnellement descriptif de l'analyse multivariée pour rejoindre celui des analyses de variance multivariées (MANOVA, Morrison 1967, ch. 5).

Dans les trois types de méthodes présentées, la prise en compte de la variabilité inter-individuelle et des dépendances entre individus, notamment par le biais des modèles markoviens et des effets aléatoires, est une convergence marquante. Cette évolution est encore en cours et les bénéfices que l'on peut en attendre pour la biologie évolutive sont loin d'être épuisés.

Les présentations et références de ce chapitre aideront, nous l'espérons, le chercheur confirmé comme le débutant à se forger une culture statistique en relation avec ses centres d'intérêt biologique. Au-delà de la bonne pratique d'un logiciel et de la sophistication propre à certains modèles, nous recommandons aux biologistes d'accorder suffisamment d'importance aux éléments de culture générale en statistique. Seule cette attitude leur donnera à la fois une autonomie suffisante dans leurs analyses statistiques, une capacité de dialogue pluridisciplinaire avec des statisticiens et des modélisateurs et une capacité à s'ouvrir aux nouveaux outils qui apparaissent sans cesse.

RÉFÉRENCES

Aitkin, M., Anderson, D., Francis, B. et Hinde, J. 1988. *Statistical modelling in GLIM*. Clarendon.

Beerli, P. et Felsenstein, J. 1999. Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**:763-773.

Bennington, C.C. et Thayne, W.V. 1994. Use and misuse of mixed-model analysis of variance in ecological studies. *Ecology* **75**:717-722.

Blomberg, S.P., Garland, T. et Ives, A.R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**:717-745.

Burnham, K.P. et Anderson, D.R. 2002. *Model selection and multimodel inference: A practical information-theoretic approach.* Springer Verlag, New York.

Cam, E. 2009. Contribution of capture-mark-recapture modeling to studies of evolutionary processes by natural selection. In *Modeling Demographic Processes in Marked Populations.* Thomson, D. L. ; Cooch, E. G. ; Conroy, M. J. (Eds.). Springer Series : Environmental and Ecological Statistics, Vol. 3, pages 83-129.

Cam, E., Link, W.A., Cooch, E.G., Monnat, J.-Y. et Danchin, E. 2002. Individual covariation in life-history traits: Seeing the trees despite the forest. *American Naturalist* **159**:96-105.

Casula, P. 2006. Evaluating hypotheses about dispersal in a vulnerable butterfly. *Ecological Research* **21**:263-270.

Chakraborty, R. et Danker-Hopfe, H. 1991. Analysis of population structure: a comparative study of Wright's fixation indices. *Statistical methods in biological and medical sciences*. C. R. Rao et R. Chakraborty. Amsterdam, North-Holland. **8**:203-254.

Charmantier, A., Perrins, C., McCleery, R.H. et Sheldon, B.C. 2006. Evolutionary response to selection on clutch size in a long-term study of the mute swan. *American Naturalist* **167**:453-465.

Choquet, R., Reboulet, A.-M., Pradel, R., Gimenez, O. et Lebreton, J.-D. 2004. M-SURGE: new software specifically designed for multistate capture-recapture models. *Animal Biodiversity and Conservation* **27.1**:207-215.

Choquet, R., Reboulet, A.M., Lebreton, J.-D., Gimenez, O. et Pradel, R. 2005. *U-CARE 2.2 User's Manual.* CEFE - CNRS, Montpellier.

Choquet, R., Rouan, L. et Pradel, R. 2009. Program E-SURGE: a software application for fitting Multievent models. In *Modeling Demographic Processes in Marked Populations*. Thomson, D. L. ; Cooch, E. G. ; Conroy, M. J. (Eds.). Springer Series: Environmental and Ecological Statistics, Vol. 3, pages 847-868.

Clobert, J. 1995. Capture-recapture and evolutionary ecology: A difficult wedding? *Journal of Applied Statistics* **22(5-6)**:989-1008.

Clutton-Brock, T.H., ed. 1988. *Reproductive Success. Studies of Individual Variation in Contrasting Breeding Systems*. University of Chicago Press, Chicago.

Conover, W.J. 1980. *Practical non parametric statistics*. New York, John Wiley & Sons.

Covas, R., Brown, C.R., Anderson, M.D. et Brown, M.B. 2002. Stabilizing selection on body mass in the sociable weaver *Philetairus socius*. *Proceedings of the Royal Society of London Series B-Biological Sciences* **269(1503)**:1905-1909.

Crawely, M.J. 1993. *GLIM for ecologists*. Blackell, Oxford.

Crow, J.F. et Kimura, M. 1970. *An introduction to population genetics theory*. Harper & Row, New York.

Dagnélie, P. 2006. *Statistique théorique et appliquée. Tome 2. Inférence statistique à une et à deux dimensions.* De Boeck et Larcier, Bruxelles.

Dagnélie, P. 2007. *Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique.* De Boeck et Larcier, Bruxelles.

Davison, A.C. et Hinkley, D.V. 1997. *Bootstrap methods and their applications.* Cambridge University Press, Cambridge, Mass.

Eisenhart, C. 1947. The asumptions underlying the analysis of variance. *Biometrics* **3**:1-21.

Ellison, A.M. 2004. Bayesian inference in ecology. *Ecology Letters* **7**:509-520.

Ewens, W.J. 1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**:87-112.

Ewens, W.J. 2004. *Mathematical population genetics I. Theoretical introduction*. Springer Verlag, New York, 2nd edn.

Fisher, R.A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**:399-433.

Frankham, R. 1995. Effective population size / adult population size ratios in wildlife: a review. *Genetical Research* **66**:95-107.

Gallais, A. 1990. *Théorie de la sélection en amélioration des plantes*. Paris, Masson.

Galton, F. 1886. Regression towards mediocrity in heredity stature. *Heredity* **15**:246-263.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. et Thompson, R. 2006. *ASReml 2 user guide.* VSN, Hemel, Hempstead.

Gimenez, O., Covas, R., Brown, C.R., Anderson, M.D., Bomberger Brown, M. et Lenormand, T. 2006. Nonparametric estimation of natural selection on a quantitative trait using capture-mark-recapture data. *Evolution* **60**:460-466.

Gimenez, O., Rossi, V., Choquet, R., Dehais, C., Doris, B., Varella, H., Vila, J.-P. et Pradel, R. 2007. State-space modelling of data on marked individuals. *Ecological Modelling* **206**:431-438.

Gimenez, O., Viallefont, A., Charmantier, A., Pradel, R., Cam, E., Brown, C.R., Anderson, M.D., Bomberger Brown, M., Covas, R., Gaillard, J.-M. 2008. The risk of flawed inference in evolutionary studies when detectability is less than one. *The American Naturalist* **172**:441-448.

Gimenez, O., Bonner, S., King, R., Parker, R.A., Brooks, S.P., Jamieson, L.E., Grosbois, V., Morgan, B.J.T. et Thomas, L. 2009. WinBUGS for Population Ecologists: Bayesian Modeling Using Markov Chain Monte Carlo Methods. In *Modeling Demographic Processes in Marked Populations*. Thomson, D. L. ; Cooch, E. G. ; Conroy, M. J. (Eds.). Springer Series: Environmental and Ecological Statistics, Vol. 3, pages 885-918.

Hein, J., Schierup, M.H. et Wiuf, C. 2005. *Gene genealogies, variation and evolution*. Oxford Univ. Press, Oxford, UK.

Hestbeck, J.B., Nichols, J.D. et Malecki, R. 1991. Estimates of movement and site fidelity using mark-resight data of wintering Canada geese. *Ecology* **72**:523-533.

Huet, S., Jolivet, E. et Messean, A. 1992. *La régression non-linéaire : méthodes et applications en biologie*. Paris, Inra.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**:187-211.

Iosifescu, M. et Tautu, P. 1973. *Stochastic processes and applications in biology and medicine. – Theory(1), Applications (2)*. Vol 3 et 4, Springer-Verlag.

Ito, P.K. 1980. *Robustness of ANOVA and MANOVA test procedures*. In Krisnaiah P.R. (ed). Handbook of statistics vol. 1. North Holland, Amsterdam, pp. 199-236.

Ives, A.R. et Zhu, J. 2006. Statistics for correlated data: Phylogenies, space, and time. *Ecological Applications* **16**:20-32.

Johnson, D.H. 1995. Statistical sirens: the allure of nonparametrics. *Ecology* **76**:1998-2000.

Kéry, M., Gregg, K.B. et Schaub, M. 2005. Demographic estimation methods for plants with unobservable life-states. *Oikos* **108**:307-320.

Kingsolver, J.G. et Smith, S.G. 1995. Estimating selection on quantitative traits using capture-recapture data. *Evolution* **49**:384-388.

Kingsolver, J.G., Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E., Hoang, A., Gibert, P. et Beerli, P. 2001. The strength of phenotypic selection in natural populations. *American Naturalist* **157**:245-261.

Konovalov, D.A. et Heg, D. 2008. A maximum likelihood estimator allowing for negative relatedness values. *Molecular Ecology Resources* **8**:256-263.

Kruuk, L.E.B. 2004. Estimating genetic parameters in natural populations using the 'animal model'. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* 359:873-890.

Lande, R. et Arnold, S.J. 1983. The measurement of selection on correlated characters. *Evolution* **37**:1210-1226.

Lebreton, J.-D. 2006. Dynamical and statistical models of Vertebrate population dynamics. *Comptes Rendus Biologies* **329**:804-812.

Lebreton, J.D. et Pradel, R. 2002. Multistate recapture models: modelling incomplete individual histories. *Journal of Applied Statistics* **29**:353-369.

Lebreton, J.D., Hines, J.E., Pradel, R., Nichols, J.D. et Spendelow, J.A. 2003. Estimation by capture-recapture of recruitment and dispersal over several sites. *Oïkos* **101**:253-264.

Lebreton, J.D., Burnham, K.P., Clobert, J. et Anderson, D.R. 1992. Modeling survival and testing biological hypotheses using marked animals – A unified approach with case studies. *Ecological Monographs* **62**:67-118.

Legay, J.-M. 1973. *La méthode des modèles, état actuel de la méthode expérimentale.* Informatique et Biosphère, Paris.

Link, W.A., Cooch, E.G. et Cam, E. 2002. Model-based estimation of individual fitness. *Journal of Applied Statistics* **29**:207-224.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. et Schabenberger, O. 2006. *SAS for Mixed Models, Second Edition*. Cary, NC, SAS Institute Inc.

Loison, A., Festa-Bianchet, M., Gaillard, J.M., Jorgenson, J.T. et Jullien, J.M. 1999. Age-specific survival in five populations of ungulates: Evidence of senescence. *Ecology* **80**:2539-2554.

Lynch, M. et Walsh, B. 1998. *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Mass.

Malécot, G. 1948. *Les mathématiques de l'hérédité*. Masson, Paris.

Maynard Smith, J. 1974. *Models in ecology*, Cambridge University Press.

McCarthy, M.A. et Masters, P. 2005. Profiting from prior information in Bayesian analyses of ecological data. *Journal of Applied Ecology* **42**:1012-1019.

McGraw, J.B. et Caswell, H. 1996. Estimation of individual fitness from life-history data. *American Naturalist* **147**:47-64.

Milligan, B.G. 2003. Maximum-likelihood estimation of relatedness. *Genetics* **163**:1153-1167.

Milner, J.M., Pemberton, J.M., Brotherstone, S. et Albon, S.D. 2000. Estimating variance components and heritabilities in the wild: a case study using the `animal model' approach. *Journal of Evolutionary Biology* **13**:804-813.

Mood, A.M., Graybill, F. et Boes, D.C. 1974. *Introduction to the theory of statistics.* 3rd edition. McGraw–Hill, New York.

Morrison, D.F. 1967. *Multivariate statistical methods*. Mc-Graw-Hill, New York.

Moyes, K., Coulson, T., Morgan, B.J.T., Donald, A., Morris, S.J. et Clutton-Brock, T.H. 2006. Cumulative reproduction and survival costs in female red deer. *Oïkos* **115**:241-252.

Newton, I., ed. 1989. *Lifetime Reproduction in Birds*. Academic Press, London.

Nichols, J.D. et Kendall, W.L. 1995. The use of multi-state capture-recapture models to address questions in evolutionary ecology. *Journal of Applied Statistics* **22**:835-846.

Nichols, J.D., Hines, J.E., Pollock, K.H., Hinz, R.L. et Link, W.A. 1994. Estimating breeding proportions and testing hypotheses about costs of reproduction with capture-recapture data. *Ecology* **75**:2052-2065.

Nichols, J.D., Hines, J.E. et Blums, P. 1997. Tests for senescent decline in annual survival probabilities of common pochards, Aythya ferina. *Ecology* **78**:1009-1018.

Nielsen, R. et Wakeley, J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**:885-896.

Nordborg, M. 2007. Coalescent theory. In Balding, D.J., Bishop, M. et Cannings, C. (eds.). *Handbook of statistical genetics.* Wiley, Chichester, 3rd edn:843-877.

Pinheiro, J.C. et Bates, D.M. 2000. *Mixed-effects models in S and S-Plus*. Springer Verlag, New York.

Pons, O. et Petit, R.J. 1995. Estimation, variance and optimal sampling of gene diversity I. Haploid locus. *Theoretical and Applied Genetics* **90**:462-470.

Potvin, C. et Roff, D. 1993. Distribution-free methods: Viable alternatives to parametric statistics. *Ecology* **74**:17-28.

Pradel, R. 2005. Multievent: an extension of capture-recapture models to uncertain states. *Biometrics* **61**:442-447.

Pradel, R., Johnson, A.R., Viallefont, A., Nager, R.G. et Cézilly, F. 1997. Local recruitment in the greater flamingo: a new approach using capture-mark-recapture data. *Ecology* **78**:1431-1445.

Raymond, M. et Rousset, F. 1995. GENEPOP (version-1.2) – Population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**:248-249.

Rousset, F. et Leblois, R. 2007. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model misspecification. *Mol. Biol. Evol.* **24**:2730-2745.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**:1219-1228.

Rousset, F. 2004. *Genetic structure and selection in subdivided populations*. Princeton University Press, Princeton, NJ.

Rousset, F. 2007. Inferences from spatial population genetics. In Balding, Bishop, D.J., M. & Cannings, C. (eds.). *Handbook of statistical genetics*. Wiley, Chichester, 3rd edn 945-979.

Searle, S.R., Casella, G. et McCulloch, C.E. 1992. *Variance components*. Wiley, New York.

Seber, G.A. et Wild, C.J. 1988. *Nonlinear regression*, Wiley, New York.

Skalski, J.R., Hoffman, A. et Smith, S.G. 1993. Testing the significance of individual- and cohort-level covariates in animal survival studies. In Lebreton, J.-D. et North, P.M. (Eds). *Marked individuals in the study of bird population*. Birkhauser, Bâle, pp. 9-28.

Sokal, R.R. et Rohlf, F.J. 1981. *Biometry. The principles and practice of statistics in biological research*. New-York, Freeman.

Staudte, R.G. et Sheather, S.J. 1990. *Robust estimation and testing*. Wiley, New York.

Stearns, S.C. 1992. *The evolution of life histories*. New York, Oxford University Press.

Sullivan, J.P., Lundberg, J.G. et Hardman, M. 2006. A phylogenetic analysis of the major groups of catfishes (Teleostei:Siluriformes) using rag1 and rag2 nuclear gene sequences. *Molecular Phylogenetics and Evolution* **41**:636-662.

Tassi, P. 1985. *Méthodes statistiques*. Economica, Paris.

Tavaré, S. 2004. Ancestral inference in population genetics. In Tavaré, S. & Zeitouni, O. (Eds). *Lectures on probability theory and statistics*. Springer, Heidelberg, pp. 1-188.

Tomassone, R., Lesquoy, E. et Millier, C. 1983. *La régression, nouveaux regards sur une ancienne méthode statistique*. Masson, Paris.

Townsend, H.M. et Anderson, D.J. 2007. Assessment of cost of reproduction in a pelagic seabird using multistate mark-recapture models. *Evolution* **61**:1956-1968.

Vaupel, J.W. et Yashin, A.I. 1985. Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics. *The American Statistician* **39**:176-185.

Wang, J. 2002. An estimator for pairwise relatedness using molecular markers. *Genetics* **160**:1203-1215.

Watts, P.C., Rousset, F., Saccheri, I.J., Leblois, R., Kemp, S.J. et Thompson, D.J. 2007. Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. *Molecular Ecology* **16**:737-751.

Weir, B.S. et Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**:1358-1370.

White, G.C. et Burnham, K.P. 1999. Program MARK: survival estimation from populations of marked animals. *Bird Study* **46**:120-139.

Whitlock, M. et McCauley, D.E. 1999. Indirect measures of gene flow - Fst does not equal 1 / (4Nm+1). *Heredity* **82**:117-125.

Williams, B.K., Nichols, J.D. et Conroy, M.J. 2002. *Analysis and management of animal populations.* Academic Press, San Diego.

Wright, S. 1931. Evolution in Mendelian population. *Genetics* **16**:97-159.

Wright, S. 1949. Adaptation and selection. In Mayr, E. (ed). Genetics, Paleontology, and Evolution. Princeton University Press, Princeton, pp. 365-389.